Statistical methods to improve understanding of the genetic basis of complex diseases

Anna Megan Hutchinson

St Catharine's College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

October 2021

For “Team Hutch”: Mum, Dad, Phil, Lucy, Granny, D, Danny, Christina and Conor.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Anna Megan Hutchinson
October 2021

Abstract

Statistical methods to improve understanding of the genetic basis of complex diseases

Anna Megan Hutchinson

Robust statistical methods, utilising the vast amounts of genetic data that is now available, are required to resolve the genetic aetiology of complex human diseases including immune-mediated diseases. Essential to this process is firstly the use of genome-wide association studies (GWAS) to identify regions of the genome that determine the susceptibility to a given complex disease. Following this, identified regions can be fine-mapped with the aim of deducing the specific sequence variants that are causal for the disease of interest.

Functional genomic data is now routinely generated from high-throughput experiments. This data can reveal clues relating to disease biology, for example elucidating the functional genomic annotations that are enriched for disease-associated variants. In this thesis I describe a novel methodology based on the conditional false discovery rate (cFDR) that leverages functional genomic data with genetic association data to increase statistical power for GWAS discovery whilst controlling the FDR. I demonstrate the practical potential of my method through applications to asthma and type 1 diabetes (T1D) and validate my results using the larger, independent, UK Biobank data resource.

Fine-mapping is used to derive credible sets of putative causal variants in associated regions from GWAS. I show that these sets are generally over-conservative due to the fact that fine-mapping data sets are not randomly sampled, but are instead sampled from a subset of those with the largest effect sizes. I develop a method to derive credible sets that contain fewer variants whilst still containing the true causal variant with high probability. I use my method to improve the resolution of fine-mapping studies for T1D and ankylosing spondylitis. This enables a more efficient allocation of resources in the expensive functional follow-up studies that are used to elucidate the true causal variants from the prioritised sets of variants.

Whilst GWAS investigate genome-wide patterns of association, it is likely that studying a specific biological factor using a variety of data sources will give a more detailed perspective on disease pathogenesis. Taking a more holistic approach, I utilise a variety of genetic and functional genomic data in a range of statistical genetics techniques to try and decipher the role of the Ikaros family of transcription factors in T1D pathogenesis. I find that T1D-associated variants are enriched in Ikaros binding sites in immune-relevant cell types, but that there is no evidence of epistatic effects between causal variants residing in the Ikaros gene region and variants residing in genome-wide binding sites of Ikaros, thus suggesting that these sets of variants are not acting synergistically to influence T1D risk.

Together, in this thesis I develop and examine a range of statistical methods to aid understanding of the genetic basis of complex human diseases, with application specifically to immune-mediated diseases.

Acknowledgements

Firstly, I would like to thank my supervisors, Dr Chris Wallace and Dr David Willé, for their continued support on my PhD journey. Chris for believing in me, which allowed me to believe in me, and David for his enthusiasm, encouragement and affable insights into industry. I would also like to thank my earlier mentors, Dr Ying Yan for introducing me to the wonderful world of biostatistics and Professor Steve Abel for his advice and companionship in Durham.

I would like to thank everyone at the MRC Biostatistics Unit for helping to create an exceptionally friendly and welcoming working environment. My colleagues, Angela Talbot for her caring support, Colin Starr for his patience and Caroline Winsor for expertly navigating a global pandemic. My colleagues-turned-friends both at the BSU and in the Wallace group, including Stephen Coleman, Isobel Barrett, Andrew Manderson, Guillermo Reales and Thomas Willis.

Thank you to Polly Machin, Jess Forsdyke and Cathy Galvin, for making my time in Cambridge unforgettable, and to the Vesey girls who have supported and cheered me on throughout. Thank you Hayley Williamson who has always been my rock and thank you Lydia Carr for the constant support, love and chat. Thank you to Izzi Mastrolonardo, my absolute saviour during the pandemic and the most kind-hearted person that I have ever had the pleasure of meeting - I am eternally grateful for your support over the past year.

Finally, the best family I could have ever wished for - “Team Hutch”. My brother Phil for his tough love, Danny and Christina for their support and my sister Lucy who always has and always will be my ultimate role model. To my grandparents Granny and D, for their affection, enthusiasm and for making me feel warm inside. To my mum and dad who are my best friends and the most perfect parents that I am so lucky to have - thank you for always supporting me and being eternally proud of me, I would not be where I am today without your guidance and love.

Thank you to Conor O’Carroll, the missing link to Team Hutch and the love of my life. Thank you for your constant encouragement, support and patience. Thank you for helping me through one of the toughest years of my life and for loving me unconditionally.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Foreword

This year marks the 20th anniversary of publications detailing the draft human genome sequence. Since then, the field of human genetics has advanced at an exhilarating pace, not only in the study of human development and evolution, but also in the study of human diseases - the focus of this thesis.

Large-scale genetics initiatives and generous funding have been instrumental in helping the scientific community to better understand the genetics underpinning complex human diseases. This is likely to have enormous implications for human welfare, including improvements to diagnosis and treatment options, whilst simultaneously highlighting concerns relating to ancestry disparities and ethics.

These large-scale genetics initiatives are generating unprecedented volumes of genetic data, necessitating robust statistical methods to dissect the aetiology of each disease. In addition, genetic data is now often complemented by functional genomic data that is generated from sophisticated, high-throughput experiments. More complex statistical methods are therefore required to deduce meaning from these heterogeneous data types.

An alliance between the generation of biological data and the development of applicable statistical methods has become essential. By utilising this methodology we can hope to gain a comprehensive understanding of complex disease pathogenesis for the future benefit of humanity.

1.2 Genome-wide association studies

1.2.1 Background

A principal goal of quantitative genetics is to identify genetic risk factors for an observable characteristic called a "phenotype" (Bush and Moore, 2012). Genome-wide association studies (GWAS) are a tool used to identify genetic variations that are common in the population and that associate with a phenotype of interest, such as a "complex disease" which is caused by both genetic and environmental risk factors. GWAS are performed by genotyping large participant cohorts at specific regions of the genome, generally those with a single base-pair ("bp") difference in the DNA sequence that is common in the population, called single nucleotide polymorphisms or "SNPs" (used interchangeably with "variants" in this thesis). These SNPs are then individually tested for an association with the phenotype of interest in a regression framework.

Following the completion of the first draft of the human genome sequence published in 2001 (International Human Genome Sequencing Consortium, 2001), the first GWAS was performed a year later by the Tanaka lab at the SNP Research Centre in Tokyo (Ozaki et al., 2002). At this time the number of genes in the human genome was estimated to be approximately $100,000$ and, naively believing that one gene was expected to be covered by one SNP (Ikegawa, 2012), the group proceeded by genotyping 92,788 SNPs in 94 patients with myocardial infarction and 658 control samples from the general Japanese population (Ohnishi et al., 2001). The researchers identified a SNP in the lymphotoxin-α (*LTA*) gene, residing in the major histocompatibility complex (MHC) region of the genome, that associated with myocardial infarction. That is, there was a significant difference in genotypes at this SNP between myocardial infarction cases and the control samples. Curiously, this study is often overlooked, and Klein and colleagues are generally accredited with the "first GWAS" for their study of age-related macular degeneration in 2005 (Klein et al., 2005; Visscher et al., 2012). The researchers genotyped 160,000 SNPs in 96 cases and 50 control samples and identified a SNP in the complement factor H (*CFH*) gene that associated with age-related macular degeneration.

In 2007, a major GWAS landmark came with the publication of the Wellcome Trust Case Control Consortium (WTCCC) paper (Wellcome Trust Case Control Consortium, 2007), which represented the first large, well-designed GWAS for complex diseases (Ikegawa, 2012). The study used an oligonucleotide SNP array (a type of DNA microarray used to genotype samples at specific SNPs) with good coverage of the human genome. The core study comprised an analysis of 3000 control samples from the British population and 2000 case samples from each of seven diseases (type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and Crohn's disease) and identified 24 independent associations throughout the genome.

The results from the WTCCC study and other subsequent GWAS initiated the development of a custom genotyping array, the "Immunochip", that provided dense coverage of SNPs residing in genomic regions with evidence of an association with immune-mediated diseases. Specifically, the developing 1000 Genomes resource (The 1000 Genomes Project Consortium, 2015) was used to obtain all available variants that were identified in European populations and that resided close to regions robustly associated with one or more autoimmune or inflammatory disease for inclusion on the chip (Cortes and Brown, 2011). The Immunochip provided a breakthrough in immunogenetic association testing in relation to cost and efficiency, albeit at the expense of reduced genomic coverage.

Numerous experimental and computational advances over the past decade, such as improvements in genotyping platforms, statistical imputation methods and increased cohort sample sizes have resulted in a plethora of genetic associations for human traits, with the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) now containing 293,040 associations from 5343 publications (accessed 2021-09-23). Following the sensation of the "GWAS-era", we are now positioned in the "post-GWAS era" where there has been a pronounced shift in focus to the functional and mechanistic understanding of the associations yielded by GWAS.

1.2.2 Genetic association testing

Genetic association testing in GWAS begins by identifying a well-defined phenotype, generally a quantitative trait that produces a continuous phenotype such as height, or a binary trait such as case or control status for a specific disease. Participants are then recruited and genotyped at pre-specified genomic regions, typically a subset of SNPs.

Humans are "diploid" because they carry paired chromosomes with one inherited from each parent. SNPs are germline substitutions of a single nucleotide at a specific position in the genome (termed a "locus"). For "biallelic" SNPs, the two possible nucleotides or "alleles" (one from each parent) are labelled as "reference" (typically the allele that is foundt in the reference genome) or "alternative" (referring to any allele other than the reference allele). At each genomic locus housing a SNP, an individual can be homozygous for the reference allele (carry two copies of the reference allele), heterozygous (carry one copy of each allele), or homozygous for the alternative allele (carry two copies of the alternative allele) which correspond to genotype counts of 0, 1 and 2 for the alternative allele, respectively.

For a quantitative trait, linear regression is used to model the relationship between the trait and each SNP i $(i = 1, ..., m)$ that is typed in the study:

$$\boldsymbol{y} = \beta_0 + \beta_i \boldsymbol{x}_i + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y} = (y_1, ..., y_n)$ is a vector of phenotype values (e.g. height in centimetres) across n samples, $\boldsymbol{x}_i = (x_1, ..., x_n)$, $x_j \in \{0, 1, 2\}$, is a vector of genotype information at SNP i across n samples (0, 1 or 2 counts of the alternative allele) and $\boldsymbol{\epsilon}$ is a normally distributed error term.

For binary traits, linear regression is not suitable since the corresponding model space allows parameters that may result in predictions outside the range $[0, 1]$. Instead, logistic regression is used:

$$\text{logit}(Pr(\boldsymbol{y} = \boldsymbol{1} | \boldsymbol{X}_i = \boldsymbol{x}_i)) = \beta_0 + \beta_i \boldsymbol{x}_i, \tag{1.2}$$

where logit represents the logit function, defined by $\text{logit}(z) = \log(\frac{z}{1-z})$, $\boldsymbol{y} = (y_1, ..., y_n)$ is a binary vector of phenotype values (e.g. $0 =$ controls from the general population, $1 =$ cases for a disease) across n samples and $\boldsymbol{x}_i = (x_1, ..., x_n)$, $x_j \in \{0, 1, 2\}$, is a vector of genotype information at SNP i across n samples.

For both quantitative and binary traits, the parameters to be estimated are β_0 and β_i. Most GWAS assume an additive mode of inheritance, thus utilising an additive model based on alleles that assumes there is a uniform, linear increase in risk for each copy of the alternative allele. Assuming an additive model for quantitative traits in equation (1.1), β_0 is the mean phenotype value for genotypes of 0 at SNP i and β_i is the effect of each copy of the alternative allele of SNP i on the mean phenotype value.

Assuming an additive model for binary traits (corresponding to additive on the log odds scale and hence multiplicative on the odds scale) in equation (1.2), β_0 is the logarithm of odds for genotype 0 at SNP i and β_i is the log odds ratio (OR) between genotype values 1 and 0 for SNP i. If we denote the event of having the alternative allele as $A = 1$ (and similarly the event of not having the alternative allele as $A = 0$) then the OR compares the odds of $Y = 1$ (e.g. having the disease) contingent on having the alternative allele with the odds of $Y = 1$ contingent on not having the alternative allele, so that

$$OR = \frac{\frac{Pr(Y = 1 | A = 1)}{1 - Pr(Y = 1 | A = 1)}}{\frac{Pr(Y = 1 | A = 0)}{1 - Pr(Y = 1 | A = 0)}}. \tag{1.3}$$

Since an additive model is assumed, 2 × β_i is the log OR between genotype values 2 and 0 for SNP i (and β_i is also the log OR between genotype values 1 and 2 for SNP i). Other genetic models representing alternative modes of inheritance are possible, such as the dominant or recessive model whereby two of the genotypes are grouped based on the dominant or recessive variant (for example if the alternative allele is dominant then genotype counts of 1 and 2 can be grouped together as they theoretically produce the same phenotype). In any case, a specific mode of inheritance needs to be considered because the OR cannot directly compare the three genotypes values that are possible in diploid organisms.

The parameters β_0 and β_i are estimated by $\hat{\beta}_0$ and $\hat{\beta}_i$, for example by using maximum likelihood estimation. For normally distributed errors, the ordinary least squares (OLS) estimator can be used:

$$(\hat{\beta}_0, \hat{\beta}_i)^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1.4}$$

where $\boldsymbol{X} = \begin{bmatrix}\mathbf{1} & \boldsymbol{x}_i\end{bmatrix}$. The estimated parameter $\hat{\beta}_i$ is typically the quantity of interest in GWAS as it assesses the evidence of an association between SNP i and the phenotype of interest.

Both linear and logistic regression approaches in GWAS typically test one SNP at a time due to the computational and statistical constraints of fitting regression models for up to millions of SNPs in up to millions of individuals. Moreover, when the number of predictors (m; the number of SNPs) exceeds the number of samples (n; the number of individuals), the OLS estimator is unworkable as $\boldsymbol{X}^T\boldsymbol{X}$ is a singular matrix and the inverse does not exist. That is, there is no unique solution to the system where the number of parameters to estimate is larger than the number of samples. Various mixed-model-based approaches have now been developed for whole-genome regression, including BOLT-LMM (Loh et al., 2015), SAIGE (Zhou et al., 2018) and REGENIE (Mbatchou et al., 2021), but the uptake of these methods by the GWAS community is slow, perhaps due to the computational complexity of the methods.

The GWAS analysis framework for quantitative and binary traits now converge, with the derivation of Z-scores for each SNP. Under the assumption that the sample size is large, $\hat{\beta}_i$ follows a normal distribution with 0 mean, so that the standardised Z-score is

$$Z_i = \frac{\hat{\beta}_i}{\sqrt{V_i}}, \tag{1.5}$$

where $V_i = var(\hat{\beta}_i)$ which can be directly estimated from the regression model if the full genotype data is available. However, due to privacy concerns researchers may not have access to the full genotype data but instead have access to study-specific and population-specific summary data, such as Z-scores and minor allele frequencies (measuring the frequency at which the less frequent alleles occur) (Shim et al., 2015). Specifically, for a quantitative trait standardised so that $var(\boldsymbol{y}) = 1$,

$$V_i \approx \frac{1}{2 \times n \times MAF_i \times (1 - MAF_i)}, \tag{1.6}$$

where n is the total sample size and MAF_i is the minor allele frequency for SNP i. Similarly, for a binary trait,

$$V_i \approx \frac{1}{2 \times n \times MAF_i \times (1 - MAF_i) \times s \times (1 - s)}, \tag{1.7}$$

where s is the proportion of case samples.

The Z-scores can then be used as the test statistics in a formal hypothesis test. For example a p-value can be derived using a two-tailed test, which is defined as the probability of seeing a Z-score as extreme or more extreme than the observed Z-score, given that the null hypothesis of no association ($\beta_i = 0$, typically denoted by H_0) is true. The null hypothesis is rejected if the p-value is below some pre-selected threshold, and the corresponding SNP is then said to associate with the phenotype of interest.

1.2.3 Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association between alleles at different loci in a specific population. Two alleles are said to be in LD if they co-occur together more, or less often than what would be expected if the loci were independent and associated randomly (Slatkin, 2008). This can occur due to recombination during meiosis, where alleles are exchanged between chromosomes. If two genetic loci are close to each other and the frequency of recombination between them is low, then they will often be inherited as a single unit and said to be in "high LD".

Definitions of LD depend on the coefficient of LD,

$$D_{AB} = p_{AB} - p_A p_B, \tag{1.8}$$

which compares the frequency with which two alleles, A and B, at different loci co-occur (p_{AB}) to the frequency with which they would co-occur if they were independent ($p_A p_B$). Although the quantity D_{AB} relates to specific alleles at specific loci, if the loci are biallelic and a and b are the other possible alleles, then $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab} = D$ (Calabrese, 2019).

Since the range of possible values for D depends on the frequency of the alleles, it is not straightforward to compare values between different pairs of alleles. A solution would be to normalise D (Lewontin, 1964) but these values have been shown to be inflated if the sample size is small or in the case of an extreme allele frequency (Teare et al., 2002). Another approach to quantify LD, and the one which will be used throughout this thesis, is analogous to the square of the Pearson correlation coefficient:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}, \tag{1.9}$$

which has a maximum value of 1 and is equal to 0 if two biallelic loci are in perfect equilibrium.

1.3 Fine-mapping genetic associations

1.3.1 Background

Specific SNPs are genotyped in GWAS, yet due to LD these SNPs serve as "proxies" for other SNPs in LD with these. That is, GWAS identify multiple statistical, but often non-causal, associations at SNPs that are in LD with the true "causal variants" (the genetic variants which are responsible for the association signal) (Spain and Barrett, 2015; Visscher et al., 2012). It is for this reason that GWAS are often interpreted as identifying "genomic regions" that associate with a trait of interest.

In GWAS, LD is convenient as it means that only a subset of "tagging SNPs" need to be genotyped in order to identify many associations, thereby saving time and resources. However, in order to identify the true underlying causal variants, these LD patterns need to be deconvoluted in an additional "fine-mapping" step. Fine-mapping is a statistical approach, and so a final follow-up step is required that directly tests the individual candidate causal variants that were prioritised by statistical fine-mapping for a functional effect. These wet-lab follow-up experiments are typically laborious and expensive.

All fine-mapping studies require that causal variants are available in the GWAS data set (either directly genotyped or imputed), with large sample sizes to enable sufficient power to distinguish between associated variants in LD and good-quality data to avoid misleading results (Spain and Barrett, 2015). The early statistical fine-mapping methods that were developed assumed any genomic region containing variants in LD with a GWAS association signal could contain at most one causal variant (Maller et al., 2012). Whilst biologically unrealistic, the approaches developed set the framework for much future work.

1.3.2 Traditional fine-mapping approaches

The SNP with the strongest statistical evidence of an association from GWAS (for example the smallest *p*-value) in a specific genomic region containing variants in LD is called the "lead variant". Due to LD, the lead variant may not be the causal variant in the region (Farh et al., 2015), and van de Bunt et al. (2015) showed that the lead variant was only the causal variant in 2.4% of simulations when using representative parameter values. Other approaches are therefore required to select sets of the most likely causal variants (although the lead variant is often still reported to represent the association signal, e.g. on the NHGRI-EBI GWAS Catalog).

The "top-k SNPs" method prioritises the top k variants with the highest evidence of an association in each associated region, however this method requires an arbitrary choice for the value of k and does not adequately account for LD between variants (Hormozdiari et al., 2014). Another approach is to prioritise the lead variant before extending the analysis to include variants in "high LD" with the lead variant, however this approach requires an arbitrary choice for the quantification of "high LD". Both of these methods also lack an objective quantification that the true causal variants are actually prioritised by the methods, leading to problems relating to the allocation of resources in the expensive functional follow-up studies that are required to elucidate the true causal variants from the prioritised sets of variants.

1.3.3 Bayesian fine-mapping

In 2012, and unusually for GWAS at the time, Bayesian approaches were developed for fine-mapping which resolved many of the drawbacks of the existing methods (Maller et al., 2012). Bayesian approaches are naturally suited to the evaluation of multiple hypotheses, each corresponding to the possibility that a different variant could be causal and responsible for the whole pattern of association across a region. In Bayesian fine-mapping, evidence for association at each SNP is summarised by a posterior probability of causality (PP), which is the probability that a given SNP is the causal SNP relative to all other SNPs in a region, given the data. Unlike p-values for which the traditional fine-mapping methods were based, posterior probabilities can be meaningfully compared both within and across studies.

The supplementary text in Maller et al. (2012) describes a method to calculate per-SNP posterior probabilities of causality using genotype data, upon which the following is based. Let $\hat{\beta}_i$ for $i = 1, ..., m$ SNPs in a genomic region be the regression coefficients from single-SNP logistic regression models, quantifying the evidence of an association between SNP i and the binary trait of interest. Assuming that there is only one causal variant per region and that this is typed in the study, if SNP i is causal, then $\beta_i \neq 0$, and β_j (for $j \neq i$) is non-zero only through LD between SNPs i and j. No parametric assumptions are required for β_i at this stage, so I write that it is sampled from some distribution, $\beta_i \sim [\ ]$. The likelihood is then,

$$\begin{aligned} Pr(\boldsymbol{X}|\beta_i \sim [\ ],\ i \text{ causal},\ \boldsymbol{y}) &= Pr(\boldsymbol{X}_i|\beta_i \sim [\ ],\ i \text{ causal},\ \boldsymbol{y}) \times Pr(\boldsymbol{X}_{-i}|\boldsymbol{X}_i,\ \beta_i \sim [\ ],\ i \text{ causal},\ \boldsymbol{y}) \\ &= Pr(\boldsymbol{X}_i|\beta_i \sim [\ ],\ i \text{ causal},\ \boldsymbol{y}) \times Pr(\boldsymbol{X}_{-i}|\boldsymbol{X}_i,\ i \text{ causal})\,, \end{aligned} \tag{1.10}$$

since $\boldsymbol{X}_{-i}$ is independent of β_i given $\boldsymbol{X}_i$. Here, and in contrast to section 1.2.2, $\boldsymbol{X}$ is the genotype information (0, 1 or 2 counts of the alternative allele per sample) for all SNPs in the genomic region and i is a SNP in the region, such that $\boldsymbol{X}_i$ and $\boldsymbol{X}_{-i}$ are the genotype data at SNP i and at the remaining SNPs across samples, respectively.

Parametric assumptions can now be placed on SNP i's true effect on disease. This is typically quantified as log OR, and is assumed to be sampled from a Gaussian distribution, $\beta_i \sim N(0, W)$, where the prior variance W is chosen to reflect the researcher's prior belief on the variability of the true OR. For example, $W = 0.2^2$ is often selected in a case-control setting, which puts probability 0.02 on ORs either above 1.5 or below 0.67 (Wellcome Trust Case Control Consortium, 2007). Other distributions have been considered by researchers, such as the Laplace distribution (Walters et al., 2019), but in joint work with my supervisor, we found that fine-mapping conclusions were similar across a range of prior distributions placed on the causal variants effect size (Hutchinson et al., 2020a).

The posterior probabilities of causality for each SNP i in an associated genomic region with m SNPs can be written as

$$PP_i = Pr(\beta_i \sim N(0, W),\ i \text{ causal} | \boldsymbol{X},\ \boldsymbol{y}), \quad i \in \{1, ..., m\}. \tag{1.11}$$

Assuming that each SNP is *a priori* equally likely to be causal, then

$$Pr(i \text{ causal}) = \frac{1}{m}, \quad i \in \{1, ..., m\} \tag{1.12}$$

and Bayes theorem can be used to write

$$PP_i = Pr(\beta_i \sim N(0, W),\ i \text{ causal} | \boldsymbol{X},\ \boldsymbol{y}) \propto Pr(\boldsymbol{X} | \beta_i \sim N(0, W),\ i \text{ causal},\ \boldsymbol{y}). \tag{1.13}$$

In Bayesian theory, the "Bayes factor" (BF) compares the likelihood under one particular hypothesis to the likelihood under another hypothesis. In Bayesian fine-mapping, the Bayes factor compares the marginal likelihood of the data at that SNP under different prior distributions for its effect on the phenotype, where the prior distributions compare an "associated" hypothesis, H_A, to a "non-associated" hypothesis, H_0. Under H_0, we assume $\beta_i = 0$ and under H_A we assume $\beta_i \sim N(0, W)$.

Dividing equation (1.13) by the probability of the genotype data given the null model of no genetic effect yields a likelihood ratio which is exactly the Bayes factor,

$$PP_i \propto \frac{Pr(\boldsymbol{X} | \beta_i \sim N(0, W),\ i \text{ causal},\ \mathbf{y})}{Pr(\boldsymbol{X} | \beta_i = 0,\ \boldsymbol{y})} = BF_i. \tag{1.14}$$

The evaluation of the likelihoods assume that genotype data is available for all study participants, but often only summary statistics (consisting of Z-scores and MAFs, or estimated effect sizes and their standard errors) are available, for example on the NHGRI-EBI GWAS Catalog. Wakefield (2009) enabled the natural extension of the Bayesian fine-mapping framework to summary statistics, describing Wakefield's Approximate Bayes Factors (ABFs). Given that $\hat{\beta}_i \sim N(\beta_i, V_i)$ and *a priori* $\beta_i \sim N(0, W)$, only the marginal Z-score for SNP i, and estimates for V_i and W are required to derive

$$PP_i \propto BF_i \approx ABF_i = \sqrt{\frac{V_i}{V_i + W}} exp\left(\frac{Z_i^2}{2}\frac{W}{(V_i + W)}\right). \tag{1.15}$$

Generally no single variant is identified as overwhelmingly likely to be causal, and so researchers prioritise credible sets of potentially causal variants. These are derived by sorting variants into descending order of posterior probability and adding variants to the set until the cumulative sum of posterior probabilities exceeds some threshold, α, to form a $(100 \times \alpha)\%$ credible set (Maller et al., 2012).

1.3.4 Modern fine-mapping approaches

Recently there have been several key contributions to the field of statistical fine-mapping, particularly with regard to removing the single causal variant assumption and simultaneous fine-mapping in multiple ancestries, of multiple traits, or by including external biological data. These are discussed in Hutchinson et al. (2020a) for which I wrote the first draft, and the text reused in this section is my own work.

In relation to the single causal variant assumption, for a locus containing m SNPs, removing this assumption and allowing for multiple causal variants in region yields a total of 2^m possible causal configurations. Earlier methods such as CAVIAR (Hormozdiari et al., 2014) and its successor CAVIARBF (Chen et al., 2015) used exhaustive search to enumerate over all possible causal configurations. Other methods such as GUESSFM (Wallace et al., 2015), FINEMAP (Benner et al., 2016), JAM (Newcombe et al., 2016) and SuSiE (Wang et al., 2020) have built more complex but scalable alternative search strategies. Briefly, GUESSFM clusters SNPs into "tag sets" to initially reduce the search space and then uses the stochastic search algorithm, GUESS (Bottolo and Richardson, 2010; Bottolo et al., 2013), to explore the multimodal model space. FINEMAP implements a shotgun stochastic search, evaluating many neighbouring models at each iteration to efficiently search the vast space of models, whilst JAM uses a formal reversible jump MCMC algorithm. The "Sum of Single Effects" (SuSiE) regression model and its associated model selection framework, iterative Bayesian stepwise selection (IBSS), offer novel deterministic algorithms for computing approximate posterior distributions without assuming a single causal variant. Briefly, the overall effect vector is constructed as a sum of multiple single-effect vectors that each have one non-zero entry for a potential causal variant. SuSiE outputs credible sets of potentially causal variants, analogous to those from the conventional single causal variant Bayesian fine-mapping approach.

When used to fine-map 84 prostate cancer susceptibility loci using summary data from eight GWAS sub-cohorts, JAM identified additional independent signals in 12 regions (Dadaev et al., 2018). Likewise, when using SuSiE to fine-map SNPs that influence splicing in 77,345 introns, 156 additional independent signals were uncovered (Wang et al., 2020). These signals would likely have been missed if using conventional fine-mapping methods that assume a single causal variant in each region, and illustrate the potential gains of modelling multiple causal variants.

Statistical power is the probability of identifying a true positive result, for example a true causal variant in statistical fine-mapping. The power for fine-mapping is influenced by factors such as the effect sizes of the causal variants, the local LD structure, the sample size and the SNP density (Schaid et al., 2018). Joint fine-mapping of multiple outcomes has the potential to improve statistical power through increased sample sizes. For example, the multinomial fine-mapping (MFM) approach (Asimit et al., 2019) fine-maps multiple diseases in data sets that share control samples, combating the high computational load of considering multiple outcomes by expressing the ABF of a joint model as a function of the individual disease ABFs, the model sizes and the sample sizes.

Allele frequencies and LD patterns vary between populations which have been geographically separated. Consequently, combining information across populations by simultaneously fine-mapping multiple ancestries not only increases fine-mapping power through sample size, but also increases resolution through exploiting differences in LD. The CAVIAR framework has been extended to consider multiple populations with different LD structures, and assumes the same causal variants between populations but allows for different effects at those variants (LaPierre et al., 2021). Alternatively, MANTRA exploits the expected similarity in allelic effects between closely related populations, but assumes a single causal variant per region (Morris, 2011).

Finally, rather than using only association statistics for fine-mapping, external biological information can also be incorporated. For example, PAINTOR (Kichaev et al., 2014) allows multivariate binary functional annotation data to influence fine-mapping by allowing the probability that a SNP is causal to vary. I found that this was most useful for distinguishing multiple associated variants in high LD (Fig 1.1). Alternatively, PolyFun (Weissbrod et al., 2020) estimates per-SNP heritabilities for groups of SNPs showing similar functional enrichment with the trait of interest. Prior probabilities of SNP causality are then specified as proportional to the average heritability for the SNP group, and these can then be used directly in existing fine-mapping approaches such as FINEMAP and SuSiE.

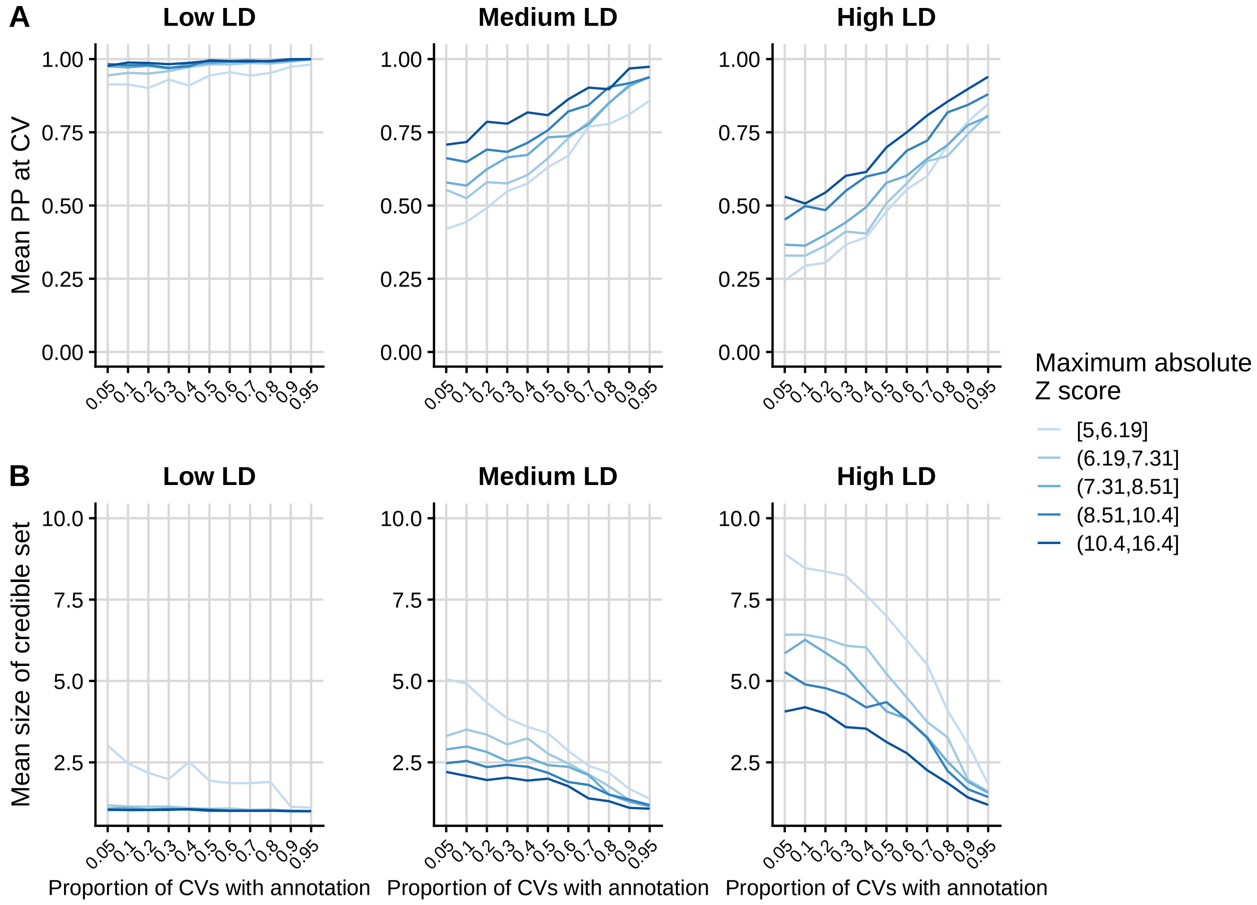


Fig. 1.1 The utility of incorporating functional annotation data for single causal variant fine-mapping using PAINTOR. I simulated summary GWAS data for 8000 regions, each with a single causal variant, and used PAINTOR to analyse sets of 100 regions, varying the proportion of causal variants carrying a specific annotation which was controlled to be present in 5% of all non-causal SNPs. As the proportion of causal variants with the annotation increases, (A) the mean posterior probability (PP) at the causal variant tended to increase and (B) the size of the 95% credible set tended to decrease. The greatest gain in using relevant functional annotation data was for regions with medium-high LD, where the enrichment of the functional annotation allowed a variant with the annotation to be picked from a set of variants showing similar levels of association. I distinguished between low, medium and high LD causal variants ($MAF > 0.05$), according to the number of other SNPs that they were in LD with ($r^2 > 0.5$): 2 or fewer, 3-10, or more than 10, respectively. Figure from Hutchinson et al. (2020a) and generated by myself.

1.4 Functional genomic data and SNP enrichment

1.4.1 Background

GWAS and fine-mapping relate to the field of quantitative genetics, which is the study of the genetic basis underlying phenotypic variation among individuals (Reshma and Das, 2021). Distinct yet related is the field of molecular biology, which is the study of "the composition, structure and interaction of cellular molecules such as nucleic acids and proteins that carry out the biological processes essential for the cells functions and maintenance" (https://www.nature.com/subjects/molecular-biology). A fundamental aim of molecular biology is to understand the mechanistic function of the human genome, and a question that lies at the intersection of quantitative genetics and molecular biology is therefore: "What are the functional effects of genetic variants across the human genome?" (Lappalainen, 2015).

Analogous to how GWAS are a key tool for quantitative genetics, functional genomic experiments are a key tool for molecular biology. Functional genomic experiments generate functional genomic data, and broadly refer to those experiments that measure many biological factors (e.g. genes or proteins) in parallel under different experimental or environmental conditions. In the past decade, the maturation of functional genomic experiments has led to the growth of publicly available large-scale, multi-tissue functional data sets, such as those from The ENCODE Project (ENCODE Project Consortium, 2012), NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), GTEx (GTEx Consortium, 2013) and BLUEPRINT (Stunnenberg et al., 2016). These data sets cover DNA methylation (where methyl groups are added to the DNA molecule which typically change the activity of the DNA region; Fig 1.2A), 3D structure (relating to the 3D configuration of the genome; Fig 1.2B) and chromatin modifications (where histone modifications regulate the physical properties of chromatin) including changes in accessibility (relating to the level of physical compaction of chromatin; Fig 1.2C).

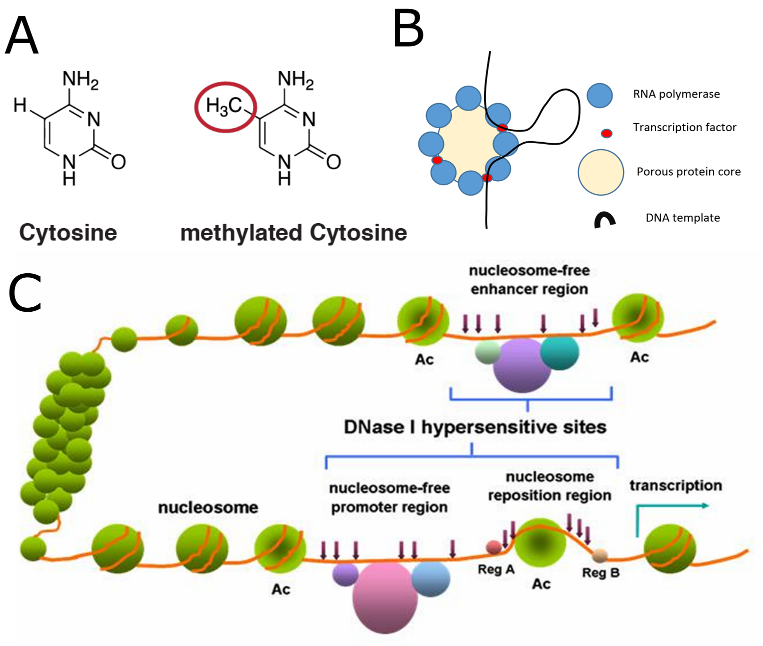


Fig. 1.2 (A) Figure showing the chemical structure of the DNA nucleotide cytosine, both unmethylated and methylated (downloaded from https://en.wikipedia.org/wiki/DNA_methylation; By Mariuswalter - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=54318073). (B) Figure depicting a transcription factor causing the formation of a chromatin loop (downloaded from https://en.wikipedia.org/wiki/File:A_transcription_factory_causing_the_formation_of_a_chromatin_loop.png; By AdrianBaily - Own work, CC BY-SA 3.0). (C) Figure showing how regions of the DNA can be accessible ("nucleosome-free region") or inaccessible due to the position of nucleosomes (downloaded from Wang Y-M, Zhou P, Wang L-Y, Li Z-H, Zhang Y-N, et al., CC BY-SA 2.5 https://en.wikipedia.org/wiki/DNase_I_hypersensitive_site, via Wikimedia Commons)

SNP enrichment methods are an exemplar of the intersection of quantitative genetics and molecular biology. They aim to elucidate the functional annotations that trait-associated SNPs overlap more frequently than expected by chance and have revealed how trait-associated SNPs are enriched in open chromatin regions of the genome (de la Torre-Ubieta et al., 2018; Finucane et al., 2015), including non-coding regulatory segments of the DNA called "enhancers" (Corradin and Scacheri, 2014). Since many functional annotations (including enhancers) are cell-type specific, it is important that the results are considered with respect to the cell type within which the functional annotation was measured, and thus SNP enrichment is generally considered in a trait-specific manner (Cano-Gamez and Trynka, 2020). The most informative annotations likely relate to the cell-type specific biological mechanism of the trait, for example annotations representing enhancer marks in pancreatic tissue may be relevant for dissecting loci associated with type 2 diabetes, but less so for (immune-mediated) type 1 diabetes, where enhancer marks in lymphoid tissues may be more relevant.

Enhancer sequences encompass transcription factor binding sites. A transcription factor is a protein that controls the rate of transcription by binding to a specific DNA sequence - the "consensus sequence". Although in practise, transcription factors are able to bind at thousands of regions across the genome, including those were the DNA sequence differs from the consensus sequence. Transcription factor binding sites make up 31% of GWAS SNPs yet only comprise 8% of the genome (ENCODE Project Consortium, 2012), and studies have shown how causal variants can functionally alter the binding sites for key transcription factors (Claussnitzer et al., 2015; Gupta et al., 2017). Together, this implies a key role for transcription factors in disease pathogenesis.

1.4.2 Protein-DNA binding

ChIP-seq

Chromatin immunoprecipitation with sequencing (ChIP-seq) is a functional genomic experiment that captures a snapshot of protein-DNA binding events and chemical modifications of histone proteins (Furey, 2012). It is generally seen as the gold standard for identifying DNA-protein interactions (Cheng et al., 2014) and so I give a brief overview of the experimental procedure. Cells are firstly cross-linked (typically with formaldehyde) to covalently stabilise protein-DNA complexes and are then lysed to extract all nuclear material. The DNA is fragmented into small pieces ($\approx$ 500-bp) and an antibody that is specific to the protein of interest is used to immunoprecipitate and isolate the target DNA pieces containing protein-DNA interactions. The cross links between the protein and DNA are then reversed and the DNA is sequenced. The conventional output of a ChIP-seq experiment are `fastq` files storing DNA sequences which can be mapped back to the reference genome. The mapped DNA sequences are then used in a peak-calling algorithm to identify genomic regions where the protein of interest was bound.

In addition to usual experimental design considerations relating to biological replicates, in ChIP-seq experiments, the antibody must work in chromatin immunoprecipitation and must also be specific for the protein of interest. However, many proteins do not yet have antibodies which satisfy these requirements, thus limiting the applicability of ChIP-seq. The results from ChIP-seq are also cell-type and cell state specific and should be interpreted with this in mind. Although ChIP-seq experiments are seen as the gold standard for identifying DNA-protein interactions and there are now several publicly available large-scale databases containing ChIP-seq data, due to antibody constraints and heterogeneity in cell types and cell states, relevant ChIP-seq data may not be readily available for specific research questions.

Position weight matrices

Position weight matrices (PWMs) are a widely adopted approach to represent motifs in biological sequences (Wasserman and Sandelin, 2004). The rows of the matrix represent each symbol in the relevant alphabet (e.g. four rows for A, T, C and G nucleotides) and the columns represent each position in the motif (e.g. a transcription factor binding motif) (Stormo et al., 1982). PWMs are generally constructed using experimental data which identify the DNA sequences that bind the desired target, such as ChIP-seq experiments for proteins (Chai et al., 2011). From the experimental data, a position frequency matrix (PFM) is constructed by counting the occurrences of each nucleotide at each position of the DNA sequence. Each column of the PFM is then normalised to generate the position probability matrix (PPM), and the final PWM is then obtained by logarithmic transformations of the PPM divided by each nucleotides background probability. To avoid bias due to small sample sizes, such as PPM entries having a value of 0 by chance (leading to negative infinities in the PWM), pseudocounts are generally added to each element of the PFM. Motif matrices are typically visualised as sequence logos, where the size of the symbol in the relevant alphabet represents the relative frequency of that symbol (Fig 1.3).

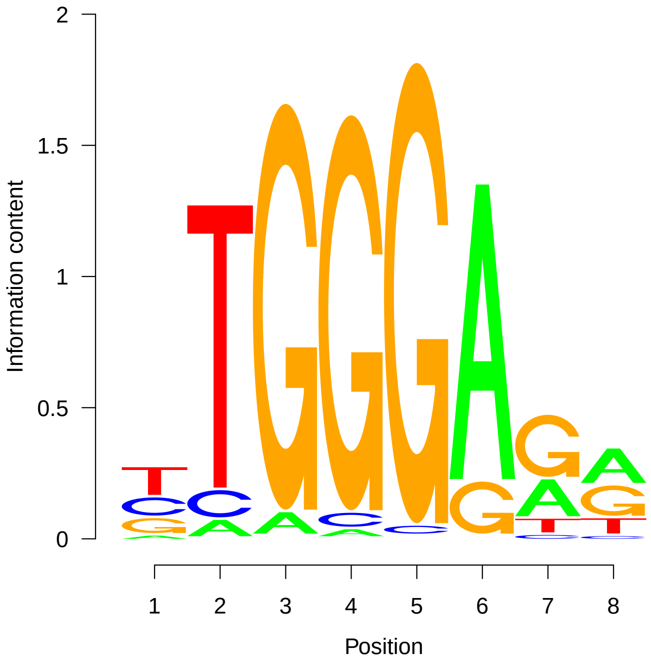


Fig. 1.3 Sequence logo for the transcription factor, Ikaros, made using the `seqLogo` R package (https://bioconductor.org/packages/release/bioc/html/seqLogo.html). The PWM was downloaded from https://hocomoco11.autosome.ru/motif/IKZF1_HUMAN.H11MO.0.C.

Predicting effects of mutations within binding motifs

Whilst PWMs are a useful tool to predict where proteins bind in the genome, they are based only on the frequency of nucleotides and therefore do not provide any information on the direction of effect of mutations within the motif (Weirauch et al., 2013). The Intragenomics Replicates (IGR) method aims to predict the effects of mutations within a motif. As an example, for transcription factor motifs, SNPs residing in a transcription factor binding site are scored based on their modulation in affinity using ChIP-chip or ChIP-seq data (Cowper-Sal·lari et al., 2012). To do this, instances of the mutated motif (i.e. the transcription factor binding motif with the SNP of interest) are searched for in open chromatin regions of the genome and ChIP-seq data is then used to see whether the transcription factor still binds to this mutated motif. To quantify the effect that the mutation has on transcription factor binding, the ChIP-seq signals over all instances of the mutated DNA sequence are averaged, and normalised to the average binding score of the original (non-mutated) sequence.

The SEMpl method is similar to the IGR method and aims to predict the magnitude and direction of effect of all possible SNPs within a transcription factor binding motif (Nishizaki et al., 2020). Firstly, the PWM for the transcription factor of interest is used to enumerate all k-mers for which the transcription factor has an increased likelihood of binding using the TFM-PVALUE method with $p < 4^{-5}$ (Touzet and Varré, 2007), and all possible SNPs are simulated for each k-mer to create lists of mutated k-mers. Each of these mutated k-mers are then aligned to the reference human genome in regions of open chromatin in the relevant cell type using `Bowtie` (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). A “ChIP-seq score” is calculated for each mutated k-mer as the highest ChIP-seq signal value over the region 50-bp before and after the aligned site in the relevant cell type. The “SEM score” for each position is the log_2 transformed average ChIP-seq signal to endogenous signal ratio for the mapped k-mers for each mutated k-mer list. This procedure is repeated using different p-value thresholds in the TFM-PVALUE method until convergence using an expectation–maximisation method, which corrects for differences arising from unique starting k-mers.

1.4.3 SNP enrichment

Overview and confounding

SNP enrichment methods test whether trait-associated variants overlap specific genomic annotations more frequently than expected by chance. However, heterogeneous distributions of trait-associated variants and genomic annotations throughout the genome implicates confounding variables. Confounding variables are defined as factors that influence both the exposure (e.g. SNP association statistics) and the outcome (e.g. genomic annotations). For example, two confounders in SNP enrichment analyses are proximity to gene and MAF (Cano-Gamez and

Trynka, 2020). Proximity to gene is a confounder because trait-associated SNPs are generally found closer to genes, and various functional annotations also tend to reside closer to genes (such as DHS sites measuring chromatin accessibility and protein binding sites). MAF is a confounder because trait-associated SNPs generally have larger MAFs (since MAF directly relates to the power to detect an association in GWAS) and important functional sites may have selective pressure against mutations (which alter function), meaning that the frequency of the alternative alleles do not rise in the population. The most influential confounder, however, is LD (Trynka et al., 2015) which needs to be accounted for in the sense that it captures the correlation and average relationships between SNPs. For example, the more SNPs a particular SNP tags, the more likely it is that one of them will overlap the annotation, even by chance. LD also captures the other sources of confounding by proxy, since if SNPs are correlated then they are more likely to be closer together and have similar MAFs and distances to genes.

Observed test statistics corresponding to the true SNP-annotation overlap are generated in the presence of confounding. This means that in order to accurately assess the significance of these test statistics, a null distribution that captures this confounding is required. One way to do this is to specify a null distribution by matching SNPs based on various confounding characteristics (in so-called "SNP matching methods"). For example, GREGOR (Schmidt et al., 2015) matches SNPs based on LD, gene proximity and MAF, and uses a permutation-based approach to calculate enrichment p-values, whilst GARFIELD (Iotchkova et al., 2019) matches SNPs based on LD and local gene density and uses a logistic regression framework to derive statistical significance. However, by pre-specifying confounding variables, hidden confounders which may bias the enrichment statistics can be missed (Trynka et al., 2015). Moreover, whilst matching based on LD is crucial in SNP matching methods, it is not always clear which other matching parameters should be used in a given analysis. Trynka et al. (2015) found that controlling for LD alone was sufficient to mitigate false positives in some instances, but many other matching parameters were required in other instances. In addition, Iotchkova et al. (2019) found that controlling for LD meant that controlling for MAF was no longer required, which is likely due to the relationship between LD and MAF (e.g. that rarer variants are less likely to be in LD with other variants).

GoShifter

GoShifter (Genomic Annotation Shifter) (Trynka et al., 2015) is a statistical method to test for SNP enrichment that uses an innovative approach to account for confounding. In the method, loci are defined as the genomic regions encompassing all variants in LD with each lead variant (e.g. $r^2 > 0.8$ using a representative reference panel) and these are then extended by twice the median size of the tested annotation to ensure that the loci also contain SNPs not linked to the lead variant. The proportion of loci containing a SNP which overlaps the annotation is calculated, and this quantity is used as the observed test statistic.

To evaluate the significance of the observed test statistic, a null distribution which preserves confounding is derived. Specifically, each locus is circularised and the annotation data is randomly "shifted" within the locus (if an annotation is shifted beyond the boundaries of the locus, it re-emerges the other side), and the proportion of loci containing a SNP which overlaps this shifted annotation is calculated. This is repeated for many iterations to generate many test statistics which comprise the null distribution. Formulating the null distribution in this way preserves key confounding parameters relating to the spatial distribution of genomic features (such as gene density and gene proximity) but does not require that these are specified in advance, or even known *a priori* at all. A p-value for SNP enrichment is then calculated as the proportion of iterations where the calculated test statistic was greater than or equal to the observed test statistic.

1.5 Multiple hypothesis testing

1.5.1 Background

Multiple hypothesis testing refers to conducting multiple statistical tests which can produce a "discovery" in parallel. Table 1.1 shows the possible outcomes when conducting m statistical tests in parallel, where an arbitrary statistical test is used to declare test statistics as significant (the null hypothesis is rejected) or not-significant (the null hypothesis is not rejected). A type I error is a false positive result (in Table 1.1 the number of false positives is V) whilst a type II error is a false negative result (in Table 1.1 the number of false negatives is T).

When considering a single hypothesis test that generates a p-value, using a significance level (or significance threshold) of α corresponds to rejecting the null hypothesis if $p \leq \alpha$. This means that the probability of a type I error is α, and so the probability of no type I errors is $1 - \alpha$. If we now consider conducting m hypothesis tests in parallel, then assuming that all hypotheses are null and the p-values are independent, the probability of no type I errors is $(1 - \alpha)^m$. For $\alpha = 0.05$ and $m = 1, 10, 100$ the probability of no type I errors is 0.95, 0.599 and 0.006 respectively, illustrating how false positive results can become more prevalent when conducting many tests in parallel, coined "the multiple testing problem".

	Non-significant	Significant	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
	m-R	R	m

Table 1.1 Table for multiple testing definitions.

1.5.2 FWER and the Bonferroni correction

The family-wise error rate (FWER) is defined as the probability of at least one false positive finding in a family of hypothesis tests, that is $FWER = Pr(V \geq 1)$ (Table 1.1). The Bonferroni correction (Bonferroni, 1936) is the most popular method used to control the FWER when conducting multiple tests in parallel (Dunnett, 1955). This approach makes the significance threshold more stringent based on the number of independent statistical tests that are conducted in parallel. Such that for m statistical tests, the Bonferroni correction reduces the significance threshold to α/m, where α is the desired overall α level.

Theorem 1. *The Bonferroni correction strongly controls the $FWER$ at level α.*

Proof.

$$\begin{aligned}
FWER = Pr(V \geq 1) &= Pr\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \\
&\leq \sum_{i=1}^{m_0}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\} \\
&= m_0\frac{\alpha}{m} \\
&\leq m\frac{\alpha}{m} \\
&= \alpha
\end{aligned}$$

utilising Boole's inequality, and where m_0 is the number of true nulls and m is the total number of tests (Goeman and Solari, 2014).

□

In GWAS, to account for conducting many tests in parallel (one for each SNP included in the analysis) a stringent "genome-wide significance threshold" of $p \leq 5e - 08$ is used to call significant associations. This value was selected by the GWAS community based on the Bonferroni correction corresponding to $\alpha = 0.05$ and the estimated effective number of independent tests in the genome if all common SNPs in HapMap European samples (The

International HapMap Consortium, 2003) were tested ($\approx 1e{+}06$) (Panagiotou et al., 2012). The Bonferroni correction corresponding to $\alpha = 0.05$ in this instance is $\alpha = 0.05/1e + 06 = 5e - 08$.

Whilst the Bonferroni method reduces type I errors compared to the conventional single-test p-value thresholding, it substantially decreases power which manifests in more type II errors. Holm (1979) described a simple extension to the Bonferroni procedure with the aim of increasing power. Rather than the standard "single-step" Bonferroni procedure (whereby all tests are compared to a single threshold value simultaneously), Holm described a "step-down" procedure which sequentially rejects hypotheses until no more can be rejected. Specifically, the "sequentially rejective Bonferroni test" first sorts the p-values into ascending order and then compares the first smallest p-value to a significance threshold of α/m, the second smallest p-value to a significance threshold of $\alpha/(m-1)$ and so on, until the p-value is larger than its significance threshold, when the procedure stops and the null hypotheses for this test and all subsequent tests are accepted. This extension still controls the FWER and has greater power than the standard Bonferroni procedure, but Lin (2005) showed that this increase in power is marginal when the number of tests is large (since $m - r \approx m$ when m is very large and r is small).

Controlling FWER tends to be over-conservative in practise, and is best suited to instances where an erroneous finding in just one test is particularly relevant. It treats a single false positive finding the same as multiple false positive findings, and it is difficult to imagine applications where this is appropriate (Benjamini and Hochberg, 1995). If researchers are willing to accept some false positives rather than strictly none at all, then other error measures such as the false discovery rate may be preferable.

1.5.3 Frequentist FDR and the Benjamini-Hochberg procedure

Sorić (1989) warned that "a large part of statistical discoveries may be wrong" and quantified the expected number of false discoveries divided by the total number of discoveries as $E[V]/R$ (where V and R are as defined in Table 1.1). Based on this observation, Benjamini and Hochberg (1995) described a new quantity which could be controlled for, the false discovery rate (FDR), which was the first alternative to the FWER to gain widespread use.

Let Q be the proportion of false discoveries amongst the set of discoveries, the FDR is defined as the expectation of Q,

$$FDR = E[Q] = E[V/R] \tag{1.16}$$

(if $R = 0$ then $Q = 0$, to prevent division by 0) (Benjamini and Hochberg, 1995).

In the special case where all m nulls are true, there cannot be any correct rejections so that $R = V$. If there are any rejections in this instance ($R \geq 1 \implies V \geq 1$) then $Q = 1$, otherwise

$Q = 0$. If we consider writing out the expectation explicitly in this case, then we see that if all null hypotheses are true, the FDR is equivalent to the FWER:

$$FDR = E[Q] = [1 \times Pr(Q = 1)] + [0 \times Pr(Q = 0)] = Pr(Q = 1) = Pr(V \geq 1) = FWER. \tag{1.17}$$

Of course, in general there will be some non-true nulls ($m_0 < m$), in which case the FDR is smaller than or equal to the FWER. To see this, consider that there is at least one false rejection ($V \geq 1$), then $V/R \leq 1$, and the indicator function that takes the value 1 if there is at least one false rejection, $\mathbf{1}_{V \geq 1}$, can never be less than $V/R = Q$. If we take expectations,

$$E(\mathbf{1}_{V \geq 1}) \geq E(Q) = FDR. \tag{1.18}$$

The expected value of an indicator function is the probability of the event, so that $E(\mathbf{1}_{V \geq 1}) = Pr(V \geq 1)$, which is the FWER (Benjamini and Hochberg, 1995).

Benjamini and Hochberg (1995) described an FDR controlling procedure (commonly called the "Benjamini-Hochberg (BH) procedure") for which the goal is to call each individual test as either significant (i.e. reject the null hypothesis) or not-significant (i.e. do not reject the null hypothesis) whilst keeping the FDR below some threshold, α. In the procedure, p-values are ranked in ascending order and the largest k is found such that $P_{(k)} \leq \frac{k}{m}\alpha$. The null is then rejected for the kth smallest p-values. This procedure can be shown graphically, whereby the p-values are ranked and plotted against their rank index (Tang, 2019). The null hypothesis is then rejected for the first k tests with p-values below the $y = \frac{\alpha}{m}x$ line (Fig 1.4).

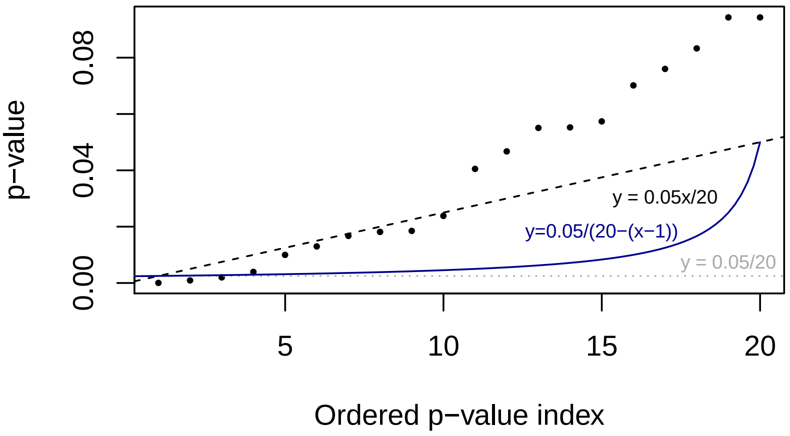


Fig. 1.4 Graphical representations of the BH, Bonferroni and sequentially rejective Bonferroni procedures, with $\alpha = 0.05$. I simulated 20 p-values ($m = 20$) below 0.1 and plotted these against their ordered rank. Dashed line shows $y = \alpha x/m = 0.05x/20$ line. The BH procedure rejects the null hypothesis for the kth smallest p-values where k is the largest value such that $P_{(k)} \leq \frac{k}{m}\alpha$ (here $k = 10$). Grey dotted line shows the standard Bonferroni adjusted significance threshold, whereby only the first three tests with the smallest p-values are called significant. The solid blue curve shows the rejection threshold for the sequentially rejective Bonferroni test ($y = \alpha/(m - (x - 1))$) which also only rejects the null for the first three tests with the smallest p-values.

When considering the number of false positive findings, the BH procedure can be interpreted as an intermediate between uncorrected testing (using $p \leq \alpha$) and corrected testing using the Bonferroni correction (using $p \leq \alpha/m$) (Genovese and Wasserman, 2002). In fact, the BH procedure actually controls the FDR at level $\frac{m_0}{m}\alpha$, meaning that the BH procedure controls the FDR at a level too low. Follow-up papers suggested an adaptive procedure which estimates the number of true nulls (m_0) and applies the BH procedure at level $\alpha \times m/m_0$, and this has been shown to be more powerful than the original approach (Benjamini and Hochberg, 2000; Benjamini et al., 2006).

1.5.4 Storey's q-value

Storey's approach (Storey, 2002; Storey and Tibshirani, 2003) extends the concepts of FDR that are introduced above. Rather than fixing α and then using this to determine which null hypotheses to accept or reject (i.e. determining the "rejection region"), Storey (2002) proposed fixing the rejection region and then estimating α, subsequently using this methodology to define the "q-value".

Recall that the FDR is the expected proportion of false discoveries amongst the set of discoveries. If we incorporate some threshold t, then we aim to estimate that FDR when calling all features significant whose p-value is less than or equal to this threshold value (Storey and Tibshirani, 2003). That is, we are fixing the rejection region whilst estimating

$$FDR(t) = E\left[\frac{F(t)}{S(t)}\right] \approx \frac{E\left[F(t)\right]}{E\left[S(t)\right]}, \tag{1.19}$$

where $F(t)$ is the number of null tests with $p_i \leq t$ (estimating the number of false positives) and $S(t)$ is the total number of tests with $p_i \leq t$. An estimate of E[S(t)] is the observed S(t), which can be calculated by simply counting the number of tests with $p_i \leq t$. Since null p-values are uniformly distributed, $m_0 \times t$ is a suitable estimate of $E[F(t)]$. However the total number of nulls (m_0) is unknown in practise.

Storey proposed estimating the *proportion* of null features $\pi_0 \equiv m_0/m$, which he quantified by

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, ..., m\}}{m(1-\lambda)}, \tag{1.20}$$

where λ is an appropriately selected tuning parameter. For example, if we assume that all tests with corresponding p-values > 0.5 are null (e.g. because the histogram of p-values is flat for $p > 0.5$), then we could set $\lambda = 0.5$. Alternatively, a series of values for λ could be considered,and the final solution could be obtained by smoothing over the resultant estimates (Storey and Tibshirani, 2003).

Storey thus defined

$$F\hat{D}R(t) = \frac{m_0 \times t}{S(t)} = \frac{\frac{m_0}{m} \times m \times t}{\#\{p_i \leq t\}} = \frac{\hat{\pi}_0(\lambda) \times m \times t}{\#\{p_i \leq t\}}, \tag{1.21}$$

which corresponds to the estimated FDR when calling all features with $p_i \leq t$ significant.

The FDR is not monotonically increasing for a sorted set of p-values (corresponding to increasing t), meaning that the FDR associated with the first l ranked tests may be bigger than that for the first $l+1$ ranked tests. Storey therefore defined the q-value as the minimum FDR attained at or above a given value, t:

$$\hat{q}(p_i) = \min_{t \geq p_i} F\hat{D}R(t). \tag{1.22}$$

The q-value is a measure of significance in terms of the FDR rather than the false positive rate. That is, whilst p-values measure significance in terms of the rate that truly null tests are called significant (the false positive rate), q-values measure significance in terms of the rate that significant tests are truly null (the FDR). The q-values are typically thresholded at level α to produce a set of significant features such that a proportion α are expected to be false positives. This approach benefits from greater power than the BH approach, which assumes $\pi_0 = 1$ (Storey and Tibshirani, 2003). The disadvantage of Storey's approach is that it requires the estimation of π_0, which is difficult in practise.

It should be noted that the q-value is technically defined as the minimum "positive FDR" (pFDR) at which the feature can be called significant (rather than the minimum FDR). The pFDR is defined as $pFDR = E(V/R|R > 0)$, but since m is typically large in multiple testing applications, $Pr(R > 0) \approx 1$ and so FDR $\approx$ pFDR (Storey and Tibshirani, 2003). A detailed explanation of the pFDR is given by Storey (2002).

1.5.5 Bayesian FDR and local FDRs

The multiple testing techniques that have been discussed so far reside in the frequentist realm since they are concerned with expected values over repeated experiments. Efron et al. (2001) described a Bayesian version of the FDR estimator, which is asymptotically equivalent (Wen, 2018) to the FDR introduced by Benjamini and Hochberg.

Let z_i denote some univariate summary statistic for the i-th comparison (for example a p-value for SNP i from GWAS) and suppose that the m statistical tests can be divided into two classes: null or non-null with prior probabilities p_0 and $(1 - p_0)$, respectively. The densities of the summary statistics are denoted by $f_0(z)$ and $f_1(z)$ respectively. Efron et al. (2001) defined the mixture density,

$$z_i \sim p_0 f_0(z_i) + (1 - p_0) f_1(z_i) = f(z_i), \tag{1.23}$$

which is also called the "two-groups model".

By applying Bayes theorem, he proposed estimating the posterior probability that a test is null, calling this the "local FDR":

$$\begin{aligned} fdr &\equiv Pr(H_0|Zi = z_i) \\ &= p_0 f_0(z_i)/f(z_i). \end{aligned} \tag{1.24}$$

The values for $f_0(z_i)$ and $f(z_i)$ can either be estimated separately, or their ratio can be estimated directly, typically utilising an empirical Bayes approach. The prior probability, p_0, is also generally estimated from the data and corresponds to π_0 in Storey's q-value approach (Storey and Tibshirani, 2003).

This approach differs from the FDR discussed so far as it considers densities rather than tail areas (Efron, 2004). The approach was therefore extended to define the Bayesian FDR:

$$\begin{aligned} FDR &\equiv Pr(H_0|Zi \leq z_i) \\ &= p_0 F_0(z_i)/F(z_i) \end{aligned} \tag{1.25}$$

where $F_0(z_i)$ and $F(z_i)$ are the corresponding cumulative distribution functions (CDFs), which can be estimated empirically.

The Bayesian FDR is therefore a "mixture average" of local FDRs over the range $Z_i \leq z_i$. That is,

$$FDR = E_f(fdr(Z_i)|Z_i \leq z_i) \tag{1.26}$$

(Efron, 2007).

1.6 Multiple hypothesis testing with covariates

1.6.1 Background

The conventional multiple testing correction procedures control some error measure by assuming that each hypothesis is equally likely *a priori* to be true or false. For example, the BH procedure provides nearly optimal control of the FDR under the condition that almost all null hypotheses are true (Lei and Fithian, 2018). But one can imagine instances where we may believe that some hypotheses are *a priori* more likely than others. For example, we may have additional "covariate" information relating to the null probability of each hypothesis. In these instances the conventional multiple testing correction procedures that obey a monotonicity property (if test statistic z_i is called significant and z_j is more extreme than z_i, then z_j is also called significant) may not be optimal. This motivated the development of a range of covariate-informed multiple testing methods which leverage auxiliary covariates with test statistics, with a consistent aim of minimising type II errors (or equivalently increasing statistical power) whilst controlling some appropriate error rate, such as the FDR (Basu et al., 2018; Bourgon et al., 2010; Cai et al., 2019; Ferkingstad et al., 2008; Genovese et al., 2006; Hu et al., 2010; Ignatiadis et al., 2016; Lei and Fithian, 2018; Li and Barber, 2017; Roeder and Wasserman, 2009; Sun et al., 2006).

1.6.2 Methods based on filtering or stratifying by covariate values

The simplest approach to adjust for multiple testing with covariates is independent filtering, whereby test statistics are first filtered based on the value of the covariate and a multiple testing procedure is then implemented on only the retained test statistics. However, this approach is simplistic, does not utilise the full information contained within the covariate values and the choice of the covariate threshold is subjective (Du and Zhang, 2014), which could potentially lead to problems similar to p-value hacking (Ignatiadis et al., 2016). Moreover, the BH procedure is often used in independent filtering methodologies, but several authors have shown that the FDR is no longer controlled by this approach (Benjamini, 2008; Bourgon et al., 2010).

Alternatively, the stratified BH approach first stratifies test statistics into groups based on their covariate values and then implements the BH procedure in each group before combining the rejections made across the strata (Sun et al., 2006). The development of this approach was motivated by the meta-analysis of five phenotypes relating to type 1 diabetes in the DCCT/EDIC (Diabetes Control and Complications Trail/ Epidemiology of Diabetes Interventions and Complications) study, where Sun et al. (2006) recognised that if the power to detect associations differed between the stratifications (for example because some phenotypes were measured in a larger number of samples or were implicated by SNPs with larger effect sizes), then power could be gained by analysing associations within their stratifications independently rather than aggregating these at the start. Compared to independent filtering, stratified BH better utilises the information contained within the covariate values, but the heterogeneity of covariate values within groups is not exploited. Moreover, this approach can lead to loss of frequentist FDR control if the proportion of true nulls is approximately 1 (Ignatiadis et al., 2016).

1.6.3 P-value weighting methods

Methods have been developed which incorporate weights for each test statistic. Generally, w_i corresponds to the weight of the p-value for hypothesis i, and a weighted p-value is derived by dividing a p-value by its weight, such that $s_i = p_i/w_i$. Higher weights should therefore be allocated to more promising hypotheses, as these would decrease the value of s_i. Conventional multiple testing procedures can then be applied to the weighted p-values.

Weighted Bonferroni and BH procedures

The Bonferroni procedure has been extended to incorporate non-negative weights relating to each hypothesis (Genovese et al., 2006). The original p-values are divided by their weights and the standard Bonferroni correction is then applied to these weighted p-values. That is, hypothesis i is rejected if $p_i/w_i \leq \alpha/m$. The weighted Bonferroni procedure controls the FWER at α if the weights are non-negative and average to 1 (Rubin et al., 2006; Wasserman and Roeder, 2006).

The BH procedure has also been extended to incorporate weights, whereby the standard procedure is applied to the set of weighted p-values (Genovese et al., 2006). As before, if the weights are non-negative and average to 1 then this weighted BH procedure provides finite sample FDR control for sets of independent p-values (Genovese et al., 2006).

Grouped BH and independent hypothesis weighting

The aforementioned weighting procedures do not describe how to transition from covariate values, which may be of any type (e.g. binary, continuous or categorical) to weights satisfying certain constraints required for error control (typically non-negative weights that average to 1). The grouped BH procedure groups test statistics based on their covariate values and then determines a weight for each group based on these covariate values (Hu et al., 2010). Each test statistic in a group is therefore allocated the same weight, and the weighted BH procedure is applied. However, the derivation of these weights is non-trivial and the procedure has been shown to not maintain FDR control (Ignatiadis et al., 2016).

The independent hypothesis weighting (IHW) method is an extension of the grouped BH procedure which assigns optimal weights to each group, defined as those which maximise the number of discoveries whilst controlling the FDR (Ignatiadis et al., 2016). The method uniquely interprets the problem as a resource allocation task, treating FDR as the resource. It allocates low weights to covariate groups with weak evidence against the null, but also shifts weight away from covariate groups with very strong evidence against the null, as the null will be rejected for these hypotheses anyway. The motivation is that it is best to focus on covariate groups with intermediate evidence against the null in order to maximise the number of true discoveries.

The main drawback of these more sophisticated p-value weighting methods is the grouping of tests required to derive group-specific weights. The same weight is typically allocated to large groups of variables, meaning that the entire dynamic range of the auxiliary data may not be fully exploited.

1.6.4 FDR regression

Approaches have been developed that extend the two-class model in equation (1.23) to incorporate covariate information, x_i, so that

$$z_i \sim p_0(x_i)f_0(z_i) + (1 - p_0(x_i))f_1(z_i) \tag{1.27}$$

(Scott et al., 2015). In a range of methods relating to “FDR regression”, linear or logistic regression frameworks are used to model the relationship between covariates and the prior probability that a given observation is a signal (Boca and Leek, 2018; Scott et al., 2015).

The concepts of FDR regression were first introduced by Scott et al. (2015) and were motivated by an application to detect interactions between pairs of neurons that fired in synchrony, using data measured in the primary visual cortex of a rhesus macaque monkey. The motivation was that prior knowledge, for example that neurons are most likely to interact if they are spatially close together, could be used to better inform the analysis.

Recall that in the absence of covariates, the local FDR is the posterior probability that a test is null given the test statistic (equation 1.24). Estimating this quantity requires the estimation of either $f_0(z_i)$ and $f(z_i)$, or their ratio. That is, the local FDR can be estimated without the explicit deconvolution of the mixture, meaning that $f_1(z_i)$ (which is difficult to estimate since the true non-nulls are unknown) does not need to be directly estimated. In contrast, the local FDR when incorporating covariates requires the direct estimation of $f_1(z_i)$, because $f(z_i)$ is no longer a common mixture distribution (since each z_i has its own associated $p_0(x_i)$). The posterior probability of interest is therefore:

$$Pr(H_0|Z_i = z_i, x_i) = \frac{p_0(x_i)f_0(z_i)}{p_0(x_i)f_0(z_i) + (1 - p_0(x_i))f_1(z_i)}. \tag{1.28}$$

Consequently, Scott et al. (2015) utilised parametric assumptions to estimate $f_0(z)$ and $f_1(z)$. Specifically, they assumed Gaussian distributions for the test statistics:

$$\begin{aligned} f_0(z) &= N(z|\mu, \sigma^2) \\ f_1(z) &= \int N(z|\mu + \theta, \sigma^2)\pi(\theta)d\theta, \end{aligned} \tag{1.29}$$

where μ and σ^2 were either assumed to be known or were estimated using existing approaches that are described in Efron (2004) and Martin and Tokdar (2012). In their empirical Bayes approach, Scott et al. (2015) estimated $\pi(\theta)$ using a predictive recursion approach (Newton, 2002), ignoring the covariate values.

In the approach, all that remains is the estimation of $p_0(x_i)$ in equation (1.28). For this, Scott et al. (2015) introduced binary latent variables for whether each test came from the null or alternative hypotheses. They used an expectation-maximisation (EM) algorithm to estimate the coefficients of a linear predictor using user-specific covariates as the regressors. Scott et al. (2015) also proposed a fully Bayes approach, where a Markov chain Monte Carlo (MCMC) algorithm was used to draw iteratively from three complete conditional distributions: for the mixing density $\pi(\theta)$, for the latent binary variables h_i and for the regression coefficients β.

Boca and Leek (2018) described a new approach to FDR regression that no longer assumes that the test statistics are Gaussian. Their approach, Boca and Leek's FDR regression (BL), extends the estimate of $\hat{\pi}_0(\lambda)$ to include covariates and ultimately derives $F\hat{D}R(x_i)$ values by multiplying the BH adjusted p-values by this estimate.

The approach begins by defining a binary indicator, $Y_i = 1(P_i > \lambda)$ such that $\#\{p_i > \lambda; i = 1, ..., m\} = \sum_{i=1}^{m} Y_i$ and so that

$$\begin{aligned} \hat{\pi}_0(\lambda) &= \frac{\sum_{i=1}^{m} Y_i}{m(1-\lambda)} \\ &= \frac{E(Y_i)}{1-\lambda}. \end{aligned} \tag{1.30}$$

This is then be extended to incorporate covariates:

$$\hat{\pi}_0^{(\lambda)}(x_i) = \frac{E(Y_i|X_i = x_i)}{1-\lambda}. \tag{1.31}$$

Boca and Leek (2018) used logistic regression to estimate $E(Y_i|X_i = x_i)$ and subsequently used this as a plug-in estimate in equation (1.31). The value for $\hat{\pi}_0^{(\lambda)}(x_i)$ was smoothed over a series of thresholds $\lambda \in (0, 1)$ to obtain $\hat{\pi}_0(x_i)$. Finally, $F\hat{D}R(x_i)$ was obtained by multiplying the BH adjusted p-values by $\hat{\pi}_0(x_i)$.

In addition to allowing for non-Gaussian distributed test statistics, BL has also been shown to outperform the approach by Scott et al. (2015) in terms of consistency (of results across scenarios) and FDR control by an independent research group (Korthauer et al., 2019).

1.6.5 Methods specifically developed in the context of GWAS

As well as substantial research in the statistical literature, covariate-informed multiple testing methods have also been extensively researched specifically in the context of GWAS (Darnell et al., 2012; Eskin, 2008; Hou and Zhao, 2013; Kichaev et al., 2019; Lu et al., 2016b; Pickrell, 2014; Roeder et al., 2007; Sveinbjornsson et al., 2016; Wen et al., 2016). For example, Boca and Leek (2018) applied their approach to GWAS data on BMI from the Genetic Investigation of ANthropometric Traits (GIANT) consortium, using information on sample size and allele frequencies as covariates. In this section I describe three covariate-informed multiple testing methods that have been developed specifically in the context of GWAS.

GenoWAP

GenoWAP (Lu et al., 2016b) is a Bayesian approach that derives posterior scores of disease-specific SNP functionality using GWAS p-values and specific covariate data. The method uses GenoCanyon scores as covariates, which aim to infer the functional potential of each position in the human genome (Lu et al., 2015). These are derived from the union of 22 computational and experimental annotations broadly falling into conservation measure, open chromatin, histone modification and transcription factor binding site categories in different cell types.

The authors define Z_D as a binary indicator of disease-specific functionality ("1 if this SNP or its surrounding region is involved in the disease pathway") and Z as a binary indicator of general functionality ("1 if this SNP or its surrounding region is active in any functional pathway"). The quantity of interest is

$$Pr(Z_D = 1|p) = \frac{f(p|Z_D = 1) \times Pr(Z_D = 1)}{f(p|Z_D = 0) \times Pr(Z_D = 0) + f(p|Z_D = 1) \times Pr(Z_D = 1)}. \tag{1.32}$$

The authors state that $\{Z_D\} \subseteq \{Z\}$ and use this to write:

$$Pr(Z_D = 1|p) = \frac{f(p|Z_D = 1) \times Pr(Z_D = 1|Z = 1)Pr(Z = 1)}{f(p|Z_D = 0) \times Pr(Z_D = 0) + f(p|Z_D = 1) \times Pr(Z_D = 1|Z = 1)Pr(Z = 1)}. \tag{1.33}$$

In the method, $Pr(Z = 1)$ is estimated by the mean GenoCanyon score of the surrounding 10-kbp. To estimate the remaining quantities, SNPs are partitioned into functional ($Z = 1$) or non-functional ($Z = 0$) subsets based on their GenoCanyon score. A threshold of 0.1 is generally used to distinguish between functional and non-functional SNPs, although there is little qualification for the use of this thresholding value. A histogram (with the number of bins chosen through cross-validation) is used to estimate $f(p|Z_D = 0)$, whilst an EM algorithm is used to estimate the remaining quantities.

FINDOR

FINDOR (Kichaev et al., 2019) uses the weighted Bonferroni procedure to derive adjusted GWAS p-values using specific covariate information on SNP functionality. In the original manuscript, the authors showed that FINDOR was generally less powerful, but superior in terms of false positive findings, to stratified FDR, grouped BH and IHW methods.

The approach groups SNPs based on how well they tag functional categories that are enriched for heritability (Finucane et al., 2015; Gazal et al., 2017) and derives group-specific weights for use in a weighted Bonferroni procedure (Genovese et al., 2006). These weights are proportional to the ratio of the estimated proportion of alternative to null SNPs in each group (Hu et al., 2010; Storey and Tibshirani, 2003). The FINDOR methodology is similar to that of the grouped BH procedure, but includes an additional step whereby the weights are normalised to average 1. This normalisation step is significant because Roeder et al. (2007) demonstrated that using a data-dependent weighting scheme with weights that average to 1 preserves control of type I error with high probability if the number of weights learned is significantly less than the number of hypothesis test performed (Kichaev et al., 2019).

Whilst grouping-based approaches are satisfactory when they capture all information provided by the covariate, as is possible in the case of categorical covariates, these approaches are limited in the case of continuous covariates or more complex multi-dimensional covariate spaces. That is, subjective thresholding and coarse binning is generally required, meaning that the entire dynamic range of the auxiliary data is often not fully exploited. For example, applying FINDOR to Biobank style data with the recommended 100 bins results in bins containing approximately 100K SNPs within which covariate values will vary.

Conditional FDR

The conditional FDR (cFDR) approach, developed and applied by Andreassen and colleagues (Andreassen et al., 2013, 2014a,b,c, 2015), is a natural extension to the Bayesian FDR in the presence of auxiliary covariates. Given a set of p-values relating to m variables for trait 1 $(p_1, ..., p_m)$, and an additional set of p-values relating to the same m variables for trait 2 $(q_1, ..., q_m)$, the Bayesian FDR can be extended to the conditional Bayesian FDR (cFDR).

Assuming that p_i and q_i (for $i = 1, ..., m$) are independent and identically distributed (iid) realisations of random variables P, Q satisfying:

$$\begin{aligned} P|H_0^p &\sim U(0,1) \\ P &\perp\!\!\!\perp Q|H_0^p, \end{aligned} \tag{1.34}$$

then the cFDR is defined as the probability that the null hypothesis for trait 1 (H_0^p) is true at a random SNP given that the observed p-values at that SNP are less than or equal to p in trait 1 and q in trait 2,

$$cFDR(p,q) = Pr(H_0^p|P \leq p, Q \leq q) \tag{1.35}$$

(Andreassen et al., 2013, 2014a). The cFDR approach was originally developed to increase power for GWAS discovery in a principal trait (trait 1) by leveraging GWAS test statistics for a related trait (trait 2), and is described in more detail in chapter 2 of this thesis.

1.7 Immune-mediated diseases

The methods presented in this thesis are applied to a variety of data sets relating to immune-mediated diseases (IMDs). Consequently, in this section I introduce the concepts relevant to these applications and also provide details for the specific diseases that feature in this thesis.

1.7.1 Background

IMDs are conditions which result from abnormal activity of the body's immune system (Adapa, 2011). In this thesis, I use IMD as an umbrella term for diseases with underlying immune-mediated pathogenesis, which includes “autoimmune” and “autoinflammatory” diseases.

Autoimmune diseases occur when an individual's immune system recognises "self" cells as foreign, which results in aberrant immune responses. Autoinflammatory diseases are clinical disorders marked by abnormally increased inflammation (Ciccarelli et al., 2013).

1.7.2 Innate and adaptive immunity

The human immune system is formed of two parts: the innate immune system and the adaptive immune system.

The latin root of innate is "existing from birth" and accordingly, the innate immune system is present from birth and refers to the non-specific first line of defence against a pathogen. Components of the innate immune system range from physical barriers preventing pathogen entry (such as skin and mucosa) to defence mechanisms (such as secretions like bile and tears) to general immune responses. The immune cells involved in the innate immune system are generally leukocytes (white blood cells), including phagocytes such as macrophages that engulf and destroy pathogens, and mast cells that secrete small proteins involved in cell signalling, called "cytokines", to initiate an inflammatory cascade.

The adaptive immune response is slower and more specific than the innate immune response, arose in evolution less than 500 million years ago and is confined to vertebrates (Alberts et al., 2002). The adaptive immune system primarily involves two types of lymphocytes (a type of white blood cell): B cells (primarily involved in "humoral immunity") and T cells (primarily involved in "cell-mediated immunity").

B cells possess membrane-bound antibodies that are specific to a particular antigen. When a B cell encounters its specific antigen it divides to become a "memory B cell" or an "effector B cell". A memory B cell is a clone of its "parent B cell" and can circulate in the body for many years waiting to encounter the same antigen, which would trigger an accelerated secondary immune response. An effector B cell (also called a "plasma cell") secretes large amounts of specific antibodies, which are transported by the blood plasma to the site of the target antigen.

T cells mature in the thymus and express both a T cell receptor and either CD4 ("CD4 T cells") or CD8 ("CD8 T cells") proteins. CD4 T cells are generally referred to as "helper T (Th) cells" and assist other components of the immune system, whilst CD8 T cells are called "cytotoxic T cells" and directly remove pathogens and kill infected host cells. Whilst antibodies are able to bind to antigens directly, T cell receptors can only recognise antigens when they are bound to certain receptor molecules, called Major Histocompatibility Complex (MHC) class 1 and class 2. Only a subset of the cells that are involved in the immune response express the MHC receptors on their surface, and T cells therefore rely on these "antigen-presenting cells" (which include dendritic cells and macrophages) to initiate a response.

1.7.3 Major histocompatibility complex

The MHC in humans (also called the human leukocyte antigen (HLA) system) is a 4.6-Mb region located on the short arm of chromosome 6 that encodes the MHC proteins found on the surface of antigen-presenting cells. Genetic variations within the MHC have been found to associate with almost every IMD, but it is extremely difficult to unravel these associations due to the extensive LD in the region (Fernando et al., 2008). Moreover, the MHC is extremely gene-rich and the genes themselves are extremely variable (for example, *HLA-B* is the most polymorphic gene known in the human genome) (Fernando et al., 2008) rendering it difficult to infer holistic biological meaning from the genetic variations in this region. For example, an association between type 1 diabetes and genetic variants in HLA class II was first identified almost half a century ago (Platz et al., 1981; Svejgaard et al., 1983; Thomsen et al., 1975), but it was only in the past few months that the biological mechanisms underpinning this association were unravelled (García et al., 2021).

Genotyping in the HLA region is complex and expensive (due to the large variability in the region) and large sample sizes are required in genetic association testing to capture differences in allele frequencies between case and control samples for the multi-allelic loci in this region. The MHC region of the genome is generally either excluded from genetic association studies or analysed independently from the rest of the genome. In this thesis I exclude the MHC region from my analyses due to its complex LD structure.

1.7.4 Asthma

Asthma is a complex disease that often starts in childhood and affects 1 in 12 adults in the UK (https://www.asthma.org.uk/about/media/facts-and-statistics/). It is a condition in which the airways of the lungs constrict in response to certain stimuli, such that following the inhalation of pathogens or particles, patients exhibit wheezing, coughing and tightness in the chest. The two cells in the airways of the lungs that are implicated in asthma pathogenesis are the epithelial cells, which initiate airway inflammation and produce mucus leading to airway obstruction, and smooth muscle cells, which contract to cause airway obstruction (Erle and Sheppard, 2014). Asthma is caused by an overactive immune response and so is classified as an IMD in this thesis. GWAS have identified 212 susceptibility loci for asthma so far (Han et al., 2020), indicating a strong genetic component. However these loci only explain approximately $8 - 9\%$ of the total heritability of asthma, suggesting that other biological phenomena, for example pertaining to low effect variants or complex interactions between genes, may be relevant in asthma pathogenesis.

1.7.5 Ankylosing spondylitis

Ankylosing spondylitis is a complex disease that often starts in teenagers and young adults and is twice as common in men than women (https://www.nhs.uk/conditions/ankylosing-spondylitis/). It is a chronic inflammatory form of arthritis that predominantly affects the spine joints, causing pain and discomfort. In severe cases, it can progress to spinal fusion, in which the vertebrae fuse together leading to back-pain, hunched-posture and inflexibility.

In 1973, the first genetic association for ankylosing spondylitis was found in the *HLA-B27* gene (Brewerton et al., 1973; Caffrey and James, 1973; Schlosstein et al., 1973). In 2013, the International Genetics of Ankylosing Spondylitis Consortium (IGAS) genotyped 10,619 ankylosing spondylitis cases and 15,145 control samples using the Immunochip and identified 24 associated loci, of which 13 were novel (International Genetics of Ankylosing Spondylitis Consortium (IGAS), 2013). More recently, a combined analysis of Immunochip data from five diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, and ulcerative colitis) identified 113 ankylosing spondylitis-associated genome-wide significant variants (Ellinghaus et al., 2016) which contribute roughly 10% of the heritability of ankylosing spondylitis (Costantino et al., 2018; Hanson and Brown, 2017). Whilst the pathogenesis of ankylosing spondylitis is still largely unclear, it is thought to be predominantly a genetic disease with *HLA-B27* remaining the most strongly associated gene (Tam et al., 2010).

1.7.6 Type 1 diabetes

Type 1 diabetes (T1D) is a complex disease that often starts in childhood and is estimated to affect 400,000 people in the UK (https://jdrf.org.uk/information-support/about-type-1-diabetes/facts-and-figures/). It is an autoimmune disease characterised by the destruction of the insulin producing beta cells in the pancreatic islet tissue. T1D patients are therefore unable to produce insulin, a hormone which promotes the absorption of glucose by cells. T1D patients rely on insulin-replacement treatment (such as insulin injections) to control their blood glucose levels. T1D has both short-term and long-term risks. In the short term, hypoglycemia (low blood glucose) is usually rapid and is fatal if untreated, whilst in the long-term hyperglycemia (high blood glucose) can lead to conditions such as cardiomyopathy, renal failure and retinopathy.

The first genomic loci to be associated with T1D was the HLA locus in 1973, for which the effect size was so large that the association was uncovered by genotyping only 50 T1D cases and 233 controls (Cudworth and Woodrow, 1975; Singal and Blajchman, 1973). By 2010, 41 genomic loci had been associated with T1D (Barrett et al., 2009) but the sparse genotyping of variants within these loci and the lack of robust imputation prohibited fine-mapping to elucidate the true causal variants. In 2015, Onengut-Gumuscu et al. (2015) identified 44 genomic loci that associated with T1D and fine-mapped each of these loci to find credible sets of putative causal variants. The most recent GWAS for T1D identified 78 susceptibility loci, of which 36 were novel (Robertson et al., 2021).

1.8 Thesis outline and contributions

The next two chapters of this thesis focus on the development of statistical methods that aim to build a better understanding of the genetic basis of complex human diseases. In chapter 2 I describe a novel methodology based on the cFDR to leverage functional genomic data with statistics from genetic association studies to boost power for GWAS discovery. I include applications to asthma and T1D, where I leverage a variety of functional genomic data to uncover new regions of the genome that associate with each of the diseases.

In chapter 3 I examine results from statistical fine-mapping and describe a novel methodology to improve the accuracy of inferences from fine-mapping analyses. This enables efficient allocation of resources in the expensive functional follow-up studies conducted on the prioritised causal variants. I include applications to ankylosing spondylitis and T1D, where my method improves the resolution of inferences from fine-mapping without using any additional data. The statistical methods described in chapters 2 and 3 are accompanied by user-oriented software and webpages to enable their widespread use in the scientific community.

In chapter 4 I then take a more focussed look at genetic data for T1D in relation to the Ikaros family of transcription factors, which have been linked to T1D for some time but have not yet been studied systematically with regard to their role in T1D pathogenesis. Recognising the relative gains of additionally exploiting functional genomic data to build on biological understanding, I explore various statistical approaches uniting these data with genetic data in an attempt to decipher biological mechanisms pertaining to Ikaros that underpin T1D pathogenesis. I conclude by summarising the contributions of this thesis and outline research directions for the future.

During the thesis I have contributed to a number of publications which are referenced in each relevant chapter. For completeness, I also list them here:

- Anna Hutchinson, Hope Watson, and Chris Wallace. Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLOS Computational Biology, 16(4):e1007829, April 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007829. **This paper relates to the methodology and application to T1D that are described in chapter 3 of this thesis. I contributed to the conceptualisation, data curation, formal analysis, investigation, methodology, resources, software, validation and visualisation, and I also wrote the original draft of the manuscript.**

- Christophe Bourges, Abigail F Groff, Oliver S Burren, Chiara Gerhardinger, Kaia Mattioli, Anna Hutchinson, Theodore Hu, Tanmay Anand, Madeline W Epping, Chris Wallace, Kenneth GC Smith, John L Rinn, and James C Lee. Resolving mechanisms of immune-mediated disease in primary CD4 T cells. EMBO Molecular Medicine, 12(5):e12112, May 2020. ISSN 1757-4676. doi: 10.15252/emmm.202012112. **This paper relates to the application to ankylosing spondylitis in chapter 3 of this thesis. I performed the analysis described in the "Fine-mapping ankylosing spondylitis (AS) association on 2p15" section of the manuscript.**

- Anna Hutchinson, Jennifer Asimit, and Chris Wallace. Fine-mapping genetic associations. Human Molecular Genetics, 29(R1):R81–R88, September 2020. ISSN 0964-6906. doi: 10.1093/hmg/ddaa148. **This review paper describes key concepts, methodologies and future opportunities in statistical fine-mapping and relates to chapters 3 and 4 of this thesis. I wrote the original draft of the manuscript.**

- Anna Hutchinson, Guillermo Reales, Thomas Willis, and Chris Wallace. Leveraging auxiliary data from arbitrary distributions to boost GWAS discovery with Flexible cFDR. bioRxiv, April 2021. doi: 10.1101/2020.12.04.411710. (Editorially accepted in PLOS Genetics) **This paper relates to the Flexible cFDR methodology and applications to asthma that are described in chapter 2 of this thesis. I contributed to the conceptualisation, data curation, formal analysis, investigation, methodology, resources, software, validation and visualisation, and I also wrote the original draft of the manuscript.**

- Anna Hutchinson, James Liley and Chris Wallace. fcfdr: an R package to leverage auxiliary data with GWAS test statistics using Flexible cFDR. (In preparation) **This paper relates to the Binary cFDR methodology, application to T1D and accompanying software that are described in chapter 2 of this thesis. I wrote the original draft of the manuscript.**

Chapter 2

Leveraging auxiliary functional data to boost GWAS discovery

This chapter introduces the conditional false discovery rate (cFDR) which was originally developed to leverage GWAS test statistics from related traits to boost statistical power for GWAS discovery. I describe two new approaches based on the cFDR, called “Flexible cFDR” and “Binary cFDR”, which extend the practicability of cFDR from leveraging only GWAS statistics to leveraging a wider variety of auxiliary data types. My approach directly enables the wealth of functional genomic data to be leveraged with GWAS test statistics in a statistically robust framework, which I show increases sensitivity whilst controlling the frequentist FDR. This is an edited version of Hutchinson et al. (2021) and there is significant textual overlap in sections 2.1 to 2.3, 2.6 to 2.9 and 2.11. I wrote the first draft of the manuscript and the text reused is my own work.

2.1 Introduction

A stringent significance threshold is required to identify robust associations in GWAS due to multiple testing constraints. The statistical power to detect associations that exceed this threshold can be increased by leveraging relevant auxiliary covariates. For example, pervasive pleiotropy throughout the genome (Sivakumaran et al., 2011) suggests that leveraging GWAS test statistics for related traits may be beneficial, whilst the non-random distribution of trait-associated SNPs across various functional categories (Schork et al., 2013) suggests that incorporating functional genomic data may also be useful. In fact, the expansive range of relevant auxiliary covariates has accumulated in a wealth of covariate-informed multiple testing methods which leverage auxiliary covariates (e.g. SNP-level data) with test statistics for variables (e.g. GWAS p-values for SNPs) to increase statistical power. These methods have been extensively researched both in the statistical literature (Basu et al., 2018; Bourgon et al., 2010; Cai et al., 2019; Ferkingstad et al., 2008; Genovese et al., 2006; Hu et al., 2010; Ignatiadis et al., 2016; Lei and Fithian, 2018; Li and Barber, 2017; Roeder and Wasserman, 2009; Sun et al., 2006) and specifically in the context of GWAS (Darnell et al., 2012; Eskin, 2008; Hou and Zhao, 2013; Kichaev et al., 2019; Lu et al., 2016b; Pickrell, 2014; Roeder et al., 2007; Sveinbjornsson et al., 2016; Wen et al., 2016) with a consistent aim of minimising type II errors (or equivalently increasing statistical power) whilst controlling some appropriate error rate, such as the FDR.

The cFDR approach is a natural extension to the FDR in the presence of auxiliary covariates. This intuitive approach mitigates many of the shortcomings of comparator methods: it does not bin variables and thus makes full use of the dynamic range of covariate values, it does not include any subjective thresholding and does not require the definition of a normalised weighting scheme. However it was designed for a very specific setting, that is to increase GWAS discovery (in the "principal trait") by leveraging GWAS test statistics from a genetically related ("auxiliary") trait.

I was interested to see whether a more general form of cFDR could address the same covariate-informed multiple testing problems as the range of methods cited above. In this chapter, I describe "Flexible cFDR" and "Binary cFDR", which are new cFDR frameworks that enjoy all of the benefits of the conventional cFDR approach but support a wider range of auxiliary data types. My computationally efficient frameworks extend the usage of cFDR beyond only GWAS to the accelerating field of functional genomics, and can be applied iteratively to incorporate additional layers of auxiliary information.

2.2 cFDR estimator

I begin by restating the definition and empirical estimator of the cFDR. Consider p-values for m SNPs, denoted by $p_1, ..., p_m$, corresponding to the null hypotheses of no association between the SNP and a principal trait (denoted by H_0^p). Let $p_1, ..., p_m$ be realisations from the random variable P. The Bayesian FDR is defined as the probability that the null hypothesis is true for a random SNP in a set of SNPs with $P \leq p$:

$$FDR(p) = Pr(H_0^p | P \leq p). \tag{2.1}$$

This Bayesian definition of a tail area FDR (Efron, 2007) is asymptotically equivalent (Wen, 2018) to the frequentist interpretation of the FDR introduced by Benjamini and Hochberg (1995), which is the expected fraction of false discoveries amongst all discoveries.

Given additional p-values, $q_1, ..., q_m$, for the same m SNPs for an "auxiliary trait", the Bayesian FDR can be extended to the cFDR by conditioning on both the principal and the auxiliary trait variables (in contrast to the standard FDR which conditions only on the principal trait variable). Assuming that p_i and q_i (for $i = 1, ..., m$) are realisations of random variables P, Q satisfying

$$\begin{aligned} P|H_0^p &\sim U(0,1) \\ P &\perp\!\!\!\perp Q|H_0^p, \end{aligned} \tag{2.2}$$

then the cFDR is defined as the probability that H_0^p is true at a random SNP given that the observed p-values at that SNP are less than or equal to p in the principal trait and q in the auxiliary trait (Andreassen et al., 2013). Using Bayes theorem,

$$\begin{aligned} cFDR(p,q) &= Pr(H_0^p|P \leq p, Q \leq q) \\ &= \frac{Pr(P \leq p|H_0^p, Q \leq q) \times Pr(H_0^p|Q \leq q)}{Pr(P \leq p|Q \leq q)}. \end{aligned} \tag{2.3}$$

The cFDR framework implicitly assumes that there is a "positive stochastic monotonic relationship" between p and q, meaning that on average SNPs with smaller p-values in the auxiliary trait are enriched for smaller p-values in the principal trait. This assumption is naturally satisfied in the typical use-case of cFDR that leverages p-values for genetically related traits.

Using Bayes theorem and standard conditional probability rules,

$$\begin{aligned} cFDR(p,q) &= \frac{Pr(P \leq p|H_0^p, Q \leq q) \times \dfrac{Pr(Q \leq q|H_0^p)Pr(H_0^p)}{Pr(Q \leq q)}}{\dfrac{Pr(P \leq p, Q \leq q)}{Pr(Q \leq q)}} \\ &= \frac{Pr(P \leq p|H_0^p, Q \leq q) \times Pr(Q \leq q|H_0^p) \times Pr(H_0^p)}{Pr(P \leq p, Q \leq q)} \end{aligned} \tag{2.4}$$

(Liley and Wallace, 2015).

It is conventional in the cFDR literature to conservatively approximate $Pr(H_0^p) \approx 1$, and this is reasonable in the GWAS setting as the proportion of true signals is expected to be very low (this may be debatable as sample sizes increase, but it is still appropriate in terms of being conservative). Given the assumptions in property (2.2), we can also approximate $Pr(P \leq p|H_0^p, Q \leq q) \approx p$, noting that this is an equality if p is correctly calibrated. The estimate of cFDR is therefore:

$$c\widehat{FDR}(p,q) = \frac{p \times Pr(Q \leq q|H_0^p)}{Pr(P \leq p, Q \leq q)}. \tag{2.5}$$

Existing methods use empirical CDFs to estimate $Pr(Q \leq q|H_0^p)$ and $Pr(P \leq p, Q \leq q)$, and typically use hard thresholding to approximate $Pr(Q \leq q|H_0^p) \approx Pr(Q \leq q|P > 0.5)$ (Liley and Wallace, 2021).

Having derived $c\widehat{FDR}$ values for each p-value-covariate pair, a simple rejection rule would be to reject $H_0^p(i)$ for any $c\widehat{FDR}(p_i, q_i) \leq \alpha$, for $0 < \alpha < 1$. Andreassen et al. (2013) used the decision rule:

$$\text{Reject } H_0^p \text{ if: } \exists\, p' \geq p_i : c\widehat{FDR}(p', q_i) \leq \alpha \tag{2.6}$$

which closely follows the BH procedure. Yet unlike the BH procedure, this rejection rule does not control frequentist FDR at α (Liley and Wallace, 2015). Liley and Wallace (2021) described a method to transform the cFDR estimates into "v-values", which are analogous to p-values and can be used to control FDR (e.g. in the BH procedure). However, this approach is currently only suited to instances where the auxiliary data may be modelled using a mixture of centred Gaussian distributions (i.e. by transforming auxiliary p-values to Z scores; $q := -\phi^{-1}(\frac{q}{2})$).

2.3 Flexible cFDR

The cFDR estimator in equation (2.5) holds in the more general setting where $q_1, ..., q_m$ are real continuous values from some arbitrary distribution that is positively stochastically monotonic in p. However, the current methods to estimate the cFDR use empirical CDF estimates which may be inaccurate for data from arbitrary distributions, suggesting that an alternative approach may be required to estimate the cFDR in these instances. For example, empirical CDFs are typically inaccurate in sparse data regions (because they are step functions), but sparse data regions are likely to be found more often in unbounded auxiliary data from arbitrary distributions (for example near the extreme observations) than auxiliary data that are p-values (and are thus bounded by $[0, 1]$). Moreover, the method used to control the frequentist FDR (Liley and Wallace, 2021) is only suitable for auxiliary data that can be modelled using a mixture of centred Gaussian distributions, rendering it unsuitable for auxiliary data from arbitrary distributions. I consequently describe a new, more versatile cFDR framework for data pairs consisting of p-values for the principal trait (p) and continuous covariates from more general distributions (q). I call my method "Flexible cFDR" and show that it is naturally suited to leveraging functional genomic data, which is not typically Gaussian.

2.3.1 Estimator

To estimate both $Pr(Q \leq q|H_0^p)$ and $Pr(P \leq p, Q \leq q)$ in equation (2.5) I first fit a bivariate kernel density estimate (KDE) using a Gaussian kernel. To do this, I transform the p-values for the principal trait (derived from a two-tailed test, as is typical in GWAS) to absolute Z-scores (Z_p; because the sign of the associated Z-scores are essentially arbitrary as they depend on which allele is designated "effect"). Since the absolute Z-scores are bounded by 0, the KDE will penalise for the lack of negative data points and may underestimate the true density in regions close to 0. To avoid this boundary effect, I mirror the absolute Z-scores onto the negative real line together with their associated Q values, but only estimate the KDE for the non-negative part of the data, akin to the "reflection technique" described by Silverman (1986). I consequently model the PDF corresponding to Z_p, Q in the usual way as

$$f(x, y) = \frac{1}{m} \sum_i \frac{1}{\sigma_p \sigma_q} \phi \left(\sqrt{\left(\frac{x - \{-\phi^{-1}(\frac{p_i}{2})\}}{\sigma_p} \right)^2 + \left(\frac{y - q_i}{\sigma_q} \right)^2} \right), \tag{2.7}$$

where ϕ is the standard normal density and the values σ_p and σ_q are the bandwidths determined using a well-supported rule-of-thumb (Venables and Ripley, 2002) which assumes independent samples. Consequently, I fit the KDE to a subset of independent SNPs in the data set, which can be readily found using a variety of software packages including LDAK (Speed et al., 2020) and `PLINK` (Chang et al., 2015) (as described in section 2.3.4). I then integrate over P and Q to estimate $Pr(P \leq p, Q \leq q)$.

Earlier cFDR methods use hard thresholding to approximate the distribution of $Q|H_0^p$ by $Q|P > 1/2$ (Liley and Wallace, 2021). In Flexible cFDR I empirically estimate this distribution utilising local FDRs, which estimate $Pr(H_0^p|P = p)$ (Efron, 2004). I approximate $Pr(H_0^p|P = p, Q = q) \approx Pr(H_0^p|P = p)$ assuming that the majority of information about H_0^p is contained in P so that

$$\begin{aligned} Pr(P = p, Q = q, H_0^p) &= Pr(H_0^p|P = p, Q = q) \times Pr(P = p, Q = q) \\ &\approx Pr(H_0^p|P = p) \times Pr(P = p, Q = q), \end{aligned} \tag{2.8}$$

where $Pr(P = p, Q = q)$ is estimated from the bivariate KDE and $Pr(H_0^p|P = p)$ is estimated using the local FDR. In order to avoid boundary effects, I mirror the absolute Z-scores onto the negative real line and extract only the local FDR values for the non-negative part of the data, utilising the `locfdr` R package (https://cran.r-project.org/web/packages/locfdr/index.html) to do this.

From equation (2.8), $Pr(Q = q, H_0^p)$ is estimated by integrating over P, and $Pr(H_0^p)$ is then estimated by integrating over Q to obtain

$$\widehat{Pr}(Q \leq q | H_0^p) = \frac{\widehat{Pr}(Q \leq q, H_0^p)}{\widehat{Pr}(H_0^p)}, \tag{2.9}$$

where I use $\widehat{}$ to denote that these are estimates under the assumption $H_0^p \perp\!\!\!\perp Q|P$.

My final cFDR estimator is therefore:

$$c\widehat{FDR}(p, q) = \frac{p \times \widehat{Pr}(Q \leq q | H_0^p)}{\int_{-\infty}^{q} \int_{z_p}^{\infty} f(x, y) dx dy}. \tag{2.10}$$

where z_p is the Z-score associated with p.

As in the conventional cFDR approach, my estimator implicitly assumes a positive stochastic monotonic relationship between p and q. However, this is not guaranteed for the more general auxiliary data that can now be leveraged with Flexible cFDR. If instead this relationship is negative (such that low p-values are enriched for high values of q), then the sign of the auxiliary data values can simply be reversed and the method can proceed as usual.

2.3.2 Mapping p-value-covariate pairs to v-values

I describe a similar approach to that in Liley and Wallace (2021) to generate v-values, but remove the restrictive parametric assumptions that are placed on the auxiliary data.

Following Liley and Wallace (2021), I define "L-regions" as the set of points with $c\widehat{FDR} \leq \alpha$ and the "L-curve" as the rightmost border of the L-region, found through calculating $c\widehat{FDR}$ values for p, q pairs defined using a two-dimensional grid of p and q values. I find the L-curve for each observed p_i, q_i pair, which corresponds to the contour of estimated $c\widehat{FDR} = c\widehat{FDR}(p_i, q_i)$. I then define the L-region from this L-curve.

I derive v-values, which are essentially the probability of a newly-sampled realisation (p, q) of P, Q falling in the L-region under H_0^p. These are readily calculable by integrating the PDF of $P, Q|H_0^p$, which is denoted by $f_0(p, q)$, over the L-region:

$$v(p, q) = Pr((P, Q) \in L(p, q)|H_0^p) = \int_{L(p,q)} f_0(p, q) dp dq \tag{2.11}$$

(Liley and Wallace, 2021). In the original method, $f_0(p, q)$ is estimated using a mixture-Gaussian distribution, but to support auxiliary data from arbitrary distributions (where the only distributional constraint is that the data is positively stochastically monotonic in p) I utilise the assumptions in property (2.2) to write $f_0(p, q) = f_0^q(q)$ (since the PDF of p conditional on

H_0^p is the standard uniform density). I estimate $f_0^q(q)$ as an intermediate step in the derivation of $\widehat{Pr}(Q \leq q|H_0^p)$.

The v-value can be interpreted as the probability that a randomly-chosen (p, q) pair has an equal or more extreme $c\widehat{FDR}$ value than $c\widehat{FDR}(p_i, q_i)$ under H_0^p, and is thus analogous to a p-value. Theorem 3.1 in Liley and Wallace (2021) shows that the v-values are uniformly distributed under the null hypothesis for $X = (p_i, q_i) \in [0, 1]^2$, and this naturally holds for Flexible cFDR where $X = (p_i, q_i) \in [0, 1] \times [q_{low}, q_{high}]$ (where q_{low} and q_{high} are the lower and upper limits of the KDE support respectively).

Deriving v-values, which are analogous to p-values, means that the output from Flexible cFDR can be used directly in any conventional error rate controlling procedure, such as the BH procedure. The derivation of v-values also allow for iterative usage, whereby the v-values from the previous iteration are used as the "principal trait" p-values in the current iteration, thus allowing users to incorporate additional layers of auxiliary data into the analysis at each iteration, akin to leveraging multi-dimensional covariates.

2.3.3 Adapting to sparse data regions

To ensure that the integral of the KDE approximated in my method equals 1, I define the limits of its support to be 10% wider than the range of the data. This however introduces a sparsity problem, whereby the data that is required to fit the KDE in or near these regions is very sparse. Adaptive KDE methods that find larger value bandwidths for these sparser regions are computationally impractical for large GWAS data sets. Instead, I opt to use left censoring whereby all $q < q_{low}$ are set equal to q_{low} and the value for q_{low} is found by considering the number of data points required in a grid space to reliably estimate the density (Fig 2.1). Note that since my method utilises cumulative densities, the sparsity of data for extremely large q is not an issue.

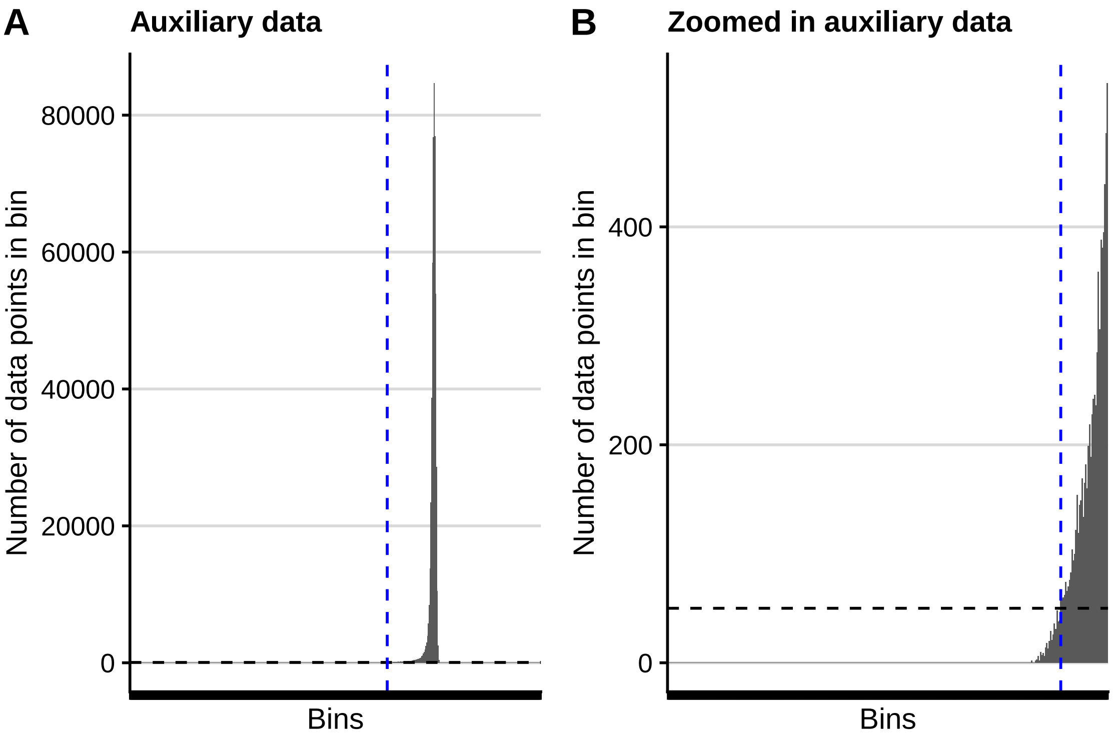


Fig. 2.1 Demonstration of the left censoring procedure. Plots showing how many data points are in each grid space (“bins”) of the auxiliary data over the support of the KDE for an example data set. (A) shows the full support of the KDE and (B) is zoomed in on the left tail. Black dashed line at $y = 50$ which is the default value of the `gridp` parameter in the `fcfdr::flexible_cfdr` function. Data points falling in grid spaces with fewer than 50 data points (those to the left of the blue dashed line) are left-censored, meaning that their value is replaced by the value of the left bound of the first grid space containing more than 50 data points. In practise, very few data points are left-censored in the approach.

Occasionally, in regions where (p, q) are jointly sparse, the v-value can appear extreme compared to the principal p-value. To avoid artifactually inflating evidence for association, I fit a spline to $log_{10}(v/p)$ against q and calculate the distance between each data point and the fitted spline, mapping the small number of outlying points back to the spline and recalculating the corresponding v-value as required (Fig 2.2).

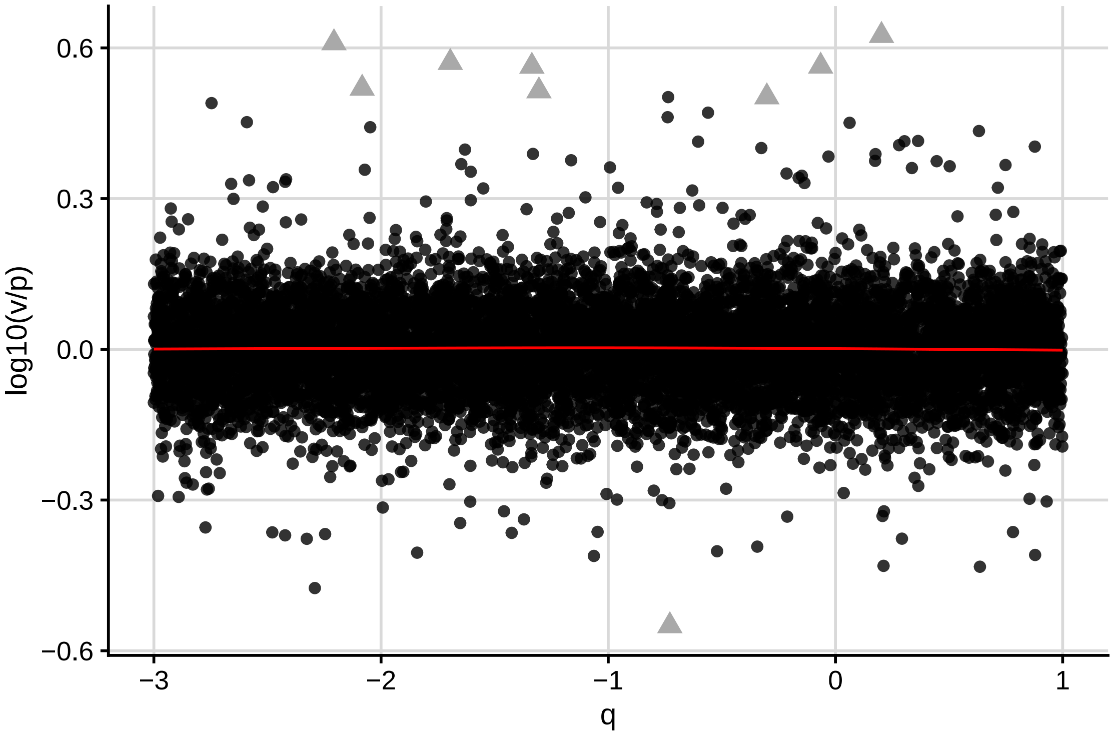


Fig. 2.2 Illustration of the spline correction procedure. A spline with 5 knots is fitted to $log_{10}(v/p)$ against q using the `bigsplines` R package (https://cran.r-project.org/web/packages/bigsplines/index.html) for an example data set. The distance between each data point and the fitted spline is calculated. If this distance is greater than the value of the `dist_thr` parameter in the `fcfdr::flexible_cfdr` function (default value is 0.5), then the data point is mapped back to the spline and the corresponding v-value is recalculated using the fitted spline. In this example, the red line shows the fitted spline and the triangular grey points are mapped back to the spline to generate new v-values.

2.3.4 Generating an independent subset of SNPs

As described earlier, the method that is used to determine the bandwidths for the KDE estimation in Flexible cFDR assumes that the samples are independent (Venables and Ripley, 2002). The method therefore requires an independent set of SNPs for the fitting of the KDE. In practise, I do this using the LDAK software (Speed et al., 2020), where I generate LDAK weights for each SNP and use the subset of SNPs with non-zero LDAK weights as the independent subset (an LDAK weight of zero means that the signal is (almost) perfectly captured by neighbouring SNPs). Other software, such as `PLINK`, can also be used to find an independent subset of SNPs.

My colleague, Thomas Willis, found that there was a confounding of LDAK weights and GWAS p-values by MAF, in that less common SNPs ($MAF < 0.05$) were over-represented among the independent subset and have, on average, larger p-values. In the method, I therefore down-sample the independent subset of SNPs to match the MAF distribution in this subset to that in the whole set of SNPs. This prevents a bias of the KDE fit towards the behaviour of rarer SNPs.

2.4 Binary cFDR

Conventional cFDR approaches, and also Flexible cFDR, do not support binary auxiliary data. This means that the cFDR approach cannot currently be used to leverage auxiliary data with a binary representation, such as whether SNPs are synonymous or non-synonymous or whether they reside in a coding region of the genome. Here, I describe a final extension to the cFDR framework for binary auxiliary data - "Binary cFDR".

2.4.1 Methods

This subsection is based on rough derivations by Dr James Liley (currently working at the MRC Human Genetics Unit, The University of Edinburgh), which I formalised into a precise statistical framework with his permission.

As before, let $p_1, ..., p_m \in (0, 1]$ be a set of p-values corresponding to the null hypotheses of no association between the SNP and the trait. Denote the null hypothesis by H_0^p and the alternative hypothesis by H_1^p. Now suppose that we also have a set of binary covariates, $q_1, ..., q_m \in \{0, 1\}$. As before, assume that (p_i, q_i) are realisations of random variables P, Q satisfying property (2.2).

Since all q are binary the support of P, Q is just two lines, and so the rejection regions (corresponding to L-regions) are of the form

$$L(p_0, p_1) = (P \leq p_0, Q = 0) \cup (P \leq p_1, Q = 1), \tag{2.12}$$

where p_0 and p_1 are unknown.

We wish to find v-values such that for all α,

$$\begin{aligned} &Pr(v_i < \alpha | H_0^p) = \alpha \\ &Pr(v_i < \alpha | H_1^p) \text{ is maximal.} \end{aligned} \tag{2.13}$$

That is, the v-values behave like p-value in that they are uniform under the null, but are also as small as possible under the alternative hypothesis. Appendix A.1 in Liley and Wallace (2021) (and also Du and Zhang (2014) and Alishahi et al. (2016), for example) show that this corresponds to rejection regions formed by the set of points for which $f_0(p, q)/f_1(p, q) < k(\alpha)$, for some k, where $f_0(p, q) = f(P = p, Q = q|H_0^p)$ and $f_1(p, q) = f(P = p, Q = q|H_1^p)$. That is,

p_0 and p_1 will satisfy the property

$$\frac{f_0(p_0,0)}{f_1(p_0,0)} = \frac{f_0(p_1,1)}{f_1(p_1,1)}. \tag{2.14}$$

Let

$$f(p,q) = f(P=p, Q=q) = \pi_0 f_0(p,q) + (1-\pi_0) f_1(p,q), \tag{2.15}$$

where $\pi_0 = Pr(H_0^p)$. Following on from equation (2.14), we have

$$\begin{aligned} f_0(p_0,0) f_1(p_1,1) &= f_0(p_1,1) f_1(p_0,0) \\ f_0(p_0,0)\frac{f(p_1,1)-\pi_0 f_0(p_1,1)}{1-\pi_0} &= f_0(p_1,1)\frac{f(p_0,0)-\pi_0 f_0(p_0,0)}{1-\pi_0} \\ \frac{f(p_1,1)-\pi_0 f_0(p_1,1)}{f_0(p_1,1)} &= \frac{f(p_0,0)-\pi_0 f_0(p_0,0)}{f_0(p_0,0)} \\ \frac{f(p_1,1)}{f_0(p_1,1)} &= \frac{f(p_0,0)}{f_0(p_0,0)}. \end{aligned} \tag{2.16}$$

I approximate

$$\begin{aligned} \frac{f_0(p_i,q_i)}{f(p_i,q_i)} &= \frac{Pr(P=p_i, Q=q_i|H_0^p)}{Pr(P=p_i, Q=q_i)} \\ &\approx \frac{Pr(P\leq p_i, Q=q_i|H_0^p)}{Pr(P\leq p_i, Q=q_i)} \\ &= \frac{Pr(P\leq p_i|Q=q_i, H_0^p)Pr(Q=q_i|H_0^p)}{Pr(P\leq p_i|Q=q_i)Pr(Q=q_i)} \\ &\approx \frac{p_i \times \widehat{Pr}(Q=q_i|H_0^p)}{|j: p_j\leq p_i, q_j=q_i|/m}, \end{aligned} \tag{2.17}$$

where $\widehat{Pr}(Q=q_i|H_0^p) = \dfrac{|j: q_j=q_i, p_j>1/2|}{|j: p_j>1/2|}$ and m is the total number of SNPs. Therefore, if $q_i=0$ then we can set $p_0=p_i$ and use this approximation to solve equation (2.14) for p_1. If $q_i=1$, then we can set $p_1=p_i$ and solve for p_0.

Specifically, if $q_i = 0$ then I set $p_0 = p_i$ and solve the following for p_1:

$$\begin{aligned}
\frac{p_i \times \frac{|j : q_j = 0, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_i, q_j = 0|/m} &= \frac{p_1 \times \frac{|j : q_j = 1, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_1, q_j = 1|/m} \\
\frac{p_i \times \frac{|j : q_j = 0, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_i, q_j = 0| \times \frac{|j : q_j = 1, p_j > 1/2|}{|j : p_j > 1/2|}} &= \frac{p_1}{|j : p_j \leq p_1, q_j = 1|}.
\end{aligned} \tag{2.18}$$

In practise, I do this using a fold-removal protocol for estimation to ensure that rejection rules are not applied to the same data on which those rules were determined. This is required since the method utilises empirical CDFs and including an observation when estimating its own L-curve causes the curve to deviate around the observed point (Liley and Wallace, 2021). I either leave out each chromosome or each LD block in turn, and use the remaining SNPs to estimate the values for the held out SNPs. This also allows me to calculate an approximation to g_0^{-1} for each fold, where

$$g_0(x) = \frac{x}{|j : p_j \leq x, q_j = 0|}. \tag{2.19}$$

Similarly, if $q_i = 1$ then I set $p_1 = p_i$ and solve the following for p_0:

$$\begin{aligned}
\frac{p_0 \times \frac{|j : q_j = 0, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_0, q_j = 0|/m} &= \frac{p_i \times \frac{|j : q_j = 1, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_i, q_j = 1|/m} \\
\frac{p_0}{|j : p_j \leq p_0, q_j = 0|} &= \frac{p_i \times \frac{|j : q_j = 1, p_j > 1/2|}{|j : p_j > 1/2|}}{|j : p_j \leq p_i, q_j = 1| \times \frac{|j : q_j = 0, p_j > 1/2|}{|j : p_j > 1/2|}},
\end{aligned} \tag{2.20}$$

and approximate g_1^{-1} where

$$g_1(x) = \frac{x}{|j : p_j \leq x, q_j = 1|}. \tag{2.21}$$

I derive the final v-values by integrating the distribution of P, Q under the null hypothesis over the rejection regions:

$$\begin{aligned}
\int_{L(p_0,p_1)} df_0 &= Pr((P,Q) \in L(p_0,p_1)|H_0^p) \\
&= Pr((P \leq p_0, Q=0) \cup (P \leq p_1, Q=1)|H_0^p) \\
&= Pr(P \leq p_0, Q=0|H_0^p) \\
&\quad + Pr(P \leq p_1, Q=1|H_0^p) \\
&= Pr(P \leq p_0|Q=0, H_0^p)Pr(Q=0|H_0^p) \\
&\quad + Pr(P \leq p_1|Q=1, H_0^p)Pr(Q=1|H_0^p) \\
&= p_0 \times (1-q_0) + p_1 \times q_0,
\end{aligned} \tag{2.22}$$

where $q_0 = \widehat{Pr}(Q=1|H_0^p)$.

2.5 fcfdr R package

I have created a user-oriented R package, `fcfdr`, to implement Flexible cFDR and Binary cFDR. The software webpage (https://annahutch.github.io/fcfdr/) contains fully reproducible vignettes that illustrate how the methods can be used to generate v-values from GWAS p-values and relevant auxiliary data, and how these can be used directly in any error rate controlling procedure (for example using the `stats::p.adjust` function in R with `method="BH"` for FDR-adjusted p-values) (Appendix A.1).

2.6 Simulation method

I used simulations to assess the performance of Flexible cFDR and Binary cFDR when iteratively leveraging various types of auxiliary data. I validated Flexible cFDR against the existing cFDR framework, which I call "Empirical cFDR" since it uses empirical CDFs (Liley and Wallace, 2021), in two use-cases where appropriate. I then evaluated the performance of Flexible cFDR in three novel use-cases where the auxiliary data was simulated from arbitrary distributions, and compared results to those when using Boca and Leek's FDR regression (referred to as BL) (Boca and Leek, 2018). I evaluated the performance of Binary cFDR when iterating over both independent and dependent binary data, and compared the results to those from BL. I included results from BL in my simulation analysis because this approach has been shown to outperform other methods by an independent research group (Korthauer et al., 2019) and because it was the only other method out of those considered that allowed for multiple covariates of this nature.

2.6.1 Simulating GWAS results (p)

I first simulated GWAS p-values for the arbitrary "principal trait" to be used as p in my simulations. I collected haplotype data for 3781 individuals from the UK10K project (REL-2012-06-02) (The UK10K Consortium, 2015) at 80,356 SNPs residing on chromosome 22 with $MAF \geq 0.05$ (to match the convention that genetic association studies identify common genetic variation). I split the haplotype data into 24 LD blocks representing approximately independent genomic regions defined by the LD detect method (Berisa and Pickrell, 2016). I then further stratified these so that no more than 1000 SNPs were present in each block, subsequently recording the LD block that each SNP resided in.

I next used the `simGWAS` R package (https://github.com/chr1swallace/simGWAS) (Fortune and Wallace, 2019) to simulate Z-scores for SNPs within each block. The `simGWAS::simulate_z_scores` function requires input for (i) the number of cases and controls (ii) the causal variants (iii) the log ORs at the causal variants and (iv) haplotype frequencies. For my simulation analysis, I selected 5000 cases and 5000 control samples, and within each block I randomly sampled 2, 3 or 4 causal variants with log OR effect sizes simulated from the standard Gaussian prior used in case-control genetic fine-mapping studies, $N(0, 0.2^2)$ (Wellcome Trust Case Control Consortium, 2007) (the mean number of simulated causal variants across all blocks in each simulation was 54). For the haplotype frequency parameter, I supplied a `data.frame` of haplotypes using the UK10K data, with a column of computed frequencies for each haplotype. I collated the Z-scores from each region and converted these to p-values representing the evidence of association between the SNPs and the arbitrary principal trait.

To generate an independent subset of SNPs required to fit the KDE, I converted the haplotype data to genotype data by summing haplotype values for each individual, and used the `write.plink` function (Chang et al., 2015) to generate the files required for the LDAK software (Speed et al., 2020). I generated LDAK weights for each of the SNPs and used the subset of SNPs with non-zero LDAK weights as an independent subset of SNPs.

2.6.2 Simulating continuous auxiliary data (q in Flexible cFDR)

I considered five use-cases of Flexible cFDR (simulations A-E), defined by (i) the distribution of the auxiliary data q (ii) the relationship between p and q and (iii) the relationship between different q in each iteration (5 realisations of q were sampled in each simulation representing multi-dimensional covariates so that cFDR could be applied iteratively) (Table 2.1). I denote the value of q at SNP i in realisation k as $q_i^{(k)}$.

In simulation A, I sampled $q_i \sim Unif(0, 1)$ to represent iterating over null p-values (Fig 2.3A). In simulation B, I investigated the standard use-case of cFDR by iterating over p-values from

Simulation	Distribution of q	Relationship between p_i and $q_i^{(k)}$	Average pairwise correlation between $q^{(k)}$ and $q^{(l)}$
A	p-values (all null)	Independent	0
B	p-values (related trait)	Shared causal variants	0.04
C	Functional	Independent	0
D	Functional	Dependent	0.08
E	Functional	Dependent	0.19

Table 2.1 Table describing relationships between variables in simulation analysis. Pairwise correlations are the Pearson correlation coefficients.

"related traits" (Fig 2.3B). To do this, I used the `simGWAS` R package (Fortune and Wallace, 2019) to simulate p-values, specifying the shared causal variants such that each pair of vectors p, q were guaranteed to share causal variants in exactly 4 of the 24 LD blocks, whilst each pair of vectors $q^{(k)}, q^{(j)}$ were expected to share causal variants in 4 of the 24 LD blocks.

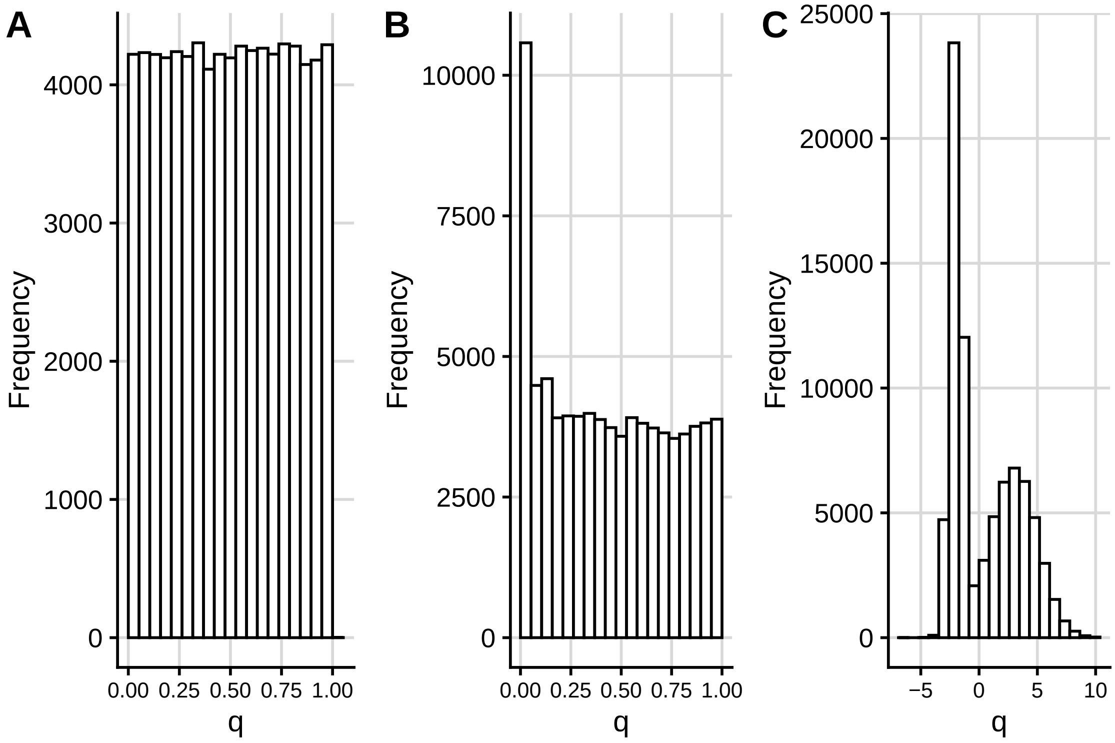


Fig. 2.3 Histograms of auxiliary data leveraged in Flexible cFDR simulation analysis. Example data leveraged in (A) simulation A (simulated from standard uniform distribution) (B) simulation B (simulated p-values for related traits) and (C) simulations C, D and E (simulated from a mixture Gaussian distribution)

In simulations C-E, I simulated auxiliary data representing functional genomic data sampled from arbitrary distributions, and which varied based on dependence structure with the principal trait p-values. In simulation C, I sampled q_i from a bimodal mixture Gaussian distribution that was independent of p_i: $q_i \sim 0.5 \times N(-2, 0.5^2) + 0.5 \times N(3, 2^2)$ (Fig 2.3C). In simulations D and E I simulated continuous auxiliary data that was dependent on p_i by first defining "functional SNPs" as causal variants plus any SNPs within 10,000-bp (to incorporate SNPs residing in the same arbitrary "functional mark"), and "non-functional SNPs" as the remainder. In simulation D, I then sampled q_i from different mixture Gaussian distributions for functional and non-functional SNPs:

$$q_i \sim \begin{cases} w \times N(\mu_1, 1) + (1-w) \times N(\mu_2, 0.5^2), & \text{if SNP } i \text{ is non-functional} \\ (1-w) \times N(\mu_1, 1) + w \times N(\mu_2, 0.5^2), & \text{if SNP } i \text{ is functional} \end{cases} \tag{2.23}$$

where $\mu_1 \in \{2.5, 3, 4\}$, $\mu_2 \in \{-1.5, -2, -3\}$, $w \in \{0.6, 0.7, 0.8, 0.9, 0.95\}$ vary across iterations.

I anticipated that my method would be used to leverage functional genomic data iteratively, and so I also evaluated the impact of repeatedly iterating over auxiliary data that captures the same functional mark. To do this, in simulation E I iterated over realisations of q that were sampled from the same distribution,

$$q_i \sim \begin{cases} N(3, 2^2), & \text{if SNP } i \text{ is non-functional} \\ N(-2, 0.5^2), & \text{if SNP } i \text{ is functional.} \end{cases} \tag{2.24}$$

2.6.3 Simulating binary auxiliary data (q in Binary cFDR)

I considered three use-cases of Binary cFDR (simulations F-H) defined by dependence on p_i and correlation between realisations of q. In simulation F I leveraged binary auxiliary data that was independent of p_i, $q_i \sim \text{Bernoulli}(0.05)$. In simulations G and H I leveraged binary data that was dependent on p_i. Defining functional SNPs as above, in simulation G I sampled

$$q_i \sim \begin{cases} \text{Bernoulli}(0.05), & \text{if SNP } i \text{ is non-functional} \\ \text{Bernoulli}(0.4), & \text{if SNP } i \text{ is functional.} \end{cases} \tag{2.25}$$

To evaluate the impact of repeatedly iterating over highly correlated auxiliary data that captures the same functional mark, such that $q_i^{(k)} \not\perp q_i^{(l)}$, in simulation H I sampled

$$q_i \sim \begin{cases} \text{Bernoulli}(0.05), & \text{if SNP } i \text{ is non-functional} \\ \text{Bernoulli}(0.8), & \text{if SNP } i \text{ is functional.} \end{cases} \tag{2.26}$$

2.6.4 Implementation of methods

I used Flexible cFDR, BL and Empirical cFDR (where applicable) to leverage the continuous auxiliary data. Flexible cFDR was implemented using the `fcfdr::flexible_cfdr` function with default parameter values. The optional `maf` parameter (that performs down-sampling of independent SNPs based on MAF) was not required here since over the restricted interval of MAF values considered ($MAF \geq 0.05$), the MAF distributions of the whole SNP set and the independent subset were largely comparable. To implement BL, I used the `lm_qvalue` function in the `swfdr` Bioconductor R package (version 1.16.0) (Leek et al., 2021), using a covariate matrix that consisted of five columns for $q^{(1)}, q^{(2)}, q^{(3)}, q^{(4)}, q^{(5)}$ to derive adjusted p-values.

Following the vignette for the Empirical cFDR software (https://github.com/jamesliley/cfdr/blob/master/vignettes/cfdr_vignette.Rmd), I first used the `cfdr::vl` function to generate L-curves. I used a leave-one-out-procedure, whereby L-curves were fit separately for data points in each LD block using data points from the other LD blocks. To ensure that the cFDR curves were strictly decreasing (preventing a complication whereby all v-values corresponding to the smallest p-values were given the same value), I reduced the value of the `gx` parameter to the minimum p-value in the LD block. I then estimated the distribution of $P, Q|H_0^p$ using the `cfdr::fit.2g` function and integrated its density over the computed L-regions using the `cfdr::il` function, specifying a mixture Gaussian distribution for the Z-scores.

Binary cFDR was implemented using the `fcfdr::binary_cfdr` function with a leave-one-out procedure based on LD block. All cFDR methods were applied iteratively 5 times in each simulation to represent leveraging multi-dimensional covariates.

2.6.5 Evaluating sensitivity, specificity and FDR control

To quantify the results from my simulations, I used the BH procedure to derive FDR-adjusted v-values from Empirical, Flexible and Binary cFDR (which I call “FDR values” for conciseness). For BL, I used the adjusted p-values as the quantity of interest. I then calculated proxies for the sensitivity (true positive rate) and the specificity (true negative rate) at an FDR threshold of $\alpha = 5e-06$, which roughly corresponds to the genome-wide significance p-value threshold of $p = 5e-08$ (Fig 2.4). I defined a subset of “truly associated SNPs” as any SNPs with $r^2 \geq 0.8$ with any of the causal variants. Similarly, I defined a subset of “truly not-associated SNPs” as any SNPs with $r^2 \leq 0.01$ with all of the causal variants. (Note that there were three non-overlapping sets of SNPs: “truly associated”, “truly not-associated” and neither of these). I calculated the sensitivity proxy as the proportion of truly associated SNPs that were called significant and the specificity proxy as the proportion of truly not-associated SNPs that were called not significant.

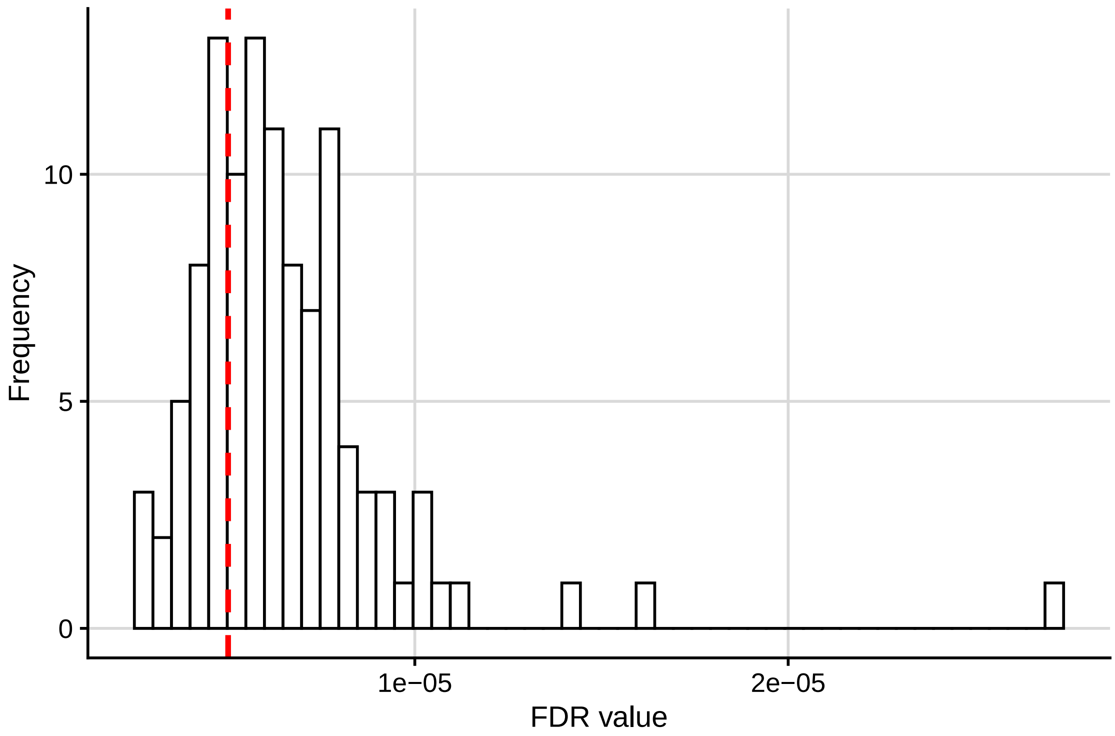


Fig. 2.4 Histogram of the maximum FDR-adjusted p-value (using BH procedure; “FDR values”) amongst SNPs with $p \leq 5e-08$ in the simulation analysis. Red dashed line at the selected FDR threshold of $FDR = 5e-06$.

To assess whether the FDR was controlled within a manageable number of simulations, I raised α to 0.05 and calculated the proportion of SNPs called FDR significant which were truly not-associated (that is, $r^2 \leq 0.01$ with all of the causal variants).

2.7 Simulation results

2.7.1 Leveraging continuous auxiliary covariates

One would expect that leveraging irrelevant data should not change the conclusions from their study. Simulations A and C showed that the sensitivity and specificity remained stable across iterations, and that the FDR was controlled at a pre-defined level, when leveraging independent auxiliary data with Flexible cFDR (Fig 2.5A, Fig 2.5C). In contrast, when leveraging relevant data, one would hope that the sensitivity improves whilst the specificity remains high. This is what I observed for Flexible cFDR in simulations B and D (Fig 2.5B, Fig 2.5D). The increase in sensitivity related to how informative the auxiliary data was, whereby the sensitivity generally increased more in simulation B than simulation D, where the average Pearson correlation coefficient between p and $q^{(k)}$ was $r = 0.07$ and $r = 0.04$ for simulations B and D, respectively.

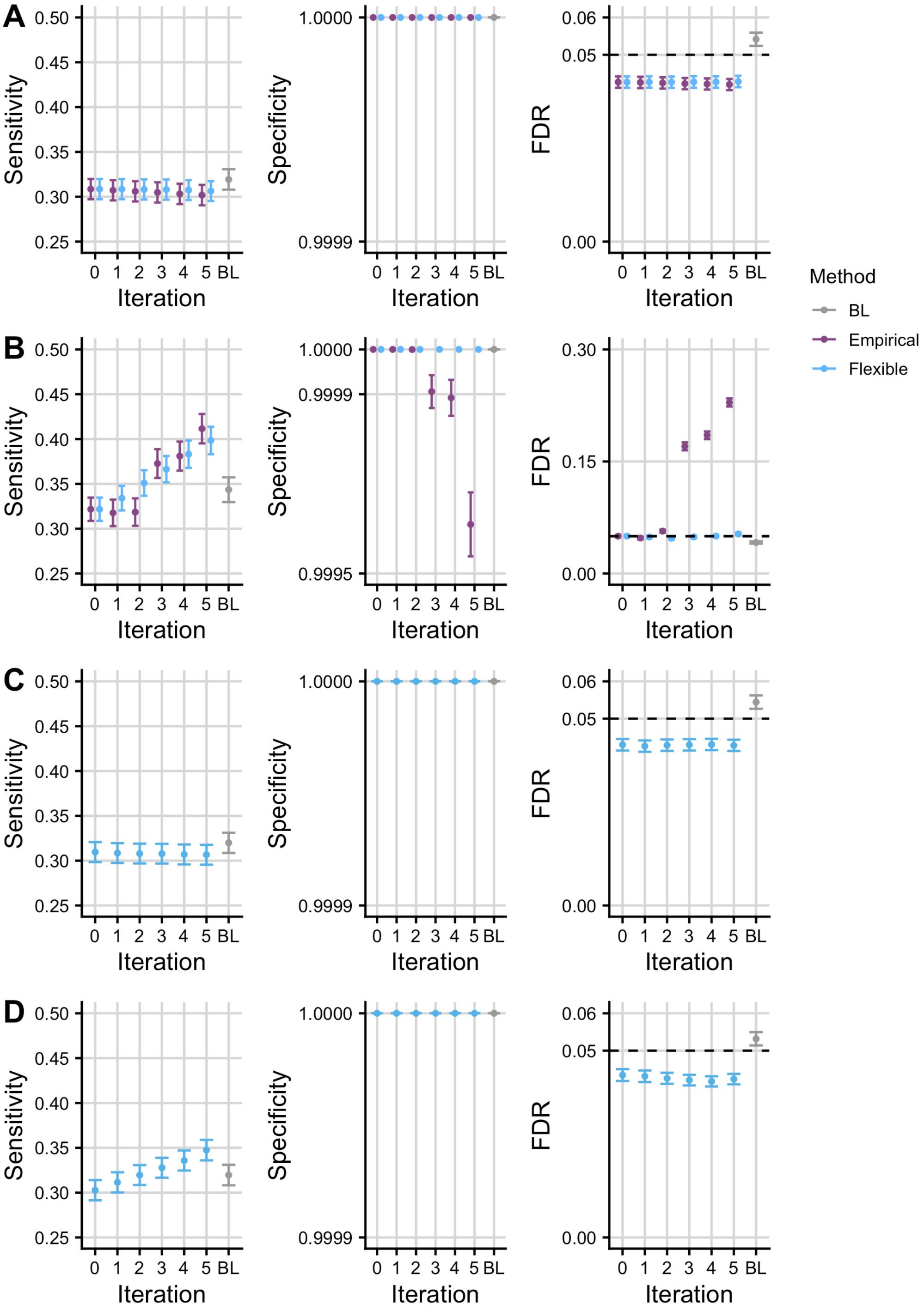


Fig. 2.5 Simulation results for Flexible cFDR, Empirical cFDR and BL. Mean +/- standard error for the sensitivity, specificity and FDR of FDR values from Empirical and Flexible cFDR when iterating over independent (A; “simulation A”) and dependent (B; “simulation B”) auxiliary data that was bounded by $[0, 1]$. Panels C and D show the results from Flexible cFDR when iterating over independent (C; “simulation C”) and dependent (D; “simulation D”) auxiliary data that was simulated from bimodal mixture Gaussian distributions. BL refers to results when using Boca and Leek’s FDR regression to leverage the 5-dimensional covariate data. Iteration 0 corresponds to the original FDR values. Results were averaged across 100 simulations.

For simulations A and B, I could compare Flexible cFDR performance to that of the current method, Empirical cFDR, since q could be transformed to a mixture Gaussian (Liley and Wallace, 2021). Performance was similar for simulation A, whilst for simulation B, the sensitivity of the two methods was comparable but Empirical cFDR exhibited a greater decrease in the specificity and failed to control the FDR in later iterations (Fig 2.5B). This contrasts with earlier results for Empirical cFDR, which showed good control of FDR (Liley and Wallace, 2021), and reflects the structure of my simulations which assume dependence between different realisations of q. Additionally, Flexible cFDR is quicker to run than Empirical cFDR, taking approximately 3 minutes compared to Empirical cFDR which takes approximately 6 minutes to complete a single iteration on 80,356 SNPs (using one core of an Intel Xeon E5-2670 processor running at 2.6GHz). Together, these findings indicate that Flexible cFDR performs no worse, and generally better, than Empirical cFDR in use-cases where both methods are supported.

When benchmarking the performance of Flexible cFDR against that of BL (Boca and Leek, 2018), I found that BL failed to control the FDR when leveraging independent covariate data (Fig 2.5A, Fig 2.5C). This may be due to the correlations between SNPs, as control of the FDR by BL was found to be worse with increasing correlation (Boca and Leek, 2018), but note that correlation between SNPs is ubiquitous in GWAS data. When leveraging dependent covariate data, BL was consistently less powerful than Flexible cFDR (Fig 2.5B, Fig 2.5D) and it failed to control the FDR in simulations leveraging dependent covariates from arbitrary distributions (Fig 2.5D), which represent the general use-case of the method.

I anticipated that Flexible cFDR would typically be used to leverage functional genomic data iteratively, and it was helpful that specificity remained high and FDR was controlled in simulation D. It is obvious that repeated conditioning on the *same* data should produce erroneous results, with SNPs with a modest p but extreme q incorrectly attaining greater significance with each iteration. For strict validity, we require $q_i^{(k)} \perp\!\!\!\perp q_i^{(l)}|H_0^p$ as the v-value from iteration k ($v_i^{(k)}$) will contain some information about $q_i^{(k)}$, and the cFDR assumes $v_i^{(k)} \perp\!\!\!\perp q_i^{(k+1)}|H_0^p$ at the next iteration. However, even when $q_i^{(k)} \not\perp\!\!\!\perp q_i^{(l)}|H_0^p$, we expect the dependence between v and q to be quite weak, hence the acceptable FDR control in simulations B and D above.

Given the wealth of functional marks available for similar tissues and cell types (for example subsets of peripheral immune cells), I wanted to assess the robustness of my procedure to more extreme dependence. I did this by repeatedly iterating over auxiliary data that was capturing the same functional mark. In simulation E, the sensitivity increased with each iteration at the expense of a drop in the specificity and loss of FDR control in later iterations (Fig 2.6). I therefore recommend that care should be taken not to repeatedly iterate over functional data that is capturing the same genomic feature, and in a real data example that follows, I average over cell types which show correlated values for functional data.

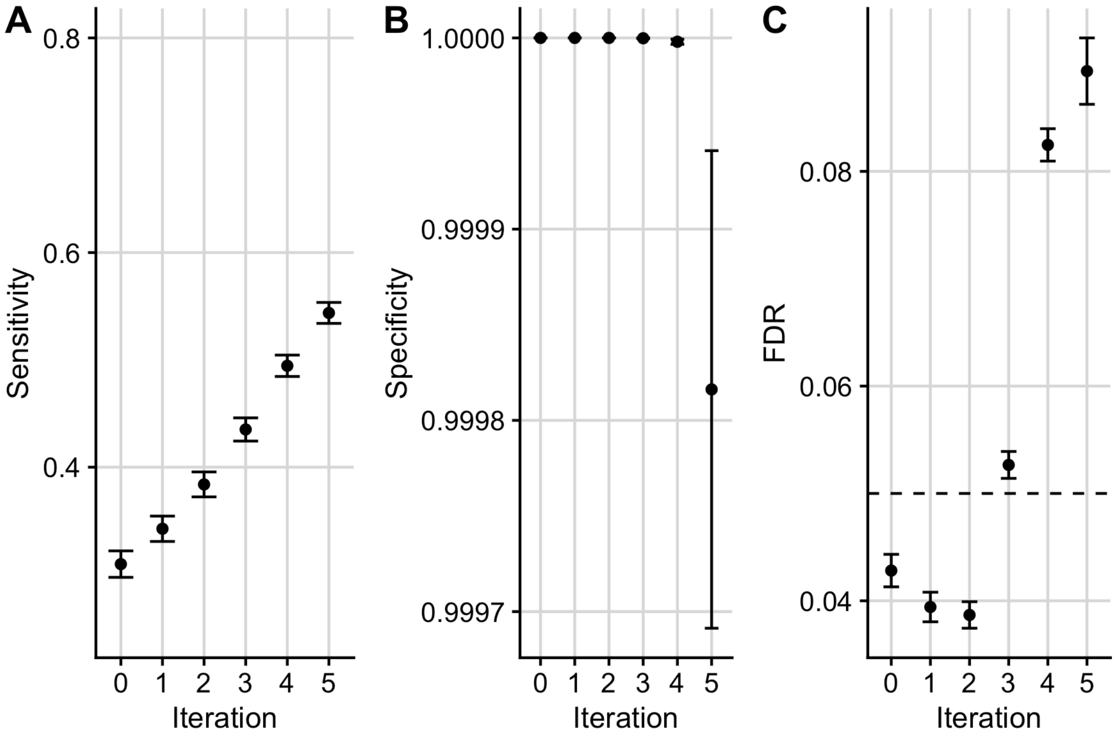


Fig. 2.6 Mean +/- standard error for the (A) sensitivity (B) specificity and (C) FDR of FDR values from Flexible cFDR when iterating over auxiliary data sampled from the same distribution ("simulation E"). Iteration 0 corresponds to the original FDR values. Results were averaged across 100 simulations.

2.7.2 Leveraging binary auxiliary covariates

Simulation F showed that the sensitivity and specificity remained stable across iterations and that the FDR was controlled at a pre-defined level when leveraging independent binary auxiliary data with Binary cFDR (Fig 2.7A). When leveraging dependent binary data in simulation G with Binary cFDR, the sensitivity increased, the specificity remained stable and the FDR remained controlled (Fig 2.7B). As expected, when leveraging highly correlated auxiliary data in simulation H, the FDR was no longer controlled (Fig 2.7C), serving as a salutary reminder to avoid iterating over functional data that is capturing the same genomic feature.

When benchmarking the performance of Flexible cFDR against that of BL, I found that BL was consistently less powerful than Binary cFDR when leveraging dependent binary auxiliary data (Fig 2.7B, Fig 2.7C). I also found that BL failed to control the FDR when leveraging both independent and dependent binary covariate data (Fig 2.7).

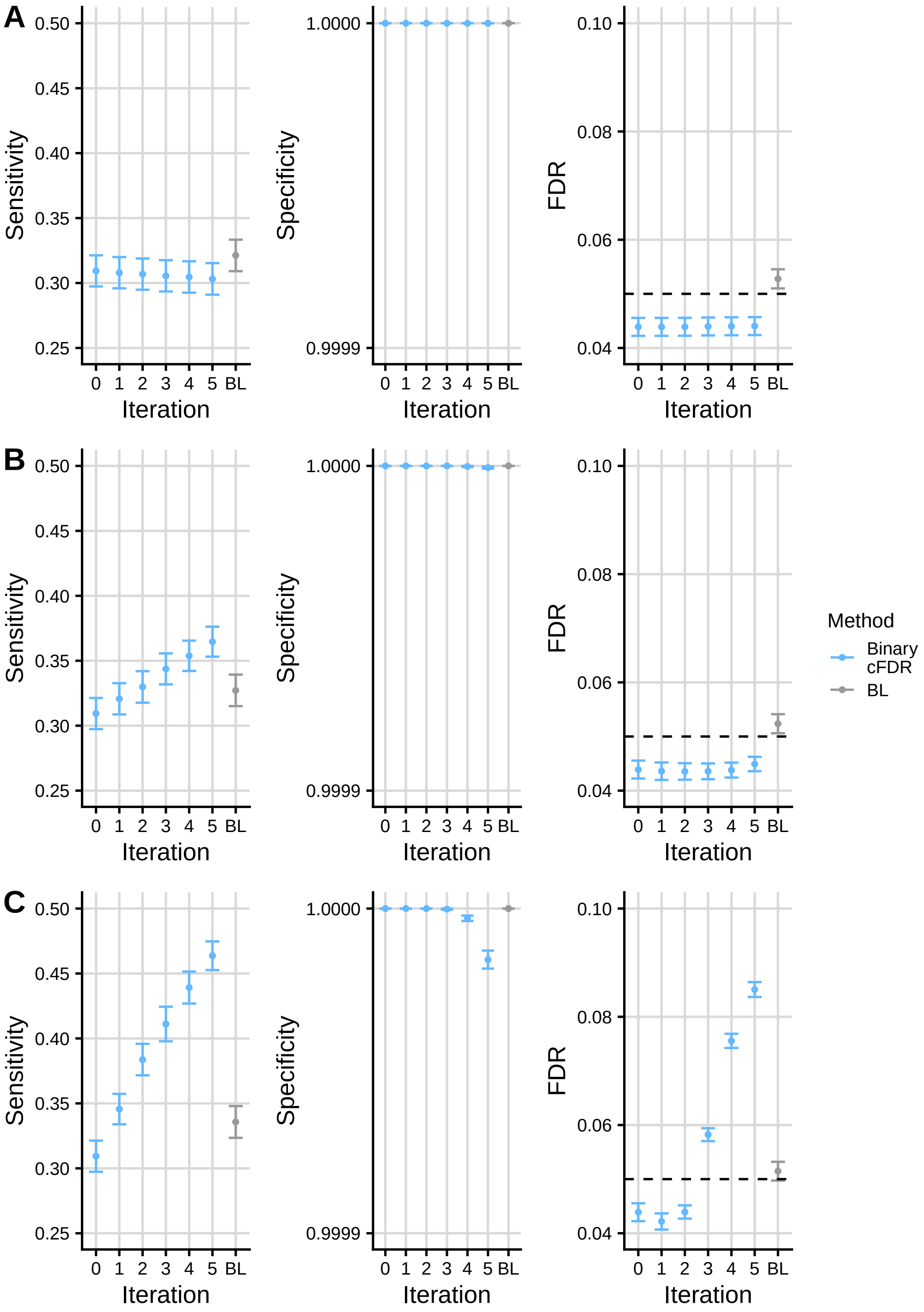


Fig. 2.7 Simulation results for Binary cFDR and BL. Mean +/- standard error for the sensitivity, specificity and FDR of FDR values from Binary cFDR when iterating over independent (A; "simulation F") and dependent (B; "simulation G" and C; "simulation H") binary auxiliary data. BL refers to results when using Boca and Leek's FDR regression to leverage the 5-dimensional covariate data. Iteration 0 corresponds to the original FDR values. Results were averaged across 100 simulations.

2.8 Application 1: Leveraging GenoCanyon scores with asthma GWAS data

I demonstrate the utility of Flexible cFDR by leveraging GenoCanyon scores measuring SNP functionality with GWAS p-values for asthma to generate functionally informed v-values. I additionally compare the performance of Flexible cFDR to that of three comparator methods: GenoWAP (Lu et al., 2016b), independent hypothesis weighting (IHW) (Ignatiadis et al., 2016) and BL (Boca and Leek, 2018).

2.8.1 Methods

Asthma GWAS data

As part of a different project, my colleague Dr Guillermo Reales, downloaded asthma GWAS summary statistics for 2,001,256 SNPs from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) for study accession GCST006862 (Demenais et al., 2018) on 2019-10-10. They extracted the p-values generated from a meta-analysis of studies from the Trans-National Asthma Genetic Consortium (TAGC) for individuals of European ancestry under a random effects model, totalling $19,954$ asthma cases and $107,715$ controls. I called this GWAS data the “discovery GWAS data set”. The genomic inflation factor for this study was $\lambda = 1.055$, implying minimal inflation of test statistics. I used the UCSC liftOver utility (Kuhn et al., 2013) to convert coordinates from GRCh38/hg38 to GRCh37/hg19, and removed those that could not be accurately converted. All genomic co-ordinates in this chapter are given with respect to human build GRCh37/hg19.

For the MAF matching step described in section 2.3.4, I used MAFs estimated from samples in the CEU sub-population (99 Utah residents with Northern and Western European ancestry) of the 1000 Genomes Project phase 3 data set (The 1000 Genomes Project Consortium, 2015). For the 639 SNPs with missing MAF values I used values randomly sampled from the empirical MAF distribution derived from the other SNPs. This reduced the independent subset of SNPs for fitting the KDE from 509,716 to 247,879 SNPs.

To identify independent associations, I began by calculating LD between SNPs using haplotype data from samples in the EUR population of the 1000 Genomes Project phase 3 data set (503 individuals of European ancestry). I then used the LD clumping algorithm in `PLINK 1.9` (Chang et al., 2015), using a 5-Mb window and an r^2 threshold of 0.01 (Kichaev et al., 2019). I called the SNP with the smallest p-value in the discovery GWAS data set in each LD clump the “lead variant”.

Dr Guillermo Reales also downloaded data from a larger GWAS performed by the Neale Lab (self-reported asthma: 20002_1111) for $41,934$ asthma cases and $319,207$ controls from UK Biobank (Sudlow et al., 2015) (URL: https://www.dropbox.com/s/kp9bollwekaco0s/20002_1111.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 downloaded on 2020-05-10). I called this GWAS data the "validation GWAS data set". I restricted my analysis to the 1,968,651 SNPs that were present in both the discovery and the validation GWAS data sets.

GenoCanyon scores

Tools have now been developed that integrate various genomic and epigenomic annotation data to quantify the pathogenicity, functionality and/or deleteriousness of both coding and non-coding GWAS variants (Boyle et al., 2012; Khurana et al., 2013; Kircher et al., 2014; Lu et al., 2015; Ritchie et al., 2014). For example, GenoCanyon scores aim to infer the functional potential of each position in the human genome (Lu et al., 2015). I downloaded GenoCanyon scores from http://zhaocenter.org/GenoCanyon_Downloads.html for each of the 1,968,651 SNPs included in my analysis.

Implementation of methods

I used the `fcfdr::flexible_cfdr` function to leverage GenoCanyon scores with asthma GWAS p-values, using the indices of the independent set of SNPs for the `indep_index` parameter and the MAF values for the `maf` parameter. Due to the positive stochastic monotonicity requirement of cFDR, the Flexible cFDR software reversed the sign of the GenoCanyon scores for its internal calculations. I used the `stats::p.adjust` function with `method="BH"` to derive FDR values from the v-values and used these as the output of interest.

The GenoWAP software (https://github.com/rlpowles/GenoWAP-V1.2) requires a `threshold` parameter defining functional SNPs according to their GenoCanyon score. For this, I used the default recommended value of 0.1, which corresponded to 40% of the SNPs in my data set being "functional". I used the `GenoWAP.py` python script to obtain posterior scores for each SNP and used these as the output of interest.

To implement BL, I used the `lm_qvalue` function in the `swfdr` Bioconductor R package (version 1.16.0) (Leek et al., 2021). I used a covariate matrix consisting of a single column of GenoCanyon scores for each SNP and derived adjusted p-values, using these as the output of interest.

To implement IHW, I used the `IHW` Bioconductor R package (version 1.18) (https://bioconductor.org/packages/release/bioc/html/IHW.html) with default parameters, specifying the level of FDR control `alpha=0.000148249`. I used the adjusted p-values as the output of interest. An FDR threshold of $FDR \leq 0.000148249$ was used to call significant results as this corresponded to the genome-wide significance p-value threshold of $p \leq 5e-08$ (0.000148249 was the maximum FDR value amongst SNPs with raw p-values $\leq 5e-08$ in the discovery GWAS data set). This FDR threshold was also used to call significant SNPs from Flexible cFDR and BL.

2.8.2 Results

Overall, 655 SNPs were FDR significant ($FDR \leq 0.000148249$) in the original asthma GWAS (Demenais et al., 2018). I found that SNPs with high GenoCanyon scores were enriched for smaller p-values in the asthma GWAS (Fig 2.8). Accordingly, FDR values from Flexible cFDR for SNPs with high GenoCanyon scores (and therefore more likely to be functional) were lower than their corresponding non-functionally informed FDR values, whilst those for SNPs with low GenoCanyon scores (and therefore less likely to be functional) were higher than their corresponding non-functionally informed FDR values (Fig 2.9; full results for applications 1 and 2 in this chapter are deposited on Zenodo: https://doi.org/10.5281/zenodo.4701287). Specifically, Flexible cFDR identified 12 newly FDR significant SNPs (rs4705950, rs6903823, rs9262141, rs1264349, rs2106074, rs3130932, rs9268831, rs3129719, rs1871665, rs16924428, rs1663687 and rs12900122) which had high GenoCanyon scores (mean GenoCanyon score = 0.77) and three SNPs were no longer FDR significant, which had low GenoCanyon scores (mean GenoCanyon score = 0.01). At the locus level no newly significant, or newly not-significant, loci were identified.

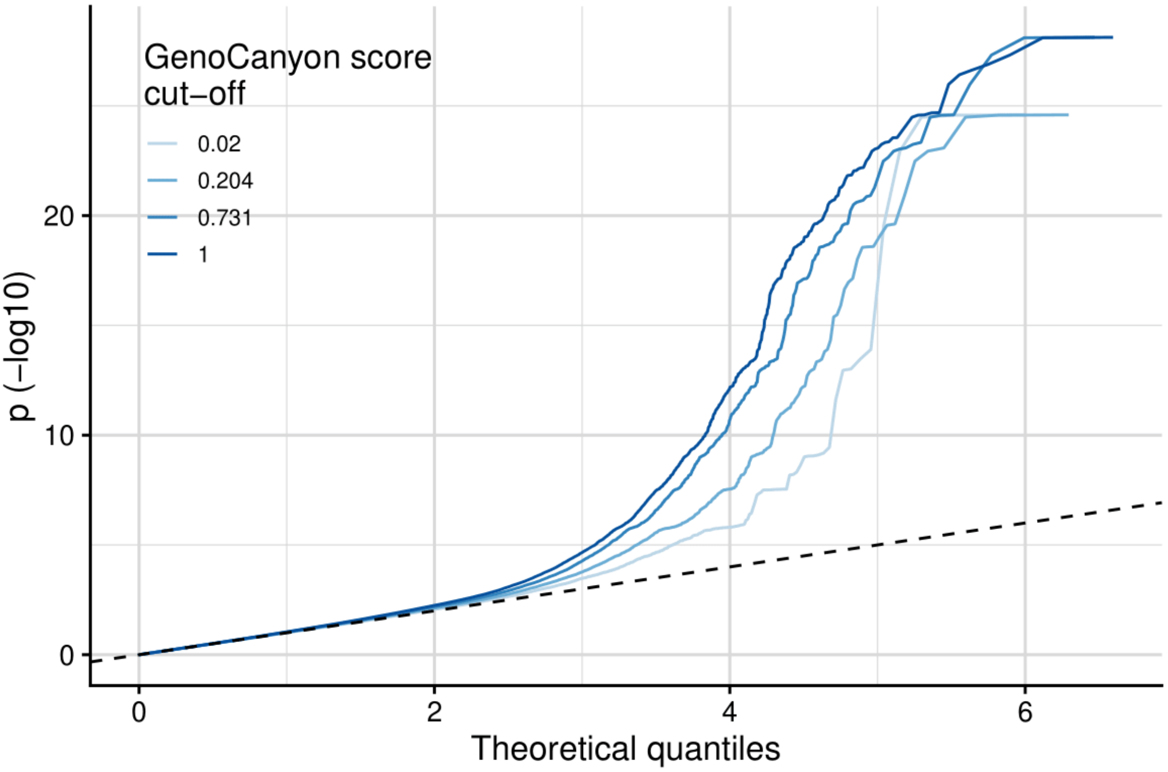


Fig. 2.8 Stratified Q-Q plot of empirical $-log_{10}$ transformed GWAS p-values for asthma against theoretical values stratified by GenoCanyon scores. The values that were used to threshold the GenoCanyon scores were the quantiles of the distribution (0.020 was the 0.25 quantile, 0.204 was the 0.5 quantile, 0.731 was the 0.75 quantile and 1 was the maximum value).

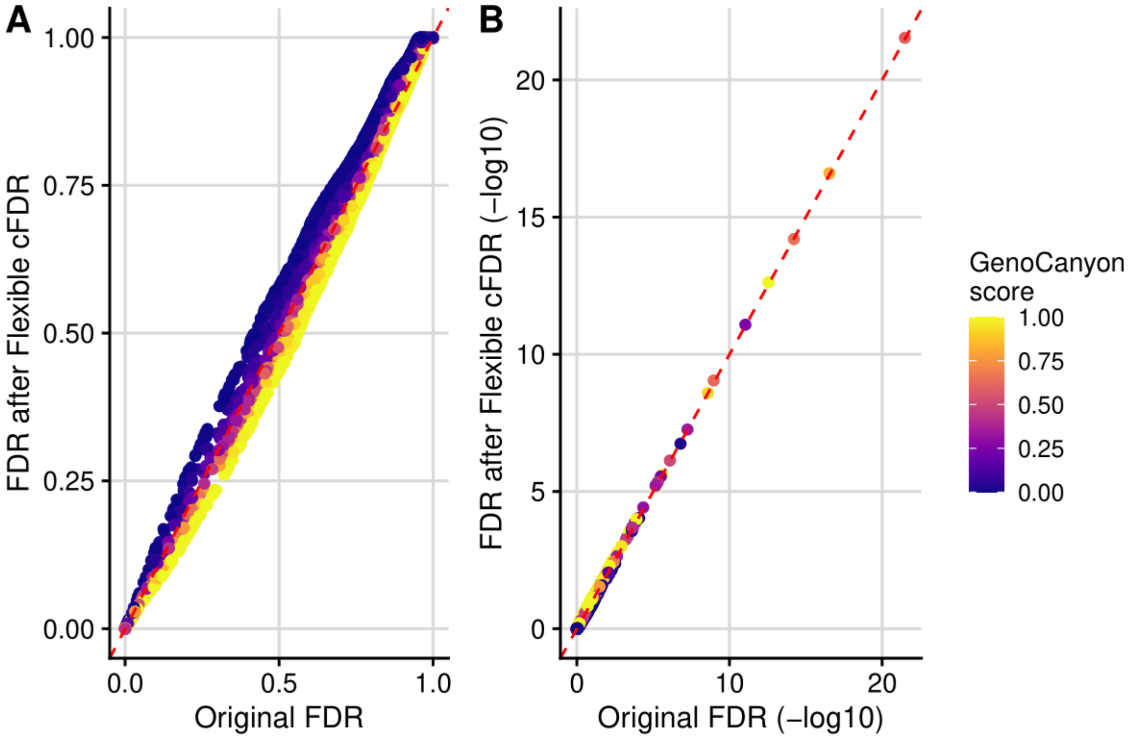


Fig. 2.9 (A) FDR values after using Flexible cFDR to leverage GenoCanyon scores with asthma GWAS p-values against original non-functionally informed FDR values coloured by GenoCanyon score. (B) as in (A) but FDR values have been $-log_{10}$ transformed.

I compared the results from Flexible cFDR to those from comparator methods when leveraging the exact same auxiliary data with the exact same GWAS data. IHW groups SNPs based on their covariate values and derives optimal group-specific weights for use in a weighted BH procedure. Interestingly, all SNPs were allocated a weight of 1 in this instance, meaning that IHW reduced to the conventional BH procedure (and so the "adjusted p-values" from IHW were identical to the original FDR values in the discovery GWAS data set).

In BL, logistic regression is used to estimate how the distribution of input p-values depend on the GenoCanyon scores, and this is then used to estimate the probability that the null hypothesis of no association is true for each SNP. When leveraging GenoCanyon scores, these probabilities ranged from 0.957 for the SNP with the largest GenoCanyon score to 0.993 for the SNP with the smallest GenoCanyon score (Fig 2.10). The consequence of the narrow range of these values is that the adjusted values from BL were very similar to the original FDR values (Fig 2.11). Specifically, BL only identified three newly FDR significant SNPs, and these were all also identified by Flexible cFDR. One of these had a very high GenoCanyon score (rs1871665 with score = 0.999) whilst the other two had medium (rs16924428 with score = 0.532) or low (rs9268831 with score = 0.224) scores. (SNP rs9268831 was likely found to be FDR significant by both Flexible cFDR and BL because its original non-functionally informed FDR value was very close to the FDR threshold ($FDR = 0.0001501209$ compared with the FDR threshold of 0.000148249) and the low GenoCanyon score was enough to push it past the significance threshold). No SNPs were identified as no longer FDR significant after applying BL and at the locus level, no newly significant, or newly not-significant, loci were identified.

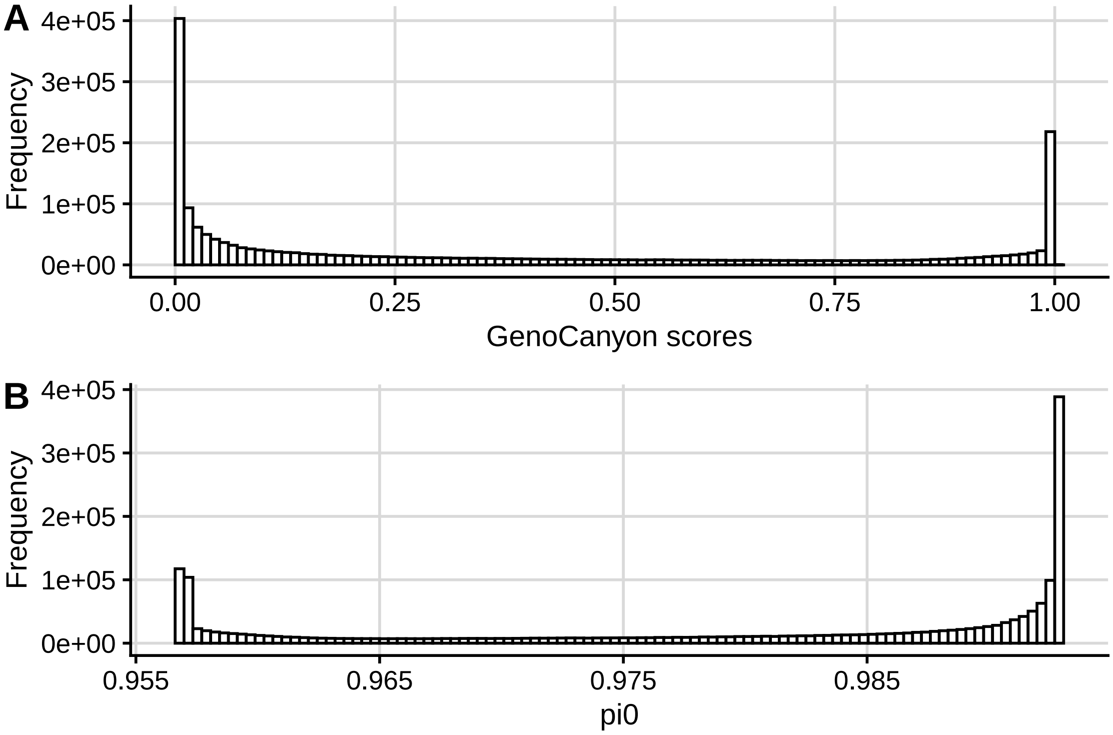


Fig. 2.10 (A) Histogram of GenoCanyon scores for SNPs in the asthma GWAS data set. (B) Histogram of estimated pi0 values (probabilities that the null hypothesis of no association is true) from BL

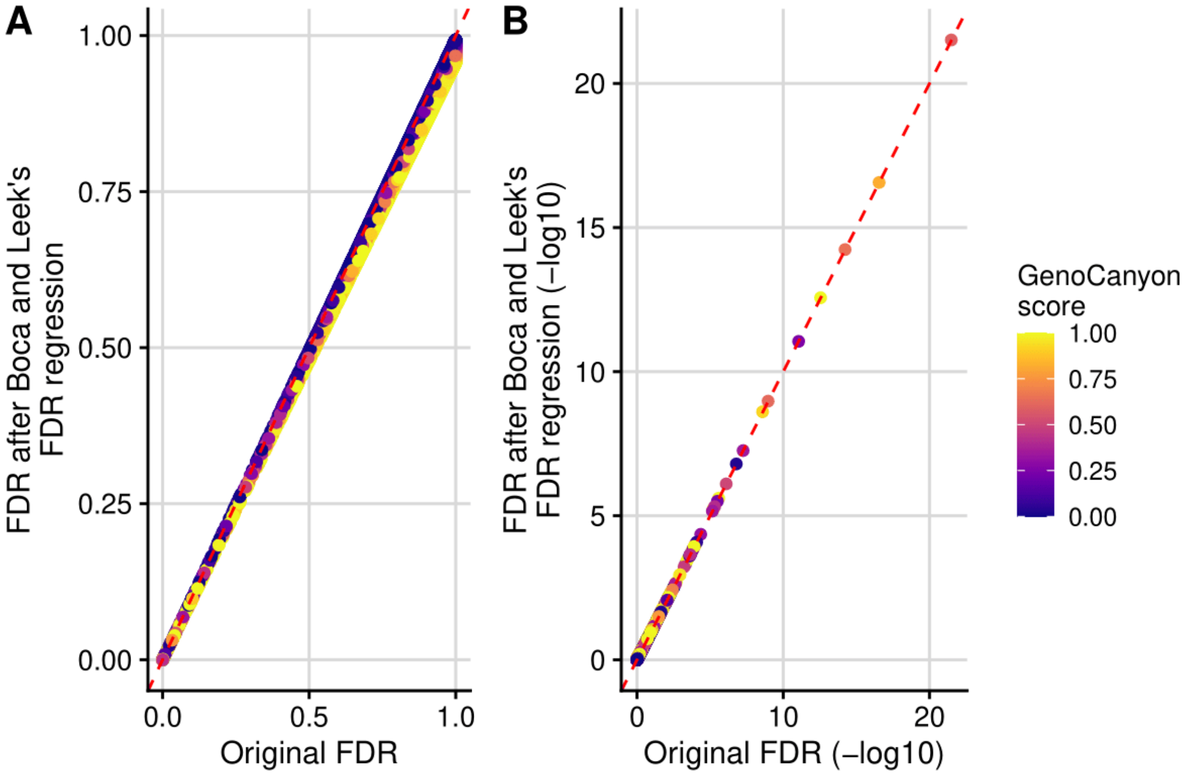


Fig. 2.11 (A) FDR values after using BL to leverage GenoCanyon scores with asthma GWAS *p*-values against original non-functionally informed FDR values coloured by GenoCanyon score. (B) as in (A) but FDR values have been $-log_{10}$ transformed.

Since GenoWAP outputs posterior probabilities rather than *p*-values, I compared the performance of Flexible cFDR and BL with GenoWAP based on the rankings of SNPs using the UK Biobank data resource (the validation GWAS data set). Firstly, at the SNP-level, for each of the 5152 SNPs that were FDR significant in the UK Biobank data, I compared the rank of the FDR value in the discovery GWAS data set with:

1. The rank of the FDR value from Flexible cFDR
2. The rank of the FDR value from BL
3. The rank of the (negative) posterior score from GenoWAP.

(IHW was not included in this comparison because the output from IHW was just the original FDR values). I found that the percentage of FDR significant SNPs in the UK Biobank data which had an improved rank after applying each of the methods was similar (Table 2.2) and that 68.5% of the SNPs that improved rank in at least one of the methods improved rank in all of the methods. Similarly, the percentage of the 1,963,499 SNPs that were not FDR significant in UK Biobank which had a decreased rank after applying each of the methods was similar

(Table 2.2) and 49.6% of the SNPs that decreased rank in at least one of the methods decreased rank in all of the methods.

	Flexible cFDR	BL	GenoWAP
UK Biobank significant which improved rank	60.3%	52.4%	61.5%
UK Biobank not-significant which decreased rank	46.8%	58.0%	44.8%

Table 2.2 Summary of SNP-level results when leveraging GenoCanyon scores with asthma GWAS p-values. Table lists the percentage of the 5152 FDR significant SNPs in UK Biobank which improved rank ("UK Biobank significant which improved rank") and the percentage of the 1,963,499 SNPs that were not FDR significant in UK Biobank which decreased rank ("UK Biobank not-significant which decreased rank") after applying Flexible cFDR, BL or GenoWAP.

Secondly, I focused on the 114 loci that were FDR significant in the UK Biobank data set. For each lead variant, I compared the rank of the FDR values in the discovery GWAS data set with the rank of the FDR value from Flexible cFDR, the rank of the FDR value from BL and the rank of the (negative) posterior score from GenoWAP (as before). I found that the percentage of UK Biobank significant SNPs (including lead variants) that improved rank after applying Flexible cFDR was greater than that for BL (Table 2.2, Table 2.3), which matched results from my simulation analysis showing that BL was generally less sensitive than Flexible cFDR. Similarly, the percentage of the 301 loci that were not FDR significant in UK Biobank which had a decreased rank after applying each of the methods was similar (Table 2.3) and 51.4% of the lead variants that decreased rank in at least one of the methods decreased rank in all of the methods.

	Flexible cFDR	BL	GenoWAP
UK Biobank significant which improved rank	42.1%	28.9%	55.3%
UK Biobank not-significant which decreased rank	40.5%	40.9%	34.6%

Table 2.3 Summary of locus-level results when leveraging GenoCanyon scores with asthma GWAS p-values. Table lists the percentage of the 114 FDR significant lead variants in UK Biobank which improved rank ("UK Biobank significant which improved rank") and the percentage of the 301 lead variants that were not FDR significant in UK Biobank which decreased rank ("UK Biobank not-significant which decreased rank") after applying Flexible cFDR, BL or GenoWAP.

In all, the results were similar for Flexible cFDR, BL and GenoWAP when leveraging GenoCanyon scores of SNP functionality with asthma GWAS p-values, but rather unexciting as no newly significant loci were identified. This could be due to the one-dimensional non-trait-specific auxiliary data that was being leveraged in this analysis, which is unlikely to capture enough disease-relevant information to substantially alter conclusions from a study. This speculation is supported both by the results from IHW, where the optimal weights were all equal to 1, and also by the intermediary results from BL, where the estimated proportions of true null hypotheses conditional on the GenoCanyon scores were almost negligible.

2.9 Application 2: Leveraging ChIP-seq data with asthma GWAS data

I used Flexible cFDR to leverage H3K27ac ChIP-seq data in asthma-relevant cell types with GWAS p-values for asthma. I additionally compared the performance of Flexible cFDR to that of two comparator methods that allow for multi-dimensional covariate data, BL (Boca and Leek, 2018) and FINDOR (Kichaev et al., 2019).

2.9.1 Methods

Asthma GWAS data

The discovery GWAS data set and the validation GWAS data set are as described in section 2.8.1.

H3K27ac ChIP-seq data

The histone modification H3K27ac is associated with active enhancers (Creyghton et al., 2010), and so SNPs residing in genomic regions with high H3K27ac counts in relevant cell types may be more likely to be associated with the trait of interest (Corradin and Scacheri, 2014).

I downloaded consolidated fold-enrichment ratios of H3K27ac ChIP-seq counts relative to expected background counts from NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) in primary tissues and cells that are relevant to asthma: immune cells and lung tissue from https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/. I mapped each SNP in my GWAS data set to its corresponding genomic region and recorded the H3K27ac fold change value for each SNP in each cell type using the `bedtools intersect` utility (Quinlan and Hall, 2010). For SNPs on the boundary of a genomic region (and therefore mapping to two regions), I randomly selected the fold change value for one of the regions.

The raw H3K27ac fold change data had very long tails and so I transformed the values: $q := log(q+1)$. I observed that the data for the different cell types roughly fell into two clusters: lymphoid cells clustered with CD56 cells whilst lung tissue clustered with monocytes (Fig 2.12). I therefore averaged the transformed H3K27ac fold change values in lymphoid and CD56 cell types to derive `q1` (the vector of auxiliary data values to be used in the first iteration of Flexible cFDR), and the transformed H3K27ac fold change values in lung tissue and monocytes to derive q2 (the vector of auxiliary data values to be used in the second iteration of Flexible cFDR). I added a small amount of noise $[N(0, 0.1^2)]$ to the latter to smooth out the discrete valued counts (Fig 2.13).

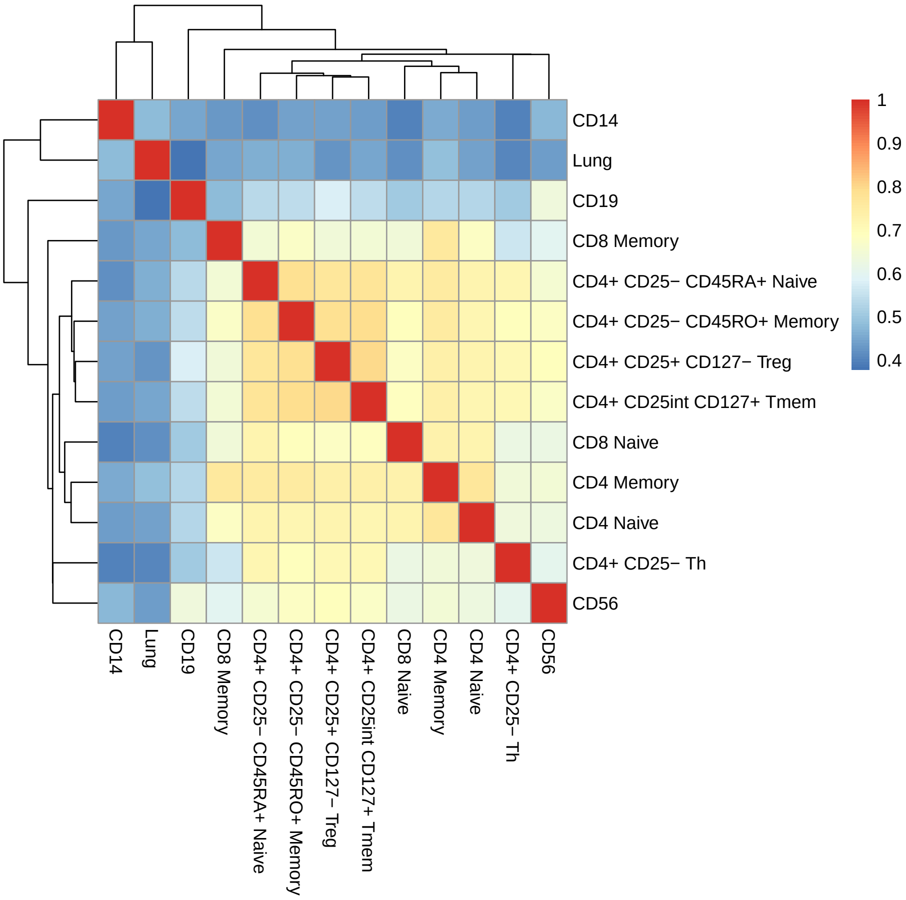


Fig. 2.12 Heatmap of the Pearson correlation coefficients between H3K27ac fold change values amongst asthma-relevant cell types downloaded from NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010). Figure generated using the pheatmap R package (https://cran.r-project.org/web/packages/pheatmap/index.html).

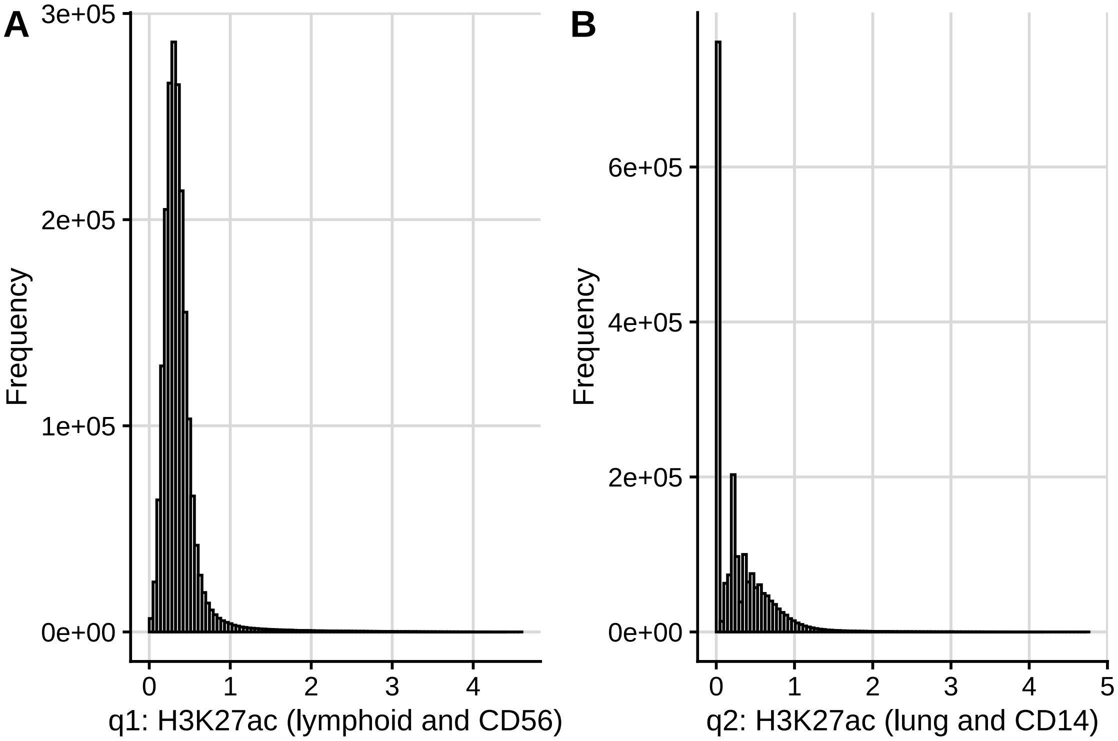


Fig. 2.13 Histograms of auxiliary data used in H3K27ac application. (A) `q1` is the average of (log transformed) H3K27ac fold change values in lymphoid and CD56 cell types (B) `q2` is the average of (log transformed) H3K27ac fold change values in lung tissue and CD14+ cells with a small amount of noise added.

Implementation of methods

I used the `fcfdr::flexible_cfdr` function to leverage the H3K27ac data with asthma GWAS p-values, using the indices of the independent set of SNPs for the `indep_index` parameter and the MAF values for the `maf` parameter. Due to the positive stochastic monotonicity requirement for cFDR, the Flexible cFDR software reversed the sign of the auxiliary data values for its internal calculations. I used the `stats::p.adjust` function with `method="BH"` to derive FDR values from the v-values and used these as the output of interest. To implement BL, I used the `lm_qvalue` function in the `swfdr` Bioconductor R package (version 1.16.0), with a covariate matrix consisting of two columns for `q1` and `q2`, to derive adjusted p-values and used these as the output of interest.

GenoWAP could not be implemented in this application because it only supports auxiliary data that is GenoCanyon scores (or tissue-specific GenoSkyline (Lu et al., 2016a) or GenoSkyline-Plus (Lu et al., 2017) scores). IHW could also not be implemented because it does not support multi-dimensional covariate vectors.

FINDOR uses the baseline-LD model from Gazal et al. (2017) for prediction, and so I was unable to directly compare the results from Flexible cFDR with those from FINDOR when leveraging the same ChIP-seq data. Instead, and as recommended, I used FINDOR to leverage the 96 annotations from the latest version (at the time of this research) of the baseline-LD model (version 2.2) with asthma GWAS p-values. Briefly, this auxiliary data contained the 75 annotations from Gazal et al. (2017) (including functional regions, histone marks, MAF bins and LD-related annotations) plus extra annotations including synonymous/ non-synonymous, conserved annotations, two flanking bivalent transcriptional start site/ enhancer annotations from NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), promoter/ enhancer annotations (Villar et al., 2015), promoter/ enhancer sequence age annotations (Marnetto et al., 2018) and 11 new annotations from Hujoel et al. (2019) (five new binary annotations and corresponding flanking annotations, and one continuous count annotation). I matched SNPs to their annotation values using rsIDs and GRCh37/hg19 coordinates.

To run FINDOR, stratified LD score regression (S-LDSC) must first be implemented to obtain annotation effect size estimates, $\hat{\tau_C}$. To run S-LDSC, I downloaded (i) partitioned LD scores from the baseline-LD model (version 2.2) (ii) regression weight LD scores and (iii) allele frequencies for available variants in the 1000 Genomes Project phase 3 data set. I then used the `munge_sumstats.py` python script in the `ldsc` package (https://github.com/bulik/ldsc) to convert the asthma GWAS summary statistics to the correct format for use in the `ldsc` software. I restricted my analysis to HapMap3 SNPs using the `-merge-alleles` flag, as recommended in the LDSC and FINDOR documentation.

I ran S-LDSC with the `-print-coefficients` flag to generate the `.result` file containing the annotation effect size estimates that are required for FINDOR. To run FINDOR, partitioned LD scores must also be supplied for the SNPs in the data set. To generate these, I downloaded the 1000 Genomes Project phase 3 `PLINK` files for EUR samples and annotation data, and followed the 'LD Score Estimation Tutorial' on the LDSC GitHub page (https://github.com/bulik/ldsc/wiki/LD-Score-Estimation-Tutorial). Partitioned LD scores could be generated for 1,976,360 (out of 2,001,256) SNPs in the asthma data set that were also present in the 1000 Genomes Project phase 3 data set. I then generated a file for the asthma GWAS data, including columns for sample sizes, rsIDs and Z-scores. I used this file, along with the computed partitioned LD scores and the `.result` file from S-LDSC to obtain re-weighted p-values for the 1,968,651 SNPs using FINDOR. I used the BH procedure to convert these to FDR values and used these as my output of interest.

As in application 1, I used an FDR threshold of $FDR \leq 0.000148249$ to call significant SNPs.

2.9.2 Results

Flexible cFDR

In agreement with reports that GWAS SNPs are enriched in regions of active chromatin (Soskic et al., 2019), I observed that H3K27ac fold change values in asthma-relevant cell types were negatively correlated with asthma GWAS *p*-values (Fig 2.14) such that SNPs with high fold change values were enriched for smaller *p*-values (Fig 2.15). Accordingly, FDR values from Flexible cFDR for SNPs with high H3K27ac fold-change counts in asthma-relevant cell types were lower than their corresponding original FDR values, whilst those for SNPs with low H3K27ac fold-change counts in asthma-relevant cell types were higher than their corresponding original FDR values (Fig 2.16).

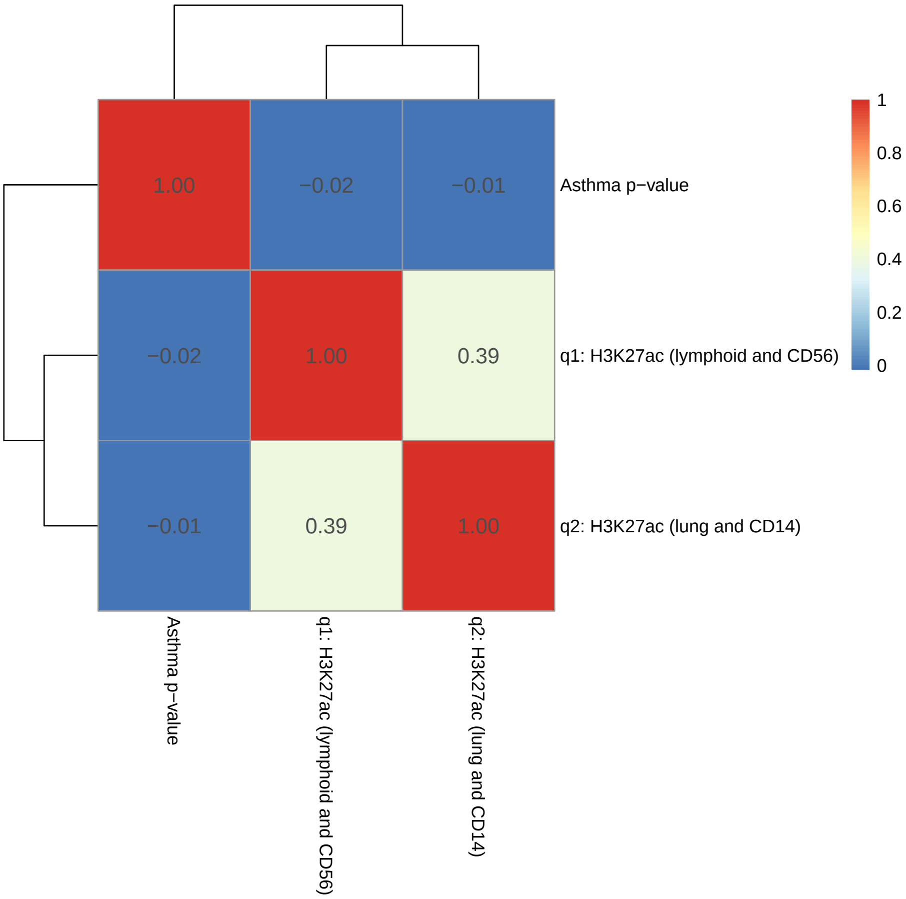


Fig. 2.14 Heatmap of Pearson correlation coefficients between log-transformed asthma GWAS *p*-values and the summarised H3K27ac fold change values leveraged by Flexible cFDR. `q1` is the average of (log transformed) H3K27ac fold change values in lymphoid and CD56 cell types. `q2` is the average of (log transformed) H3K27ac fold change values in lung tissue and CD14+ cells.

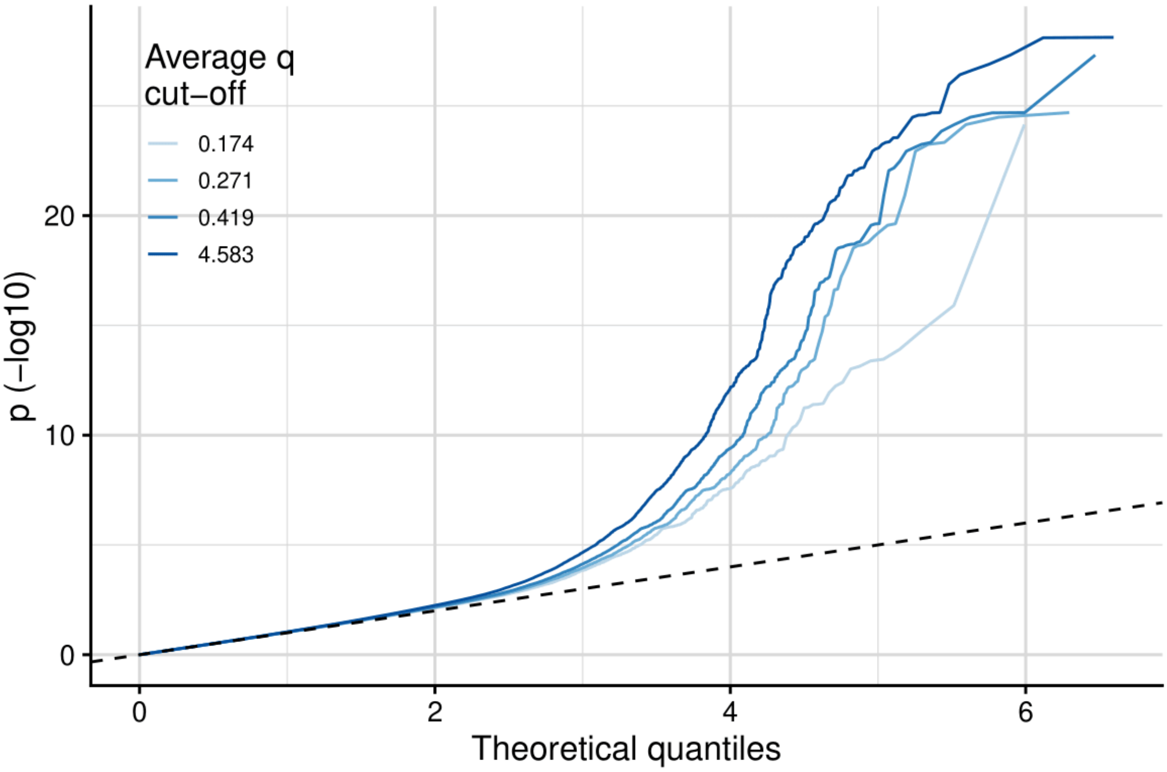


Fig. 2.15 Stratified Q-Q plot of empirical $-log_{10}$ transformed GWAS p-values for asthma against theoretical values stratified by average H3K27ac fold change values in asthma-relevant cell types. The values that were used to threshold q (average H3K27ac fold change value) were the quantiles of the distribution (0.174 was the 0.25 quantile, 0.271 was the 0.5 quantile, 0.419 was the 0.75 quantile and 4.583 was the maximum value).

The 655 SNPs that were FDR significant ($FDR \leq 0.000148249$) in the original asthma GWAS (Demenais et al., 2018) had strong replication p-values in the UK Biobank data set that was used for validation (Fig 2.16D; Iteration 0). By leveraging H3K27ac data, Flexible cFDR identified weaker signals that were not significant in the original data, but that had reassuringly small p-values in the UK Biobank data (median p-value in UK Biobank data for these SNPs was $4.65e-21$; Fig 2.16D). Specifically, Flexible cFDR identified 51 newly significant SNPs when leveraging average H3K27ac fold change values in lymphoid and CD56 cell types (Fig 2.16D; Iteration 1) and 24 newly significant SNPs when subsequently leveraging average H3K27ac fold change values in lung tissue and monocytes (Fig 2.16D; Iteration 2). The maximum p-value for the 69 newly significant SNPs (6 SNPs newly significant after iteration 1 were no longer significant after iteration 2) in the discovery GWAS data set was $2.15e-06$, and the maximum UK Biobank p-value for these SNPs was 0.02. The newly significant SNPs had relatively small estimated effect sizes (Fig 2.17), implying that there may be many more regions associated with asthma with increasingly smaller effect sizes that are missed by current GWAS sample sizes.

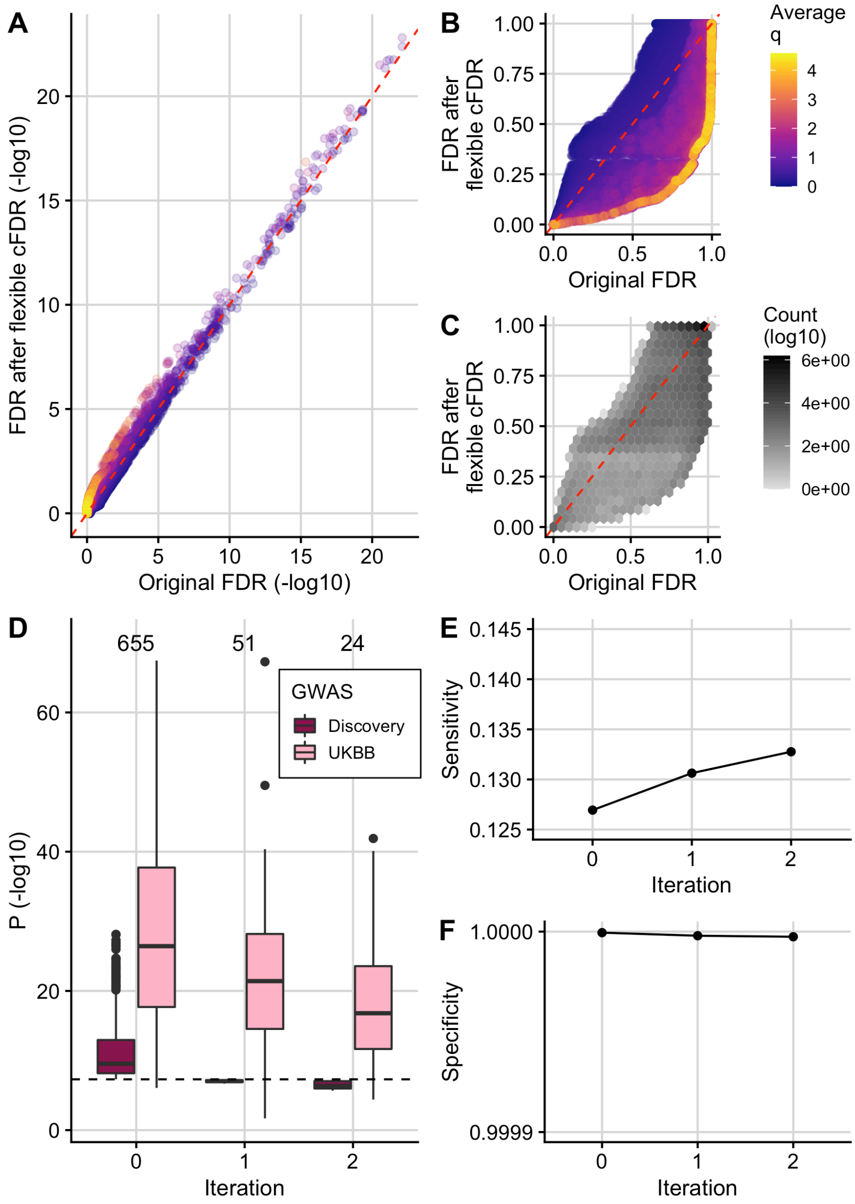


Fig. 2.16 Using Flexible cFDR to leverage H3K27ac data with asthma GWAS p-values. (A) ($-log_{10}$ transformed) FDR values after 2 iterations of Flexible cFDR leveraging H3K27ac counts in relevant cell types against raw ($-log_{10}$ transformed) FDR values coloured by the average value of the auxiliary data across iterations. (B) As in (A) but non-log-transformed FDR values. (C) As in (B) but coloured by log_{10} transformed counts of data points in each hexbin. (D) Box plots of ($-log_{10}$ transformed) p-values in the discovery GWAS and the UK Biobank data set for the 655 SNPs that were FDR significant in the original GWAS (Iteration 0), 51 SNPs that were newly FDR significant after iteration 1 of Flexible cFDR (leveraging average H3K27ac fold change values in lymphoid and CD56 cell types) and 24 SNPs that were newly FDR significant after iteration 2 of Flexible cFDR (subsequently leveraging average H3K27ac fold change values in lung tissue and CD14$^+$ cells). Black dashed line at genome-wide significance ($p = 5e-08$). (E) Sensitivity proxy and (F) specificity proxy for the H3K27ac application results. Sensitivity proxy was calculated as the proportion of SNPs that were FDR significant in the UK Biobank data set that were also FDR significant in the original GWAS (iteration 0), after iteration 1 of Flexible cFDR or after iteration 2 of Flexible cFDR. Specificity was calculated as the proportion of SNPs that were not FDR significant in the UK Biobank data set that were also not FDR significant in the original GWAS (iteration 0), after iteration 1 of Flexible cFDR or after iteration 2 of Flexible cFDR.

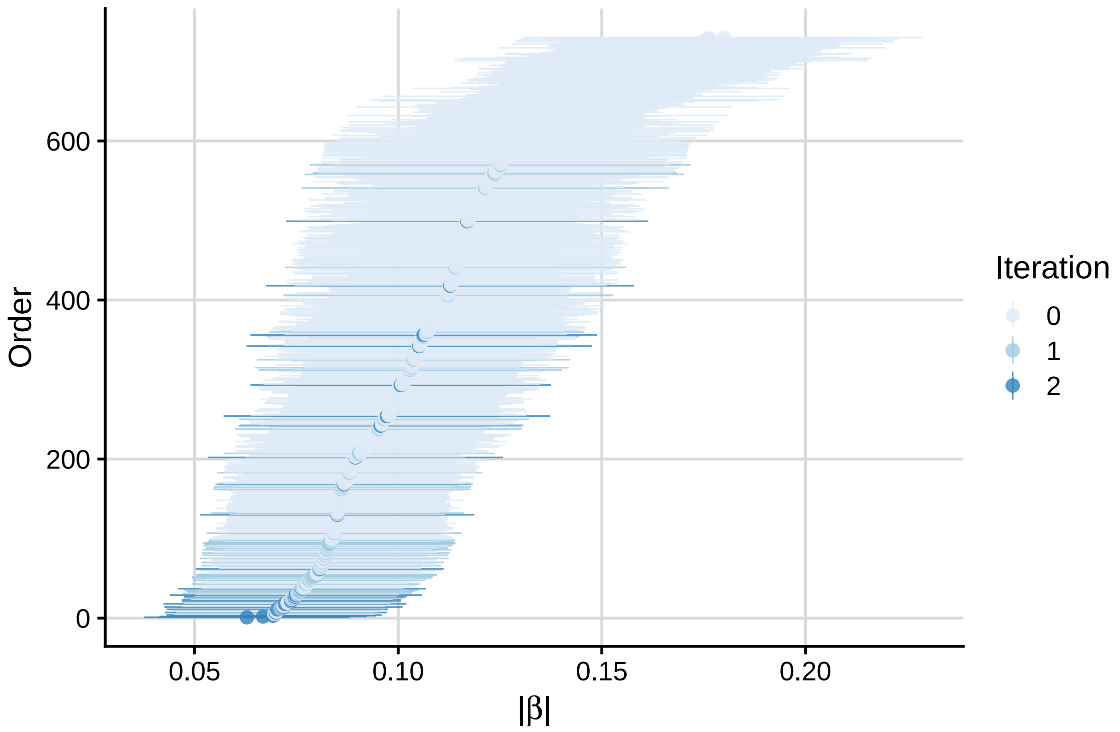


Fig. 2.17 Absolute estimated effect sizes ($|\beta|$; log OR) $+/-$ 1.96× standard error of SNPs significantly associated ($FDR \leq 0.000148249$) with asthma in the original discovery GWAS data set ("iteration 0") and those newly significant after iteration 1 and 2 of Flexible cFDR.

As a proxy for sensitivity, I calculated the proportion of FDR significant SNPs in the UK Biobank data set that were also found to be FDR significant both before ("iteration 0") and after each iteration of Flexible cFDR. I found that the sensitivity increased from 0.127 to 0.131 after iteration 1 (leveraging average H3K27ac fold change values in lymphoid and CD56 cell types) and to 0.133 after iteration 2 (leveraging average H3K27ac fold change values in lung tissue and monocytes) (Fig 2.16E). As a proxy for specificity, I calculated the proportion of SNPs not FDR significant in the UK Biobank data set that were also not FDR significant both before ("iteration 0") and after each iteration of Flexible cFDR, finding that the specificity remained close to 1 (≥ 0.9999975) (Fig 2.16F). I also found that the order of which I iterated over the auxiliary data had minimal impact on the results (Fig 2.18).

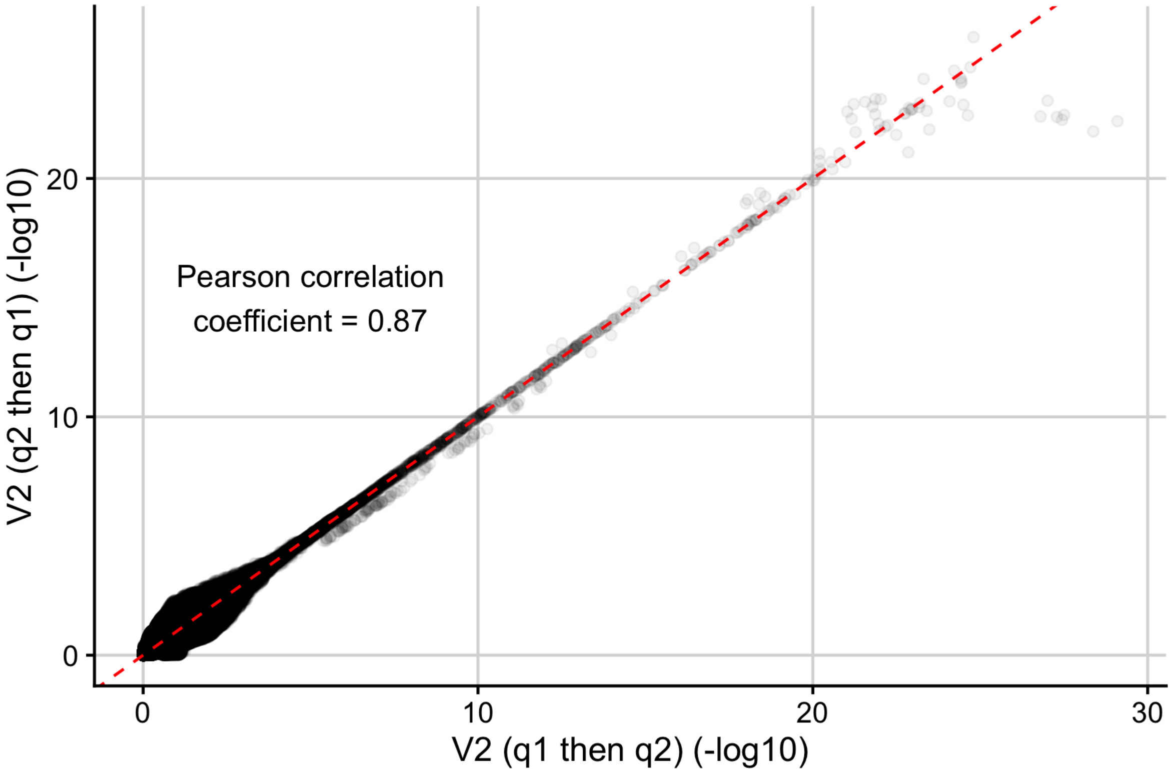


Fig. 2.18 ($-log_{10}$ transformed) v-values after 2 iterations of Flexible cFDR leveraging H3K27ac data when iterating over `q2` and then `q1` against ($-log_{10}$ transformed) v-values when iterating over `q1` then `q2`.

At the locus level, 18 loci were FDR significant in the original asthma GWAS (Demenais et al., 2018). Flexible cFDR identified four additional significant loci with lead variants: rs9501077 (chr6:31167512), rs4148869 (chr6:32806576), rs9467715 (chr6:26341301) and rs167769 (chr12:57503775) (Fig 2.19; Table 2.4). Three of the four (rs4148869, rs9467715 and rs167769) validated in the UK Biobank data set at Bonferroni corrected significance (for four tests the Bonferroni corrected significance threshold corresponding to $\alpha = 0.05$ is $0.05/4 = 0.0125$). One locus was found to be no longer FDR significant, with lead variant rs12543811 (chr8:81278885) (Fig 2.19).

SNPs rs9501077 and rs4148869 reside in the MHC region of the genome, which is renowned for its long-range LD structures that make it difficult to dissect genetic architecture in this region. SNPs rs9501077 and rs4148869 are in linkage equilibrium ($r^2 = 0.001$), and are in very weak LD with the lead variant for the whole MHC region (rs9268969 with $FDR = 7.35e - 15$; $r^2 = 0.005$ and $r^2 = 0.001$ with rs9501077 and rs4148869 respectively). SNP rs9501077 had relatively high H3K27ac counts in asthma-relevant cell types (mean percentile was 90th) and Flexible cFDR used this extra information to increase the significance of this SNP beyond the

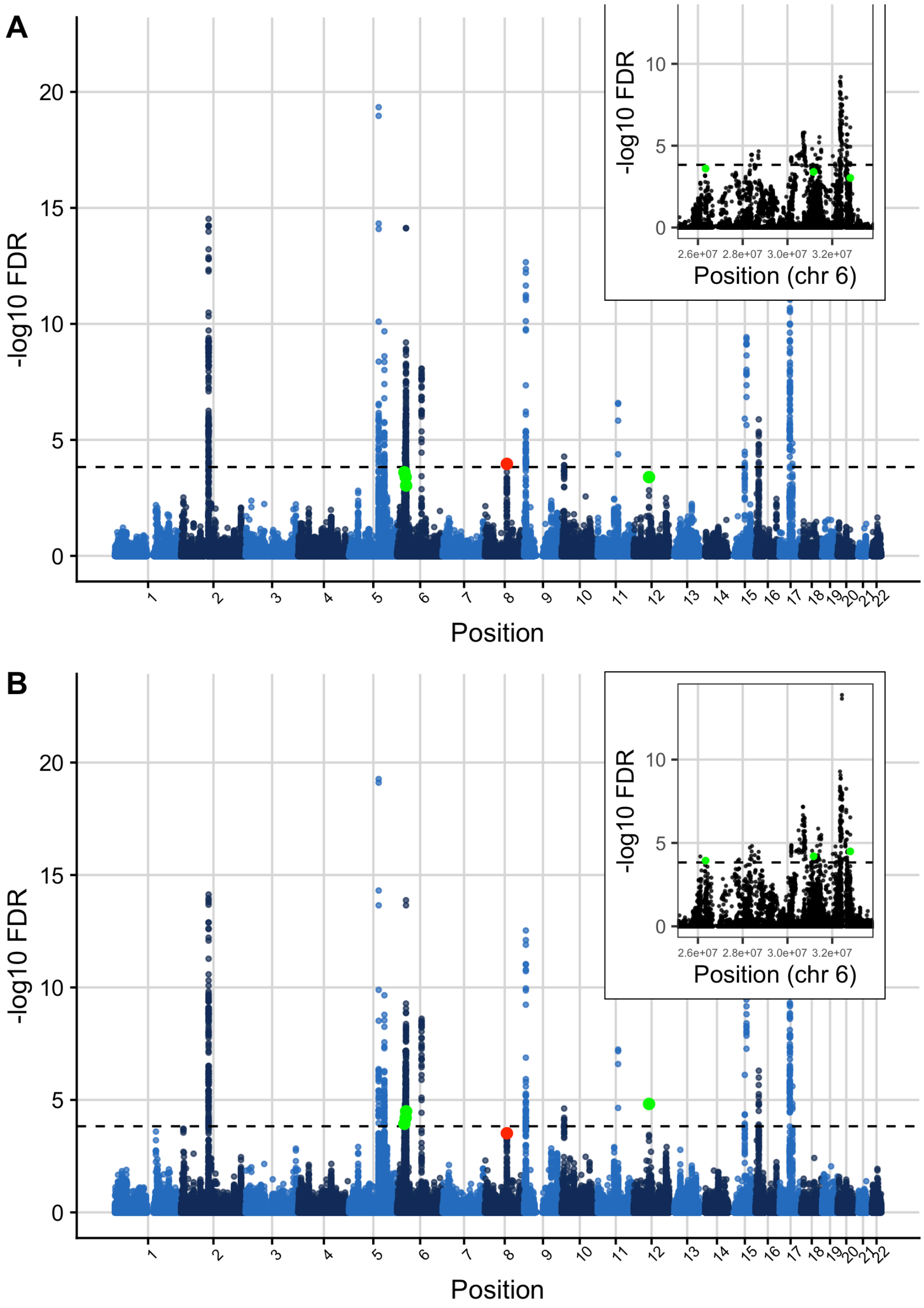


Fig. 2.19 Manhattan plots of $-log_{10}$ transformed FDR values (A) before and (B) after applying Flexible cFDR to leverage H3K27ac counts in asthma-relevant cell types. Points are coloured by chromosome and green points indicate the four lead variants that were identified as newly FDR significant after applying Flexible cFDR (rs167769, rs9467715, rs9501077 and rs4148869) whilst the red point indicates the single lead variant that was identified as newly not FDR significant after applying Flexible cFDR (rs12543811). Black dashed line at FDR significance threshold ($FDR = 0.000148249$). Zoomed in panel shows the three independent newly significant signals residing on chromosome 6.

significance threshold (FDR before Flexible cFDR = $3.99e-04$, FDR after Flexible cFDR = $6.26e-05$; Table 2.4). This SNP is found in the long non-coding RNA (lncRNA) gene, *HCG27* (HLA Complex Group 27), which has been linked to psoriasis (Villarreal-Martínez et al., 2016). However this SNP did not replicate in the UK Biobank data (UK Biobank $p = 0.020$).

SNP rs4148869 had very high H3K27ac fold change values in asthma-relevant cell types (mean percentile was 99.6th) and so Flexible cFDR decreased the FDR value for this SNP from $9.28e-04$ to $3.22e-05$ when leveraging this auxiliary data (Table 2.4). This SNP is a 5' UTR variant in the *TAP2* gene. The protein TAP2 assembles with TAP1 to form a "transporter associated with antigen processing" (TAP) complex. The TAP complex transports foreign peptides to the endoplasmic reticulum where they attach to MHC class I proteins, which are in turn trafficked to the surface of the cell for antigen presentation in order to initiate an immune response (Hewitt, 2003). Studies have found *TAP2* to be associated with various immune-related disorders, including autoimmune thyroiditis and T1D (Carvalho-Silva et al., 2019; Tomer et al., 2015), and also pulmonary tuberculosis in Iranian populations (Naderi et al., 2016). Recently, Ma et al. (2020) identified three cis-regulatory eSNPS for *TAP2* as candidates for childhood-onset asthma risk (rs9267798, rs4148882 and rs241456). One of these (rs4148882) was present in the asthma GWAS data set used in my analysis ($FDR = 0.12$) and was in weak LD with rs4148869 ($r^2 = 0.4$).

SNP rs9467715 is a regulatory region variant with a raw FDR value that was very nearly significant in the original GWAS (FDR = $2.49e-04$). This SNP had moderate H3K27ac fold change values in asthma-relevant cell types (mean percentile was 67.9th) so that when these values were leveraged using Flexible cFDR, the new FDR value for this SNP just exceeded the FDR significance threshold (FDR after Flexible cFDR = $1.15e-04$; Table 2.4).

SNP rs167769 had a borderline FDR value in the original GWAS discovery data set ($FDR = 4.04e-04$) but was found to be significant in the multi-ancestry analysis in the same manuscript ($FDR = 1.61e-05$) (Demenais et al., 2018). This SNP had very high H3K27ac fold change values in asthma-relevant cell types (mean percentile was 98.4th) and Flexible cFDR decreased the FDR value for this SNP to $1.51e-05$ when leveraging this auxiliary data (Table 2.4). rs167769 is an intron variant in *STAT6*, a gene that is activated by cytokines IL-4 and IL-13 (Takeda et al., 1996a,b) to initiate a Th2 response and ultimately inhibit the transcription of innate immune response genes (Albanesi et al., 2007; Ohmori and Hamilton, 2000). Transgenic mice over-expressing constitutively active *STAT6* in T cells are predisposed towards Th2 responses and allergic inflammation (Bruns et al., 2003; Kaplan et al., 2007) whilst *STAT6*-knockout mice are protected from allergic pulmonary manifestations (Kuperman et al., 2002). Accordingly, rs167769 is strongly associated with *STAT6* expression in the blood (Grundberg et al., 2012; Liang et al., 2013; Westra et al., 2013) and lungs (Hao et al., 2012) and is also associated with

SNP	Chr	BP	Ref	Alt	beta	SE	H3K27ac percentile	p	FDR (p)	p (UKBB)	v	FDR (v)
rs167769	12	57503775	C	T	7.87e-02	1.50e-02	98.4th	1.55e-07	4.04e-04	4.69e-24	3.75e-09	1.51e-05
rs9467715	6	26341301	T	C	-8.61e-02	1.61e-02	67.9th	8.96e-08	2.49e-04	5.93e-04	3.83e-08	1.15e-04
rs9501077	6	31167512	A	G	-8.06e-02	1.54e-02	90.5th	1.53e-07	3.99e-04	2.01e-02	1.91e-08	6.26e-05
rs4148869	6	32806576	C	T	7.03e-02	1.39e-02	99.6th	4.03e-07	9.28e-04	1.49e-15	9.07e-09	3.22e-05

Table 2.4 Summary of newly significant lead variants for asthma when using Flexible cFDR to leverage H3K27ac data. Details of lead variants that became newly FDR significant ($FDR \leq 0.000148249$) after using Flexible cFDR to leverage H3K27ac fold change values with asthma GWAS p-values. Table contains the rsIDs (SNP), genomic positions (Chr: chromosome, BP: base pair given in GRCh37/hg19), reference (Ref) and alternative (Alt) alleles, log ORs (beta), standard errors (SE) and p-values from the discovery GWAS (p), mean percentile of H3K27ac fold change values across asthma-relevant cell types, p-values from UK Biobank and resultant v-values from Flexible cFDR. For the original p-values (and v-values), the corresponding FDR values are also given, calculated using the BH procedure.

increased risk of childhood atopic dermatitis (Howell et al., 2011; Lee et al., 2015), which often progresses to allergic airways diseases such as asthma in adulthood. No genetic variants in the *STAT6* gene region (chr12:57489187-57525922) were identified as significant in the original GWAS, and only rs167769 was identified as significant after leveraging H3K27ac data using Flexible cFDR.

One lead variant that was significant in the original discovery data set was no longer significant after applying Flexible cFDR. SNP rs12543811 is located between genes *TPD52* and *ZBTB10* and had moderate H3K27ac fold change values in asthma-relevant cell types (mean percentile was 52th). This SNP only just exceeded the FDR significance threshold in the original GWAS (FDR $= 1.08e-04$) but by leveraging its H3K27ac fold change values using Flexible cFDR, the resultant FDR value fell just below the significance threshold (FDR after Flexible cFDR $= 3.04e-04$). This SNP is in strong LD with rs7009110 ($r^2 = 0.79$) which has previously been associated with asthma plus hay fever but not with asthma alone (Ferreira et al., 2014). Conditional analyses show that these two SNPs represent the same signal which is likely to be associated with allergic asthma (Demenais et al., 2018). SNP rs12543811 was found to be significant in the UK Biobank data (UK Biobank $p = 1.42e-19$).

BL

The estimated probabilities that each SNP was null (not associated) from BL ranged from 0.746 to 1 and were negatively correlated with H3K27ac fold change values in asthma-relevant cell types (Fig 2.20). In total, BL identified five SNPs as newly FDR significant, which replicated in the UK Biobank validation data set (rs4705950 UK Biobank $p = 7.2e-23$, rs9268831 UK Biobank $p = 1.3e-42$, rs17533090 UK Biobank $p = 4.4e-41$, rs1871665 UK Biobank $p = 5.5e-24$ and rs16924428 UK Biobank $p = 3.6e-40$) (Fig 2.21). These SNPs were a subset of the 69 newly significant SNPs that were identified by Flexible cFDR, except for rs16924428 which had very low H3K27ac fold change values in asthma-relevant cell types (mean percentile was 4.6th). The sensitivity increased slightly from 0.127 to 0.128 after applying BL (compared to 0.133 after Flexible cFDR) (Fig 2.21C) and the specificity remained stable at 0.9999995 (Fig 2.21D). No SNPs were found to be FDR significant in the discovery data set and no longer FDR significant after applying BL, and no new loci were found to be newly FDR significant (or newly not FDR significant) after applying BL.

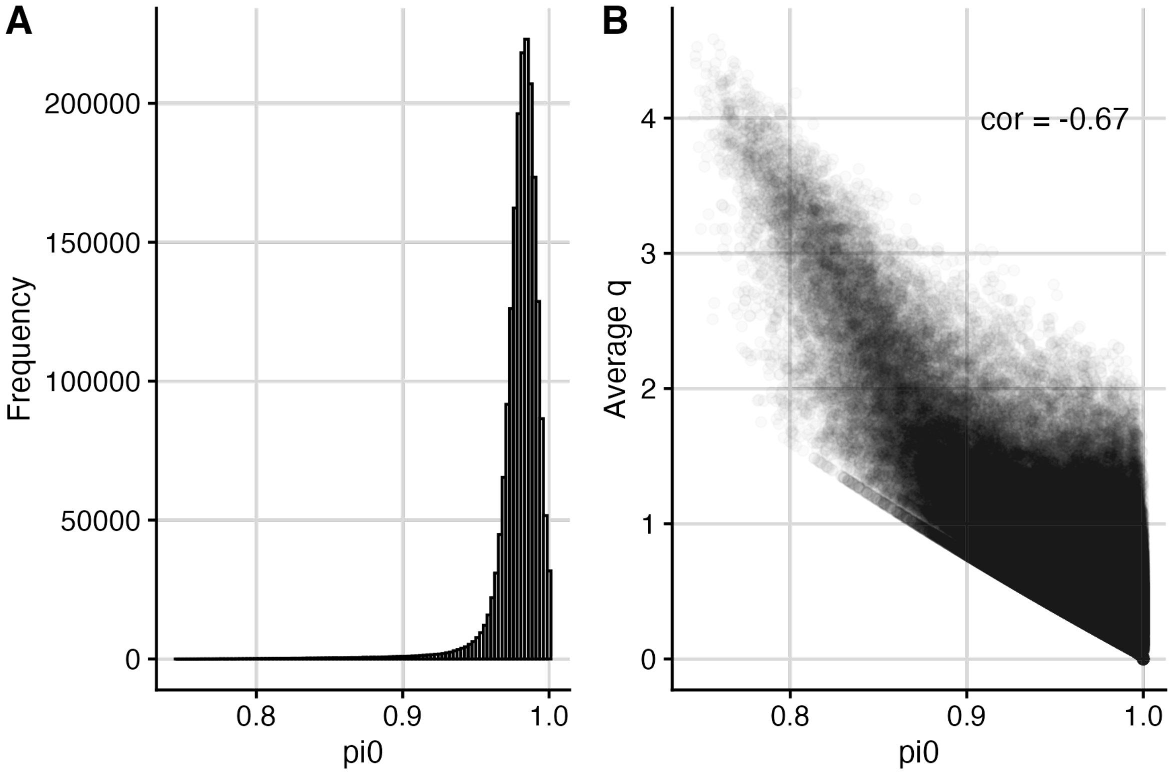


Fig. 2.20 (A) Histogram of estimated probabilities that the null hypothesis is true (“pi0”) for all 1,968,651 SNPs (B) Average H3K27ac fold change values in asthma-relevant cell types (q) against estimated probabilities. (The pi0 values are estimated as an intermediate step in BL.) “Cor” is the Pearson correlation coefficient between average q and pi0 values.

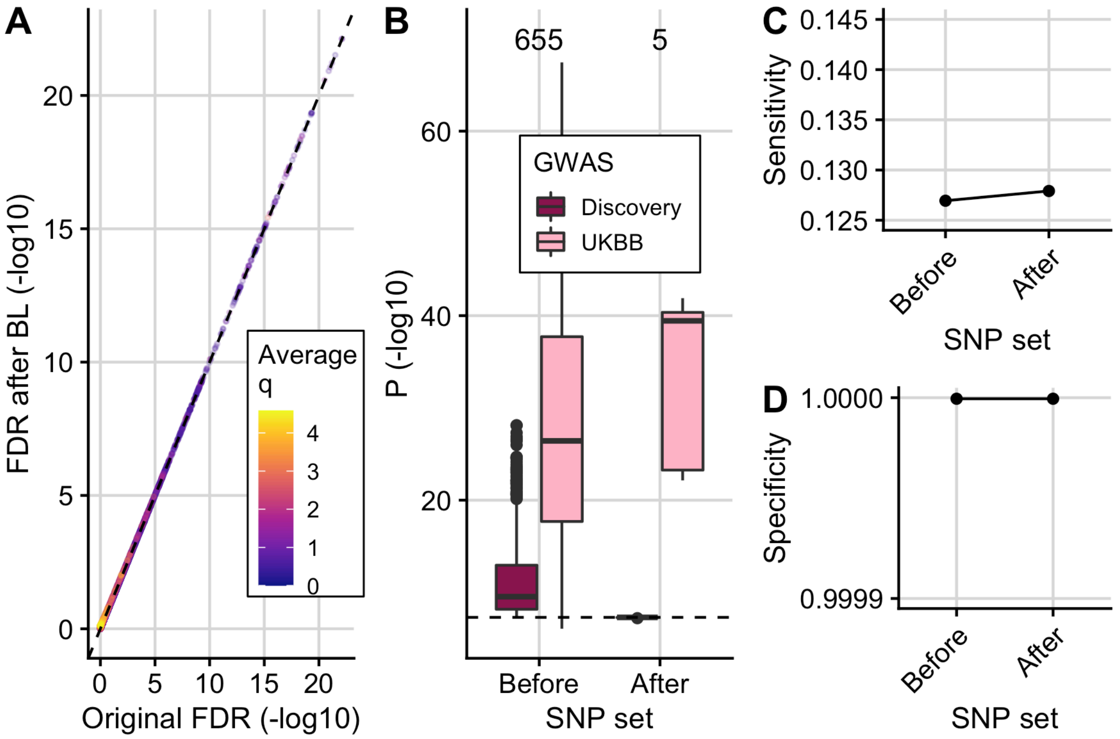


Fig. 2.21 Using Boca and Leek's FDR regression to leverage H3K27ac data with asthma GWAS p-values. (A) ($-log_{10}$ transformed) adjusted p-values from BL against raw ($-log_{10}$ transformed) FDR values coloured by average value of q (H3K27ac fold change value). (B) Box plots of ($-log_{10}$ transformed) p-values in the discovery GWAS data set and the UK Biobank data set for the 655 SNPs that were FDR significant in the original GWAS ("before") and five newly significant SNPs after applying BL ("after"). Black dashed line at genome-wide significance threshold ($5e-08$). (C) Sensitivity and (D) specificity proxies for the results. Sensitivity proxy was calculated as the proportion of SNPs that were FDR significant in the UK Biobank data set that were also FDR significant in the original GWAS or after applying BL. Specificity was calculated as the proportion of SNPs that were not FDR significant in the UK Biobank data set that were also not FDR significant in the original GWAS or after BL.

FINDOR

FINDOR identified 119 newly FDR significant SNPs which had a median p-value of $4.44e-15$ in the UK Biobank validation data, but the maximum UK Biobank p-value for these 119 SNPs was 0.98 (Fig 2.22A; Fig 2.22B). The proportion of FDR significant SNPs in the UK Biobank data set that were also FDR significant in the discovery GWAS data set increased from 0.127 to 0.146 (compared to 0.128 after BL and 0.133 after Flexible cFDR) (Fig 2.22C) and the specificity remained high (Fig 2.22D). The increase in sensitivity from FINDOR was greater than that of Flexible cFDR and BL, which may reflect the information gain in leveraging 96 annotations rather than a single histone mark.

At the locus level, FINDOR identified two newly FDR significant lead variants: rs13018263 (chr2:103092270; original $FDR = 6.79e-04$, new $FDR = 1.00e-04$) and rs9501077

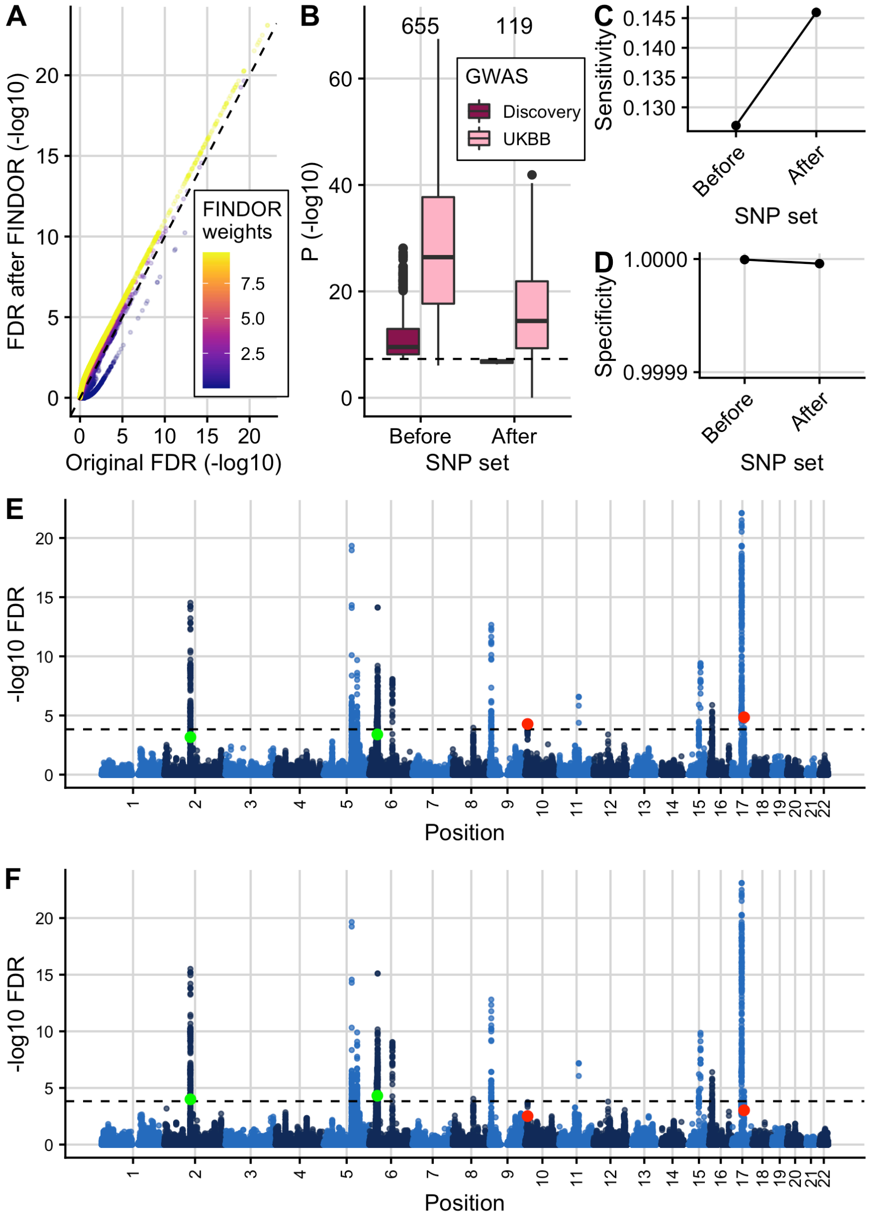


Fig. 2.22 (A) ($-log_{10}$ transformed) FDR values from FINDOR against ($-log_{10}$ transformed) original FDR values, coloured by FINDOR weights. (B) Box plots of ($-log_{10}$ transformed) p-values in the discovery GWAS data set and the UK Biobank data set for the 655 SNPs that were FDR significant in the original GWAS ("before") and 119 newly significant SNPs after re-weighting using FINDOR ("after"). Black dashed line at genome-wide significance threshold ($p \leq 5e-08$). (C) Sensitivity and (D) specificity proxies for the FINDOR results. Sensitivity proxy was calculated as the proportion of SNPs that were FDR significant in the UK Biobank data set that were also FDR significant in the original GWAS or after p-value re-weighting using FINDOR. Specificity was calculated as the proportion of SNPs that were not FDR significant in the UK Biobank data set that were also not FDR significant in the original GWAS or after p-value re-weighting using FINDOR. Manhattan plots of FDR values before (E) and after (F) re-weighting by FINDOR. Green points indicate the two lead variants that were newly identified as FDR significant by FINDOR [rs13018263 (chr2:103092270) and rs9501077 (chr6:31167512)]. Red points indicate the two lead variants that were newly identified as not FDR significant by FINDOR [rs2589561 (chr10:9046645) and rs17637472 (chr17:47461433)]. Black dashed line at FDR significance threshold ($FDR = 0.000148249$).

(chr6:31167512; original $FDR = 3.99e-04$, new $FDR = 4.86e-05$) (Fig 2.22E; Fig 2.22F). SNP rs13018263 is an intronic variant in *SLC9A4* and was strongly significant in the UK Biobank validation data set ($p = 4.78e-31$). Ferreira et al. (2017) highlighted rs13018263 as a potential eQTL for *IL18RAP*, a gene which is involved in IL-18 signalling which in turn mediates Th1 responses (Hedl et al., 2014), and is situated just upstream of *SLC9A4*. Genetic variants in *IL18RAP* are associated with many immune-mediated diseases, including atopic dermatitis (Hirota et al., 2012) and T1D (Smyth et al., 2008). Interestingly, although different auxiliary data was leveraged using Flexible cFDR and FINDOR in our analyses, both methods found lead variant rs9501077 to be newly significant, but this SNP did not validate in the UK Biobank data (UK Biobank $p = 0.020$).

Two additional lead variants were found to be no longer significant after re-weighting by FINDOR: rs2589561 (chr10:9046645; original $FDR = 5.25e-05$, new $FDR = 3.06e-03$) and rs17637472 (chr17:47461433; original $FDR = 1.42e{-}05$, new $FDR = 9.42e{-}04$). However, both of these SNPs were strongly significant in the UK Biobank validation data set ($p = 2.09e-29$ and $p = 1.75e-14$ respectively).

SNP rs2589561 resides in a gene desert that is 929-kb from *GATA3*. *GATA3* encodes a transcription factor that is involved in the Th2 pathway, which mediates the immune response to allergens (Demenais et al., 2018; Shrine et al., 2019). Hi-C data in hematopoietic cells showed that two proxies of rs2589561 ($r^2 > 0.9$) are located in a region that interacts with the *GATA3* promoter in CD4 T cells (Javierre et al., 2016), suggesting that rs2589561 could function as a distal regulator of *GATA3* in this cell type (which is relevant to asthma). SNP rs2589561 had relatively high H3K27ac fold change values in the asthma-relevant cell types that were leveraged by Flexible cFDR (mean percentile was 85th) and Flexible cFDR decreased the FDR value from $5.25e-05$ to $2.41e-05$.

SNP rs17637472 is a strong cis-eQTL for *GNGT2* in whole blood (Grundberg et al., 2012; Liang et al., 2013; Westra et al., 2013; Zeller et al., 2010). *GNGT2* encodes a protein that is involved in NF-κB activation (Vibhuti et al., 2011). This SNP had moderate H3K27ac fold change values in relevant cell types (mean percentile was 62th) and the FDR values for this SNP were similar both before and after using Flexible cFDR to leverage the H3K27ac data (original $FDR = 1.42e-05$, new $FDR = 1.46e-05$).

2.10 Application 3: Leveraging a variety of relevant data with T1D GWAS data

In a final all-encompassing example, I used Flexible cFDR and Binary cFDR iteratively to leverage a variety of relevant genetic and genomic data with GWAS p-values for T1D.

2.10.1 Methods

T1D GWAS data

I downloaded GWAS summary statistics for T1D from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) for study accession GCST005536 (Onengut-Gumuscu et al., 2015) on 2020-01-10. Onengut-Gumuscu et al. (2015) genotyped 6,670 T1D cases and 12,262 controls using the Immunochip, and I called the data from this study my "discovery GWAS data set". To find an independent subset of SNPs, I used the LDAK software to obtain an LDAK weight for each SNP and defined my independent SNP set as the SNPs given a non-zero LDAK weight. For the MAF matching step described in section 2.3.4, I used MAFs estimated from the CEU sub-population of the 1000 Genomes Project phase 3 data set (The 1000 Genomes Project Consortium, 2015).

I used data from release 5 of the FinnGen project (Borodulin et al., 2018; FinnGen, 2021) as a validation data set. I downloaded GWAS summary statistics for the "T1D strict definition" endpoint (6,692 T1D cases and 212,100 controls) from https://storage.googleapis.com/finngen-public-data-r5/summary_stats/finngen_R5_T1D_STRICT.gz and matched these to SNPs from my discovery GWAS data set by rsID.

To identify independent loci for locus-level results, I used `PLINK 1.9` LD-clumping algorithm with default parameter values, using haplotype data from the 503 individuals of European ancestry from 1000 Genomes Project phase 3 as a reference panel to calculate LD between SNPs.

Auxiliary data

I downloaded GWAS p-values for rheumatoid arthritis (RA) from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) for study accession GCST005569 (Eyre et al., 2012) on 2020-09-10. I mapped each SNP in my T1D GWAS data set to its p-value for RA (using genomic positions and rsIDs) to generate `q1`.

I downloaded SNP-level annotations for all 1000 Genomes SNPs from the baseline-LD model (version 2.2) described in Gazal et al. (2017). I used the binary annotation "DGF_ENCODE" which quantifies sites of regulatory factor occupancy. Briefly, this annotation is derived from merging all DNase I digital genomic footprinting (DGF) regions from the narrow-peak classifications across 57 cell types (ENCODE Project Consortium, 2012; Gusev et al., 2014). DGF regions (corresponding to DGF annotation values of 1) are expected to precisely map sites where regulatory factors bind to the genome (Neph et al., 2012b). I matched each SNP in my T1D GWAS data set to its binary DGF annotation (using genomic positions) to generate `q2`.

I downloaded consolidated fold-enrichment ratios of H3K27ac ChIP-seq counts relative to expected background counts from NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) in naive CD4 T helper cells (believed to be a key cell type involved in T1D pathogenesis) from https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E043-H3K27ac.fc.signal.bigwig. I mapped each SNP in my T1D GWAS data set to its corresponding genomic region and recorded the H3K27ac fold change value. For SNPs on the boundary of a genomic region (and therefore mapping to two regions) I randomly selected a fold change value from one of the regions. I transformed the fold change values ($q := log(q + 1)$) to deal with long tails and consequently derived `q3`.

Implementation of methods

I used the `fcfdr::flexible_cfdr` and `fcfdr::binary_cfdr` functions to leverage the auxiliary data with T1D GWAS p-values iteratively. For the `group` parameter in `fcfdr::binary_cfdr` I used the chromosome for which each SNP resided, and I used the optional `maf` parameter in the `fcfdr::flexible_cfdr` function to supply the MAF values that are required for the MAF matching procedure. I used the `stats::p.adjust` function with `method="BH"` to derive FDR values from the v-values (after the 3 iterations) and used these as the output of interest.

For BL, I used the `lm_qvalue` function in the `swfdr` Bioconductor R package (version 1.16.0) to derive adjusted p-values and used these as the output of interest (the covariate matrix consisted of three columns for `q1`, `q2` and `q3`). I used an FDR threshold of $FDR \leq 5e-06$ to call significant SNPs, which roughly equates to the genome-wide significance threshold $p \leq 5e-08$ ($FDR = 5e-06$ was rounded up from $4.5e-06$ which was the maximum FDR value amongst SNPs with raw p-values $\leq 5e-08$ in the discovery GWAS data set).

The other comparator methods that I considered could not be applied here since GenoWAP, Empirical cFDR and FINDOR do not support auxiliary data of this type, and IHW cannot be applied to multi-dimensional auxiliary data.

2.10.2 Results

At the SNP level, my cFDR approach identified 58 SNPs as newly FDR significant ($FDR \leq 5e-06$) (Fig 2.23). These had relatively small p-values for RA (median $p = 0.002$ compared with median $p = 0.423$ in full data set), were more likely to be found in regulatory factor binding sites (mean binary value was 0.345 compared to 0.23 in full data set) and had larger H3K27ac fold change values in naive CD4 T helper cells (median fold change relative to expected background counts was 1.76 compared with 0.97 in full data set). Similarly, 44 SNPs were identified as newly not FDR significant which had relatively high p-values for RA (median $p = 0.631$), were less likely to be found in regulatory factor binding sites (mean binary value was 0.182) and had smaller H3K27ac fold change values in naive CD4 T helper cells (median fold change relative to expected background counts was 0.447). Reassuringly, the 58 SNPs that were found to be newly FDR significant had smaller p-values in the FinnGen validation data set (median $p = 0.023$) than the 44 SNPs found to be newly not significant by the method (median $p = 0.353$).

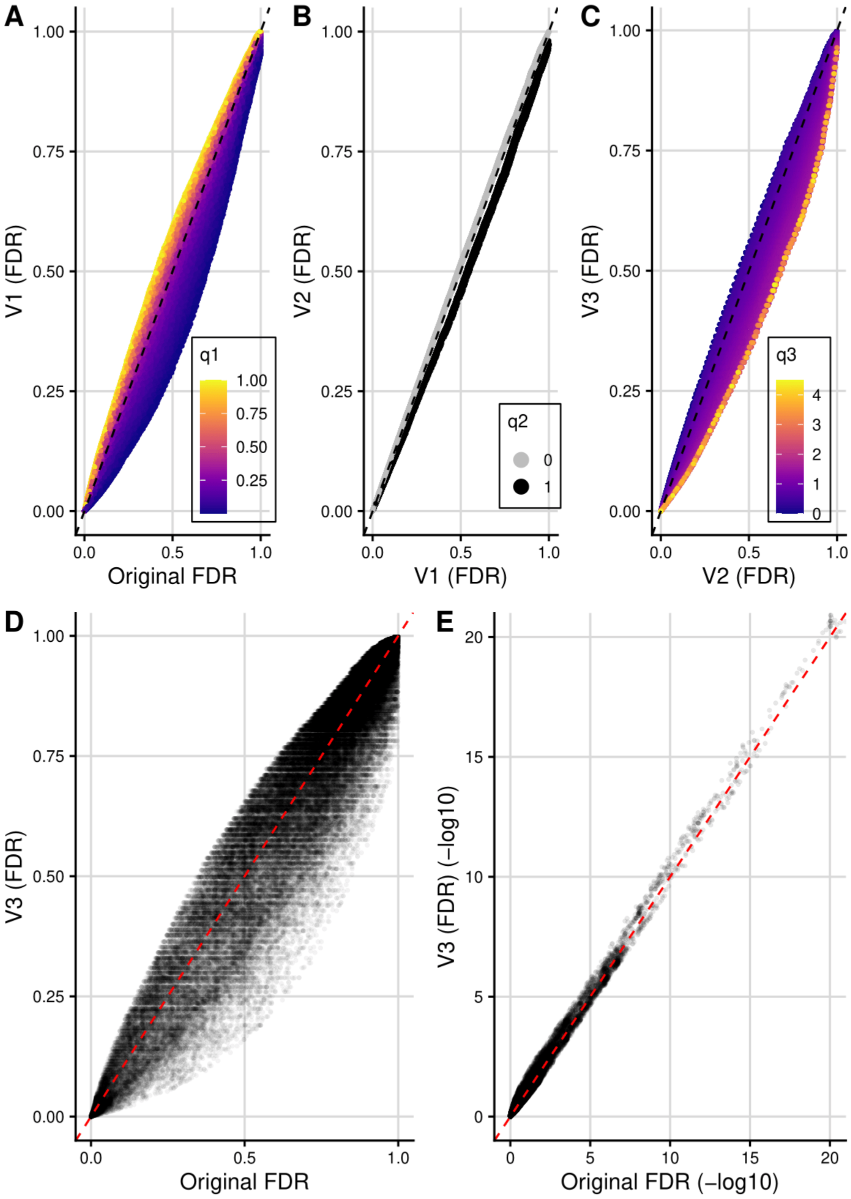


Fig. 2.23 Summary of cFDR results for T1D application. Top panel shows FDR values before and after each iteration: (A) for iteration 1 (B) for iteration 2 and (C) for iteration 3. Points are coloured by auxiliary data leveraged in that iteration (q1 is p-values for rheumatoid arthritis, q2 is a binary indicator of general regulatory factor binding sites and q3 is H3K27ac count data). Bottom panel shows final FDR values after the three iterations against original FDR values (axes truncated in panel (E)).

Based on my definition of genomic loci, the original GWAS identified 154 loci with a significant lead variant. My implementation of cFDR identified an additional four loci with a newly significant lead variant (rs2386841, rs12150079, rs1893592 and rs7839768) (Fig 2.24). These SNPs had relatively small p-values for RA (median $p = 0.002$) and had larger H3K27ac fold change values in naive CD4 T helper cells (median fold change relative to expected background counts was 1.83). However, only one of the four lead variants (rs12150079) overlapped a regulatory factors binding site. The nearest genes to these newly significant lead variants were *IL2RA* (for rs2386841), *ZPBP2* (for rs12150079), *UBASH3A* (for rs1893592) and *TMEM68* (for rs7839768). Genes *IL2RA*, *ZPBP2* and *UBASH3A* have previously been linked to T1D (Garg et al., 2012; Ge et al., 2017; Syreeni et al., 2021) but the lead variants did not validate in the FinnGen validation data set (rs2386841 Finngen $p = 0.165$, rs12150079 Finngen $p = 0.748$, rs1893592 Finngen $p = 0.0004$ and rs7839768 Finngen $p = 0.642$).

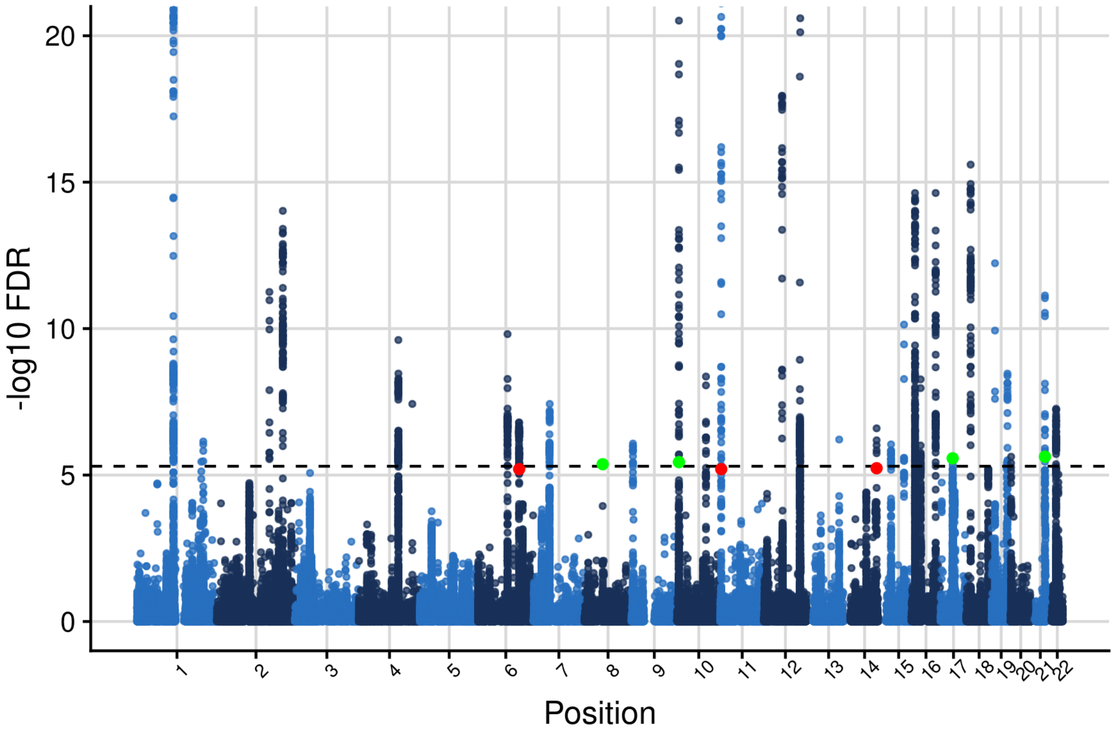


Fig. 2.24 Manhattan plot of ($-log_{10}$ transformed) FDR values after using Flexible cFDR and Binary cFDR to leverage various types of auxiliary data with T1D GWAS p-values. Green points indicate the four lead variants that were newly FDR significant after cFDR and red points indicate the three lead variants that were newly not FDR significant after cFDR. Black dashed line at FDR significance threshold ($FDR = 5e-06$). The y axis has been truncated to aid visualisation.

Similarly, my implementation of cFDR identified three lead variants which were no longer significant (rs2651830, rs2400941 and rs853970) (Fig 2.24). These SNPs had relatively high

p-values for RA (median $p = 0.44$) and had smaller H3K27ac fold change values in naive CD4 T helper cells (median fold change was 0). But two of the three lead variants (rs2400941 and rs853970) were predicted to overlap regulatory factor binding sites. These results suggest that the continuous auxiliary data was more informative than the binary data leveraged in this example. These SNPs had reassuring large p-values in the FinnGen validation data set (rs2651830 Finngen $p = 0.102$, rs2400941 Finngen $p = 0.015$, and rs853970 Finngen $p = 0.361$).

At the SNP level, BL identified 43 SNPs as newly FDR significant (Fig 2.25). These SNPs had slightly smaller p-values for RA (median $p = 0.279$ compared with median $p = 0.423$ in full data set), were more likely to be found in regulatory factor binding sites (mean binary value was 0.35 compared to 0.23 in full data set) and had slightly larger H3K27ac fold change values in naive CD4 T helper cells (median fold change relative to expected background counts was 1.34 compared with 0.97 in full data set). These SNPs had relatively small p-values in the FinnGen data set used for validation (median Finngen $p = 0.0062$) and 21 of the 43 SNPs (48.9%) were also found to be newly FDR significant in the cFDR analysis. No SNPs were identified as newly not FDR significant after applying BL (Fig 2.25). At the locus level, one lead variant was found to be newly significant (rs6043409; missense variant in *SIRPG*) but this did not validate in the Finngen data set (Finngen $p = 0.02381$).

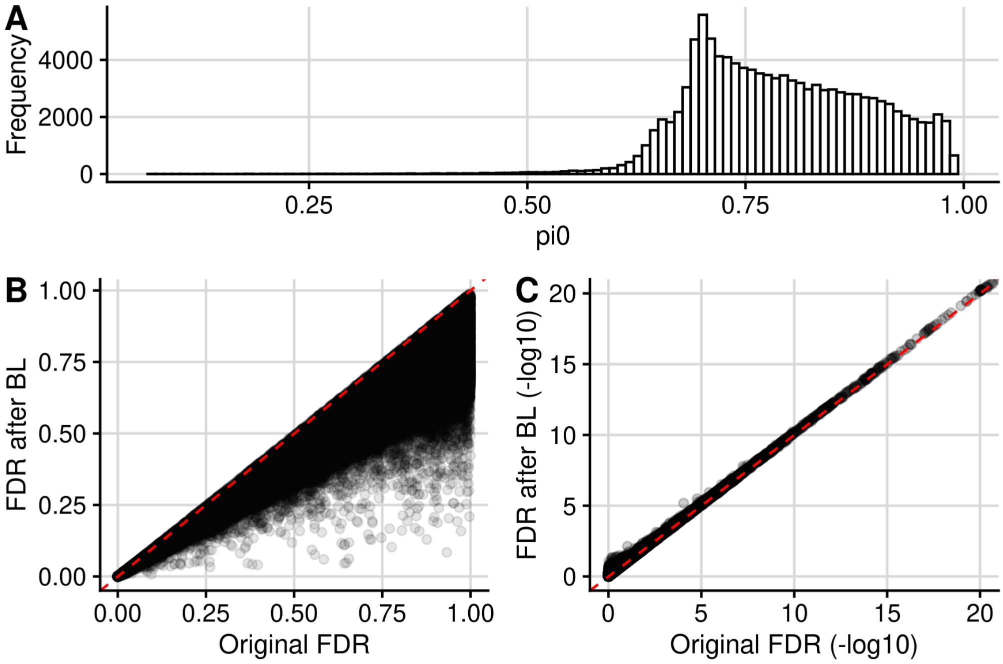


Fig. 2.25 Summary of BL results for T1D application. (A) Histogram of estimated probabilities that the null hypothesis is true (pi0) (B) FDR values after applying BL against original non-functionally informed FDR values (C) same as (B) but $-log_{10}$ transformed data (axes truncated).

2.11 Discussion

Developments in molecular biology have enabled researchers to decipher the functional effects of various genomic signatures. We are now in a position to prioritise sequence variants associated with various phenotypes not just by their genetic association statistics but also based on our biological understanding of their functional role. Binary cFDR and Flexible cFDR provide statistically robust frameworks to leverage functional genomic data with genetic association statistics to boost power for GWAS discovery.

Where appropriate, I compared the performance of Flexible cFDR and Binary cFDR to that of four comparator methods: GenoWAP, IHW, BL and FINDOR. I also tried to compare results from Flexible cFDR to those from AdaPT (Lei and Fithian, 2018), which is a popular statistical method for multiple hypothesis testing with auxiliary information. AdaPT uses a p-value masking procedure which takes many iterations of optimisation and can be computationally expensive (Zhang et al., 2019). Consequently, I found AdaPT to be too computationally demanding for large-scale GWAS data, and previous studies suggest that a SNP pre-filtering stage is required (Yurko et al., 2020). I did not include a comparison with AdaPT as it would be too difficult to compare findings. Indeed, one of the benefits of cFDR (and the other comparator methods that I considered) is that it is computationally efficient enough to be run genome-wide.

Of the methods considered, I found that only BL was as versatile as my cFDR frameworks. Specifically, IHW currently only supports univariate covariates and, unlike Flexible cFDR and Binary cFDR, cannot be applied iteratively to leverage multi-dimensional covariates. In GenoWAP, the prior probabilities that are used in the model are calculated as the mean GenoCanyon score (or tissue-specific GenoSkyline (Lu et al., 2016a) or GenoSkyline-Plus (Lu et al., 2017) score) of the surrounding $10,000$-bp, thereby restricting its utility to leveraging only these scores (which I found were unlikely to capture enough disease-relevant information to substantially alter the conclusions from a study). In FINDOR, SNPs are binned based on how well they tag heritability enriched categories and this requires the estimation of χ^2 statistics (i.e. the tagged variance) for each SNP using a range of functional annotations, which are generally those in the baseline-LD model. Users are thus required to run LD-score regression prior to running FINDOR, and this two-step approach may limit the accessibility of the method. Whilst I found that BL was as versatile as my cFDR frameworks, I found that it failed to control the frequentist FDR in some simulations, and that it was also less powerful than cFDR. Whilst FINDOR was shown to be the most powerful method in application 2, this may reflect the information gain in leveraging 96 annotations rather than a single histone mark. Although this emphasises the importance of being able to iterate over different auxiliary measures, and suggests that a fruitful area of extension for cFDR would be to increase the robustness of FDR control for dependent q across multiple iterations.

Flexible cFDR and Binary cFDR have several key advantages over competing methods. They do not bin variables and do not rely on subjective thresholding or normalised weighting schemes, which hinder many of the existing methods (Benjamini et al., 2006; Bourgon et al., 2010; Genovese et al., 2006; Hu et al., 2010; Ignatiadis et al., 2016; Kichaev et al., 2019; Lu et al., 2016b; Sun et al., 2006). Whilst LD between SNPs is often a concern (e.g. because methods such as KDE assume independence between observations), in Flexible cFDR I fit the KDE to a subset of LD-independent SNPs but then generate v-values for the full set of SNPs, thereby benefiting from computational efficiency but also facilitating downstream analyses which typically require the full set of SNPs, such as fine-mapping or meta-analysis. In Binary cFDR, I use a leave-one-out procedure to ensure that rejection rules are not applied to the same data on which those rules were determined, although investigating the impact of shorter range LD between SNPs in Binary cFDR is an area for future study. LD means that the v-values will be positively correlated, so I appeal to the established robustness of the BH FDR estimation to positive dependency (Benjamini and Yekutieli, 2001).

Whilst larger case and control cohort sizes will also boost statistical power for GWAS discovery, incorporating functional data provides an additional layer of biological evidence that an increase in sample sizes alone cannot provide. There are also instances in the rare disease domain where case sample sizes are restricted by the number of cases that are available for recruitment, for example in primary immunodeficiency disorder (Thaventhiran et al., 2020). My method has potential utility in these instances as it provides an alternative approach to increase statistical power. The choice of functional data to use in the approach may be guided by prior knowledge, or in a data driven manner using a method such as GARFIELD (Iotchkova et al., 2019) to quantify the enrichment of GWAS signals in different functional marks. Moreover, my method intrinsically evaluates the relevance of the auxiliary data by comparing the joint probability density of the test statistics and the auxiliary data to the joint density assuming independence, and can therefore be used to inform researchers of relevant functional signatures and cell types.

I detail four key advances enabling the extension of the cFDR framework to the functional genomics setting. Firstly, in Flexible cFDR I derive an estimator based on a 2-dimensional KDE of the bivariate distribution rather than empirical estimates, making my method considerably faster than earlier empirical approaches. Secondly, the cFDR framework requires the estimation of $q|H_0^p$, which Liley and Wallace (2021) approximate by $q|p > 1/2$. In contrast, Flexible cFDR utilises the local FDR to empirically evaluate the influence of specific p-value quantities on the null hypothesis and uses these in the estimation of $q|H_0^p$. Binary cFDR approximates $q|H_0^p$ by $q|p > 1/2$ and appears robust to this approximation (perhaps due to the fact that q can only take on two values), but examining this in more detail is an area for future study. Thirdly, in Flexible cFDR I remove the assumption that $q|H_0^p$ can be transformed to a mixture of centred Gaussians, and instead integrate over the previously estimated KDE, which relaxes the distributional assumptions placed on the auxiliary data. Finally, Flexible cFDR and Binary cFDR are supported by user-oriented software documented on an easy-to-navigate website (https://annahutch.github.io/fcfdr/). The website features several fully reproducible vignettes which illustrate how the method can be applied to a particular data set at the desired level of error control.

One can see that the scale on which the continuous auxiliary data is measured may impact the performance of Flexible cFDR. Usual concerns about KDE apply, including that fits may be poor if there are regions with very sparse data. The optimal scale for the auxiliary data is likely to depend on the relationship between the principal p-values and the auxiliary data, and is not something I have explored here, but as usual, data visualisation is likely to be helpful to confirm that the scale for the auxiliary data is sensible. By default, the Flexible cFDR software returns a plot of the fitted 2D KDE and the estimated density of the auxiliary data overlaid onto the real data values, enabling users to visually examine the fit of the KDE to their data.

My cFDR frameworks have several limitations. Firstly, care must be taken to ensure that the auxiliary data to be leveraged iteratively is capturing distinct disease-relevant features to prevent multiple adjustment using the same auxiliary data. The definition of "distinct disease-relevant features" to leverage is at the user's discretion and sparks an interesting philosophical discussion. For example, leveraging data iteratively from various genomic assays measuring the same genomic feature at different resolutions may be deemed invalid for some researchers but valid for others, since if the mark is repeatedly identified by different assays then it is more likely to be reliably present. Whilst I show that my method is robust to minor departures from $q_i^{(k)} \perp\!\!\!\perp q_i^{(l)} | H_0^p$, this does not extend to strongly related q. I would argue that the conservative approach would be to average over correlated auxiliary data, to ensure that the q vectors are not strongly correlated.

Secondly, for continuous auxiliary data the cFDR framework assumes a positive stochastically monotonic relationship between the test statistics and the auxiliary data: specifically, low p-values are enriched for low values in the auxiliary data. Flexible cFDR automatically calculates the correlation between p and q and if this is negative then the auxiliary data is transformed to $q := -q$. However, if the relationship is non-monotonic (for example low p-values are enriched for both very low and very high values in the auxiliary data) then the cFDR framework cannot simultaneously shrink v-values for these two extremes. This non-monotonic relationship is unlikely when leveraging single functional genomic marks, but may occur if, for example, multiple marks were decomposed via PCA. I therefore recommend that users use the `corr_plot` and `stratified_qqplot` functions in the `fcfdr` R package to visualise the relationship between the two data types. Note that this restriction could be removed if I used density instead of distribution functions, and worked at the level of local FDR (Efron, 2004) as described earlier, but this would in turn reduce the robustness my method has to data sparsity in the (p, q) plane.

Finally, in application 3 many of the findings from cFDR did not validate in the Finngen data set. This could be due to the similar case sample sizes in the discovery GWAS data set (6670 T1D cases) and the Finngen data set (6692 T1D cases). The findings from application 3 should be further scrutinised as larger T1D GWAS data sets become available.

I have described novel implementations of the cFDR framework that support a wide variety of auxiliary data types. I have also introduced an R package to implement the approaches and have described three applications to demonstrate their versatility. I hope that my extensions to cFDR and the accompanying software will be a useful and practical tool, enabling researchers to boost power for GWAS discovery by leveraging various types of auxiliary data.

Chapter 3

Credible set coverage in genetic fine-mapping

This chapter examines genetic fine-mapping, an approach that is used to pinpoint the specific sequence variants that are causal for a phenotype of interest. I begin by systematically analysing the empirical calibration of an existing fine-mapping approach, and go on to develop an adjustment to improve the empirical calibration of inferences from the approach. In practise, my method is shown to narrow down the set of prioritised sequence variants that are likely to be causal, which in turn enables efficient allocation of resources in the expensive functional follow-up studies that are used to elucidate both the true causal variants and the associated biological mechanisms. This is an edited version of Hutchinson et al. (2020b) and Hutchinson et al. (2020a) and there is significant textual overlap in all sections of this chapter. I wrote the first draft of both of those papers and the text reused is my own work.

3.1 Introduction

At the time of this research, the fine-mapping method described by Maller et al. (2012) (which assumes a single causal variant per associated region) to construct credible sets of putative causal variants was the dominant fine-mapping method, and the construction of credible sets lay the framework for current state-of-the-art fine-mapping methods. The robust Bayesian methodology underpinning the approach enabled intuitive quantification of the resultant credible sets, whereby they can be interpreted as containing the true causal variants with some coverage probability. Whilst coverage is typically understood to refer to that probability taken over all possible realisations of the data, fine-mapping is conventionally only performed on a subset of genomic regions that contain at least one SNP with a p-value of association below some threshold (for example the genome-wide significance threshold). I therefore define "conditional coverage" as the probability that the causal variant is contained in the credible set, conditional on some aspect of the observed data, for example that it has been selected for fine-mapping.

In the literature, the conditional coverage estimates for these Bayesian credible sets are based on frequentist properties that they are not designed to satisfy. For example, researchers often state that a $(100 \times \alpha)\%$ credible set contains the causal variant with $(100 \times \alpha)\%$ probability (the "threshold coverage estimate") (Demontis et al., 2019; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, 2014; Fritsche et al., 2016; Gormley et al., 2016; The EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, 2015) or with probability $\geq (\alpha \times 100)\%$ (Huang et al., 2017; The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, 2015; Wang et al., 2020). More specifically, they may be interpreted as containing the causal variant with probability equal to the sum of the posterior probabilities of the variants in the credible set (the "claimed coverage estimate") (Spain and Barrett, 2015), for which the threshold forms a lower bound. Previously, van de Bunt et al. (2015) found that the conditional coverage of the causal variant in credible sets varied with the power to detect the signal (S1 Fig in van de Bunt et al. (2015)). This implied that inferring the frequentist conditional coverage estimates of these Bayesian credible sets may not be as straightforward as the literature suggests, but no formal statistical analysis into the validity of these conditional coverage estimates had been conducted at the time of this research.

3.2 Simulation method

3.2.1 Constructing genomic loci

To simulate GWAS summary statistics for fine-mapping, I constructed genomic regions for a variety of LD patterns using haplotypes from the 1000 Genomes Project phase 3 data set (The 1000 Genomes Project Consortium, 2015) and the UK10K Project (REL-2012-06-02) (The UK10K Consortium, 2015). I constructed low, medium and high LD regions using the 1000 Genomes data. For the low and medium LD regions, I used African ("AFR") and European ("EUR") haplotypes, respectively, in the genomic region chr10:6030000-6220000. For the high LD region I used European haplotypes in the genomic region chr10:60969-431161. All genomic co-ordinates in this chapter are given with respect to human build GRCh37/hg19. Using template code from my supervisor, Dr Chris Wallace, I obtained the reference haplotype data in these genomic regions and calculated the MAF for each SNP, subsequently removing low frequency SNPs with $MAF < 0.01$ (to match the convention that genetic association studies identify common genetic variation). This resulted in genomic regions containing 714 SNPs in the low and medium LD simulations and 726 SNPs in the high LD simulations. I computed the squared Pearson correlation coefficient between each genetic variant to obtain the LD matrix.

To ensure that any findings were robust to different LD structures and haplotype measurements derived from a larger number of individuals, I also simulated genomic regions for a wider variety of LD patterns using the UK10K data (the UK10K data is for 3781 European individuals, whereas the 1000 Genomes data is for 503 European individuals). Using template code from my supervisor, I randomly sampled LD blocks on chromosome 22 bounded by recombination hotspots, which were defined using the LD detect method (Berisa and Pickrell, 2016), subsequently removing low frequency SNPs as above. I then randomly selected a starting point within the LD block and selected 100 adjacent genetic variants. To include both more correlated and less correlated genetic variants in the analysis, I also selected a second starting point from the other end of the LD block and selected an additional 100 adjacent genetic variants. I then proceeded as above to obtain the LD matrices for these 200 SNP regions with differing LD structures.

The results that I present herein are based on simulations using the 1000 Genomes data, unless stated otherwise.

3.2.2 Simulating GWAS summary statistics

I used the `simGWAS` R package (Fortune and Wallace, 2019) to simulate Z-scores for SNPs in the genomic regions described above. To ensure that my simulations were representative of real fine-mapping analyses, I randomly selected the number of cases and the number of controls from 5000, 10000 or 50000 and randomly selected a variant to be causal from all of the variants in the region. For the haplotype frequency parameter, I supplied a `data.frame` of haplotypes, with a column of computed frequencies for each haplotype.

I selected the log OR at the causal variant, β, using several methods. Firstly, and so that my simulations reflected the underlying Bayesian model exactly, I simulated the log OR from the prior distribution, $\beta \sim N(0, W)$ with $W = 0.2^2$ (Wellcome Trust Case Control Consortium, 2007). However, when investigating the distribution of lead variant effect sizes for associations listed on the NHGRI-EBI GWAS Catalog (Buniello et al., 2019), I found that this prior distribution was a poor fit to the data (Fig 3.1). This observation is supported by Nishino et al. (2018) who found that effect size distributions varied substantially across phenotypes when estimated using their semi-parametric hierarchical mixture model approach. Similarly, when modelling GWAS top hits for breast cancer, Walters and colleagues (Walters et al., 2019, 2021) advocated that a Laplace distribution with an expectation of 0 provided a better fit for the distribution of effect sizes, and suggested using maximum likelihood approaches to estimate the rate parameter, using either truncated (i.e. ignoring variants with very small effect sizes) or censored (i.e. assuming a fixed number of variants are yet to be discovered due to the limited power to detect their small effect size) distributions for the effect sizes.

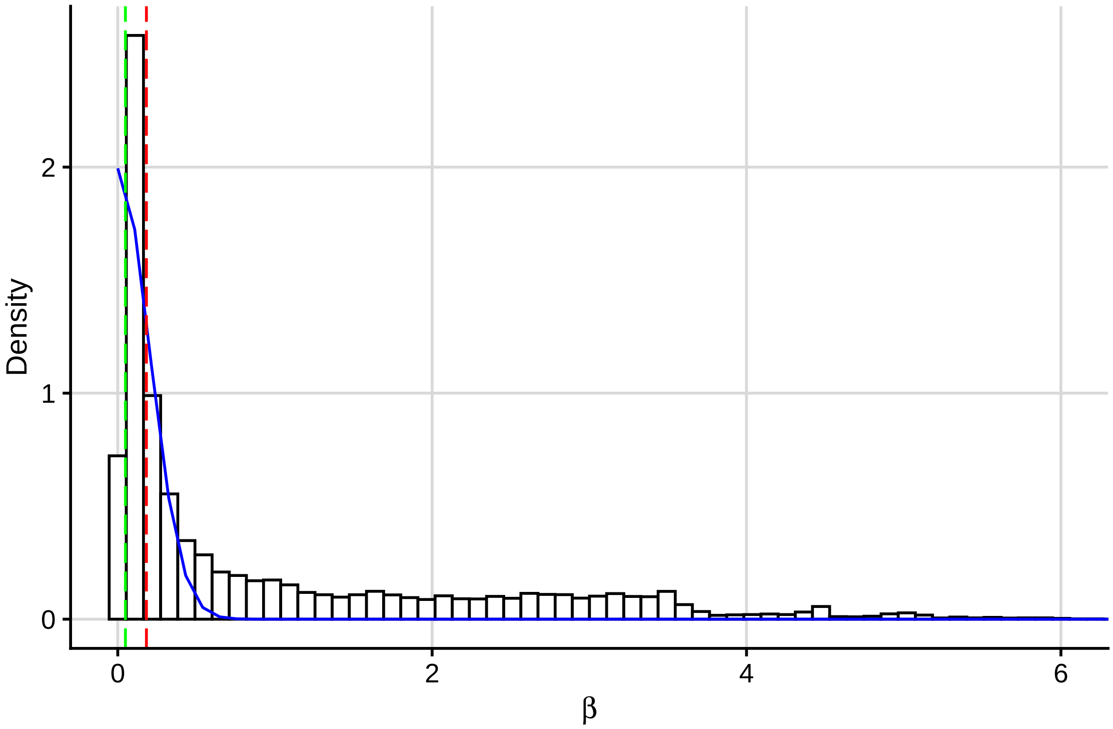


Fig. 3.1 Histogram showing the distribution of lead variant log OR values (β) for significant associations ($p \leq 5e{-}08$) deposited on the NHGRI-EBI GWAS Catalog. All GWAS associations were downloaded from the GWAS Catalog (v1.0.2 - with added ontology annotations, GWAS Catalog study accession numbers and genotyping technology; 213,519 studies accessed on 2020-10-27). Associations were downsampled to only include those from case-control GWAS studies performed using a "genome-wide genotyping array" (25,170 studies). Green dashed line at $\beta = log(1.05)$ and red dashed line at $\beta = log(1.2)$. The blue line shows the distribution of $\beta \sim N(0, 0.2^2)$, truncated at $\beta = 0$.

Rather than using a parametric distribution which may be based on computational convenience rather than empirical results, I considered the effect size at the causal variant from a single study to be sampled from a point distribution. Therefore, in addition to simulations where the observed effect size was directly sampled from the prior distribution, I also simulated data where the observed effect size was fixed at representative values of $OR = 1.05$ or $OR = 1.2$ (Fig 3.1).

3.2.3 Deriving posterior probabilities and credible sets

I used the simulated Z-scores generated using the `simGWAS` R package in equation (1.15) to derive ABFs for each SNP, setting $W = 0.2^2$ and calculating V_i using equation (1.7).

I then derived posterior probabilities of causality for each SNP i $(i = 1, .., m)$,

$$PP_i = \frac{ABF_i}{\sum_{i=1}^{m} ABF_i}. \tag{3.1}$$

I constructed credible sets of putative causal variants by sorting variants into descending order of posterior probability and adding variants to the set until the cumulative sum of posterior probabilities exceeded some threshold, α, using a variety of threshold values ($\alpha \in \{0.5, 0.9, 0.95, 0.99\}$) (Maller et al., 2012). I recorded both the "threshold coverage estimate" (α value used to construct the credible set) and the "claimed coverage estimate" (the sum of the posterior probabilities of the variants in the set) for each simulated credible set, and recorded whether the simulated causal variant was contained within the credible set or not.

3.3 Empirical calibrations in Bayesian fine-mapping

3.3.1 Empirical calibration of posterior probabilities

I began by examining the empirical calibration of posterior probabilities of causality from Bayesian fine-mapping. I used my simulated data to investigate to what extent the posterior probabilities corresponded to the probability of being causal, by binning posterior probabilities into 10 equally sized bins and calculating the proportion of simulated causal variants in each bin.

A common feature of Bayesian statistical inference is that if one simulates from the prior used in the underlying model, then the resultant posterior inferences will be accurate. This was reflected in my results, whereby the posterior probabilities were well calibrated in simulations where the effect size at the causal variant was sampled from the prior normal distribution (Fig 3.2A).

However, some bias was introduced when effect sizes were sampled from point distributions, which is arguably more akin to a real-world fine-mapping analysis (Fig 3.2B). Specifically, the posterior probabilities tended to be anti-conservatively biased in low effect sizes (i.e. they tended to overestimate the true probabilities) and conservatively biased in higher effect sizes (i.e. they tended to underestimate the true probabilities). These results imply that the posterior probabilities from real fine-mapping analyses may be biased, with the direction of the bias determined by the true effect size at the causal variant (which is unknown in practise).

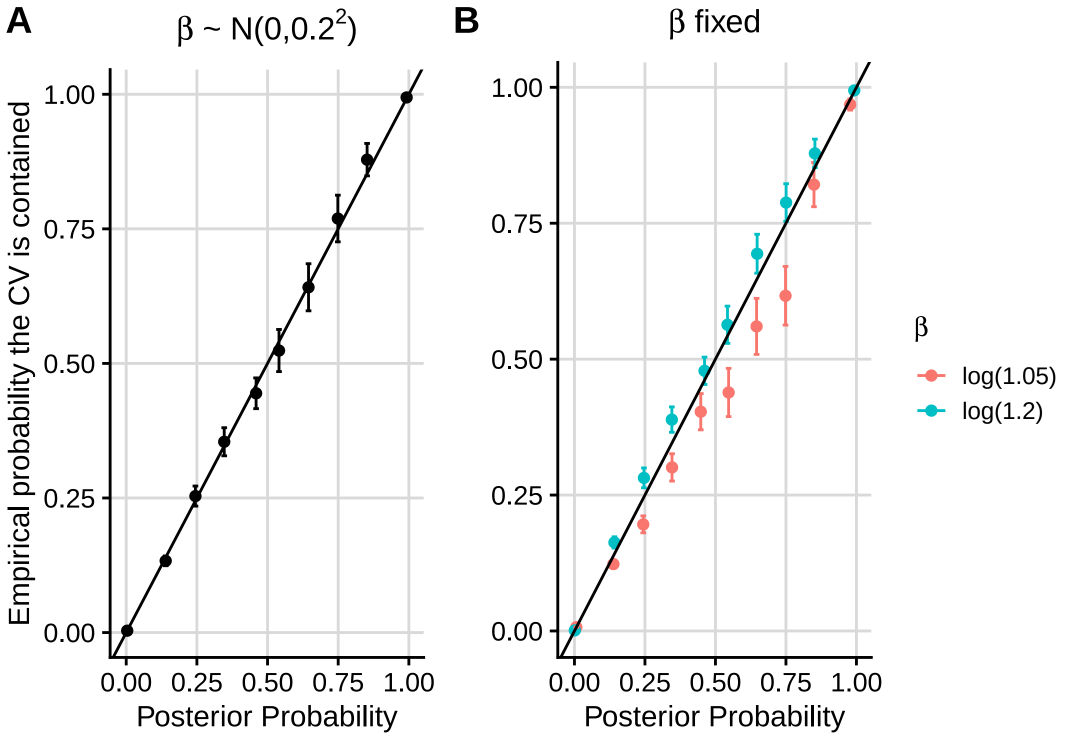


Fig. 3.2 Calibration of posterior probabilities for simulations where (A) $\beta \sim N(0, 0.2^2)$ and (B) $\beta = log(1.05)$ or $\beta = log(1.2)$. Posterior probabilities from 13000 simulations were binned into 10 equally sized bins. The y-axis is the proportion of causals in each posterior probability bin, reflecting the empirical probability that the causal variant was contained within that bin, with mean $+/-1.96 \times$ standard error shown.

3.3.2 Empirical calibration of credible set coverage

The coverage of the causal variant in a single credible set is a binary measure, since the causal variant is either present in the credible set (coverage $= 1$) or not (coverage $= 0$). To estimate the underlying coverage probabilities for credible sets created under specific simulation conditions, I fixed all parameter values to those used to simulate a given credible set, and simulated 5000 additional credible sets (for the simulations where $\beta \sim N(0, 0.2^2)$ I fixed β at its simulated value). I then calculated the "empirical coverage" of the simulated credible set as the proportion of the additional 5000 credible sets that contained the simulated causal variant.

When averaging results over all simulations, I found that threshold coverage estimates tended to be conservatively biased for credible sets, even when effect sizes were sampled from the underlying prior distribution, meaning that the threshold values generally comprised a lower bound for the true coverage (Fig 3.3). The claimed coverage estimates were slightly systematically biased when effect sizes were sampled from point distributions, but were unbiased when effect sizes were sampled from the prior distribution (Fig 3.3). The results shown are for simulated 90%

credible sets, but the results for credible sets constructed using different threshold values were similar (Fig B.1).

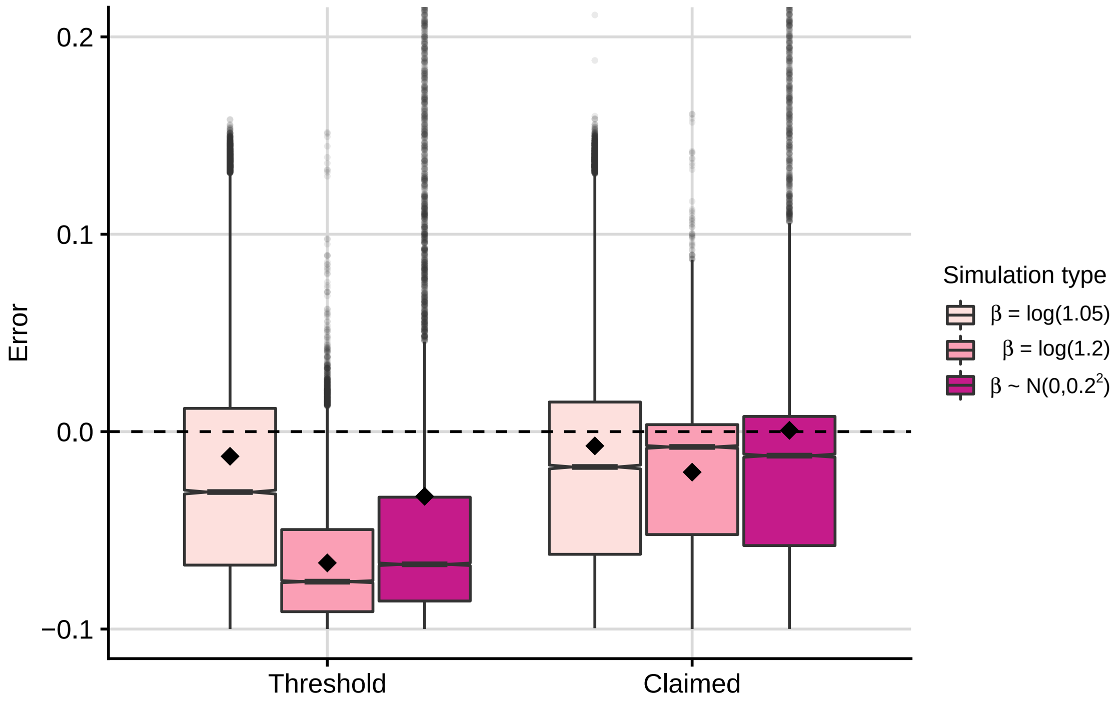


Fig. 3.3 Box plots (median and interquartile range (IQR); black diamond shows mean) showing error in threshold and claimed coverage estimates for 90% credible sets when averaged across all 5000 simulations. Error is defined as estimated conditional coverage − empirical conditional coverage, and empirical conditional coverage is the proportion of 5000 replicate credible sets that contained the causal variant.

I next interrogated the coverage of credible sets by faceting simulations by sample size and LD structure. Focussing on low effect sizes ($\beta = log(1.05)$) to begin with, I found that in smaller sample sizes (number of cases = number of controls = 5000), low effect sizes could lead to anti-conservatively biased threshold and claimed coverage estimates, meaning that they generally overestimated the true coverage (Fig 3.4; palest pink box in left hand column facet). Contrastingly, in higher sample sizes (number of cases = number of controls = 50000) low effect sizes generally resulted in conservatively biased coverage estimates (Fig 3.4; palest pink box in right hand column facet). The anti-conservatively biased estimates seen for lower sample sizes and the conservatively biased estimates seen for higher sample sizes therefore “cancel each other out” when averaging results over all simulations, resulting in the close-to-zero error in coverage estimates for low effect sizes observed in Fig 3.3.

On the other hand, I found that larger effect sizes ($\beta = log(1.2)$) consistently resulted in conservatively biased threshold coverage estimates, but that the claimed coverage estimates became less biased as the sample sizes increased (Fig 3.4; medium pink box). This describes the consistently conservative coverage estimates that were observed for larger effect sizes when averaging results over all simulation in Fig 3.3.

Finally, even when effect sizes were sampled from the prior distribution ($\beta \sim N(0, 0.2^2)$), threshold coverage estimates were consistently conservatively biased and claimed coverage estimates only became non-biased in the largest sample sizes (Fig 3.4; darkest pink box). The results were broadly similar across different LD structures (Fig 3.4).

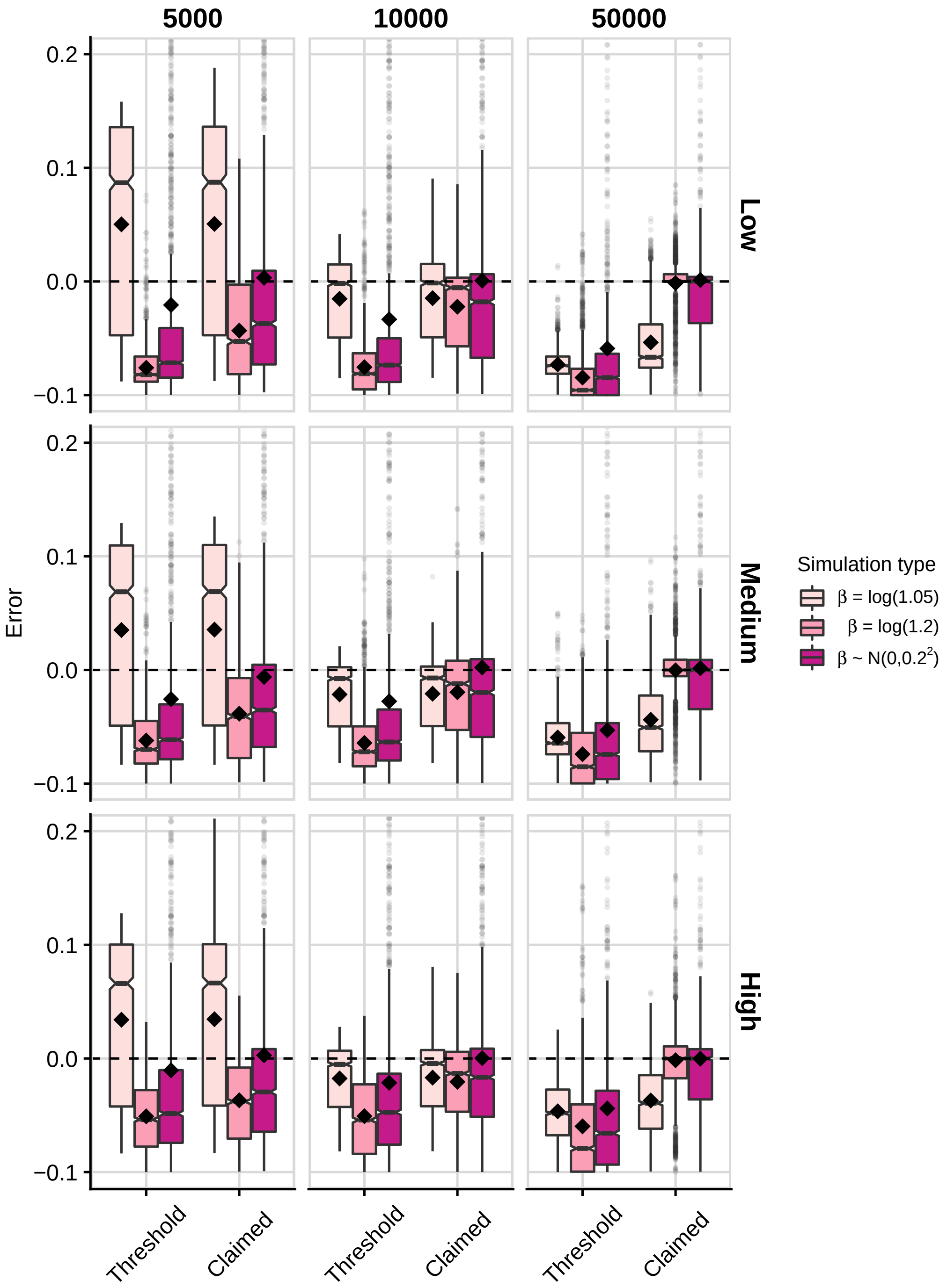


Fig. 3.4 Box plots (median and IQR; black diamond shows mean) showing error in threshold and claimed coverage estimates for 90% credible sets faceted by sample size (columns; values show the number of cases, which equals the number of controls) and LD structure (rows). Error is defined as estimated conditional coverage – empirical conditional coverage, and empirical conditional coverage is the proportion of 5000 replicate credible sets that contained the causal variant.

The systematic bias observed for coverage estimates in simulations where the causal variant's effect size was sampled from the underlying prior (Fig 3.4) is due to different parts of the prior effect size distribution being sampled in each facet, such that the distribution of effect sizes for each facet no longer resembles the underlying prior. Thus, the Bayesian paradigm that if one simulates from the prior that is used in the underlying model then the resultant posterior inferences will be accurate, no longer holds.

Alarmingly, this generalises to the routine fine-mapping approach whereby researchers select regions to fine-map based on the significance of the corresponding association signal. Generally, researchers only fine-map those associations which are found to be strongly significant in GWAS, for example those that reach genome-wide significance ($p \leq 5e-08$). If we only consider those associations that are strongly significant, then the distribution of the effect sizes is now bimodal, and does not represent the underlying model, implying that posterior inferences may be inaccurate (Fig 3.5A).

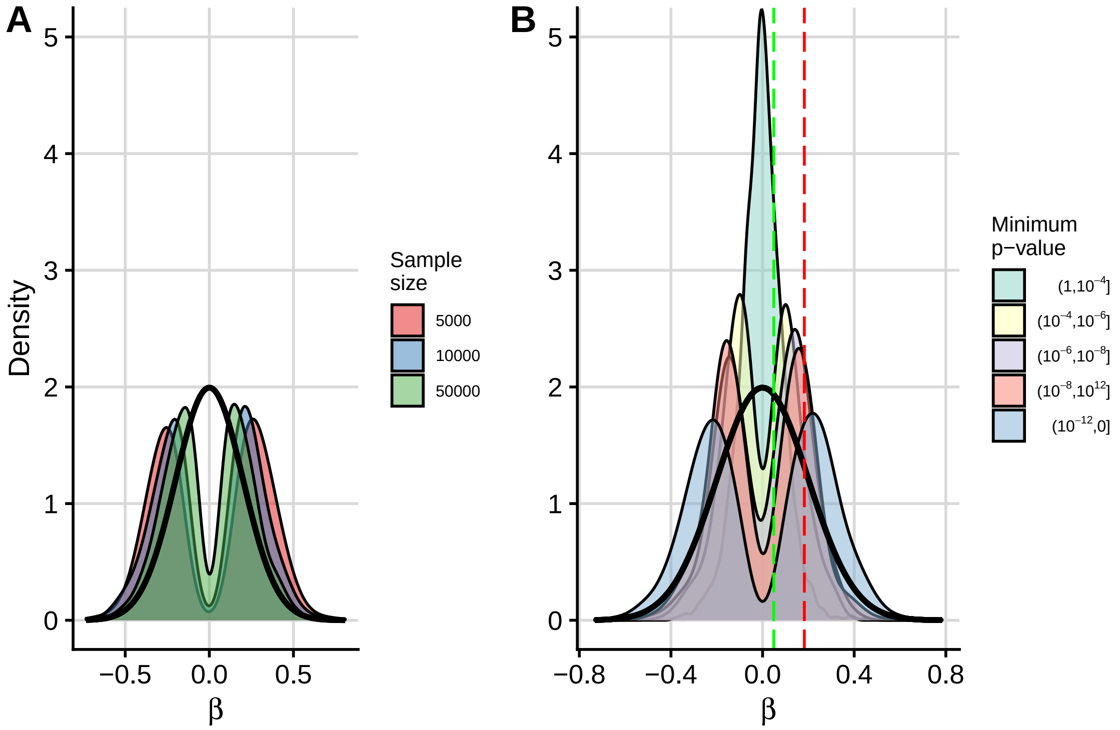


Fig. 3.5 Distribution of log OR (β) values in simulations where $\beta \sim N(0, 0.2^2)$ for (A) simulations where $P_{min} \leq 1e-08$ that have been stratified by sample size (P_{min} is the minimum p-value in the region) or (B) simulations that have been stratified by p-value bin. Thick black curve is density function for $N(0, 0.2^2)$, green dashed line at $\beta = log(1.05)$ and red dashed line at $\beta = log(1.2)$.

Since researchers typically select regions to fine-map depending on the strength of evidence of an association in that genomic region which is quantified by a p-value, I further scrutinised the accuracy of the conditional coverage estimates by binning simulations by minimum p-value (P_{min}) in the region. Similarly to faceting significant associations by sample size, each p-value bin encompassed simulations which sampled different parts of the prior effect size distribution, such that the distribution of effect sizes for each p-value bin no longer resembled the conventional $N(0, 0.2^2)$ prior (Fig 3.5B).

From here on, I focus on the claimed coverage estimates rather than the threshold coverage estimates as these are more widely used in the literature, and because I have also shown that these out-perform the threshold coverage estimates in Fig 3.3 and Fig 3.4 (in terms of error). I now also include results for simulations that were based on UK10K haplotypes to ensure that any findings were robust to a wider range of LD structures over a larger population.

I found that claimed coverage estimates were systematically biased in representatively powered scenarios where fine-mapping is usually performed, regardless of the effect size sampling method (Fig 3.6). This bias is mostly conservative, but may be anti-conservative in low powered studies (Fig B.2). Notably, even when the effect sizes were sampled from the underlying prior distribution, the claimed coverage estimates were anti-conservatively biased in low powered simulations ($P_{min} > 1e-04$), conservatively biased in intermediately powered simulations ($1e-12 < P_{min} < 1e-04$) and unbiased in very high powered simulations ($P_{min} < 1e-12$) (Fig 3.6).

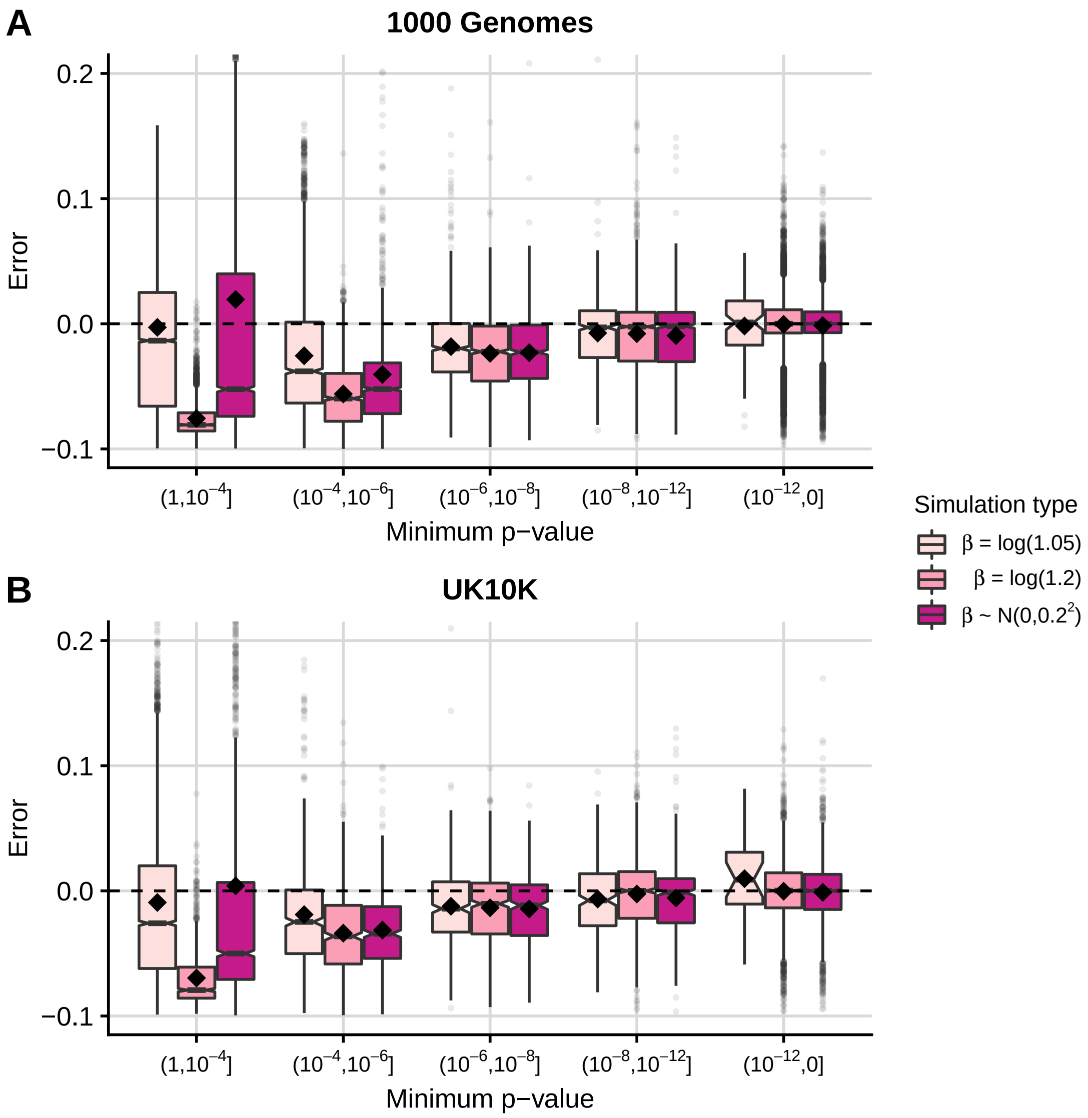


Fig. 3.6 Box plots (median and IQR; black diamond shows mean) showing error in claimed coverage estimates for 90% credible sets for simulations using (A) 1000 Genomes or (B) UK10K haplotype data. Simulations have been binned by the minimum *p*-value in the region. Error is defined as estimated conditional coverage − empirical conditional coverage, and empirical conditional coverage is the proportion of 5000 replicate credible sets that contained the causal variant.

In conclusion, the probabilities that the causal variant is contained within the credible set in intermediately powered fine-mapping studies are typically too low, and researchers can afford to be "more confident" that they have captured the true causal variant in their credible set. I have demonstrated that this bias is due to researchers selecting a biased sample of regions to fine-map, and that this bias is increased when considering causal variant effect sizes as being sampled from various point distributions.

3.4 Developing an adjusted coverage estimate

Having found that conventional coverage estimates are typically over-conservative in Bayesian fine-mapping, I decided to develop an adjusted coverage estimate to provide researchers with a more accurate estimate that the true causal variant is contained in their credible set. Analogous to my simulation analysis, where I obtained an empirical coverage estimate by simulating many credible sets and calculating the proportion of simulations where the causal variant appeared in the set, my method to derive an adjusted coverage estimate involves considering potential causal variants in turn to simulate many credible sets and empirically calculating their coverage.

3.4.1 Distribution of marginal Z-scores

In order to simulate credible sets, I first simulated Z-scores. To do this, one needs to consider the distribution from which marginal Z-scores in a GWAS are sampled. Associations between a SNP and a trait are usually tested for using single-SNP models, such that marginal Z-scores are derived. In contrast, if the SNPs in the region are jointly modelled, then joint Z-scores can be derived. Under the assumption of a single causal variant per region, the expected joint Z-score vector is

$$Z_J = (0, \dots, 0, \mu, 0, \dots, 0)^T \, , \tag{3.2}$$

where Z_J has length equal to the number of SNPs in the region, and all elements take the value 0 except at the position of the causal variant, which takes the value μ (i.e. μ is the joint Z-score at the causal variant).

Given Z_J, the expected marginal Z-scores can be written as

$$E(Z) = \Sigma \times Z_J \, , \tag{3.3}$$

where Σ is the SNP correlation matrix (Hormozdiari et al., 2014). The asymptotic distribution of these marginal Z-scores is then multi-variate normal (MVN) with variance equal to the SNP correlation matrix,

$$Z \sim MVN(E(Z), \Sigma) \tag{3.4}$$

(Hormozdiari et al., 2014).

3.4.2 Estimating Z-scores at causal variants

In order to derive an estimate of the joint Z vector I first needed to estimate the value of μ, which is unknown in genetic association studies. I used the marginal Z-scores that are available from GWAS to do this. Firstly, I considered using the absolute Z-score at the lead variant as an estimate for μ, but found this to be too high in low powered scenarios (Fig 3.7A). This is because $E(|Z|) > 0$ even when $E(Z) = 0$, and thus $E(|Z|) > E(Z)$ when $E(Z)$ is close to 0. Note that the sign of the Z-score does not matter here because we are only considering a single causal variant (in contrast to methods that model multiple causal variants, where knowing the sign of the Z-score is essential since the expected Z-score is a weighted sum of multiple Z-scores at multiple causal variants).

To obtain a more accurate estimate I considered using all available Z-scores, weighted so that the Z-scores for variants that were more likely to be causal contributed more to the estimate. For a region comprising m SNPs,

$$\hat{\mu} = \sum_{i=1}^{m} |Z_i| \times PP_i \tag{3.5}$$

where Z_i is the marginal Z-score for SNP i and PP_i is the posterior probability of causality for SNP i.

This estimate had small absolute error at larger values of μ ($\mu > 5$) representing regions where fine-mapping is typically performed (corresponds to $p < 5.7e - 07$) (Fig 3.7). It also performed better than $\hat{\mu} = max_{i \in \{1,...,m\}}(|Z_i|)$ at small μ ($\mu < 5$) representing regions that are less likely to be fine-mapped (Fig 3.7).

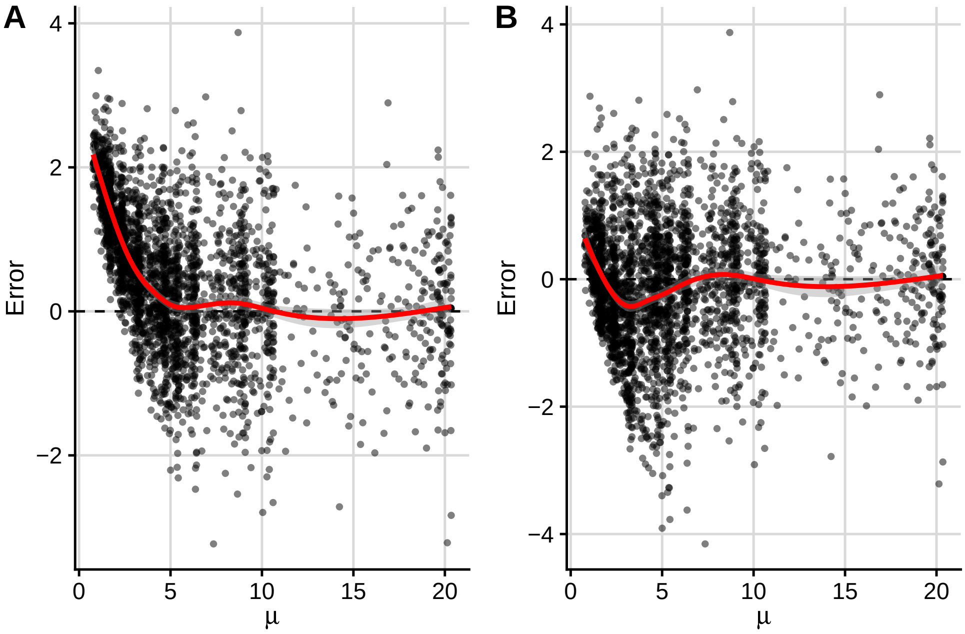


Fig. 3.7 Error of μ estimates calculated as $\hat{\mu}_X - \mu$. The x-axis is the joint Z-score at the causal variant, which was computed using the `simGWAS::expected_z_score` function. Line is fitted using a GAM as the smoothing function (`ggplot2::geom_smooth`). (A) $\hat{\mu} = max_{i \in \{1,...,m\}}(|Z_i|)$ (B) $\hat{\mu} = \sum_{i=1}^{m} |Z_i| \times PP_i$.

3.4.3 Adjusted coverage estimate method

Having found an estimate for μ, we are still not able to construct the joint Z vector because the true causal variant is unknown (and thus we don't know at what position μ appears in the vector). I therefore considered each SNP i in the region as a potential causal variant in turn, and constructed the joint Z vector as

$$\hat{Z}_J[j] = \begin{cases} 0 & j \neq i \\ \hat{\mu} & j = i. \end{cases} \tag{3.6}$$

In order to derive a proportion across many simulations, I then simulated $N = 1000$ marginal Z-score vectors for each SNP i considered causal,

$$\mathcal{Z}^*_{N=1000} = \{Z^*_1, \ldots, Z^*_{1000}\} \overset{iid}{\sim} MVN(\Sigma \times \hat{Z}_J, \Sigma). \tag{3.7}$$

Each element in the simulated Z^* vectors was then converted to a posterior probability using the ABF approach in equation (1.15), and setting $W = 0.2^2$ (see Appendix B.2 for detailed reasoning on the choice of value for W). Credible sets were then derived using the standard sort and sum method, and the proportion of the $N = 1000$ simulated credible sets that contained SNP i (where SNP i was the SNP considered causal in the simulations), $prop_i$, was then calculated.

In the method, this procedure is implemented for each SNP in the genomic region with $PP > 0.001$ considered as causal in turn. This ensures that only simulated scenarios that are realistic contribute to the final coverage estimate. The final adjusted coverage estimate is then calculated by weighting each of these proportions by the posterior probability of the SNP that is considered causal,

$$\text{Adjusted Coverage Estimate} = \frac{\sum_{i:PP_i>0.001} PP_i \times prop_i}{\sum_{i:PP_i>0.001} PP_i}. \tag{3.8}$$

Intuitively, proportions obtained from realistic scenarios (SNPs with high posterior probabilities considered as causal) are up-weighted and proportions obtained from unrealistic scenarios (SNPs with low posterior probabilities considered as causal) are down-weighted. Note that I am not attempting to reweight the posterior probabilities for inference, only to calibrate the adjusted coverage estimate.

3.5 Accuracy of adjusted coverage estimates

3.5.1 Adjusted coverage estimates improve calibration of credible sets

I found that the adjusted coverage estimates were better empirically calibrated than the claimed coverage estimates in simulations that were representative of those considered for fine-mapping (Fig 3.8). Particularly, the median and mean error of the adjusted coverage estimates decreased for $P_{min} \geq 1e-12$, and the variability between estimates also decreased even in simulations where the claimed coverage estimates were unbiased ($P_{min} < 1e-12$).

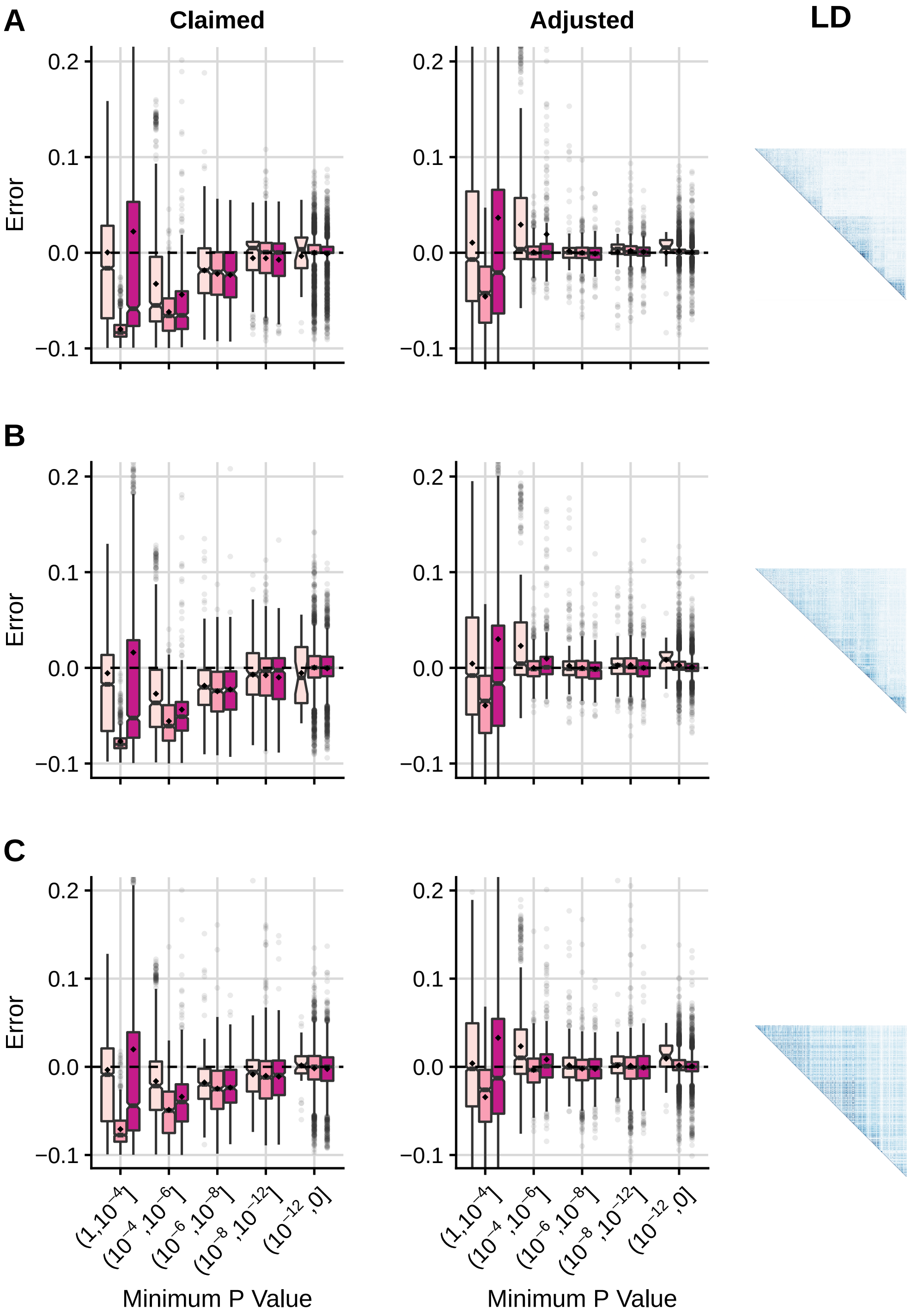


Fig. 3.8 Box plots (median and IQR; black diamond shows mean) showing error in claimed and adjusted coverage estimates for 90% credible sets where simulations have been binned by the minimum p-value in the region. Error is defined as estimated conditional coverage − empirical conditional coverage, and empirical conditional coverage is the proportion of 5000 replicate credible sets that contained the causal variant. LD column shows a graphical display of SNP correlations for (A) low (B) medium and (C) high LD regions, generated using the `corrplot` R package (darker colours for larger r^2 values) (https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html). The colour legend is as in Fig 3.3, Fig 3.4 and Fig 3.6.

3.5.2 Adjusted coverage estimates are robust to estimated MAF and LD

My method relies on MAF and SNP correlation data to simulate GWAS summary statistics that are representative of the observed GWAS data. So far, I have assumed that this information is available from the GWAS samples, but due to privacy concerns this is generally not the case. I therefore evaluated the performance of my adjustment method when using independent reference data to estimate MAFs and SNP correlations. I applied my method to credible sets simulated from the European 1000 Genomes data using either MAFs and SNP correlations from the original (1000 Genomes) data (Fig 3.9B) or MAFs and SNP correlations estimated from a larger reference panel of European samples (UK10K) (Fig 3.9C). I found that the adjusted coverage estimates remained accurate in either case.

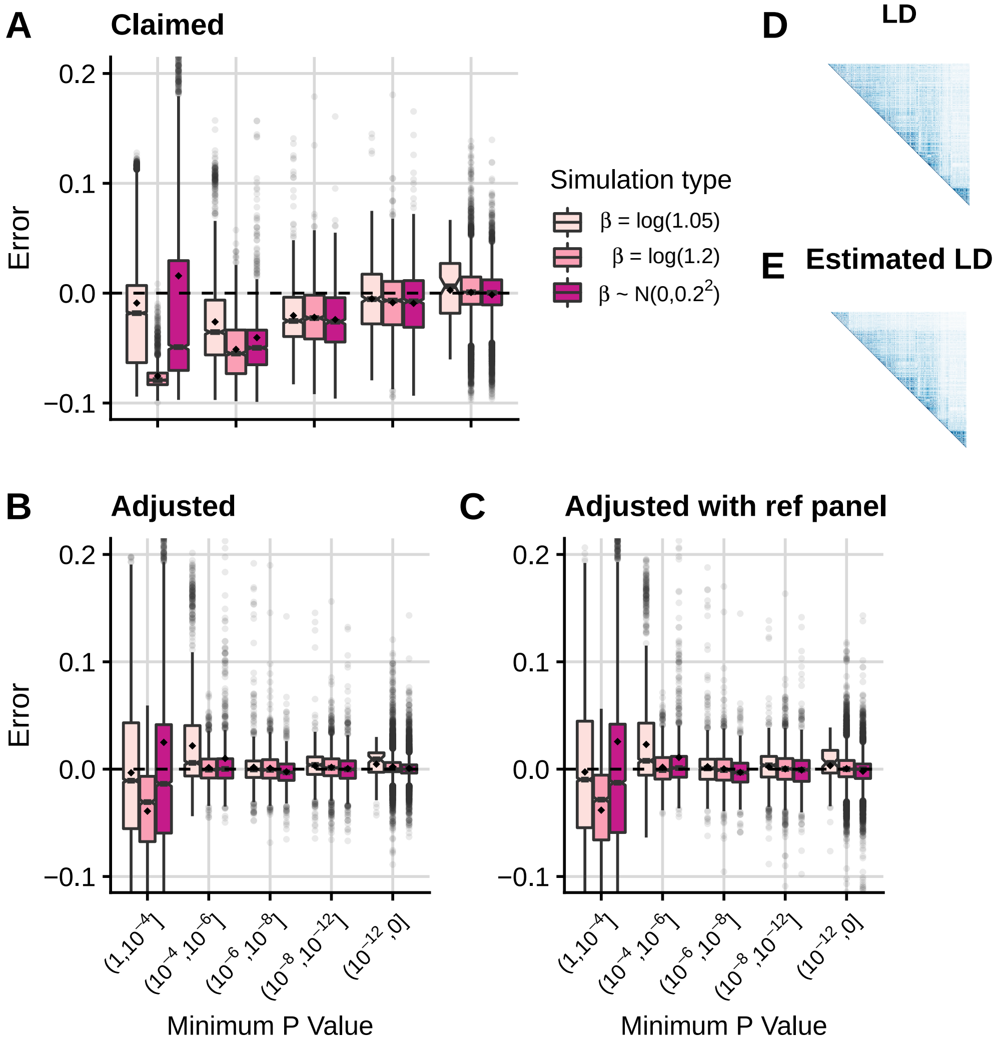


Fig. 3.9 Box plots (median and IQR; black diamond shows mean) showing error in claimed and adjusted coverage estimates for 90% credible sets where simulations have been binned by the minimum p-value in the region. Error is defined as estimated conditional coverage − empirical conditional coverage, and empirical conditional coverage is the proportion of 5000 replicate credible sets that contained the causal variant. (A) Claimed coverage estimate (the sum of the posterior probabilities of causality for the variants in the credible set) (B) Adjusted coverage estimate using MAFs and SNP correlations from the original (1000 Genomes) data (C) Adjusted coverage estimate using UK10K data to approximate MAFs and SNP correlations (D) Graphical display of SNP correlations in 1000 Genomes data (E) Graphical display of the estimated SNP correlations using UK10K data. Graphical displays of correlations were generated using the `corrplot` R package.

3.5.3 Adjusted coverage estimates are robust to departure from single causal variant assumption

The Bayesian approach for fine-mapping described by Maller et al. (2012) assumes a single causal variant per genomic region, which may be unrealistic (Asimit et al., 2019). Multi-SNP fine mapping methods should be used if one suspects that there are multiple causal variants in a region, yet it can be expected that single causal variant fine-mapping techniques have been incorrectly used in these instances, for example due to their relatively simple methodologies and data requirements (methods modelling multiple causal variants generally require information on the direction of effects and are more computationally demanding due to the larger space of possible causal configurations to consider). I therefore investigated how my adjustment method performed when the single causal variant assumption was violated.

If there are multiple causal variants in a region, multiple possible definitions of "coverage" exist. For example "coverage" could be defined as the probability that all causal variants are captured in the set (for example CAVIAR's p-causal sets in Hormozdiari et al. (2014)), the probability that a causal variant chosen at random is captured in the set or the probability that at least one of the causal variants is captured in the set. Each definition has its own caveats, for example if the latter definition is adopted then this assumes that a credible set containing all the causal variants is equal, in terms of credible set coverage performance, to a credible set containing only one of the causal variants. Moreover, as the number of causal variants increases, the chance of capturing at least one of these in the credible set increases. Since these discrepancies grow as the number of causal variants in the region increases, I limited my analysis to two causal variants in a genomic region.

I simulated genomic regions with two causal variants that either had the same effect size or different effect sizes. Defining conditional coverage as the frequency with which a credible set contained at least one causal variant, I found that the adjusted coverage estimates tended to have smaller error than the claimed coverage estimates for causal variants in low LD ($r^2 < 0.01$, Fig 3.10A). When the two causal variants were in high LD ($r^2 > 0.7$), the adjusted coverage estimates were still generally more accurate than the claimed coverage estimates, although both tended to underestimate the true conditional coverage (and were thus conservative) (Fig 3.10B).

Whilst I do not recommend using single causal variant fine-mapping techniques (and therefore my adjustment) when the single causal variant assumption is violated, it is reassuring that if this is the case then my method would perform no worse than the standard method, when considering capturing at least one causal variant as most relevant.

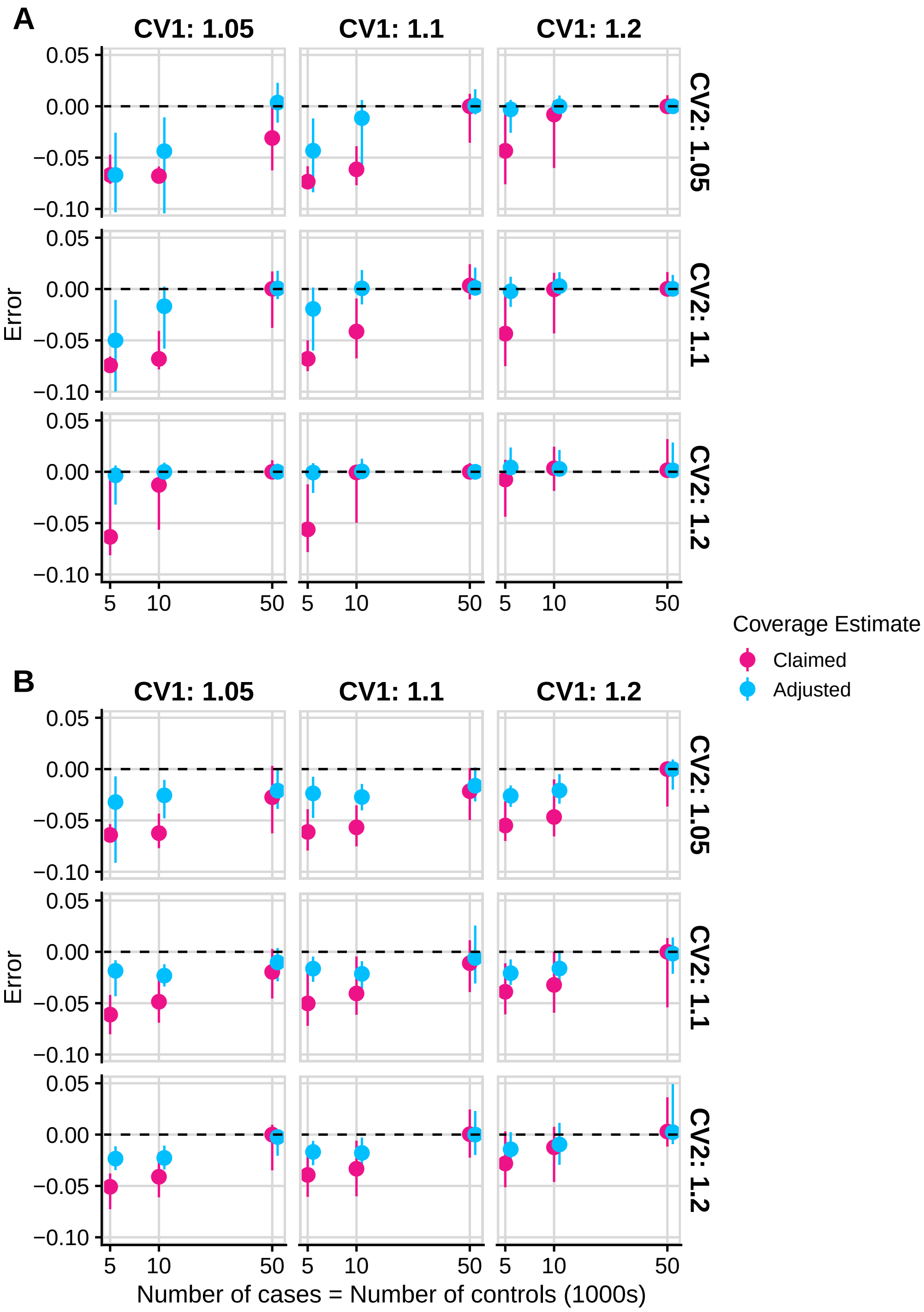


Fig. 3.10 The median error and IQR of claimed and adjusted coverage estimates for 90% credible sets from 5000 simulated regions with two causal variants that are (A) in low LD ($r^2 < 0.01$) or (B) in high LD ($r^2 > 0.7$). Error is calculated as estimated conditional coverage − empirical conditional coverage, where empirical conditional coverage is the proportion of 5000 additional simulated 90% credible sets that contained at least one of the two causal variants and estimated conditional coverage is the claimed or adjusted coverage estimate as defined in the text. Results are faceted by odds ratio values at the two causal variants (CV1 and CV2).

3.6 Adjusted credible sets

Obtaining an accurate coverage estimate that the causal variant appears in the credible set is useful in its own right, but it may also be beneficial to obtain an “adjusted credible set” - that is, the smallest set of variants required such that the adjusted coverage estimate of the resultant credible set achieves some desired conditional coverage of the causal variant. For example, discovering that a 90% credible set actually has 99% conditional coverage of the causal variant is useful, but an obvious follow-up question is “what variants do I need such that the conditional coverage is actually 90%?”. This is especially useful in instances where the conventional coverage estimates are over-conservative (which I have demonstrated is generally the case in real-world fine-mapping analyses) as the new credible sets would likely contain fewer variants. That is, adjusted credible sets could improve the resolution of credible sets reported in the literature, potentially saving time and resources in the expensive functional follow-up studies that are based on the prioritised sets of variants.

I explored this using an example simulated GWAS across 200 SNPs with the effect size at the causal variant fixed at $\beta = log(1.2)$. The 90% credible set constructed using the standard Bayesian approach contained eight variants and had a claimed coverage estimate of 0.903. The adjusted coverage estimate of this credible set was 0.969 and the estimated empirical coverage was 0.972 (Fig 3.11).

I used the root bisection method (Greene, 1992) to iteratively search for the smallest threshold value that yielded a credible set with accurate conditional coverage of the causal variant. In this example, I found that an adjusted 90% credible set could be constructed using a threshold value of 0.781. This adjusted credible set had an adjusted coverage estimate of 0.905 (empirical estimated conditional coverage of 0.907) and reduced in size from eight to four variants, with the four variants removed from the credible set holding a small proportion of the total posterior probability (Fig 3.11B, Fig 3.11C).

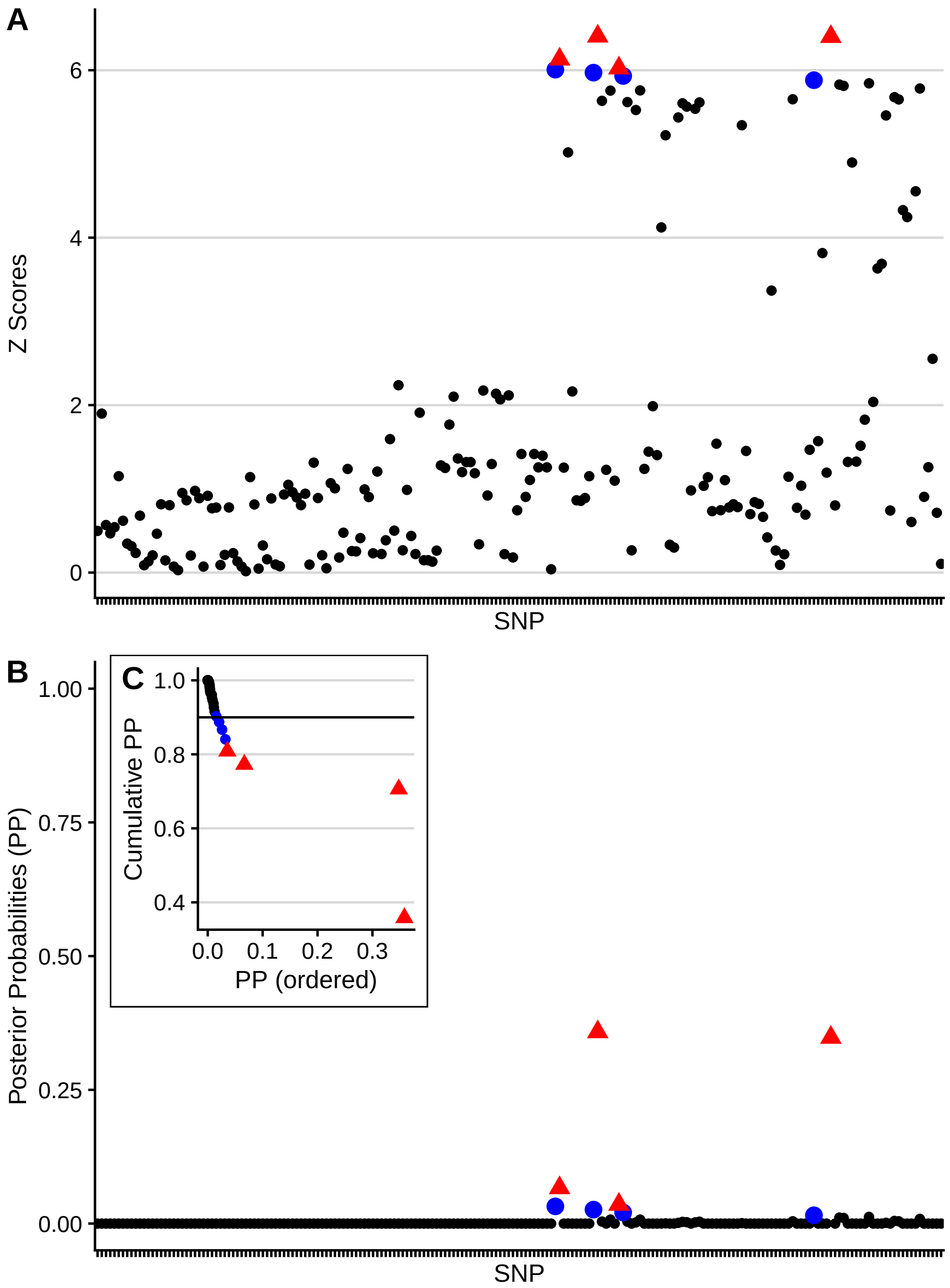


Fig. 3.11 A simple example to illustrate the results of the adjustment method. (A) The absolute Z-scores of the SNPs. (B) The posterior probabilities of the SNPs. (C) As in the fine-mapping procedure, variants are sorted into descending order of posterior probability and summed. Starting with the SNP with the largest posterior probability (far right) the cumulative sum (size) of the credible set is plotted as each SNP is added to the set. Red SNPs are those in the adjusted 90% credible set and blue SNPs are those that only appear in the original 90% credible set. The 90% credible set formed of the red SNPs has an adjusted coverage estimate of 0.905 and the credible set formed of both the blue and red SNPs has an adjusted coverage estimate of 0.969.

3.7 corrcoverage R package

I have created a CRAN R package, `corrcoverage` (https://annahutch.github.io/corrcoverage/; https://cran.r-project.org/web/packages/corrcoverage/index.html), that uses marginal GWAS summary statistics to derive adjusted coverage estimates and adjusted credible sets. The functions used to calculate adjusted coverage estimates are computationally efficient, taking approximately 1 minute for a 1000 SNP region (using one core of an Intel Xeon E5-2670 processor running at 2.6GHz). The functions used to derive adjusted credible sets require only the GWAS summary statistics and auxiliary information that is needed to derive the adjusted coverage estimate (i.e. Z-scores, sample sizes, MAFs and the SNP correlation matrix, where the latter two can be estimated from a representative reference panel) plus some user-specified desired coverage. Users are able to customise the optional arguments to suit both their accuracy requirements and computational constraints. The algorithm then works iteratively such that the threshold and the adjusted coverage estimate of each tested credible set is displayed, until the smallest set of variants with the desired conditional coverage is established, offering researchers an easy tool to improve the resolution of their credible sets.

The package web-page contains several comprehensive vignettes guiding the user through fully reproducible typical GWAS and fine-mapping analyses, with the emphasis placed on how the `corrcoverage` R package can be used as an extra step in the fine-mapping pipeline (Appendix B.3). As the users of the package may be from a range of disciplines, I also created an interactive flowchart to enhance user experience and to ensure that users are able to easily navigate the package as it suits their specific problem (Fig 3.12, interactive version available online: https://annahutch.github.io/PhD/package_flowchart.html).

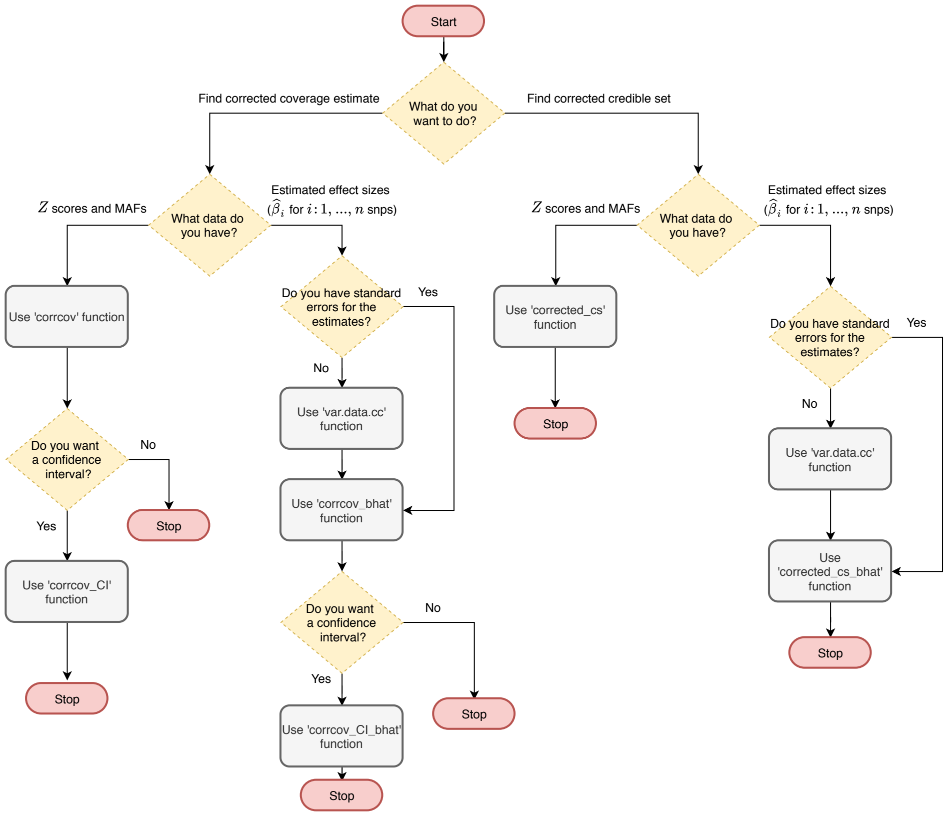

Start
What do you want to do?
Find corrected coverage estimate
Find corrected credible set
Z scores and MAFs
What data do you have?
Estimated effect sizes ($\widehat{\beta}_i$ for i: 1, ..., n snps)
Use 'corrcov' function
Do you want a confidence interval?
No
Stop
Yes
Use 'corrcov_CI' function
Stop
Do you have standard errors for the estimates?
Yes
No
Use 'var.data.cc' function
Use 'corrcov_bhat' function
Do you want a confidence interval?
No
Yes
Stop
Use 'corrcov_CI_bhat' function
Stop
Z scores and MAFs
What data do you have?
Estimated effect sizes ($\widehat{\beta}_i$ for i: 1, ..., n snps)
Use 'corrected_cs' function
Stop
Do you have standard errors for the estimates?
Yes
No
Use 'var.data.cc' function
Use 'corrected_cs_bhat' function
Stop

Fig. 3.12 Flowchart to guide users through the `corrcoverage` R package. Interactive version available at https://annahutch.github.io/PhD/package_flowchart.html

3.8 Application to immune-mediated diseases

In this section, I apply my adjustment method to two GWAS data sets. Firstly to fine-map Immunochip regions that were found to associate with T1D by Onengut-Gumuscu et al. (2015) and secondly to fine-map a genomic region associating with ankylosing spondylitis, for which the results were used to develop a functional genomic assay to elucidate the true causal variant from the prioritised set of variants (Bourges et al., 2020).

3.8.1 Application to T1D

Introduction

As introduced in section 2.10, Onengut-Gumuscu et al. (2015) genotyped 6,670 T1D cases and 12,262 controls using the Immunochip. The T1D SNP and indel content on the Immunochip was chosen on the basis of the 41 T1D-associated regions known at the time (February 2010) and 3,000 additional SNPs that tagged candidate genes or other SNPs with suggestive evidence of association with T1D (Onengut-Gumuscu et al., 2015). Of the 186 densely genotyped regions on the chip, 44 were found to associate with T1D at $p \leq 3.23e-07$ (Immunochip Bonferroni-corrected $p < 0.05$) of which six were novel. The standard Bayesian fine-mapping approach was then used to derive 99% credible sets of putative causal variants for T1D in each of the 44 associated regions. I derived adjusted coverage estimates for these credible sets, and where applicable I also constructed adjusted credible sets. I primarily focussed on 95% credible sets, as these best illustrate the utility of my method due to greater margin for error. I also present results for 99% credible sets for direct comparison with the approach used in the original manuscript.

Method

I downloaded the publicly available GWAS summary statistics for T1D (Onengut-Gumuscu et al., 2015) from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) (study accession GCST005536). The original study reported 44 genomic regions that associated with T1D, for which five showed evidence of multiple independent signals residing in the region (from a conditional regression approach). To ensure that the key assumption in single causal variant Bayesian fine-mapping was satisfied, I limited my analysis to the 39 regions that showed no evidence of harbouring more than one causal variant.

Using the published β and V values for all available variants in each of the 39 associated regions, I constructed 95% and 99% credible sets using the standard Bayesian fine-mapping method. For each credible set, I calculated the claimed and adjusted coverage estimates, using $W = 0.2^2$ and LD data which was calculated from genotype data collated across controls from multiple European Immunochip studies by my supervisor, Dr Chris Wallace.

I calculated 95% confidence intervals for the adjusted coverage estimates by calculating 100 adjusted coverage estimates and taking the 2.5th and 97.5th percentiles. If 0.95 (or 0.99 for 99% credible sets) did not fall within this confidence interval, then I used my adjustment method to derive an adjusted 95% (or 99%) credible set. That is, I only derived adjusted credible sets in regions where there was evidence that the standard credible sets were sufficiently miscalibrated. I omitted regions that both (i) already contained only a single causal variant and (ii) had an adjusted coverage estimate that was greater than 0.95 (or > 0.99 for 99% credible sets), as these sets cannot be made any smaller. Specifically, I used the `corrected_cs` function in the `corrcoverage` R package with optional parameter values: `acc=0.0001; max.iter=70` to find adjusted credible sets that had an adjusted coverage estimate within 0.0001 of the threshold value, or as close as possible after 70 iterations of the internal root bisection algorithm.

Adjusted versus claimed coverage

Similar to the patterns observed in my simulation analysis, I found that the claimed coverage estimates were generally lower than the adjusted coverage estimates of the 95% credible sets, although there were instances where the reverse was true (Fig 3.13A). This supports my general finding that claimed coverage estimates for credible sets derived using the standard framework are typically over-conservative.

The p-value at the lead variant for the 39 T1D-associated regions varied from $1e-07$ to $1e-100$. There were 16 regions with $p < 1e-12$, and where the results from my simulation analysis indicated that the claimed and adjusted coverage estimates were mostly unbiased, but that the variance was tighter for the adjusted coverage estimates. Of these 16 regions, six (37.5%) had claimed coverage estimates that were well calibrated (within the 95% confidence interval for the adjusted coverage estimate), five (31.25%) had claimed coverage estimates that were greater than the adjusted coverage estimates and another five (31.25%) had adjusted coverage estimates that were greater than the claimed coverage estimates (Fig 3.13A). For the remaining 23 regions with $p \geq 1e-12$, and where my simulation results suggested that my method may be particularly useful, only one region had a well calibrated claimed coverage estimate, one additional region had a claimed coverage estimate that was greater than the adjusted coverage estimate, and the remaining 21 regions (91.3%) had adjusted coverage estimates that were greater than the claimed coverage estimates (Fig 3.13A). These patterns matched those observed in my simulation analysis and provide further evidence that the error in claimed coverage estimates is negatively correlated with power (where power is quantified by the p-value at the lead variant).

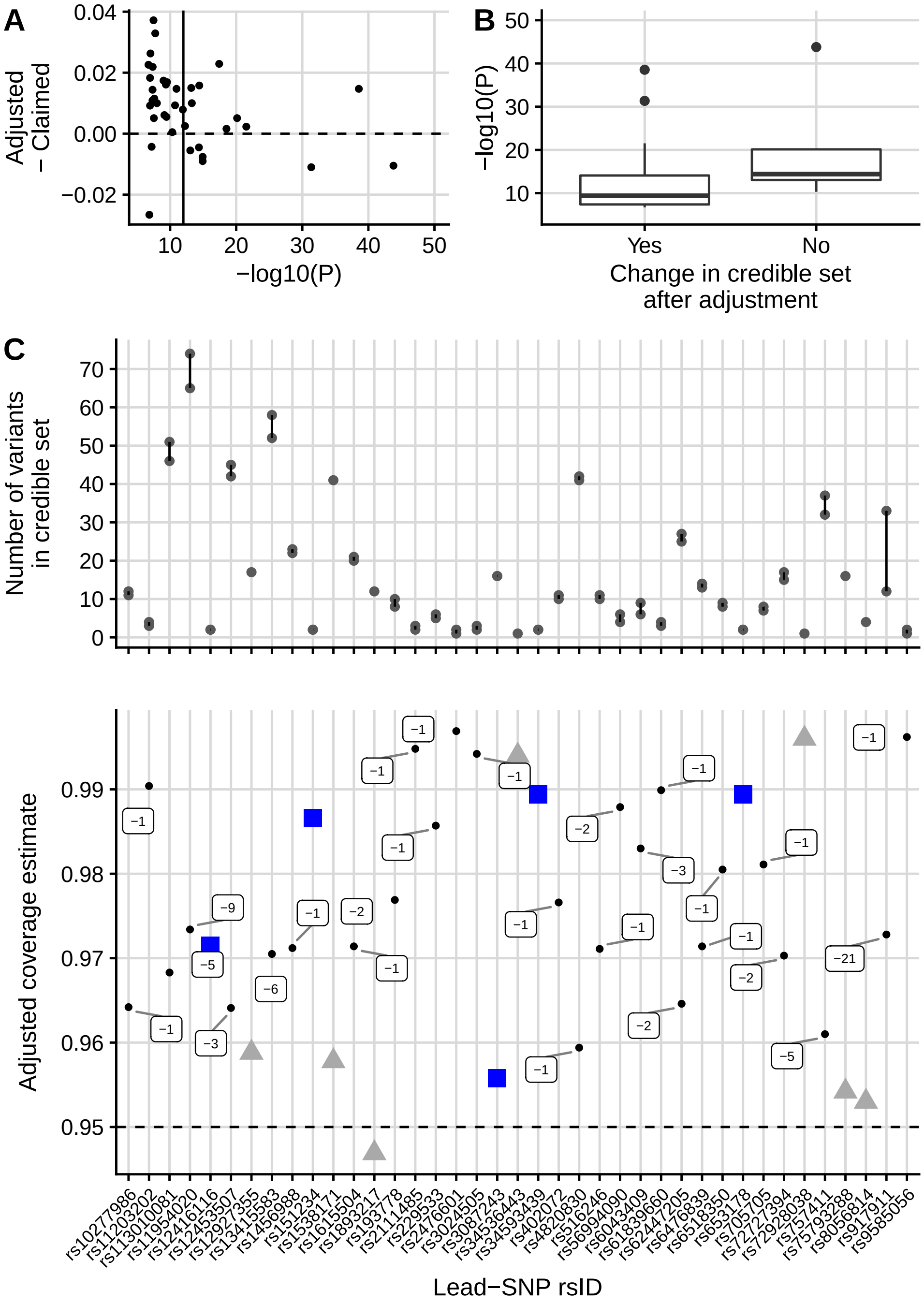


Fig. 3.13 (A) Adjusted - claimed coverage estimates against $-log_{10}(p)$ for the T1D analysis. Horizontal dashed line where adjusted = claimed coverage and vertical line at $-log_{10}(p) = 1e-12$. (B) Box plots showing the distribution of $-log_{10}(p)$ values for credible sets that either changed upon adjustment or did not change. Note that in panels A and B, results for the PTPN22 region with lead variant rs2476601 and $p < 1e-100$ has been omitted to aid visualisation. (C) Top panel: The decrease in size of the credible set after adjustment. Bottom panel: The adjusted coverage estimates of 95% Bayesian credible sets for T1D-associated genomic regions. Black points represent regions where the credible set changed after the adjustment and the "-" values represent the decrease in the number of variants from the standard to the adjusted 95% credible set. Blue square points represent regions where the credible set did not change after the adjustment and grey triangular points represent regions where the credible set did not need to be adjusted since the threshold (0.95) was contained in the 95% confidence interval of the conditional coverage estimate, or because the credible set already contained only a single variant.

Adjusted credible sets

I constructed adjusted credible sets for regions where the threshold value was not within the 95% confidence interval for the adjusted coverage estimate. Upon adjustment, credible sets for regions containing a lead variant with a larger p-value were generally more likely to be subject to change than those with smaller lead variant p-values (Fig 3.13B), a pattern matched by my simulations. Of the 16 high powered regions ($p < 1e-12$), ten credible sets were adjusted and six of these changed after adjustment (60%). Of the 23 low powered regions ($p \geq 1e-12$), 22 were adjusted and 21 of these changed after adjustment (95.5%). This provides further empirical evidence that my method can be beneficial in improving the resolution of fine-mapping, especially in lower and intermediately powered regions.

Of the six credible sets that changed upon adjustment in the high powered regions ($p < 1e-12$), a single variant was dropped in each instance. I observed that this happened in two ways: (1) in three cases, where the adjusted coverage was greater than the claimed coverage, a variant could be dropped from the set whilst keeping the adjusted coverage estimate greater than the target coverage (the threshold value of 95%) (2) in three cases, where the claimed coverage was greater than the adjusted coverage, both of the coverage estimates were so far above the target coverage (the threshold value of 95%) that it was possible to find a smaller set with claimed and adjusted coverage estimates closer to the target. Of the 22 credible sets that were adjusted in the low powered regions ($p \geq 1e-12$), the mean decrease in variants in the credible set after adjustment was 3.2 and the largest decrease in variants was for the genomic region with lead variant rs917911, for which the adjusted credible set contained 21 fewer variants (33 variants in the credible set before adjustment, 12 after adjustment). Overall, my method was able to reduce the number of putative causal variants for T1D in the 95% credible sets from 658 to 582 (Fig 3.13C).

Fine-mapping to single base resolution

Fine-mapping to single base resolution is often used as a measure of fine-mapping success. Two of the original 95% credible sets contained only a single variant: rs34536443 (missense in *TYK2*) and rs72928038 (intronic in *BACH2*). After applying my adjustment, two additional 95% credible sets were narrowed down from two variants to a single variant. First, rs2476601 (missense variant R620W in *PTPN22*) was selected, dropping rs6679677 which is in high LD with rs2476601 ($r^2 = 0.996$). These SNPs had high posterior probabilities ($PP = 0.856185774$ for rs2476601 and $PP = 0.143814226$ for rs6679677) and the adjusted credible set containing only rs2476601 had an adjusted coverage estimate of 0.9501 with a 95% confidence interval of (0.9392, 0.9613). It is likely that the missense variant R620W (rs2476601) is indeed the causal variant of *PTPN22*, perhaps through controlling the frequency of regulatory T cells in peripheral blood during T1D pathogenesis (Valta et al., 2020). Second, rs9585056 was selected, dropping rs9517719 ($r^2 = 0.483$). SNP rs9517719 is intergenic, whilst rs9585056 is in the 3' UTR of the lncRNA AL136961.1, and has also been shown to regulate expression of *GPR183*, which in turn regulates an IRF7-driven inflammatory network (Heinig et al., 2010). Whilst it is likely that R620W is indeed the causal variant at *PTPN22*, there is no conclusive data to evaluate whether rs9585056 is more likely to be causal compared to rs9517719. Nonetheless, the enrichment for missense variants was encouraging.

99% credible sets

The original manuscript reports 99% credible sets. For completeness, I briefly report the results from my analysis deriving 99% credible sets rather than 95% credible sets, noting the similarity of the results across threshold values.

Of the 16 regions with $p < 1e - 12$, four (25%) had claimed coverage estimates that were well calibrated (within the 95% confidence interval for the adjusted coverage estimate), five (31.25%) had claimed coverage estimates that were greater than the adjusted coverage estimates and seven (43.75%) had adjusted coverage estimates that were greater than the claimed coverage estimates. Of the 23 remaining regions with $p \geq 1e - 12$, only two regions (8.7%) had well calibrated claimed coverage estimates, and the remainder (91.3%) had adjusted coverage estimates that were greater than the claimed coverage estimates (Fig 3.14A).

Adjusted credible sets were constructed for regions where the threshold value (0.99) was not within the 95% confidence interval for the adjusted coverage estimate. As before, upon adjustment, credible sets for regions containing a lead variant with a larger p-value were generally more likely to be subject to change than those with smaller lead variant p-values (Fig 3.14B). Owing to the fact that there are more variants in a 99% credible set than the corresponding 95% credible set (due to larger coverage probability requirements), the decrease in the number of variants between the non-adjusted and the adjusted credible sets was higher for the 99% credible sets than the 95% credible sets. Of the 21 credible sets that were adjusted in the low powered regions ($p \geq 1e - 12$), the mean decrease in variants in the credible set after adjustment was 10.7 and the largest decrease in variants was again for the genomic region with lead variant rs917911, for which the adjusted credible set now had 80 fewer variants (144 variants in the credible set before adjustment, 64 after adjustment). Overall, my method was able to reduce the number of putative causal variants for T1D in the 99% credible sets from 1124 to 709 (Fig 3.14C).

In the original study, only one of the 39 regions contained a single variant in the 99% credible set (rs34536443; missense in *TYK2*). After my adjustment, the region with lead variant rs72928038 (intronic in *BACH2*) was also narrowed down to a single causal variant, dropping rs6908626.

Note that this 99% credible set is identical to the 95% credible set for this genomic region. Recently, Robertson et al. (2021) found that rs72928038 was likely to be the true causal variant in the region from to their own fine-mapping analysis (using GUESSFM). Moreover, their functional experiments showed that rs72928038 decreased enhancer accessibility and *BACH2* expression in T cells, potentially describing the associated biological mechanism of rs72928038 in T1D pathogenesis.

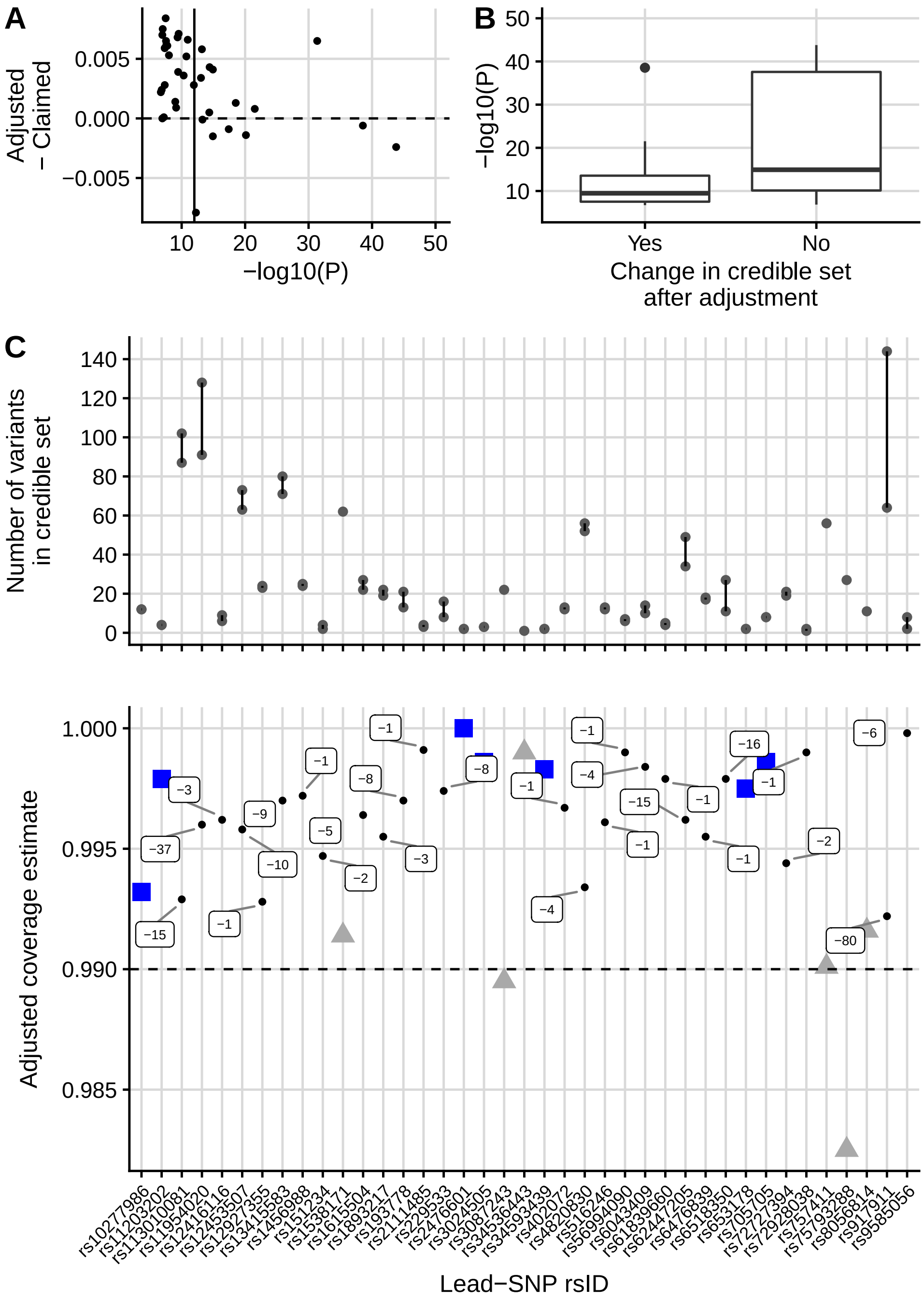


Fig. 3.14 (A) Adjusted - claimed coverage estimates against $-log_{10}(p)$ for the 99% credible set T1D analysis. Horizontal dashed line at adjusted = claimed coverage and vertical line at $-log_{10}(p) = 1e-12$. (B) Box plots showing the distribution of $-log_{10}(p)$ values for credible sets that either changed upon adjustment or did not change. Note that in panels A and B, results for the PTPN22 region with lead variant rs2476601 and $p < 1e-100$ has been omitted to aid visualisation. (C) Top panel: The decrease in size of the credible set after adjustment. Bottom panel: The adjusted coverage estimates of 99% Bayesian credible sets for T1D-associated genomic regions. Black points represent regions where the credible set changed after the adjustment and the "-" values represent the decrease in the number of variants from the standard to the adjusted 99% credible set. Blue square points represent regions where the credible set did not change after the adjustment and grey triangular points represent regions where the credible set did not need to be adjusted since the threshold (0.95) was contained in the 95% confidence interval of the conditional coverage estimate, or because the credible set already contained only a single variant.

3.8.2 Application to ankylosing spondylitis

Introduction

My collaborator (Dr James Lee, Department of Medicine, University of Cambridge, now at The Francis Crick Institute) was interested in fine-mapping "gene desert" regions (defined as genomic regions that were entirely located within intergenic regions) that associated with ankylosing spondylitis. The candidate causal variants identified from fine-mapping would then be used in massively parallel reporter assays (MPRAs), which the group had adapted for use in primary cells, to pinpoint the genetic variants that had a functional effect (in terms of transcription) in the relevant cell type. Gene deserts were selected because (i) less is known about how these predispose to disease compared with regions containing candidate genes, (ii) other non-coding mechanisms (such as splicing effects) are unlikely to account for these associations, and (iii) many of these genomic regions contain epigenetic marks consistent with enhancer activity (Bourges et al., 2020; Hnisz et al., 2013). Due to the expensive nature of the MPRAs, it was important that the fine-mapping results were of high resolution and were as accurate as possible.

Fine-mapping the 2p15 gene desert region

I downloaded the publicly available GWAS summary statistics for ankylosing spondylitis (International Genetics of Ankylosing Spondylitis Consortium (IGAS), 2013) from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) (study accession GCST005529). My collaborators were interested in the 2p15 gene desert region with lead variant rs6759298, which significantly associated with ankylosing spondylitis (GWAS $p = 5e - 47$). I defined the genomic region of interest to span from 50-kb upstream to 50-kb downstream of the lead variant (chr2:62518445-62618445) and extracted all the SNPs in this region that were available in the Immunochip data set.

For each SNP, I calculated the ABF using equation (1.15) (setting $W = 0.2^2$) and converted these to posterior probabilities using equation (3.1). I then used the standard fine-mapping method (Maller et al., 2012) to construct the 99% credible set. This set contained four SNPs: rs6759298, rs4672505, rs13001372 and rs6759003 (Fig 3.15) and had a claimed coverage estimate of 0.9999965. However, the adjusted coverage estimate of this set was 0.9985499 implying that a smaller credible set that still achieves the desired 99% coverage of the causal variant could potentially be found.

I used my adjustment method to construct the adjusted 99% credible set, which was found to contain only three variants: rs6759298, rs4672505 and rs13001372, dropping rs6759003 (Fig 3.15). This adjusted credible set had claimed coverage 0.9732124 and an adjusted coverage estimate of 0.9933525.

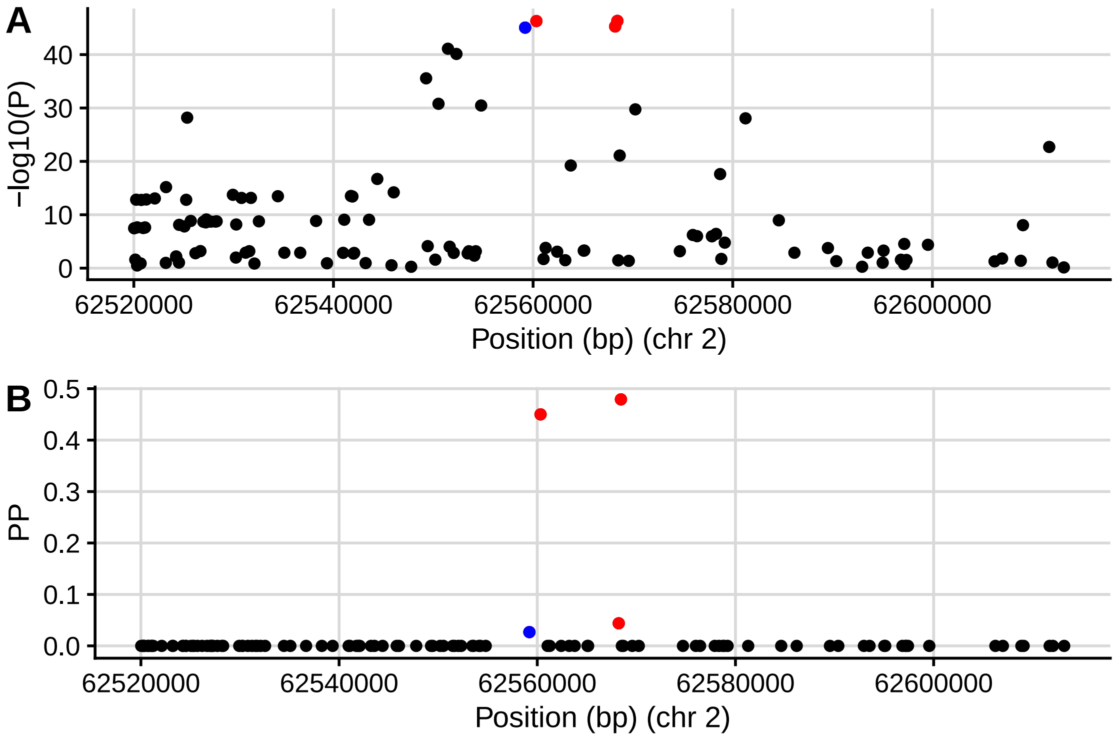


Fig. 3.15 (A) Manhattan plot of the 110 SNPs in the 2p15 genomic region. (B) Posterior probabilities for the 110 SNPs in the 2p15 genomic region. SNPs in the adjusted 99% credible set (rs6759298, rs4672505 and rs13001372) are coloured in red, and rs6759003 which only featured in the standard 99% credible set is coloured in blue.

The three SNPs in the adjusted 99% credible set were included for functional follow up in the MPRA analysis. My collaborator, Dr James Lee, performed the MPRA analysis and found that only one of these SNPs (rs6759298; $PP = 0.479$) had a significant effect on transcription in stimulated CD4 T cells, whereby the risk allele for ankylosing spondylitis (C) consistently increased transcription (Fig 3.16). This analysis was therefore successful at deciphering the likely causal variant for ankylosing spondylitis at the 2p15 locus.

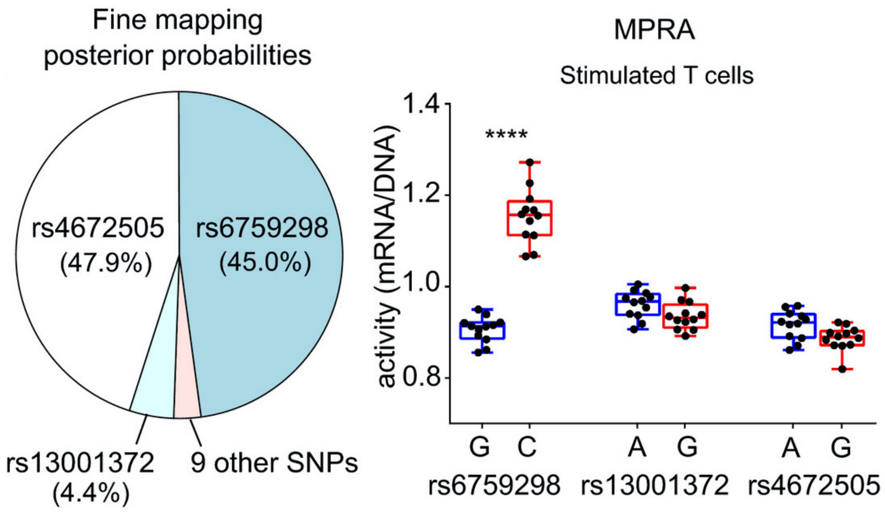


Fig. 3.16 Adapted from Fig 3 of Bourges et al. (2020) with permission. Pie chart depicting Bayesian fine-mapping results for an ankylosing spondylitis associated locus on 2p15 (left panel). MPRA results in stimulated T cells (right panel) showing that rs6759298 has a significant expression-modulating effect (the strongest of any variant at this locus) while the other candidate SNPs have negligible effects. Dr James Lee performed the MPRA analysis and created this figure.

3.9 Discussion

Statistical fine-mapping prioritises putative causal variants for functional validation. My adjustment method can be used to ensure that credible sets of putative causal variants from statistical fine-mapping are well empirically calibrated, which in turn enables efficient allocation of resources in the expensive functional follow-up studies. Statistical fine-mapping alone does not elucidate the mechanisms by which the causal variants operate to cause disease, and functional genomic techniques are required to dissect these mechanisms. For example, my method was used to reduce the number of variants in a 99% credible set for an ankylosing spondylitis locus from four variants to three, but an MPRA assay was then used to find which of the three remaining variants was most likely to be causal (Bourges et al., 2020).

The standard Bayesian approach for fine-mapping (Maller et al., 2012), and therefore my adjustment method, are limited in that they do not model multiple causal variants. Fine-mapping approaches that jointly model SNPs have been developed, such as GUESSFM (Wallace et al., 2015) which uses genotype data, and FINEMAP (Benner et al., 2016) and JAM (Newcombe et al., 2016) which attempt to reconstruct multivariate SNP associations from univariate GWAS summary statistics, differing both in the form they use for the likelihood and the method used to stochastically search the model space. The output from these methods are posterior probabilities for various configurations of causal variants, and therefore the grouping of SNPs to distinct association signals must often be performed post-hoc to obtain similar inferences to that of single causal variant fine-mapping. Newer software versions (e.g. FINEMAP v1.3 and v1.4) however are now able to produce lists of credible SNPs for the best multiple causal variant configurations. Thus, an interesting extension to this research would be to investigate whether the credible sets from multiple causal variant fine-mapping are subject to the conditional coverage biases exhibited in single causal variant fine-mapping.

The sum of single effects (SuSiE) method (Wang et al., 2020) removes the single causal variant assumption and groups SNPs to distinct association signals in the analysis, such that it aims to find as many credible sets of variants that are required so that each set captures a causal variant, whilst also containing as few variants as possible (similar to "signal clusters" in the DAP-G method (Lee et al., 2018)). This sophisticated approach has great potential, and the authors also state that "since our credible sets are Bayesian credible sets, 95% credible sets are not designed, or guaranteed, to have a frequentist coverage of 0.95". Their simulation analysis showed that the 95% credible sets formed using both the SuSiE and DAP-G methods "typically had coverage slightly below 0.95, but usually > 0.9", suggesting that a method analogous to that presented here could potentially be used to improve the empirical calibration of these credible sets. Alternatively, SuSiE could be used as an initial step for fine-mapping, and my adjustment method could then be used to improve the resolution of credible sets in the regions with evidence of a single causal variant.

Whilst my method does not address all the limitations of single causal variant fine-mapping, it improves upon the common inferences that are reported in the literature by researchers. I recommend that my adjustment is used as an extra step in the single causal variant fine-mapping pipeline, to obtain an adjusted coverage estimate that the causal variant is contained within the credible set and if required, to derive an adjusted credible set of putative causal variants.

Chapter 4

The role of the Ikaros family of transcription factors in T1D pathogenesis

This chapter systematically examines evidence for the role of the Ikaros family of transcription factors in T1D pathogenesis. The Ikaros family of transcription factors are well known for their importance in immune cell development (Powell et al., 2019), and genomic regions that contain the genes encoding the transcription factors have been found to associate with T1D in genetic association studies (Barrett et al., 2009; Todd et al., 2007). Yet beyond this statistical evidence there has been insufficient research examining the role of the Ikaros family in T1D pathogenesis. I begin by reviewing the genetic evidence from association studies linking members of the Ikaros family to T1D, and then inspect results from fine-mapping studies aiming to elucidate the causal variant(s) in the Ikaros gene region. Incorporating both published and unpublished ChIP-seq data from T1D-relevant cell types, I develop a modified version of an existing SNP enrichment method to examine whether genetic variants residing in genome-wide binding sites of Ikaros influence T1D risk. Finally, I use a case-only test for interaction to examine whether the causal variant in the Ikaros gene region acts synergistically with variants in Ikaros binding sites to influence T1D risk. I culminate this analysis by investigating whether Flexible cFDR can be used to detect these SNP-SNP interactions genome-wide whilst retaining high power by leveraging marginal GWAS test statistics.

4.1 Introduction

So far this thesis has been concerned with GWAS which investigates genome-wide patterns of association, and fine-mapping within these associated regions, but it is likely that studying a specific biological factor using a variety of data sources will give a more detailed perspective on disease pathogenesis than using genetic data alone. This chapter takes a more focused look at GWAS for T1D in relation to the Ikaros family of transcription factors, which have been linked to T1D for some time (Barrett et al., 2009; Todd et al., 2007) but to the best of my knowledge, have not been studied systematically in relation to their role in T1D pathogenesis. The goals of this chapter are to investigate the genetics relating the members of the Ikaros family of transcription factors to T1D at the genes encoding the transcription factors themselves, at the binding sites of the transcription factors in T1D-relevant cell types and at both the genes and the binding sites together.

Due to the explosion of GWAS research over the past decade and a half, there are now multiple overlapping genetic association studies for T1D, with the two largest GWAS for T1D published in *Nature* journals within one month of each other earlier this year (Chiou et al., 2021; Robertson et al., 2021). It is now common practice for researchers to supplement their genetic association results with a fine-mapping analysis aiming to elucidate the underlying causal variants responsible for the association signals. It is therefore useful to check whether GWAS and fine-mapping results are consistent across studies which vary in both the sample cohorts and methodologies that were used. I investigated whether this was the case in the genomic region containing the founding member of the transcription factor family, Ikaros.

Results from GWAS and fine-mapping alone do not generally implicate underlying biological mechanisms. For complex diseases in particular, it is difficult to envisage how associated loci with weak effect sizes that are widespread throughout the genome cohere to constitute biological mechanisms that are involved in the pathophysiology of a disease (which has been coined "the coherence problem") (Reimers et al., 2019; Turkheimer, 2011, 2012). Data from functional genomic experiments constitute an additional layer of information that can help towards the goal of interpreting association signals in terms of disease biology. For example, ChIP-seq data reveals genome-wide protein-DNA binding sites in specific cell types and cells states, and SNP enrichment approaches can be used to see if disease-associated loci overlap these sites more often than would be expected by chance, thus implicating potential disease-relevant proteins and cell types or cell states.

SNP enrichment analyses present their own challenges due to confounding variables. As introduced in section 1.4.3, GoShifter (Genomic Annotation Shifter) is a statistical method that tests for GWAS SNP enrichment within functional genomic annotations (Trynka et al., 2015), such as genomic regions corresponding to ChIP-seq peaks. The GoShifter method begins by identifying all variants that are in LD with the lead variants and defines a locus as the region between the furthest linked variants plus twice the median size of the tested annotation. Although capturing a large number of genetic variants, this protocol excludes some variants from the analysis, which may limit power (Cano-Gamez and Trynka, 2020). The GARFIELD method (Iotchkova et al., 2019), which uses a user-specified GWAS p-value threshold to determine SNP inclusion, found that some enrichments were only uncovered when using a more liberal p-value inclusion criteria (e.g. $p < 1e - 05$). It is possible that these SNPs with more modest association statistics may be missed by the SNP inclusion protocol in GoShifter. Moreover, it is generally acknowledged that the majority of trait-associated SNPs remain undiscovered (Visscher et al., 2017), and as the statistical power to detect associations increases, we can expect that many more enrichments may be uncovered by methods that use association statistics to determine SNP inclusion. I extended the GoShifter method so that it could be applied to all GWAS SNPs by defining a new test statistic. I then used this to test for the enrichment of T1D-associated variants in Ikaros binding sites in T1D-relevant cell types, using both published and unpublished ChIP-seq data.

When considering how genetic variants exert their effect, it is straightforward to assume an additive model whereby different loci act independently and cumulatively on disease risk. Yet due to the known complexity of biological mechanisms relating to disease pathogenesis, it is more intuitive to assume that loci may be acting synergistically on disease risk, for example when considering loci containing genes in similar pathways, with similar or redundant biological functions, or that code for proteins which function within the same protein complex (Seidl et al., 2016; TURNER and BUSH, 2011). The term "epistasis" is generally interpreted as the interaction between loci, although multiple definitions and assumptions of epistasis in the statistical and biological literature have led to widespread confusion in the field (Cordell, 2002). Here, I refer to epistasis as the scenario where the combined effect of two alleles on disease risk is greater than the sum of their marginal effects on disease risk (and use this term interchangeably with "synergy" and "SNP-SNP interactions").

The term epistasis was first defined in 1909 by Bateson (albeit to describe the specific scenario of one allele preventing an allele at another locus from manifesting its effect) (Bateson, 1909). Unfortunately, there has been limited success in the field of epistasis even over a century later, despite there being evidence for epistasis both at the molecular scale, for example from artificially induced mutations (Costanzo et al., 2010) and at the evolutionary scale, for example in fitness adaptation and speciation (Breen et al., 2012; Hemani et al., 2014). This lack of success has primarily been attributed to the huge search space to consider, which leads to limitations relating to statistical power. For example, testing for all possible pairwise interactions in a GWAS of one million SNPs would require $5e + 11$ statistical tests, which would correspond to (for example) an extremely stringent Bonferroni corrected significance threshold of $0.05/5e + 11 = 1e - 13$. The magnitude of this problem increases exponentially with increasing numbers of SNPs, and ignoring these interaction effects has, in part, been attributed to the "missing heritability problem" (Zuk et al., 2012). Approaches to increase the power to detect SNP-SNP interactions will ultimately help researchers to dissect genetic susceptibility to complex human diseases.

In the hope that a hypothesis-driven limitation of the search space would give greater power to detect SNP-SNP interactions, when examining potential interactions with the causal variant in the Ikaros gene region I used functional genomic data on Ikaros binding to limit my search space. In a secondary analysis, I investigated whether the Flexible cFDR approach could be used to increase power for a genome-wide SNP-SNP interaction exploration (Moore et al., 2020). Any significant findings from my analyses would not only be interesting in relation to the involvement of Ikaros in T1D pathogenesis, but they would also provide support for narrowing down the hypothesis search space or using Flexible cFDR as a means to increase power in other such studies examining interaction effects.

A popular method to test for SNP-SNP interactions is a case-only test, which is based on the premise that if the SNP genotypes are not associated in the general population, but are associated in case samples for the phenotype of interest, then this may indicate that the SNPs are interacting to implicate the phenotype of interest (Cordell, 2009). A χ^2 test is generally used to test for associations between genotypes, but when considering multiple GWAS studies for the same trait, a secondary step is required to meta-analyse the χ^2 statistics from the different studies. The Cochran-Mantel-Haenszel (CMH) test is essentially an extension of the χ^2 test for multiple contingency tables (Kuritz et al., 1988), and thus could be used to test for an association between the genotypes at the SNPs whilst controlling for the stratification by study. However the χ^2 test, and thus the CMH test, are low powered, requiring four degrees of freedom.

Alternatively, if an additive model is assumed and genotypes are treated as ordinal variables, such that the effect of a genotype value of 2 is twice the effect of a genotype value of 1 (as is standard in GWAS), then the generalised CMH test (Landis et al., 1978) is applicable. Briefly, the generalised CMH test is used to derive a CMH correlation statistic against the null hypothesis that there is no correlation between the genotypes at the SNPs. Similarly, if the genotypes are treated as continuous variables then a meta-analysis on the correlations between genotypes can be conducted using a fixed-effect or random-effect model. A fixed-effect model assumes that there is a true underlying correlation between the genotypes, and that any deviations from this true value are due to estimation error, whilst the random-effects model assumes that each study has its own true underlying correlation, and thus aims to estimate both the study-specific true correlations and the "grand" correlation underlying all studies (Bell et al., 2019).

4.2 Evidence linking members of the Ikaros family to T1D

In this section I introduce each member of the Ikaros family of transcription factors and then discuss the multiple lines of evidence linking members of the family to T1D, with a focus on genetic evidence from genetic association studies.

4.2.1 Background on the Ikaros family of transcription factors

The Ikaros zinc finger (IkZF) family of transcription factors are key regulators of immune and blood cell development (Powell et al., 2019). The founding member of the family, Ikaros (encoded by the gene *IKZF1*) was identified in 1992 (Georgopoulos et al., 1992), and this was subsequently followed by the identification of the remaining members of the family, Aiolos (encoded by *IKZF3*) (Morgan et al., 1997), Helios (encoded by *IKZF2*) (Hahm et al., 1998; Kelley et al., 1998), Eos (encoded by *IKZF4*) and Pegasus (encoded by *IKZF5*) (Perdomo et al., 2000).

The protein Ikaros is highly expressed in whole blood, the spleen and EBV-transformed lymphocytes (Fig 4.1). It is involved in hematopoietic stem cell development (Yoshida et al., 2006), where it modulates the expression of lymphoid genes by re-positioning chromatin remodelling complexes (Bottardi et al., 2014), and also acts as a tumour suppressor in T cells (Wang et al., 1996; Winandy et al., 1995). Complete Ikaros knockout mouse models lack B cells and T cells, and also have severe deficiencies in NK cells and dendritic cells (Nichogiannopoulou et al., 1998; Schmitt et al., 2002; Tonnelle et al., 2002). To investigate the role of Ikaros in mature immune cell differentiation, Lyon de Ana et al. (2019) developed a conditional knockout mouse model with Ikaros expression deleted specifically in mature T cells. They found that CD4 T cells were able to attain Th1, Th2 and Th17 cell fates, but not iTreg cell fate, implying that Ikaros may be required for iTreg differentiation. Moreover, the Th17 cells that were produced were qualitatively different from their wild-type counterparts, instead exhibiting a pathological phenotype associated with inflammation and autoimmunity, indicating that Ikaros may act to suppress this pathogenic phenotype (Lyon de Ana et al., 2019).

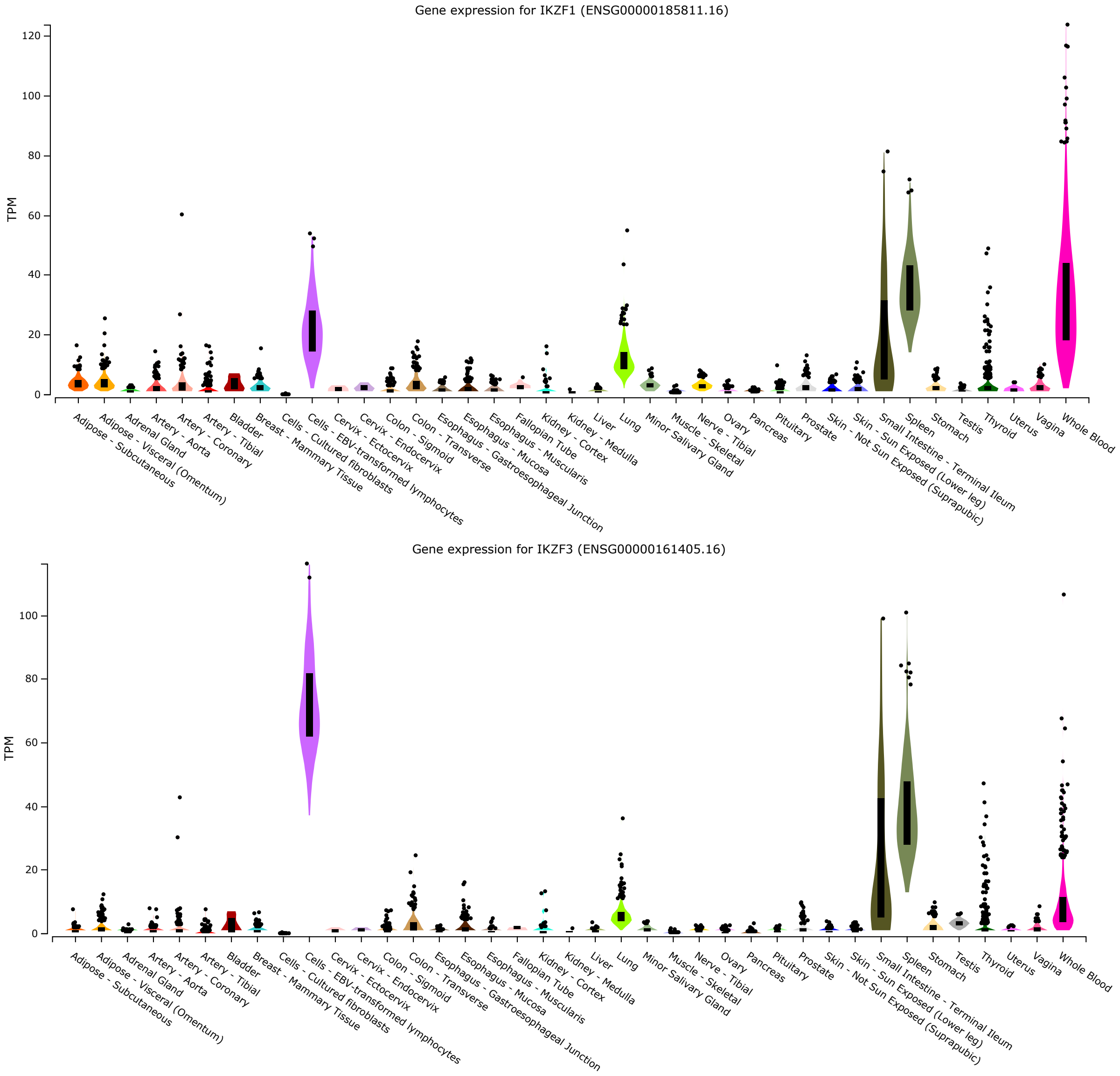


Fig. 4.1 *IKZF1* and *IKZF3* gene expression in tissues from GTEx RNA-seq of 17382 samples from 948 donors (V8, Aug 2019) (GTEx Consortium, 2013). Downloaded from https://www.gtexportal.org/home/gene/IKZF1 and https://www.gtexportal.org/home/gene/IKZF3 on 2021-06-16. Brain tissues were removed to aid visualisation (all $TPM < 10$). There was no data available on GTEx for *IKZF2*, *IKZF4* and *IKZF5*.

The second member of the family, Aiolos, is highly expressed in whole blood, the spleen and especially EBV-transformed lymphocytes (Fig 4.1). Aiolos is a transcriptional repressor and plays a key role in B-cell activation and differentiation (Morgan et al., 1997; Wang et al., 1998). Knockout of the Aiolos protein in mouse models results in chronic activation of B cells with increased levels of autoantibodies and the development of B-cell lymphomas (Vyse and Graham, 2020; Wang et al., 1998). Moreover, the expression of Aiolos is significantly up-regulated in chronic lymphocytic leukaemia cells, suggesting a tumour suppressor role (Duhamel et al., 2008; Nückel et al., 2009). Beyond B-cells, Aiolos is also implicated in T cell differentiation, where it directly silences *IL2* expression thereby promoting Th17 differentiation (Quintana et al., 2012).

The protein Helios is expressed throughout the T cell lineage, but is not expressed in most B lineage cells. Using transgenic mouse models, Dovat et al. (2005) showed that the silencing of Helios is critical for normal B cell function. The gene encoding Eos (*IKZF4*) is a paralog of Helios, and is preferentially expressed in Treg cells (Stelzer et al., 2016). Eos has been shown to interact directly with Foxp3 to regulate gene expression of key Treg-associated genes (Gokhale et al., 2019), and may therefore be involved in peripheral tolerance mechanisms that operate through Treg cells to dampen activation in secondary lymphoid tissues.

The role of the final member of the transcription factor family, Pegasus, was rather elusive until recently, when a role was established for Pegasus in megakaryocytopoiesis (Leinoe et al., 2021; Lentaigne et al., 2019). Germline mutations in *IKZF5* were shown to cause autosomal dominant inherited thrombocytopenia (a condition characterised by abnormally low levels of platelets in the blood), however no evidence of immune dysregulation was found by Lentaigne et al. (2019).

In all, four of the five members of the Ikaros family of transcription factors (Ikaros, Helios, Aiolos and Eos) have well-documented roles in the development of immune cell populations (Powell et al., 2019). Consequently, several researchers have examined the role of specific members of the family in the pathogenesis of specific IMDs (Cai et al., 2014; Chen et al., 2020; Hu et al., 2013; Lempainen et al., 2013; Qu et al., 2017; Sriaroon et al., 2019), with the majority of the relevant literature scrutinising the role of Ikaros in systemic lupus erythematosus (SLE) (Cai et al., 2014; Chen et al., 2020; Hu et al., 2013). For example, searching for “Ikaros SLE” on PubMed yields 25 results whereas “Ikaros T1D” yields only three results (accessed 2021-06-22). The research relating Ikaros to SLE is also newer than that for T1D, with the the most recent of the 25 articles being published this year by Rivellese et al. (2021) whom experimentally evaluated the potential of Ikaros and Aiolos as therapeutic targets in SLE. The first of the three related T1D articles revealed the enrichment of T1D GWAS signals in genome-wide targets of Aiolos (Burren et al., 2014), the second found an association between a polymorphism in *IKZF4* and insulin autoantibody positivity at the time of T1D diagnosis (Lempainen et al., 2013) and the third revealed positive results pertaining to expanding Treg cells in diabetics as a potential treatment (Du et al., 2013). However these articles are from 2013 and 2014, suggesting that now would be a suitable time to revive the field of research linking the Ikaros family of transcription factors to T1D pathogenesis.

4.2.2 Evidence linking Eos to T1D

The first genomic region that contained a member of the Ikaros family of transcription factors that associated with T1D in genetic association studies was cytoband 12q13, which contains the gene encoding Eos (*IKZF4*) (Hakonarson et al., 2008; Wellcome Trust Case Control Consortium, 2007). The lead variants reported in the region were rs11171739 ($OR = 1.34$, $p = 9.71e - 11$) (Wellcome Trust Case Control Consortium, 2007) and rs1701704 ($OR = 1.25$, $p = 9e - 10$) (Hakonarson et al., 2008) ($r^2 = 0.65$ between rs11171739 and rs1701704). (All SNP correlations reported in this chapter are calculated from European samples in the 1000 Genomes Project phase 3 data set.) The associated regions corresponding to these lead variants (a 450-kb block in Wellcome Trust Case Control Consortium (2007) and a 250-kb block in Hakonarson et al. (2008)) encompassed several other genes, and the limited functional genomic technologies and relatively small sample sizes that were used at this time of these studies (1963 T1D cases and 3000 controls in Wellcome Trust Case Control Consortium (2007) and 563 T1D cases and 1146 controls in Hakonarson et al. (2008)) meant that the causal variants and genes were not elucidated.

Five years later, Lempainen et al. (2013) sequenced 1554 children diagnosed with T1D before the age of 15, and found an intron variant in *IKZF4* (rs1701704) whereby the susceptible allele (C) was inversely associated with the presence of insulin autoantibody positivity at the time of T1D diagnosis (OR not published, $p = 0.00044$). These findings suggest that the *IKZF4* polymorphism is involved in an alternative biological pathway to that characterised by insulin autoantibody positivity in T1D. Due to the suggestive role of *IKZF4* in Foxp3-dependent gene signalling in Tregs (Gokhale et al., 2019), it has been postulated that this gene exerts its influence on T1D risk through Treg instability affecting peripheral tolerance mechanisms (Keene et al., 2012; Lempainen et al., 2013).

4.2.3 Evidence linking Aiolos to T1D

The next chromosomal region encompassing a member of the Ikaros family that associated with T1D was cytoband 17q12 containing the gene encoding Aiolos (*IKZF3*) (Barrett et al., 2009). Barrett et al. (2009) combined results from two previously published GWAS for T1D with their own genetic association study (total sample size of 7514 T1D cases and 9045 controls) and identified 41 genomic regions with suggestive evidence of an association with T1D ($p < 1e - 06$). The SNP rs2290400 in the 17q12 region was found to strongly associate with T1D ($OR = 1.15$, $p = 6e - 13$) but the causal gene was not elucidated in the study (rs2290400 is an intron variant in *GSDMB* and is 45-kb downstream of the transcriptional start site of *IKZF3*). The 17q12 genomic region also associates with many other IMDs, including ulcerative colitis, Crohn's disease and rheumatoid arthritis (Jostins et al., 2012; Stahl et al., 2010), suggesting that there may be an immune-mediating component underlying the associations.

Burren et al. (2014) developed a SNP enrichment method called "VSEAMS" and used this to test for the enrichment of T1D association signals near gene targets of specific transcription factors, including Aiolos (the other members of the Ikaros family were not tested in the study). Using data that identified significantly differentially expressed genes between controls and specific transcription factor knockout models in LCLs (i.e. those likely to be the targets of the specific transcription factors) (Cusanovich et al., 2014), the authors found that genes perturbed by 3 of 59 transcription factors in knockdown experiments were enriched for T1D association signals, including *IKZF3* ($p = 1.1e - 04$) (the other two genes were *BATF* ($p = 4.4e - 04$) and *ESRRA* ($p = 8e - 04$)).

Inshaw et al. (2020) stratified T1D patients by age of diagnosis (< 7 years and ≥ 13 years) and found that the genomic region containing *IKZF3* was differentially associated between these two groups of T1D patients. Specifically, the risk variants in the region had lower OR values (implying more protection from T1D) in the < 7 partition than in the ≥ 13 partition. In an attempt to find the underlying causative gene(s), the authors used a colocalization analysis with whole-blood eQTLs. They identified three genes whose expression was significantly altered by the variants in the region: *IKZF3* (whereby the variants in the region which decreased T1D risk associated with decreased *IKZF3* expression) and *GSDMB* and *ORMDLL3* (whereby the variants in the region which decreased T1D risk associated with increased gene expression). The results from this study suggests that risk variants in *IKZF3* may affect age of onset of T1D through the regulation of (at least) *IKZF3*, *GSDMB* or *ORMDLL3* expression in blood.

4.2.4 Evidence linking Ikaros to T1D

Variants in the Ikaros gene sequence have been linked to various IMDs, including asthma (Igartua et al., 2015), SLE (Cunninghame Graham et al., 2011; Han et al., 2009) and Crohn's disease (Barrett et al., 2008). Yet in terms of linking members of the Ikaros family of transcription factors to human diseases more generally, the best known example is that of Ikaros and acute lymphoblastic leukaemia (ALL), whereby certain SNPs in the *IKZF1* gene predispose patients to various types of leukaemia, including T-cell ALL (rs11978267; $OR = 1.69$, $p = 8.8e - 11$ in Treviño et al. (2009)) and childhood ALL (rs4132601; $OR = 1.69$, $p = 1.20e - 19$ in Papaemmanuil et al. (2009) and $OR = 1.66$, $p = 2.0e - 29$ in Wiemels et al. (2018)) ($r^2 = 0.986$ between rs11978267 and rs4132601). Strikingly, deletions in *IKZF1* were reported in 28.6% of childhood ALL affected adults (Martinelli et al., 2009) and in more than 70% of *BCR-ABL1* childhood ALL patients (those with the philadelphia chromosome caused by a translocation event affecting chromosome 22) (Marke et al., 2018; Mullighan et al., 2008).

Barrett et al. (2009) identified two SNPs with evidence of an association with T1D in the genomic region 7p12.2 where *IKZF1* resides: rs11980379 (T>C; $p = 2.5e-06$) which is close to the 3' UTR of *IKZF1* and rs10272724 (T>C; $p = 1.4e-06$) which is approximately 105-kb downstream of the transcriptional start site of *IKZF1*. These SNPs are both in very high LD with each other ($r^2 = 0.968$) and also the SNPs found to associate with ALL listed above (minimum r^2 between rs11980379, rs10272724, rs11978267 and rs4132601 is 0.955) suggesting that they may tag the same causal variant. SNPs rs11980379 and rs10272724 did not reach the significance threshold set by Barrett et al. (2009) for follow up ($p < 1e-06$) and the effect sizes were not available for these SNPs in the study data deposited on the NHGRI-EBI GWAS Catalog (Buniello et al., 2019).

Due to the immunological importance of *IKZF1*, its strong association with childhood ALL susceptibility and its moderate association with T1D, Swafford et al. (2011) further investigated the association of *IKZF1* with T1D by genotyping SNP rs10272724 in 8333 T1D cases, 9947 controls and 3997 families. The authors found that the same allele which conferred susceptibility to childhood ALL (the alternative allele, C) was protective for T1D ($OR = 0.87$, $p = 4.8e-09$). This pattern of opposite effect of variants associating with childhood ALL and T1D was also observed by Wiemels et al. (2018) at the associated SNP in the Aiolos gene region (rs2290400). In an attempt to give biological meaning to this seeming paradox, Wiemels et al. (2018) suggested that since rs2290400 disrupts a binding motif for the transcription factor B lymphocyte maturation-induced protein 1 (*BLIMP1*), and that *BLIMP1* is a known repressor of T-cell activation (and is expressed in early stages of B-cell development), it could be that the childhood ALL risk allele at this SNP favours B-cell differentiation whilst the alternative allele favours T-cell differentiation (childhood ALL starts in early forms of B cells whilst T1D is a T-cell mediated disease). That is, the SNP may exert its effect at key hematopoietic differentiation junctures. Swafford et al. (2011) found that the alternative allele of rs10272724 did not correlate with levels of two transcripts of *IKZF1* in peripheral blood mononuclear cells (i.e. there was no evidence of allele-specific expression) but stated that there were more than 30 reported interaction partners of the Ikaros family members, ultimately suggesting that the role of Ikaros in T1D could be "subtle yet far reaching".

4.2.5 Summary

Overall, genomic regions encompassing three members of the Ikaros family of transcription factors have been linked to T1D in genetic association studies: Ikaros (7p12.2), Aiolos (17q12) and Eos (12q13.2) (Barrett et al., 2009; Todd et al., 2007). Whilst there is no evidence of Helios and Pegasus being linked to T1D, the former has been linked to SLE (Ferreira et al., 2019) and the latter has been linked to ulcerative colitis (Anderson et al., 2011), suggesting that there may be an immune-mediating component.

4.3 Fine-mapping the Ikaros gene region

Several attempts have been made to pinpoint the specific causal variant(s) for T1D in the *IKZF1* gene region. In this section I investigate whether these fine-mapping results were consistent across a variety of studies that utilised different fine-mapping methodologies with different genetic association data.

4.3.1 Methods

The first GWAS for T1D with sufficient SNP coverage and sample sizes to attempt robust fine-mapping in the genomic region containing *IKZF1* was by Onengut-Gumuscu et al. (2015). This study has been described in the previous two chapters of this thesis. Recently, Asimit et al. (2019) used an alternative fine-mapping method, GUESSFM (Wallace et al., 2015), to fine-map the *IKZF1* gene region using the case-control data from Onengut-Gumuscu et al. (2015).

Barrett et al. (2009) conducted a GWAS for T1D in 5913 case samples and 8828 control samples by meta-analysing three study cohorts: (i) T1DGC (Rich et al., 2009) (ii) GoKinD/NIMD (Cooper et al., 2008) and (iii) WTCCC (Wellcome Trust Case Control Consortium, 2007). The data was re-analysed by Cooper et al. (2017), who excluded the USA samples and used IMPUTE2 (Howie et al., 2011) to impute unmeasured genotypes using the 1000 Genomes Project phase 3 cohort as the reference data. As part of a different project, my supervisor Dr Chris Wallace used SuSiE (Wang et al., 2020) to fine-map LD detect regions using the Cooper et al. (2017) data, generating 95% credible sets of putative causal variants (Table 4.1).

Two larger GWAS for T1D, that both include complementary fine-mapping analyses, have been published in the past year. Firstly, Robertson et al. (2021) conducted a large T1D GWAS with a focus on dense coverage of SNPs residing in Immunochip regions. The authors genotyped $140,333$ Immunochip SNPs in $16,159$ European T1D cases and $25,386$ European controls and then imputed up to $715,000$ SNPs using the TOPMed reference panel (Freeze 5) (Taliun et al., 2021). For fine-mapping, they considered all SNPs within a 1.5-Mb window of the lead variants from their GWAS, and used the GUESSFM method (Wallace et al., 2015) to prioritise likely causal configurations of SNPs (Table 4.1).

Chiou et al. (2021) described the largest T1D GWAS to date, a meta-analysis of $18,942$ T1D patients and $501,638$ control participants of European ancestry from 9 cohorts. The authors used the TOPMed Imputation Server (Das et al., 2016) to impute genotypes into the TOPMed panel (version R2) (Taliun et al., 2021), resulting in an analysis of $61,947,369$ variants spanning the whole genome. For fine-mapping, they considered variants within 1-Mb of each lead variant and used FINEMAP (Benner et al., 2016) to construct 99% credible sets of putative causal variants (Table 4.1).

GWAS reference	Number of cases	Number of controls	Fine-mapping method	IKZF1 fine-mapping region
Onengut-Gumuscu et al. (2015)	6670	12,262	Standard Bayesian approach	chr7:50508528-50558528
Cooper et al. (2017)	5913	8828	SuSiE	chr7:49172682-51607626
Robertson et al. (2021)	16,159	25,386	GUESSFM	chr7:49656053-51156053
Chiou et al. (2021)	18,942	501,638	FINEMAP	chr7:49875028-50875028

Table 4.1 Table containing information on the four studies that fine-mapped the *IKZF1* gene region, including the number of case and control samples used in the fine-mapping analysis. The standard Bayesian approach refers to that described by Maller et al. (2012), SuSiE is described by Wang et al. (2020), GUESSFM is described by Wallace et al. (2015) and FINEMAP is described by Benner et al. (2016). Genomic co-ordinates given with respect to human build GRCh38/hg38.

4.3.2 Results

The Onengut-Gumuscu et al. (2015) data consists of case and control samples, as well as 2601 affected sibling pairs and 69 trio families. The authors identified rs62447205 (A>G, $OR = 0.890$, $p = 2.5e-08$) as the lead variant in the *IKZF1* gene region when using their meta-analysis including affected sibling pairs and trio families, and found no evidence of any additional independent signals when using a conditional regression approach. To fine-map the *IKZF1* gene region, they used only their case-control data (i.e. excluding affected sibling and trio family samples) and considered all SNPs within a 50-kb window of rs62447205 (corresponding to the genomic region chr7:50508528-50558528). All genomic co-ordinates in this chapter are given with respect to human build GRCh38/hg38. When using only their case-control data, the authors identified a new lead variant in the *IKZF1* gene region (rs11770117), for which the alternative allele was now susceptible rather than protective for T1D (Fig 4.2). SNP rs11770117 (A>T, $OR = 1.137$, $p = 5.8e-09$) resides close to the 3' UTR region of *IKZF1* and is only in moderate LD with rs62447205 ($r^2 = 0.41$). The 99% credible set of putative causal variants in this region contained 43 variants, and my adjustment method described in the previous chapter of this thesis narrowed down this set to 28 variants. The SNP with the largest posterior probability was rs11770117 ($PP = 0.103$), but there were other variants contained within the credible set with relatively high posterior probabilities (Fig 4.3), and so we cannot conclusively say that fine-mapping has resolved the likely causal variant in this instance. Of note, SNP rs62447205 also resided in the credible set but only had $PP = 0.025$.

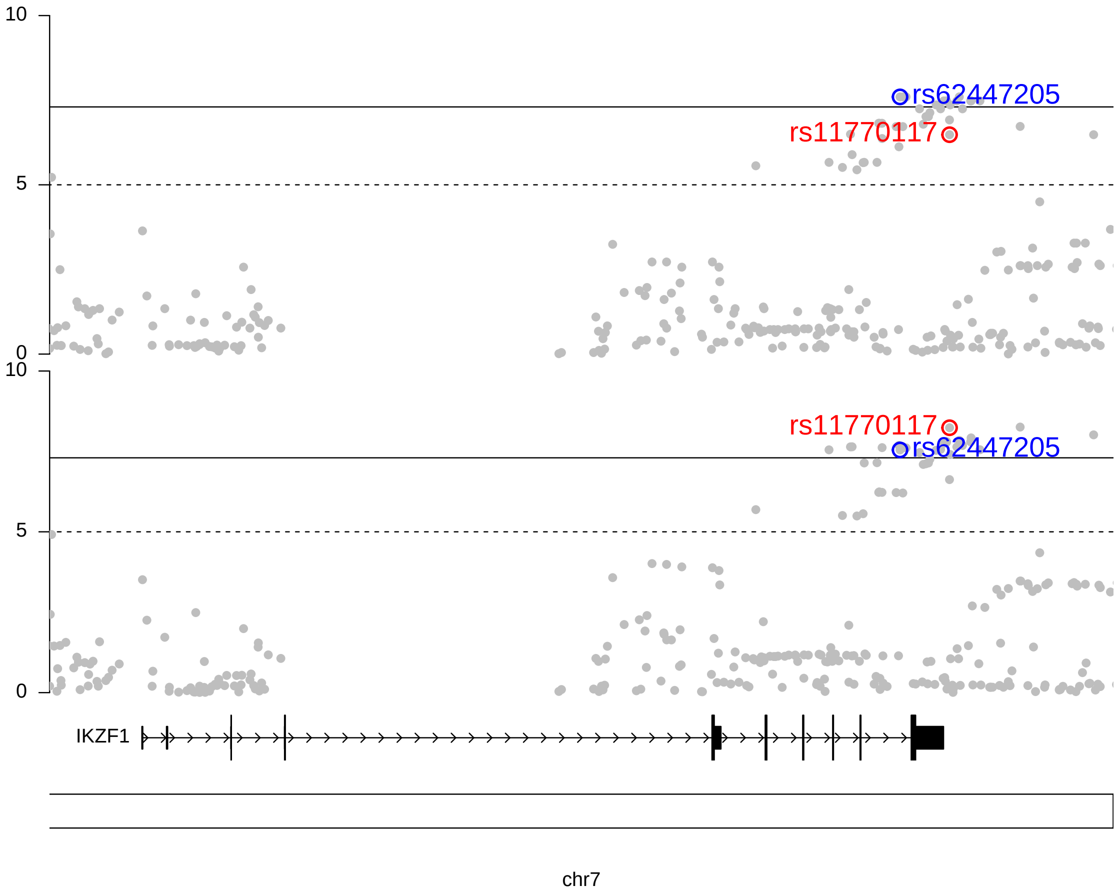


Fig. 4.2 Zoomed in Manhattan plot of T1D GWAS results from Onengut-Gumuscu et al. (2015) for the genomic region containing *IKZF1* (chr7:50290000-50500000). The y-axis is the $-log_{10}(p)$ values and the x-axis is the genomic position. The top panel is the GWAS results for the meta-analysis including affected sibling pairs and trio families, the bottom panel is the GWAS results for the case-control GWAS used for fine-mapping. rs62447205 is labelled in blue and rs11770117 is labelled in red. Dashed line at $p = 1e - 05$ and solid line at $p = 5e - 08$. Figure generated using summary statistics deposited on the NHGRI-EBI GWAS Catalog (accession GCST005536) and the `karyoploteR` R package (http://bioconductor.org/packages/release/bioc/html/karyoploteR.html).

When using SuSiE for fine-mapping, Cooper et al. (2017) identified two 95% credible sets in the LD detect region containing *IKZF1*, suggesting two independent signals. However, owing to the relatively large size of the LD detect regions compared to the window-based region definitions in the other methods (the LD detect region containing *IKZF1* was approximately 2-Mb whereas Onengut-Gumuscu et al. (2015) only considered variants within a 50-kb region), one of the credible sets identified by SuSiE resided 500-kb downstream of the *IKZF1* gene and contained intronic variants in *GRB10*, which have no evidence of being linked to *IKZF1* (using information on the Open Targets Genetics portal) (Carvalho-Silva et al., 2019). I therefore focussed on the

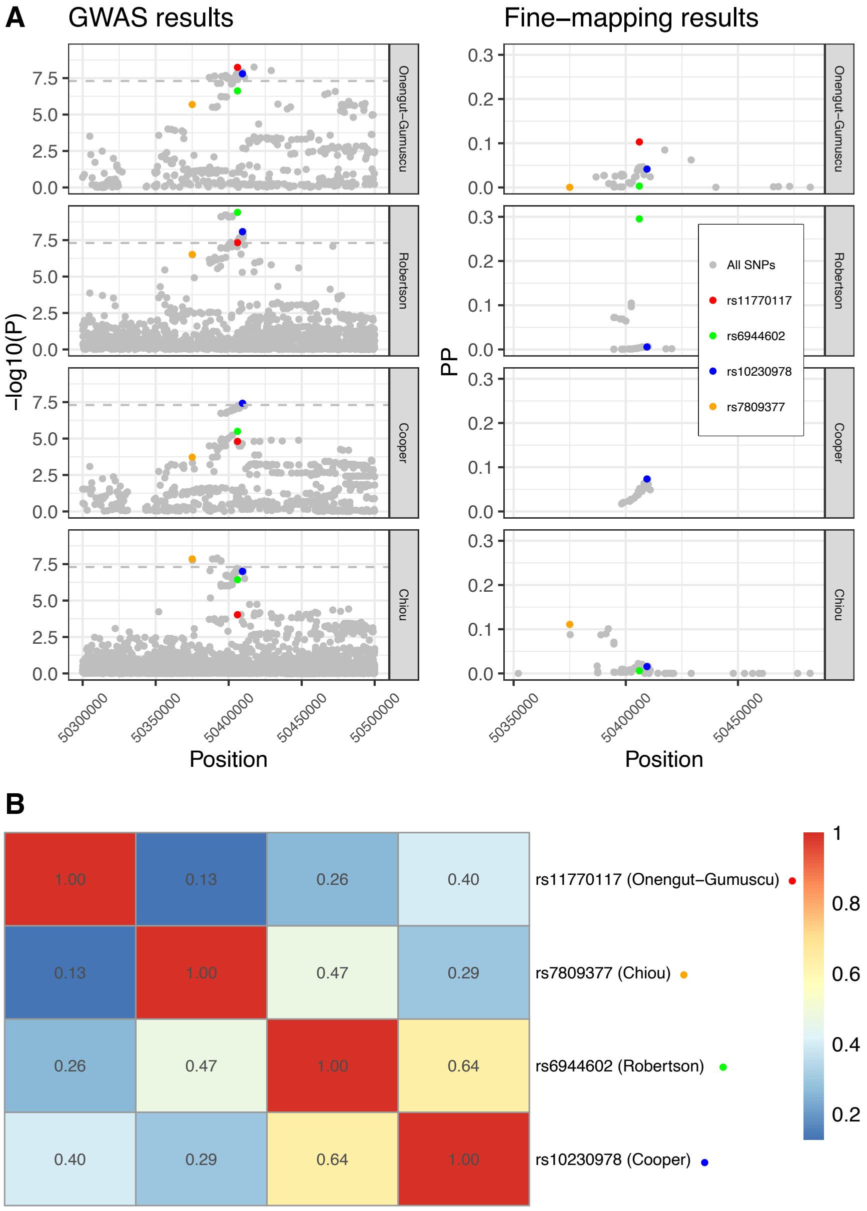


Fig. 4.3 (A) GWAS and fine-mapping results for the chr7:50300000-50500000 region containing the 3' UTR of *IKZF1* faceted by study. Left-hand column shows Manhattan plots for the GWAS results and the right-hand column shows plots of the posterior probabilities from fine-mapping. The lead variant in each study is coloured. GWAS results for Onengut-Gumuscu et al. (2015), Robertson et al. (2021) and Chiou et al. (2021) were downloaded from the NHGRI-EBI GWAS Catalog (accessions GCST005536, GCST90013445 and GCST90014023, respectively) and GWAS results for Cooper et al. (2017) were curated by a member of our group, Dr Guillermo Reales, as part of a larger project. Fine-mapping results for Onengut-Gumuscu et al. (2015), Robertson et al. (2021) and Chiou et al. (2021) were downloaded from the supplementary material of the relevant manuscripts (Supplemental Table 1, Supplemental Table 11 and Supplemental Data 1, respectively) and fine-mapping results for Cooper et al. (2017) were generated by my supervisor, Dr Chris Wallace. (B) Pairwise correlation coefficients between the lead variants identified in the four studies. Adjacent coloured dot relates to the colour of the relevant SNP in panel (A). The correlation values were obtained from the LDmatrix Tool in the LDlink suite of web-based applications (https://ldlink.nci.nih.gov/?tab=home), where data from the 1000 Genomes Project is used to derive pairwise correlation coefficients for a list of SNPs (Machiela and Chanock, 2015) (for this I specified European samples only, since the studies were conducted in Europeans).

remaining credible set which overlapped my specific region of interest (the 3' UTR of *IKZF1*). This 95% credible set contained 23 variants, for which SNP rs10230978 (G>A, $OR = 0.857$, $p = 3.80e-08$) held the largest posterior inclusion probability ($PIP = 0.0735$) (Fig 4.3). However, this SNP was only in moderate LD with the SNP prioritised in Onengut-Gumuscu et al. (2015) ($r^2 = 0.40$ with rs11770117) (Fig 4.3B) and there were other SNPs allocated high posterior inclusion probabilities in the region (for example rs17133805 with $PIP = 0.0730$). Of note, rs10230978 was in high LD with the lead variant identified in the meta-analysis including family samples described in Onengut-Gumuscu et al. (2015) ($r^2 = 0.955$ between rs10230978 and rs62447205).

When using GUESSFM for fine-mapping, Robertson et al. (2021) identified six credible sets in the large 1.5-Mb region containing the 3' UTR region of *IKZF1*. However, only two of these had high confidence of actually existing - that is, they contained variants with posterior probabilities that summed to > 0.99 (the maximum sum of the other four credible sets was 0.14). One of these high confidence sets resided 500-kb downstream of *IKZF1* (containing intronic variants in *GRB10*) and so was excluded from my analysis, whilst one resided close to the 3' UTR region of *IKZF1*. Using a case-control meta-analysis across four ancestry clusters, Robertson et al. (2021) identified rs6944602 (G>A) as the lead variant in this region ($OR = 0.901$, $p = 3.78e-09$). When combining this ancestry meta-analysis with another meta-analysis across five family-based ancestry clusters, the p-value for this SNP reduced to $p = 3.963e-10$ (Fig 4.3). The relevant credible set contained 37 variants and SNP rs6944602 held the majority of the posterior probability ($PP = 0.3$) (Fig 4.3). However, this SNP was only in moderate LD with the SNPs prioritised in the other stuides (Fig 4.3B).

Chiou et al. (2021) identified two independent signals in the large 1-Mb region containing the 3' UTR region of *IKZF1*. I omitted the signal that was 100-kb upstream of *IKZF1* from my analysis. Close to the 3' UTR of *IKZF1*, rs7809377 (G>A) was identified as the lead variant ($OR = 0.883$, $p = 1.4e-08$) and had the largest posterior probability of association in the fine-mapping analysis ($PPA = 0.110915$) (Fig 4.3). This variant was in low to moderate LD with the SNPs prioritised in the previous studies (Fig 4.3B).

The studies explored so far have used different GWAS data for fine-mapping, which could explain some of the discrepancies in the results for the *IKZF1* gene region. Asimit et al. (2019) used the GUESSFM approach to fine-map the *IKZF1* gene region using data from Onengut-Gumuscu et al. (2015) (note that Asimit et al. (2019) describes the multinomial fine-mapping (MFM) approach but since the *IKZF1* gene region did not associate with any other immune-related traits included in their study, this reduced to fine-mapping using GUESSFM). The authors found evidence of two physically overlapping sets of SNPs responsible for two independent association signals close to the 3' UTR of *IKZF1* (Fig 4.4). The lead variant identified by

Onengut-Gumuscu et al. (2015) (rs11770117) was a member of one of the credible sets of SNPs, but this set had a small overall posterior probability (≈ 0.25). None of the lead variants from Cooper et al. (2017), Robertson et al. (2021) or Chiou et al. (2021) appeared in either of the credible sets (even though they were genotyped in the study), but the lead variant identified from the meta-analysis including affected sibling and trio family samples in Onengut-Gumuscu et al. (2015) (rs62447205) was in the larger credible set with overall posterior probability ≈ 0.75 (even though Asimit et al. (2019) used the case-control data rather than the data including affected sibling and trio family samples for fine-mapping).

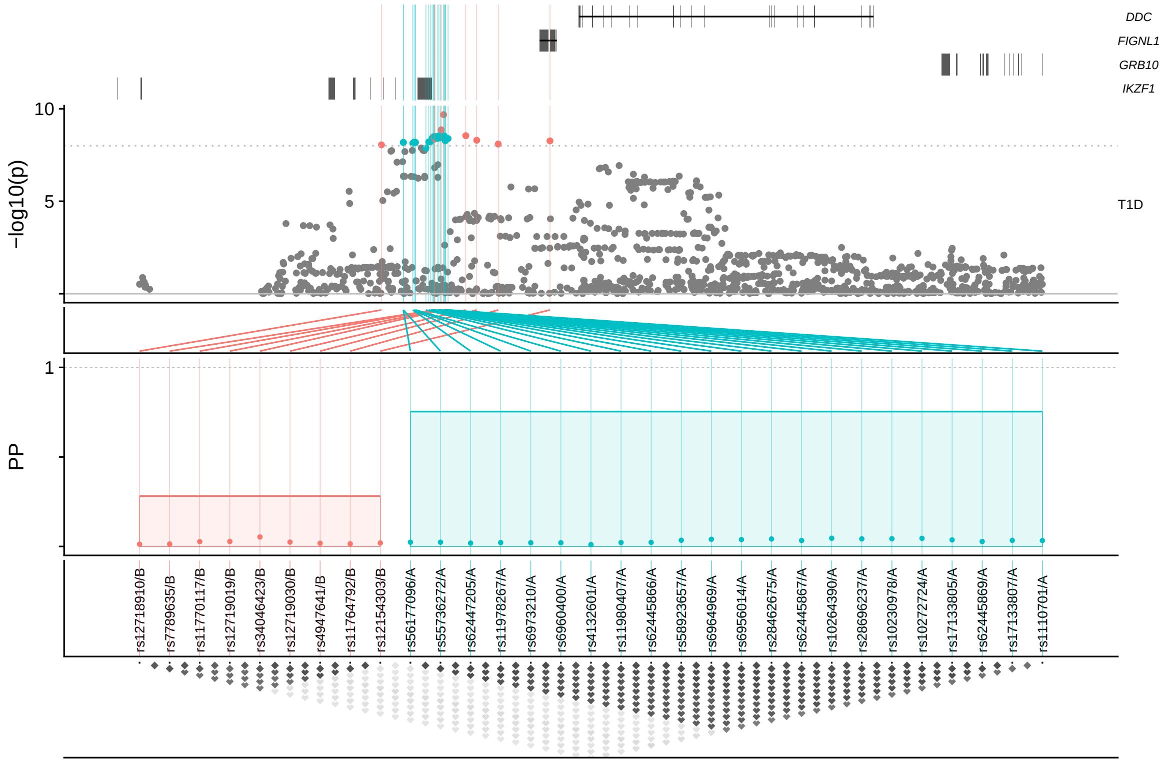


Fig. 4.4 Figure downloaded from https://chr1swallace.github.io/MFM-output/ and generated by Dr Chris Wallace showing evidence for two independent physically overlapping association signals for T1D in the *IKZF1* gene region when using GUESSFM to fine-map the case-control GWAS data from Onengut-Gumuscu et al. (2015) (Asimit et al., 2019).

The secondary signal close to the 3’ UTR of *IKZF1* may have been missed in the original study (Onengut-Gumuscu et al., 2015) due to the fine-mapping approach that was used, which has a restrictive single causal variant assumption (although when using a conditional regression approach, Onengut-Gumuscu et al. (2015) found no evidence of any additional signals in the region). That being said, this secondary signal was not identified by any of the other fine-mapping approaches that allowed for multiple causal variants (e.g. SuSiE and FINEMAP). Interestingly, Robertson et al. (2021) also used the GUESSFM method (albeit on different

GWAS data) but found no evidence of this secondary physically overlapping signal, even though the GWAS sample sizes were considerably larger (that is, there were no overlapping sets of variants near to the 3' UTR region of *IKZF1*, even when considering all six credible sets). Formally investigating the performance of different fine-mapping methods was outside the scope of this project, but my analysis has highlighted that this is an important area of future research.

Due to the differing results across studies, individual study conclusions on the likely causal variant(s) in the *IKZF1* gene region should be interpreted cautiously. That being said, the GWAS signal was strongest for the data from Robertson et al. (2021) and this study also identified a SNP with a large posterior probability of causality relative to the other SNPs (Fig 4.3). Therefore, albeit with caveats relating to replicability and reproducibility, SNP rs6944602 identified by Robertson et al. (2021) may be a good candidate for the causal variant in the region. Using the Open Targets Genetics portal (Carvalho-Silva et al., 2019), I found that there was eQTL evidence relating to this SNP and *IKZF1* expression in whole blood, and also various blood cell types relating to T1D, whereby the alternative allele (which is protective for T1D) is predicted to decrease *IKZF1* expression ($\beta = -0.443$, $p = 3.3e - 310$ in whole blood, $\beta = -0.494$, $p = 1.8e - 35$ in CD4 T cells and $\beta = -0.523$, $p = 5.9e - 30$ in CD19 B cells).

4.4 Investigating T1D susceptibility in target binding regions of Ikaros

I have used genetic data to investigate T1D susceptibility in the gene regions of Ikaros. I now expand my analysis to include various functional genomic data sets relating to Ikaros binding sites in T1D-relevant cell types to examine genetic susceptibility for T1D in genome-wide binding sites of Ikaros.

4.4.1 Methods

Existing Ikaros ChIP-seq data

I used the ReMap2020 database (Chèneby et al., 2020) to search for existing publicly available Ikaros ChIP-seq data sets (accessed 2020-06-20). Of the 9 Ikaros ChIP-seq data sets for experiments conducted in humans that were publicly available at the time of this research, three were conducted in LCLs from B-lymphocytes (GM12878), one in liver cell lines (Hep-G2), two in cancer cell lines (K-562), one in hematopoietic stem and progenitor cells (HSPCs) and two in BCR-ABL1 positive cell lines (whereby the BCR/ABL gene is introduced into the TF-1 leukemia cell line to generate TF-1 BCR/ABL cells). As lymphocytes are known to express Ikaros and HSPCs can differentiate into lymphocytes, I selected the ChIP-seq data for the experiments conducted in these cell types for my analysis. I also included the Ikaros ChIP-seq data conducted in liver cell lines to use as a negative control.

The first publicly available Ikaros ChIP-seq data set for LCLs (ENCODE experiment accession: ENCSR000EUJ) was generated in 2010, but the ENCODE audit indicated that the quality of the data was poor. Specifically, the ENCODE data audit produced warnings relating to both the control data and the experimental data, indicating that they both had very low sequencing depth and read length. The number of peaks confidently called in this experiment was also very low (≈ 1000) and I therefore decided to exclude this data set from my analysis.

The two remaining Ikaros ChIP-seq data sets for LCLs (ENCODE experiment accessions: ENCSR874AFU and ENCSR441VHN) were generated in 2016 and 2017 respectively, and each contained paired-end data for two isogenic replicates. The ENCODE audit suggested that the quality of this data was satisfactory overall, however there were warnings for “mild to moderate bottlenecking” and “antibody characterised with exemption”. The former relates to the PCR bottleneck coefficient, defined as the fraction of genomic locations with *exactly* one unique read versus those covered by *at least one* unique read, and thus is a measure of the library complexity. The antibody characterisation warning indicates that the antibody did not pass its primary characterisation test (an immunoprecipitation followed by a Western blot) but that it did pass the second, which in the case of Ikaros was a motif enrichment analysis showing that there was enrichment of the Ikaros binding motif in the ChIP-seq experiment (https://www.encodeproject.org/antibodies/ENCAB590IRI/). The number of peaks called from these two experiments were satisfactory ($\approx 85,000$). I downloaded the optimal irreproducible discovery rate (IDR) thresholded peaks for these data sets, which are the largest set of peaks derived from an analysis measuring the consistency between replicates (Li et al., 2011).

The Ikaros ChIP-seq data set for HSPCs (GEO experiment accession: GSE26014) was generated in 2011, only had a very low number of peaks (1,682) and had no indication of data quality, and so I decided to exclude this data from my analysis. The Ikaros ChIP-seq data for liver cell lines (ENCODE experiment accession: ENCSR278JQG) was generated in 2016 and the ENCODE audit produced only a single warning for borderline replicate concordance, meaning that there may be low reproducibility between replicates. Since this warning was only “borderline” rather than “low” or “insufficient” I decided to proceed with this data as a negative control.

Existing data on predicted Ikaros binding sites

An alternative approach to ChIP-seq experiments to examine protein-DNA binding events is to combine genomic footprinting with known DNA binding motifs. Funk et al. (2020) developed a framework to predict the binding sites of 1515 transcription factors, including Ikaros, in 27 human tissues. They used the HINT (HMM-based identification of transcription factor footprints) (Gusmao et al., 2014, 2016) and Wellington (Piper et al., 2013) algorithms to identify genomic footprints using many DNase-seq data sets, and then used the FIMO (Find Individual Motif Occurrences) method (Grant et al., 2011) to predict which transcription factor occupied each footprint. To prevent the very high false positive rates ($FDR > 0.9$) that arise when using the approach naively, the authors recommend using only those results that have a footprint score > 200 (derived using the HINT algorithm with seed length 20, the “HINT20 score”).

I downloaded the motif-to-transcription factor mappings derived using the HINT20 method for lymphoblast tissues from http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas/bed/. This data was generated from 21 biosamples (for a mixture of FAIRE-seq and DNase-seq experiments) in LCLs, and these were used to predict regions of the genome that were bound by a protein. The FIMO method was then used to predict which transcription factor was most likely to bind at each footprint. I extracted the rows relating to the Ikaros binding motif in humans (“Hsapiens-HOCOMOCOv10-IKZF1_HUMAN.H10MO.C”) and kept only those mappings with a HINT20 score > 200, as recommended by the authors. For any duplicates, I kept the row with the largest HINT20 score.

A related manuscript by Vierstra et al. (2020) described transcription factor occupancy using DNase I footprints in over 240 human cell types and cell states. However, the prediction of transcription factor binding occupancy was done at the archetype level, and the archetype containing the Ikaros motif in humans also contained six other motifs (Fig 4.5). I did not include this data in my analysis because the data was only available at the archetype level, but it was useful to note that the Ikaros motif is a subset of motifs for other transcription factors, and I revisit this observation in section 4.6.

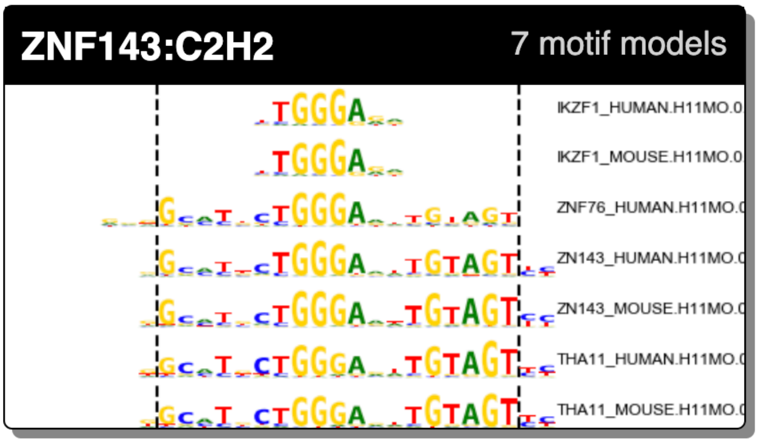


Fig. 4.5 Figure showing the transcription factor archetype containing Ikaros (IKZF1) downloaded from https://resources.altius.org/~jvierstra/projects/motif-clustering/releases/v1.0/cluster_viz.html (Vierstra et al., 2020).

Analysis of raw Ikaros ChIP-seq data conducted in activated CD4 T cells

My collaborator, Dr Antony Cutler (JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, University of Oxford) conducted a ChIP-seq experiment for Ikaros binding in activated CD4 T cells using samples from a healthy donor. The human biological samples were sourced ethically and their research use was in accordance with the terms of the informed consents under an IRB/EC approved protocol. Dr Antony Cutler generated `fastq` files for the raw single-ended reads. I then used `FastQC` (http://www.bioinformatics.babraham.ac.uk/projects/fastqc; version 0.11.4) to check the quality of the raw ChIP-seq reads, and have made this QC report publicly available via the following link: https://htmlpreview.github.io/?https://github.com/annahutch/thesis-things/blob/main/IKZF_S2_L001_R1_001_fastqc.html. This QC report showed that the data had already been trimmed (the sequence length varied between 35 to 51, if it hadn't yet been trimmed then all sequence lengths would be 51) and that there was no adapter contamination. There was a warning for per base sequence content, but this is to be expected in ChIP-seq data because the transposase used is not completely random (hence the bias in the positions at which the reads start) and because the reads primarily reside in open chromatin regions (hence the slight AT bias throughout the length of the read). The bias in sequence content at the end of the read is likely a result of the adapter trimming but should not affect peak calling results. I used `Bowtie2` (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; version 2.4.2) to map the reads back to the reference human genome (GRCh38.p13, obtained from the GENCODE project; Harrow et al. (2012)), for which there was a 97.87% overall alignment rate (and a 98.60% overall alignment rate for the mapped input sample). I converted the mapped data to `BAM` format using `samtools` (Li et al., 2009).

I used `SeqMonk` (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/; version 1.47.2) for peak calling, quantitation and annotation, and created a "vistory" describing each step of the analysis, which is publicly available via the following link: https://htmlpreview.github.io/?https://github.com/annahutch/thesis-things/blob/main/chipseq-analysis.html. As recommended in the documentation and on the "Analysis of mapped NGS data with SeqMonk" training course that I attended (a University of Cambridge Bioinformatics Training Course), upon data input to `SeqMonk` I extended read lengths by 250-bp (since they were single-ended reads) and filtered reads to only include high quality reads (mapping quality > 20). I then used the `MACS` (Model-Based Analysis for ChIP-seq) algorithm (Zhang et al., 2008) within `SeqMonk` for peak calling with $p = 1e - 05$, a sonicated fragment size of 300-bp and the ChIP-seq input sample as control. I defined coverage outliers as those regions in the input data set with a read count greater than 50 and excluded these regions from my analysis. I extracted the remaining 4910 peaks using "the annotated probe report" feature of `SeqMonk`.

Overview of relevant data on Ikaros binding

Overall, I acquired five genomic annotation data sets relating to Ikaros binding: (1) Ikaros ChIP-seq peaks in activated CD4 T cells (2) Ikaros ChIP-seq peaks in LCLs (ENCODE accession: ENCSR441VHN) (3) Ikaros ChIP-seq peaks in LCLs (ENCODE accession: ENCSR874AFU) (4) Predicted Ikaros binding sites in LCLs and (5) Ikaros ChIP-seq peaks in liver cell lines (ENCODE accession: ENCSR278JQG). Encouragingly, when comparing the distribution of peaks across chromosomes for these data sets, they were similar for the immune-related cell types (activated CD4 T cells and LCLs) but not for the liver cell line, which I included as a negative control (Fig 4.6).

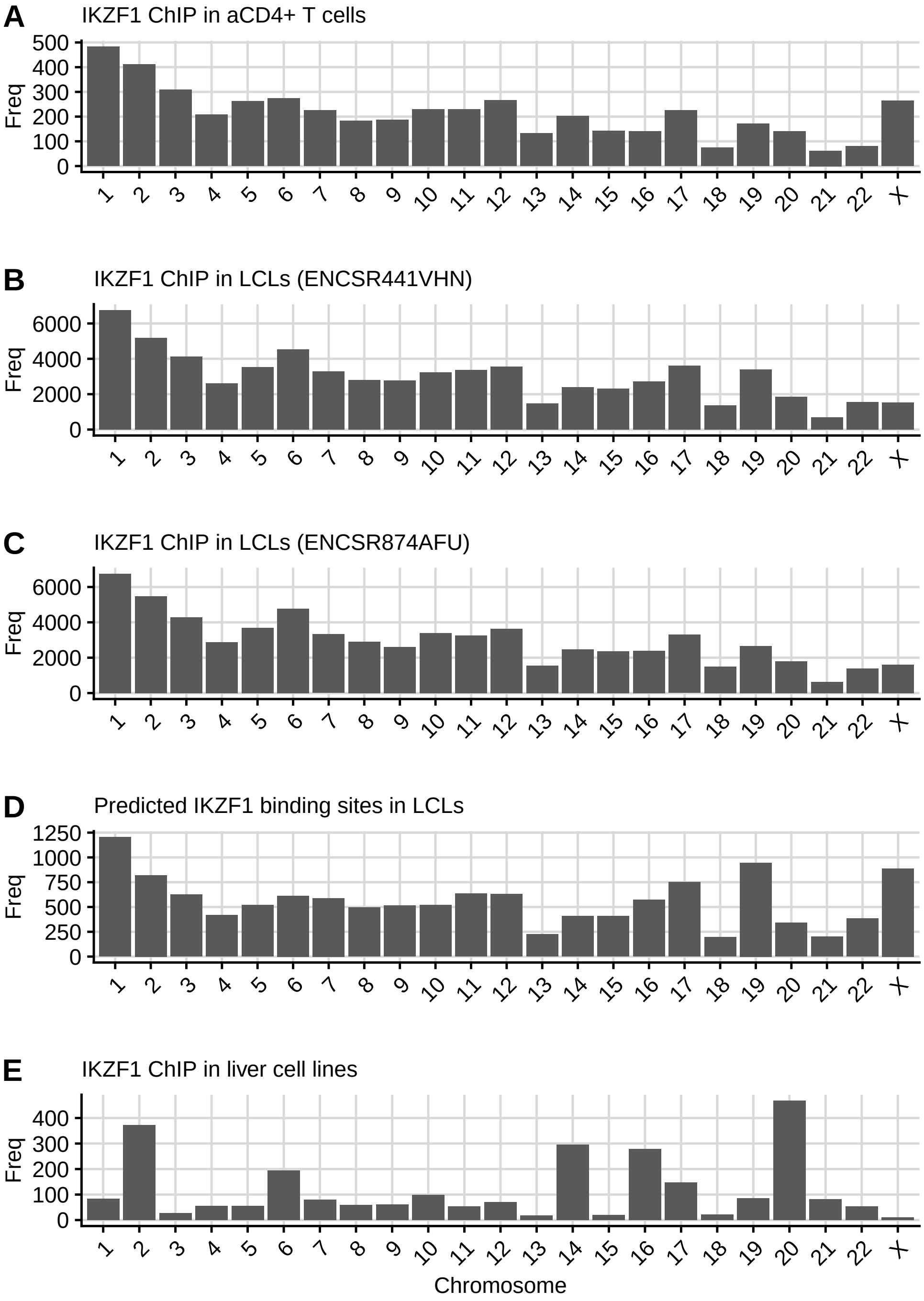


Fig. 4.6 Bar charts showing the number of peaks present on each chromosome for the five Ikaros-related functional genomic data sets. (A) Peaks called from the Ikaros ChIP-seq data conducted in activated CD4 T cells. (B) Peaks called from the Ikaros ChIP-seq data conducted in LCLs (ENCODE accession: ENCSR441VHN). (C) Peaks called from the Ikaros ChIP-seq data conducted in LCLs (ENCODE accession: ENCSR874AFU). (D) Predicted Ikaros binding sites in LCLs. (E) Peaks called from the Ikaros ChIP-seq data conducted in liver cell lines (ENCODE accession: ENCSR278JQG).

Extending the GoShifter methodology

I used the GoShifter methodology to test for enrichment of T1D-associated SNPs in target binding sites of Ikaros, extending the approach so that it could be applied to all SNPs included in a GWAS. To do this I defined a new test statistic which compared the mean ($-log_{10}$ transformed) p-value for SNPs residing within the annotation (p_1) and the mean ($-log_{10}$ transformed) p-value for SNPs not residing within the annotation (p_0). Let n_1 be the number of SNPs overlapping the annotation and n_0 be the number of SNPs not overlapping an annotation, then the overall test statistic is

$$X = \frac{\frac{\sum_{i=1}^{N} p_{1i} \times n_{1i}}{\sum_{i=1}^{N} n_{1i}}}{\frac{\sum_{i=1}^{N} p_{0i} \times n_{0i}}{\sum_{i=1}^{N} n_{0i}}}, \tag{4.1}$$

where N is the total number of SNP stratifications (which are described in the following subsection) and i denotes the specific SNP stratification. Thus, higher values of the overall test statistic imply that SNPs residing within the annotation have on average smaller p-values than those not residing in the annotation. As in the original method, the overall test statistic is computed for the observed SNP-annotation overlap to derive the observed test statistic.

To accurately determine the significance of the observed test statistics derived from my method, which were generated in the presence of confounding, it is imperative that the null distribution also captures this confounding. As in the original GoShifter method, I computed test statistics comprising the null distribution by circularising and randomly permuting the annotation data within each SNP stratification. Values for p_0 and p_1 were calculated for this annotation-permuted configuration, and used in equation (4.1) to derive the value of the test statistic (n_0 and n_1 were already calculated when deriving the observed test statistic). The magnitude of the permutation was generated by randomly sampling an integer between 1 and the number of SNPs in the SNP stratification inclusive. This was repeated 1000 times for each SNP stratification. To generate an estimate of the overall test statistic under the null distribution, I took a random sample of a permutation for each SNP stratification and used the calculated values for p_1 and p_0 to derive X. I did this $10,000$ times to generate $10,000$ test statistics computed under the null distribution. I have made my code publicly available via the following link: https://github.com/annahutch/genomewide-GoShifter.

Defining SNP stratifications

To derive SNP stratifications, I partitioned the genome into approximately independent LD blocks defined by the LD detect method (Berisa and Pickrell, 2016). These LD blocks are relatively large (mean size is 1.5-Mbp) and so I hypothesised that permuting within these blocks to generate a null distribution of test statistics would not capture all of the confounding that is present. This is because other known confounders, such as gene density and MAF, could vary substantially within these large LD blocks. I therefore investigated the impact of additionally stratifying SNPs by gene proximity and MAF.

For gene proximity, I downloaded gene locations from GENCODE (V15; GRCh38) (Harrow et al., 2012) in GTF format and converted these to sorted BED files using the `gtf2bed` function in `BEDOPS` (Neph et al., 2012a). The `closest-feature` utility, also in `BEDOPS`, was used to extract the closest gene and its distance in base pairs for each of the SNPs in my data set. I classified SNPs as either within a gene (gene distance = 0), close to a gene (gene distance was below the median of distances for SNPs not within genes) or far from a gene (gene distance was above the median of distances for SNPs not within genes). For MAF, I matched SNPs by genomic position to the Barrett et al. (2009) GWAS data set, which contains MAFs for SNPs analysed on the Illumina or Affymetrix technologies. SNPs were classified as low frequency (MAF was below the median MAF of the T1D GWAS SNPs) or high frequency (MAF was above the median MAF of the T1D GWAS SNPs). The SNPs for which the MAF was not available were randomly allocated to an MAF bin.

Applying my new GoShifter methodology

I used the T1D GWAS data from Cooper et al. (2017) for my analysis. I excluded variants that resided in LD detect blocks overlapping the MHC region of the genome (chr6:28017819-35455756), resulting in 8,557,894 SNPs in my analysis. Using each of the five genomic annotation data sets relating to Ikaros, I examined whether each SNP resided in a genomic region corresponding to an Ikaros binding site. That is, I converted the annotation data into a binary SNP overlap vector, q ($q = 1$ if SNP-annotation overlap, $q = 0$ if not). I then used my code (https://github.com/annahutch/genomewide-GoShifter) to derive observed test statistics and test statistics that were computed under the null for each of the five genomic annotation data sets.

4.4.2 Results

Ignoring confounding variables in a SNP enrichment analysis can lead to spurious results. I first investigated the impact that adjusting for various known confounding variables had on my results, using the Ikaros ChIP-seq data conducted in activated CD4 T cells as a test data set. Note that the conventional way to investigate potential confounders would be to regress these together with the annotation against ($-log_{10}$ transformed)) p-values in a multiple regression framework and examine their statistical significance. However, the p-values used as the response variable are correlated (due to LD) and so standard estimation approaches which assume independent observations would yield a smaller standard error, leading to smaller p-values measuring significance, and potentially more type I errors.

Firstly, I derived null test statistics when ignoring confounding by randomly permuting the annotation genome-wide rather than in SNP stratifications. In this instance, the test statistic reduced to the ratio of the mean ($-log_{10}$ transformed) p-value for SNPs overlapping, and SNPs not overlapping the annotation, and the null distribution of test statistics was roughly centred on 1 because the average p-values are expected to be equal for SNPs overlapping and not overlapping a randomly permuted annotation (Fig 4.7). Since the null distribution was generated ignoring confounding variables whilst the observed test statistic was computed in the presence of these confounding variables, the observed test statistic appeared highly significant.

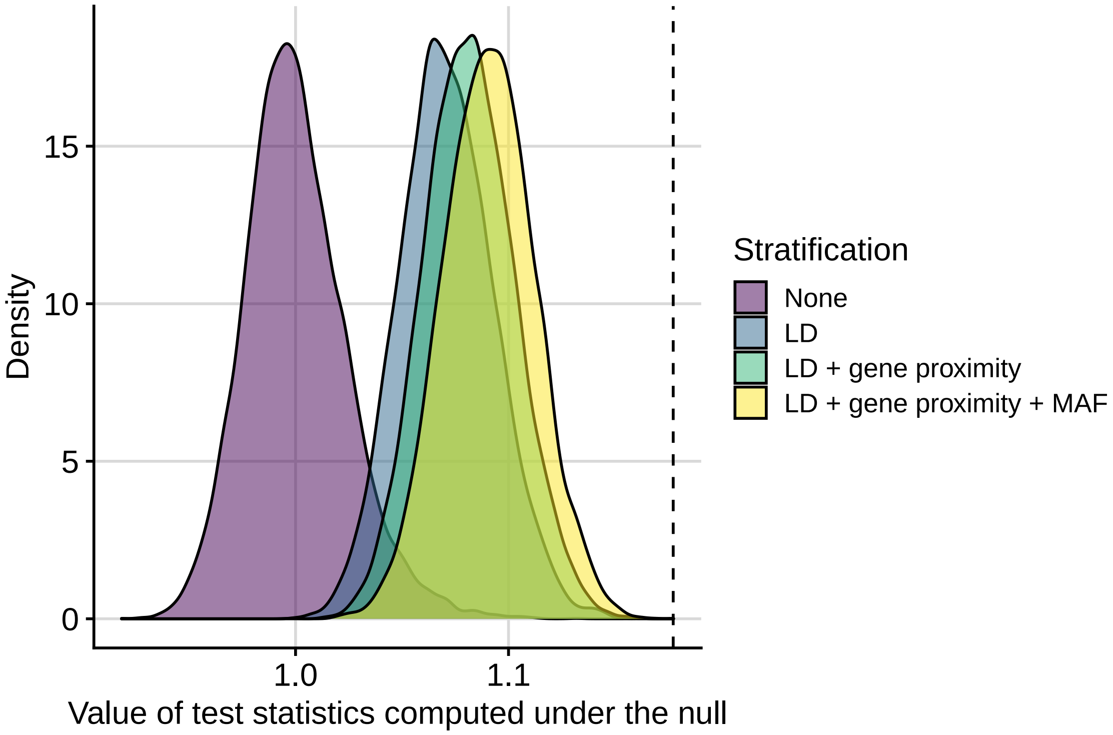


Fig. 4.7 Density plots of test statistics computed under the null hypothesis, for the Ikaros ChIP-seq data conducted in activated CD4 T cells, when using my extended GoShifter method. “None” refers to no stratification, where the annotation vector is randomly permuted 10,000 times genome-wide. “LD” refers to stratifying SNPs into LD blocks and randomly permuting within these, “LD + gene proximity” refers to stratifying SNPs into LD blocks and gene proximity bins and randomly permuting within these, “LD + gene proximity + MAF” refers to stratifying SNPs into LD blocks, gene proximity bins and MAF bins and randomly permuting within these. Vertical black dashed line shows the observed test statistic computed using SNP stratifications for LD, gene proximity and MAF.

I then derived the null test statistics when adjusting for confounding by randomly permuting the annotation within SNP stratifications relating to the confounding variables. The null distribution shifted to the right (closer to the observed test statistic) so that the significance of the observed test statistic reduced (Fig 4.7). This implied that confounding was present and that my null distribution was now capturing (at least some of) this. The null distribution had a mean value > 1 because correlated SNPs with small p-values would be contained within the same SNP stratification and the mean p-value for SNPs overlapping the randomly permuted annotation within these SNP stratifications would be smaller than expected, thus resulting in larger values of the test statistic. The shift was largest when adjusting for LD, implying that LD was the most influential confounder, which is supported by the relevant literature (Iotchkova et al., 2019; Trynka et al., 2015).

I next examined whether variants residing in Ikaros binding sites was enriched for smaller T1D GWAS p-values. When stratifying SNPs based on LD, gene proximity and MAF, I found that T1D-associated SNPs were strongly enriched ($p < 1e - 04$, since none of the $10,000$ simulated null test statistics were greater than or equal to the observed test statistic) in Ikaros ChIP-seq peaks in activated CD4 T cells and LCLs (Fig 4.8). The null and observed overall test statistics were comparable for the two Ikaros ChIP-seq data sets measured in LCLs which was encouraging, and the observed test statistic was slightly more extreme, relative to its null distribution, for Ikaros ChIP-seq peaks in activated CD4 T cells than LCLs (Fig 4.8). These findings suggest that activated CD4 T cells and B-cells (with the caveat that I am generalising results from LCLs to B-cells) may be key cell types in T1D pathogenesis relating to Ikaros.

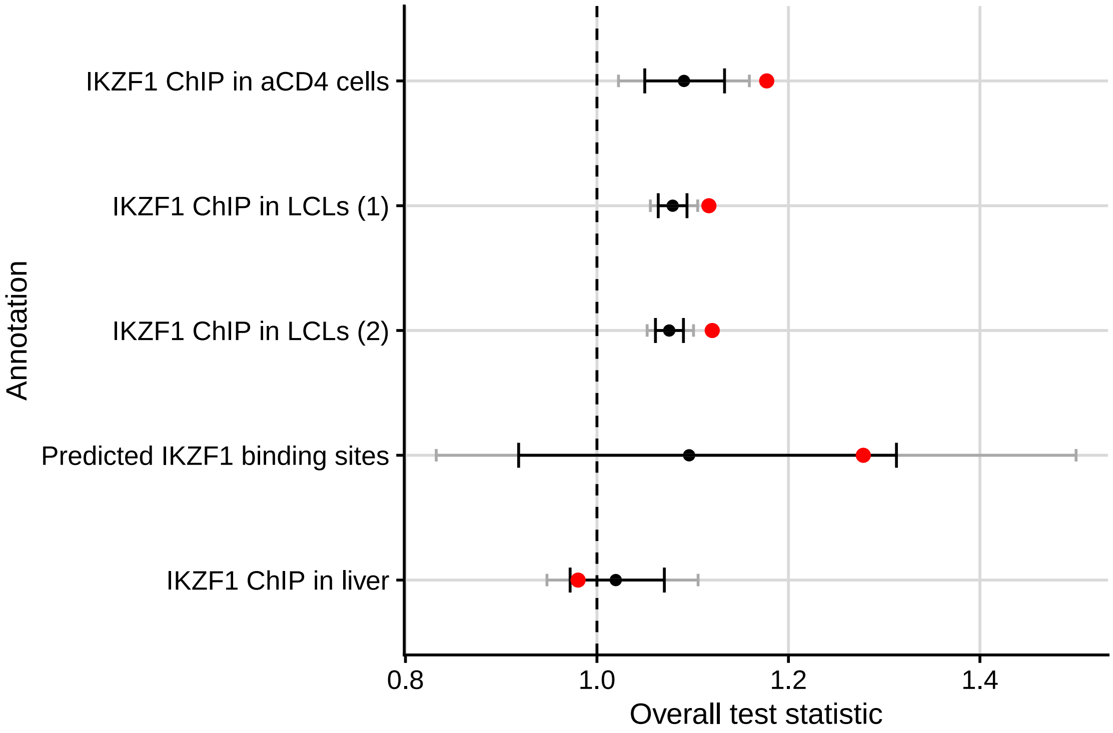


Fig. 4.8 SNP enrichment results using my extended GoShifter method to test for the enrichment of T1D-associated SNPs in various functional genomic annotations relating to Ikaros binding. Red points represent the observed overall test statistic and black points show the mean of the null distribution of test statistics with black error bars showing where 95% of the null distribution lies and grey error bars showing where 99.9% of the null distribution lies. Ikaros ChIP-seq in LCLs (1) and Ikaros ChIP-seq in LCLs (2) are for experiments with ENCODE accessions ENCSR441VHN and ENCSR874AFU, respectively. “aCD4” refers to activated CD4 T cells.

The null distribution was extremely variable for the predicted Ikaros binding site annotation (Fig 4.8). This is because these sites were extremely rare throughout the genome (208 annotation-SNP overlaps in the whole data set, which corresponds to $\approx 0.002\%$ SNP-annotation overlap), resulting in a very uncertain null distribution. My negative control shows that Ikaros ChIP-seq peaks in liver cells were not enriched for T1D-associated SNPs ($p = 0.0527$) (Fig 4.8), which was reassuring because liver cells are not thought to be relevant in T1D pathogenesis.

4.5 SNP-SNP interactions implicating T1D

Having examined variants associating with T1D in the Ikaros gene region and also variants associating with T1D in Ikaros binding sites, a natural follow-up question is whether the alleles of these sets of variants are acting synergistically or additively in T1D risk. In this section I use functional genomic data on Ikaros binding to narrow down the hypothesis search space in a case-only test for interaction. In a secondary analysis, I use Flexible cFDR to leverage marginal GWAS p-values with p-values corresponding to interaction effects, in an attempt to test for genome-wide interaction effects whilst maintaining high power.

4.5.1 Methods

Genotype data

The T1D GWAS data used in the previous section (Cooper et al., 2017) was from a meta-analysis of three cohorts: (i) T1DGC (Rich et al., 2009) (ii) GoKinD/NIMD (Cooper et al., 2008) and (iii) WTCCC (Wellcome Trust Case Control Consortium, 2007). I required raw genotype data for my analysis, but this was not publicly available. Instead, my supervisor Dr Chris Wallace, provided me with access to genotype data for individuals from the T1DGC and WTCCC cohorts, which she used as part of a different study (Wallace et al., 2010).

Some, but not all of the quality control (QC) steps listed in the original study had been implemented on the genotype data, and so I followed the QC steps outlined in Wallace et al. (2010) and also included some additional QC steps. Using sample IDs, I began by checking for overlapping samples between the two cohorts, since samples from the British 1958 Birth Cohort (C58) (Power and Elliott, 2006) are sometimes used as control samples in both WTCCC and T1DGC data sets. I found that there were no overlapping samples between the two cohorts - the C58 control samples were included in the T1DGC data and had been removed from the WTCCC data (and replaced with bipolar disorder control samples in order to maintain large sample sizes, as described in Barrett et al. (2009)).

To remove the need for special handling of male X heterozygous calls, I use the `-split-x` flag in `PLINK` to define a separate 'XY' chromosome for SNPs residing in the pseudo-autosomal region of chromosome X (corresponding to 151 SNPs in WTCCC and 10 SNPs in T1DGC). I removed three remaining problematic heterozygous haploid SNPs in the WTCCC study and four in the T1DGC study (those fagged by `PLINK` as heterozygous in males but not residing in the pseudo-autosomal region of chromosome X).

I then checked the sex of the samples based on their X chromosome inbreeding coefficients. These are equal to the X chromosome kinship coefficient between parents, which equal 0 for an outbred female and 1 for males regardless of inbreeding (because males inherit a single X chromosome from their mother). In WTCCC, three samples were listed as females in the input data set but "ambiguous" based on X chromosome inbreeding coefficients, and one sample was listed as female in the input data set but male based on X chromosome inbreeding coefficients (Fig 4.9A). In the T1DGC data, the sex was missing for 1427 samples and so I used the imputed sex based on X chromosome inbreeding coefficients using the `-impute-sex` flag in `PLINK`. In addition, five samples were listed as females in the input data set but "ambiguous" based on X chromosome inbreeding coefficients, and six samples were listed as female in the input data set but male based on X chromosome inbreeding coefficients (Fig 4.10A). I removed the four problematic samples in WTCCC and the 11 problematic samples in T1DGC.

I next searched for samples with high proportions of genotypes missing (as this may indicate genotyping errors), or those with extreme proportions of homozygous variants (since low homozygosity proportions may indicate low sample quality whilst high values may indicate inbreeding). To define relevant study-specific sample exclusion thresholds, I plotted homozygosity proportions against SNP missingness for each sample, and chose thresholds based on when one became informative for the other (e.g. where samples with more (or fewer) homozygous variants were more likely to have many genotypes missing). This corresponded to removing samples in WTCCC with either $> 0.02\%$ of SNPs missing or a proportion of homozygosity outside the range of $0.672 - 0.688$ (Fig 4.9B), and removing samples in T1DGC with either $> 0.003\%$ of SNPs missing or a proportion of homozygosity outside the range of $0.657 - 0.668$ (Fig 4.10B).

I then checked for any related individuals in each cohort by first pruning the variants so that no pair of SNPs within 50-kb were correlated ($r^2 > 0.2$) (using a step size of 5). I then used the `-genome` flag in `PLINK` to derive pairwise IBD proportions (Fig 4.9C, Fig 4.10C). I used the `-rel-cutoff` flag in `PLINK` to exclude one member of each pair of samples that had $IBD > 0.125$ (corresponding to samples that were closer than 3rd degree relatives). This method in `PLINK` attempts to maximise the final sample size, and removed seven samples in the WTCCC cohort and 11 samples in the T1DGC cohort.

As described in Wallace et al. (2010), SNPs with high missingness (> 0.05) (Fig 4.9D, Fig 4.10D) or low frequencies ($MAF < 0.01$) (Fig 4.9E, Fig 4.10E) had already been removed from the data set. Deviations from Hardy-Weinberg equilibrium (HWE) are likely to be the consequence of genotyping error, inbreeding or population stratification (Wigginton et al., 2005). Consequently (and as in Wallace et al. (2010)), I also removed SNPs with a HWE $p < 5.7e - 07$ (Fig 4.9F, Fig 4.10F) (where the p-value is computed in control samples only, because SNPs that strongly associate with the trait of interest are not expected to be in HWE in case samples).

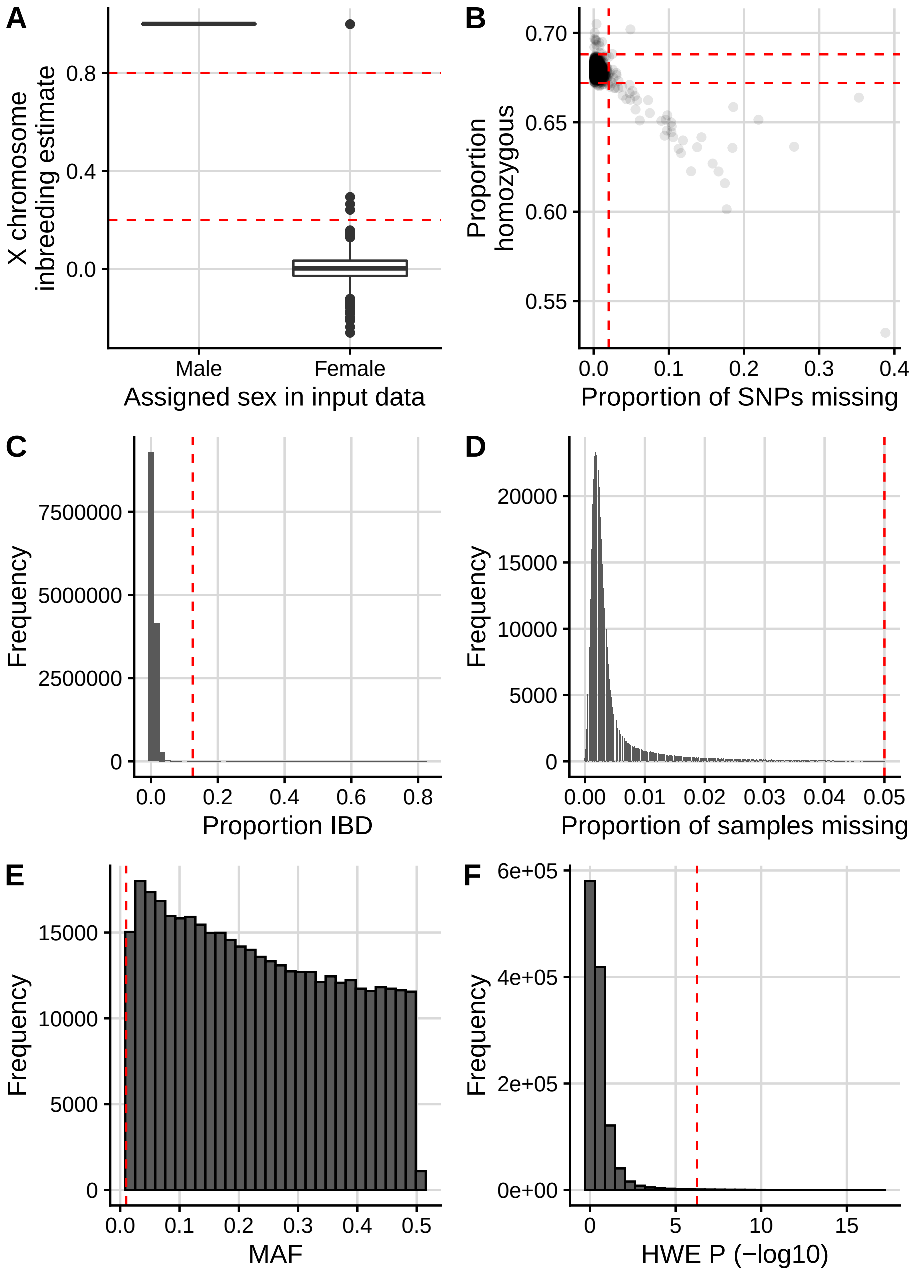


Fig. 4.9 Plots to facilitate pre-imputation QC for WTCCC data. (A) Box plots of the actual X chromosome homozygosity estimates (“F-statistics” in PLINK) for samples labelled as male and female in the input data. A male call is made if the X chromosome homozygosity estimate > 0.8 and a female call is made if this value < 0.2 (red dashed lines). (B) Per-sample homozygosity proportions against the proportions of SNPs missing in each sample, with red dashed lines showing the thresholds used to remove samples (samples removed if proportion homozygosity is outside the range $0.672 - 0.688$ or if the proportion missing > 0.02). (C) Histogram of proportion IBD values ($Pr(IBD = 2) + 0.5 \times Pr(IBD = 1)$) for each pair of samples used to determine related pairs of individuals. Red dashed line at $IBD = 0.125$. (D) Histogram of per-SNP missingness with a red dashed line at proportion missing $= 0.05$. (E) Histogram of MAF values with a red dashed line at $MAF = 0.01$. (F) Histogram of HWE p-values computed in control samples with a red dashed line at $p = 5.7e - 07$.

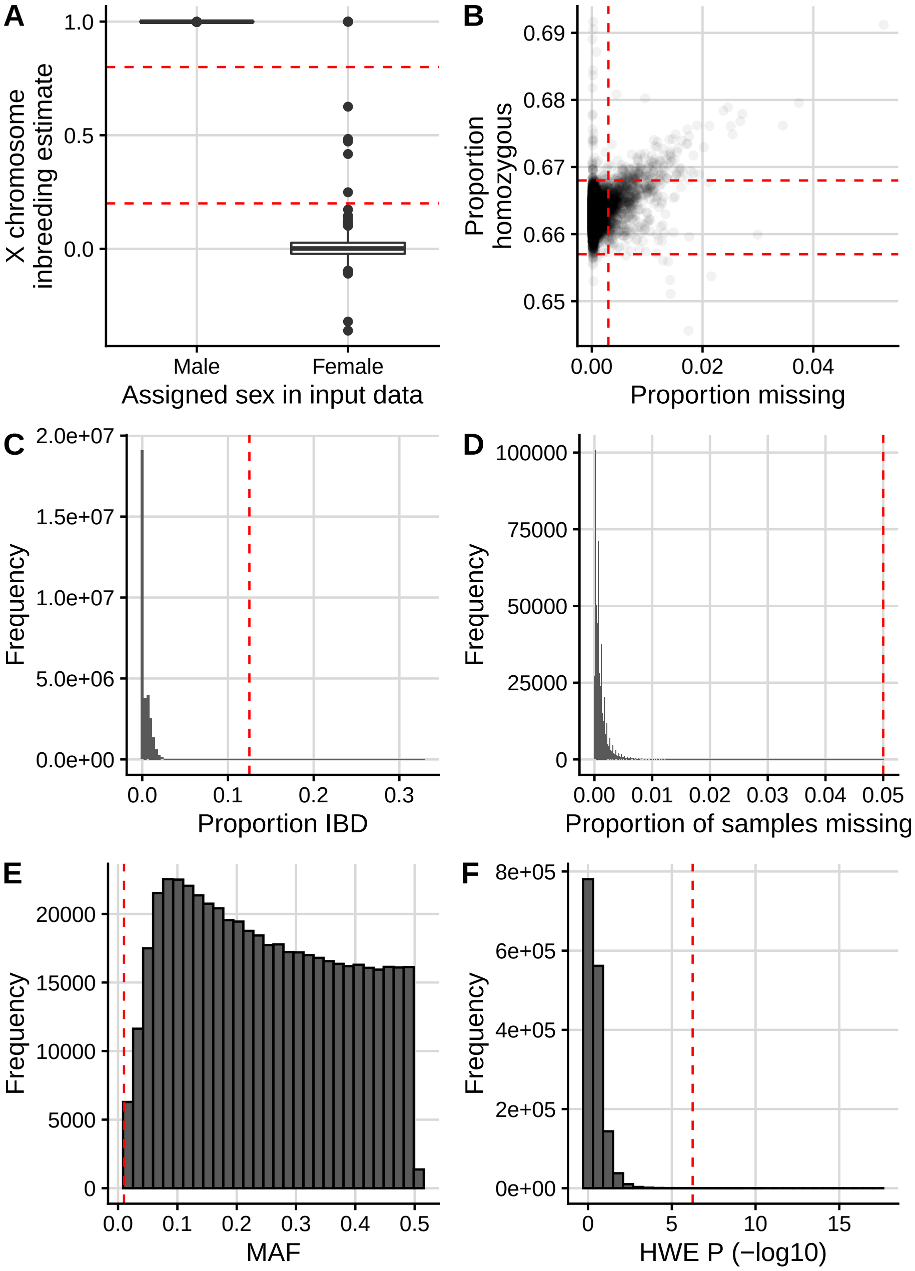


Fig. 4.10 Plots to facilitate pre-imputation QC for T1DGC data. (A) Box plots of the actual X chromosome homozygosity estimates (“F-statistics” in PLINK) for samples labelled as male and female in the input data. A male call is made if the X chromosome homozygosity estimate > 0.8 and a female call is made if this value < 0.2 (red dashed lines). (B) Per-sample homozygosity proportions against the proportions of SNPs missing in each sample, with red dashed lines showing the thresholds used to remove samples (samples removed if proportion homozygosity is outside the range $0.657-0.668$ or if the proportion missing > 0.003). (C) Histogram of proportion IBD values ($Pr(IBD = 2) + 0.5 \times Pr(IBD = 1)$) for each pair of samples used to determine related pairs of individuals. Red dashed line at $IBD = 0.125$. (D) Histogram of per-SNP missingness with a red dashed line at proportion missing $= 0.05$. (E) Histogram of MAF values with a red dashed line at $MAF = 0.01$ and (F) Histogram of HWE p-values computed in control samples with a red dashed line at $p = 5.7e - 07$.

Finally, I checked for any non-European samples using the 1000 Genomes phase 3 data set as a reference panel. Following the "Ancestry estimation based on reference samples of known ethnicities" vignette (https://cran.r-project.org/web/packages/plinkQC/vignettes/AncestryCheck.pdf) from the `plinkQC` R package, I firstly pruned my study data to exclude known regions of high LD and so that no pair of SNPs with 50-kb were correlated ($r^2 > 0.2$) (using a step size of 5). I then filtered the reference data to contain only the pruned SNPs in the study data set, and then matched SNPs between the reference data and the study data based on genomic position and alleles. I merged the two data sets and performed PCA using the `-pca` flag in `PLINK`. I then used the `plinkQC::evaluate_check_ancestry` function in R to estimate the ancestry of the study samples from the PCA results. Briefly, the function uses principal components 1 and 2 to find the centre of the known European reference samples, and any study samples whose Euclidean distance from the centre falls outside the radius specified by the maximum Euclidean distance of the reference samples multiplied by the chosen `europeanTh` value (default = 1.5) are labelled as non-European. This analysis identified 12 samples as non-European in WTCCC and 29 samples as non-European in T1DGC that I subsequently excluded from my analysis.

I used the Michigan Imputation Server (Das et al., 2016) to impute missing genotypes, using the Haplotype Reference Consortium (HRC) panel of 64,976 human haplotypes at 39,235,157 SNPs as the reference data (version r1.1 2016) (McCarthy et al., 2016). To prepare the data for imputation I used the `HRC-1000G-check-bim.pl` script written by Will Rayner (https://www.well.ox.ac.uk/~wrayner/tools/), which updates the strands, alleles and genomic coordinates of the variants to reflect HRC/hg19, and also removes variants with large differences in allele frequencies compared to HRC (> 20%) or ambiguous alleles (AT/GC variants with MAF> 40%). I then used the `VcfCooker` tool (https://genome.sph.umich.edu/wiki/VcfCooker) to convert the data to per-chromosome bgzip `VCF` files for input into the imputation server, specifying that the samples were of European ancestry and that any variants with low imputation quality ($R^2 < 0.3$) should be discarded.

I ran the imputation job on the server in the "QC and imputation mode" with phasing using Eagle version 2.4 (Loh et al., 2016). After completion, I downloaded the imputed data and used the `ic.pl` script written by Will Rayner (https://www.well.ox.ac.uk/~wrayner/tools/) to check the quality of the results. The post-imputation QC reports show that the imputed data is of high quality and are publicly available via the following links: https://htmlpreview.github.io/?https://github.com/annahutch/thesis-things/blob/main/WTCCC.html and https://htmlpreview.github.io/?https://github.com/annahutch/thesis-things/blob/main/T1DGC.html). I subsequently removed samples with extreme proportions of homozygous variants after imputation, defined as those with proportion of homozygosity outside the range of $0.910 - 0.914$ in WTCCC (Fig 4.11A) or with proportion of homozygosity < 0.917 in T1DGC (Fig 4.11B).

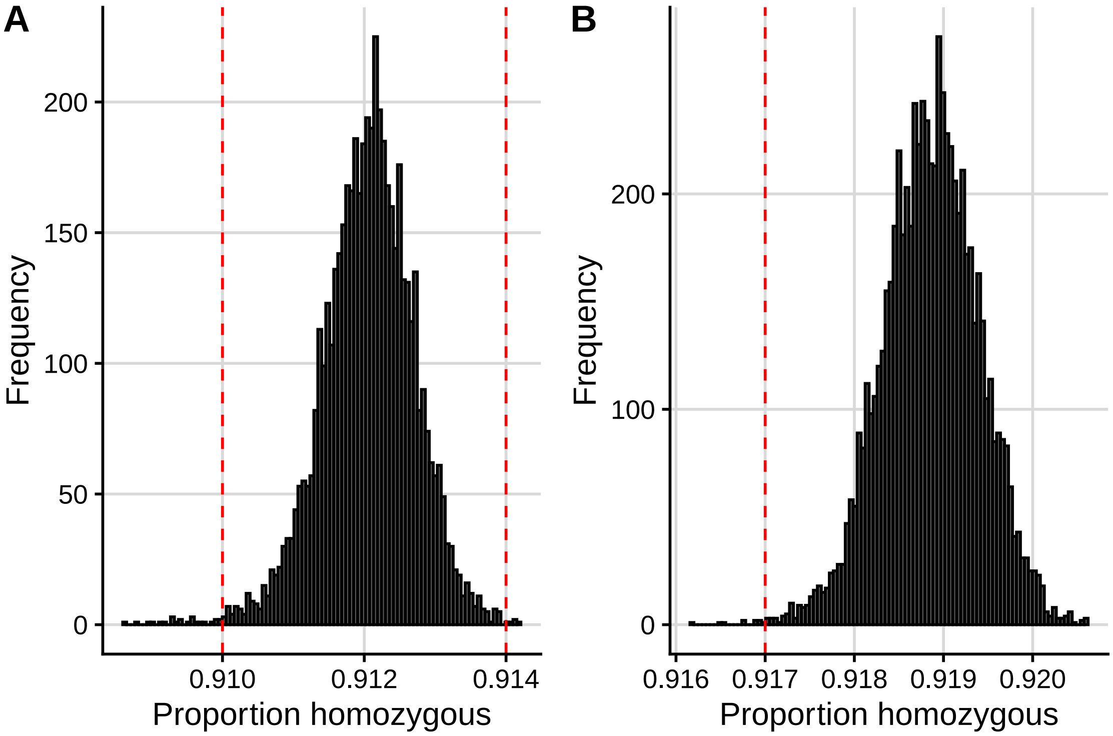


Fig. 4.11 Histogram of per-sample homozygosity proportions for (A) WTCCC and (B) T1DGC data after imputation. Red dashed line for exclusion thresholds. Samples with proportion of homozygosity outside the range of $0.910 - 0.914$ were excluded in WTCCC and those with proportion of homozygosity less than 0.917 were excluded in T1DGC.

Quantifying the predicted impact of SNPs on Ikaros binding

I used the SEMpl algorithm (Nishizaki et al., 2020) to predict the consequences that SNPs residing in Ikaros binding motifs would have on Ikaros binding in relevant cell types. The SEMpl method requires:

1. A PWM for the transcription factor of interest
2. DNase-seq data in the relevant cell type
3. ChIP-seq data for the transcription factor of interest in the relevant cell type.

For my analysis, I was limited to using LCLs as the relevant cell type as this was the only immune-relevant cell type with both DNase-seq data and Ikaros ChIP-seq data publicly available.

From ENCODE, I downloaded the DNase-seq data (accession number: ENCFF001UVC) and the ChIP-seq data (accession number: ENCSR441VHN), using the `bigWig` files for signal *p*-value for the latter (file accession: ENCFF176OVL). I downloaded the PFM for Ikaros from hocomoco (Kulakovskiy et al., 2018) (model: IKZF1_HUMAN.H11MO.0.C) and used the `universalmotif` R package (https://bioconductor.org/packages/release/bioc/html/universalmotif.html) to convert the motif from hocomoco to transfac format. I also converted the numerical values to integers, as required by the SEMpl software. I then used the `SEMpl` software (https://github.com/Boyle-Lab/SEMpl) to generate the SNP effect matrix, which quantifies the predicted effect that each possible nucleotide in the Ikaros binding motif has on Ikaros binding affinity.

Case-only test for interaction

I used a case-only test to search for SNP-SNP interactions. I treated the genotypes as continuous variables (for imputed SNPs, I used expected genotype values) and used the Pearson correlation coefficient as the statistic of interest. Since the genotypes were from two separate cohorts (WTCCC and T1DGC), I used a fixed effect meta-analysis to jointly analyse the correlation coefficients between genotypes in cases and genotypes in controls. I chose a fixed-effect model rather than a random-effect model because I only had two cohorts from which to estimate the parameter values, and this small quantity may lead to imprecise estimates for the additional parameters that need to be estimated in a random-effect model (Bell et al., 2019). My case-only test ultimately yielded p-values against the null hypothesis that there was no overall correlation (across cohorts) between the genotypes at the SNPs in case and control samples separately. I used the `metacor` R package (https://cran.r-project.org/web/packages/metacor/index.html) for this analysis.

Selecting pairs of potentially interacting SNPs

Section 4.3 explored various attempts at pinpointing the causal variant for T1D in the *IKZF1* gene region, but only Robertson et al. (2021) identified a single potential causal variant with high probability relative to the other SNPs in the region (rs6944602). The alternative allele (A) of rs6944602 is protective for T1D and is predicted to decrease *IKZF1* expression in whole blood and also various blood cell types relating to T1D (Carvalho-Silva et al., 2019). I therefore selected rs6944602 as the proxy causal variant in the *IKZF1* gene region for my SNP-SNP interaction analysis. Note that it suffices that rs6944602 serves as a proxy for the true causal variant in this analysis, because I am searching for interactions over large genomic distances that associate with T1D, rather than interactions that are causal between nearby SNPs (although the better that my selected SNP serves as a proxy for the true causal variant (i.e. the higher r^2), the greater the power to detect interactions). My results refer to SNP rs6944602 specifically, but come with the disclaimer that I am referring to any SNP that is in LD with rs6944602 and is causal for T1D.

Rather than testing all SNPs genome-wide for an interaction with rs6944602 and losing power when correcting for performing multiple statistical tests in parallel, I derived a short-list of SNPs with evidence of residing in an Ikaros binding site in a T1D-relevant cell type. I identified 96 SNPs that overlapped both an Ikaros ChIP-seq peak in LCLs (ENCODE accessions: ENCSR441VHN and ENCSR874AFU; median peak size of 636-bp) and, to refine results, a predicted Ikaros binding site in LCLs (all 7-bp) (Funk et al., 2020) (these data sets are described in section 4.4.1). I chose LCLs as the relevant cell type due to the availability of the data on predicted binding sites that was used to refine the results.

Since marginal association statistics from GWAS would also incorporate the effect of SNP-SNP interactions, I next used the marginal GWAS p-values for association with T1D to refine my SNP selection. Firstly, I selected the 9 SNPs out of the 96 that had evidence of a marginal association with T1D (GWAS $p \leq 0.05$ in Cooper et al. (2017)) and examined the local association statistics using Manhattan plots (Fig 4.12). Only two SNPs appeared to reside in genomic regions with strong association signals: rs1534430 ($p = 1e - 07$; Fig 4.12A) and rs58825580 ($p = 1e - 27$; Fig 4.12E). I therefore selected rs1534430 and rs58825580 as the candidate SNPs to test for an interaction with rs6944602.

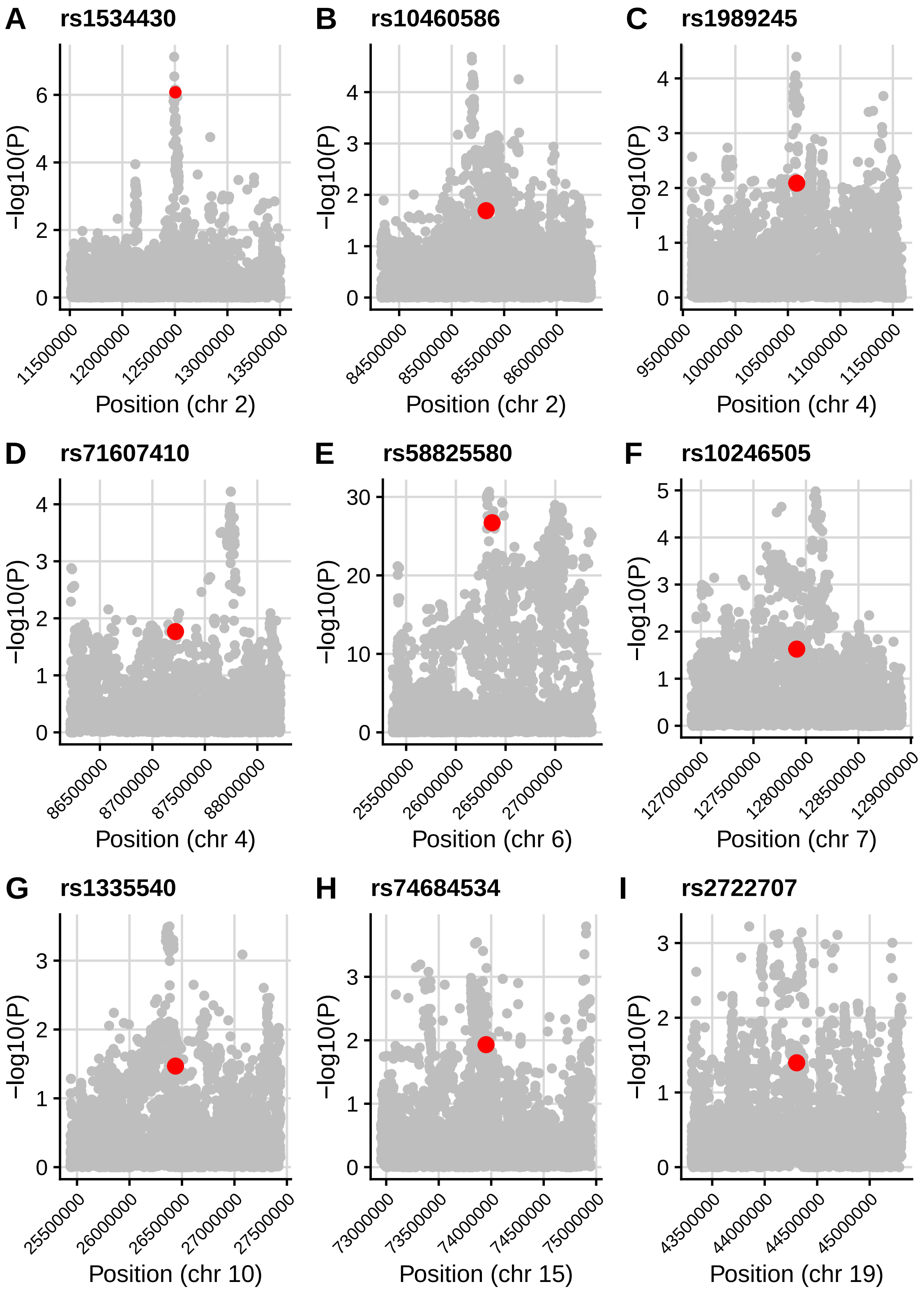


Fig. 4.12 Regional Manhattan plots for the 1-Mb genomic regions either-side of the 9 short-listed SNPs for the interaction analysis (which are coloured in red). GWAS summary statistics from Cooper et al. (2017) were curated by a member of our group, Dr Guillermo Reales, as part of a larger project.

When examining the imputation quality, MAF and HWE p-values for the SNPs included in my analysis, I found that only rs6944602 in T1DGC was directly genotyped but that the imputation quality was high ($R^2 > 0.95$) for rs1534430 and rs58825580 in both cohorts (and rs6944602 in WTCCC). The similar MAF values that were estimated for the two cohorts was reassuring, and the large HWE p-values implied that there was no evidence of genotyping errors, inbreeding or population stratification (Table 4.2).

SNP	Imputation R^2		MAF		HWE p-value	
	WTCCC	T1DGC	WTCCC	T1DGC	WTCCC	T1DGC
rs6944602	0.96750	1 (genotyped)	0.1882	0.1883	0.4654	1
rs1534430	0.95152	0.96881	0.383	0.3791	0.6316	0.2288
rs58825580	0.99807	0.99307	0.1527	0.1579	0.3465	0.3595

Table 4.2 Table containing auxiliary information for the three SNPs included in the interaction analysis. Imputation R^2 values were outputted by the Michigan imputation server, MAF values were calculated using the `-freq` flag in `PLINK` and HWE p-values were calculated using the `-hardy` flag in `PLINK` (HWE p-values were calculated in control samples only).

Upon further inspection of the genotype data, I found that each individual had been predicted a genotype with probability 1 at the imputed SNPs, meaning that the calculated expected genotypes did not discard any useful information on the uncertainty in the genotype calls.

Leveraging marginal p-values with interaction p-values using Flexible cFDR

Rather than using marginal GWAS p-values as a separate step for SNP inclusion in my analysis, I considered an alternative approach to directly incorporate these test statistics into my analysis. Specifically, I used Flexible cFDR to leverage marginal GWAS p-values for T1D with p-values corresponding to the null hypothesis of no interaction with rs6944602. Since the Flexible cFDR method utilises KDEs for estimation, it is recommended that a large number of data points are used in order to increase the accuracy of the estimated quantities. The predicted Ikaros binding sites from Funk et al. (2020) are small, very rare throughout the genome and did not show convincing enrichment of variants associated with T1D in section 4.4.2. I therefore chose to exclude this data in my analysis. Instead, I identified which SNPs in the Cooper et al. (2017) data set overlapped a ChIP-seq peak in at least one of the three ChIP-seq experiments conducted in T1D-relevant cell types, that are described in section 4.4.1. I found that 86,027 SNPs overlapped peaks, and in an attempt to avoid spurious results due to imputation errors, I subsequently limited my analysis to those SNPs which were imputed with high quality in both cohorts ($R^2 > 0.9$), which resulted in a total of 58,021 SNPs included in the analysis.

The case-only test for interaction assumes that genotypes are not correlated in control samples, but by definition SNPs that are in LD will have correlated genotypes in all samples. I therefore selected a distance window based on the estimated r^2 with rs6944602 in control samples, and excluded SNPs residing in this region from my analysis. To select a distance window, I first used the `r2` utility in `PLINK` with the `-ld-window-r2` flag set to 0.2, the `-ld-snp` flag set to rs6944602 and the `-filter-controls` flag turned on (that is, to search for SNPs with $r^2 > 0.2$ with rs6944602 using control samples only). I did this separately for the WTCCC and T1DGC data, and then determined the largest genomic region encompassing all the correlated SNPs (which was chr7:50423963-50661953), subsequently excluding the 11 SNPs that resided in this region from my analysis.

For the remaining 58,010 SNPs, I derived "interaction p-values" corresponding to the null hypothesis of no correlation between genotypes at rs6944602 and each of the SNPs in case samples using the fixed-effect model described earlier. I then used the `fcfdr::flexible_cfdr` function with default parameter values to leverage the marginal GWAS p-values from Cooper et al. (2017) with the interaction p-values. To define an independent subset of SNPs for the fitting of the KDE (corresponding to the `indep_index` parameter in `fcfdr::flexible_cfdr`), I used the LDAK procedure to obtain LDAK weights for each of the SNPs and used the subset of SNPs with non-zero weights as the independent subset. I used the BH procedure to derive FDR-adjusted v-values from Flexible cFDR and used an FDR threshold of $5e-06$ to call significant results.

4.5.2 Results

I first inspected the two SNPs that resided in Ikaros binding sites and that had strong marginal association statistics for T1D. Firstly, SNP rs1534430 (chr2:12504610) is an intron variant in *MIR3681HG* (a long non-coding RNA gene) and associates with a number of hypothyroid associated traits in UK Biobank (Sudlow et al., 2015) and FinnGen cohorts (FinnGen, 2021). SNP rs58825580 (chr6:26365451) is an intron variant in *BTN3A2* and the alternative allele which increases T1D risk (G) has strong eQTL evidence of decreasing expression of *BTN3A2* in a range of blood cell types (Carvalho-Silva et al., 2019). This SNP has also been linked to a range of IMDs, which is unsurprising as it resides in the juxta-telomeric region of the MHC class 1 locus (Stelzer et al., 2016). Both SNPs were found to reside in peaks called from all three of my T1D-relevant Ikaros ChIP-seq data sets, yet to the best of my knowledge, neither SNP has been investigated in relation to Ikaros before.

I used the SNP effect matrix for Ikaros binding in LCLs generated by SEMpl (Fig 4.13) to estimate the effect that the alternative alleles of rs1534430 and rs58825580 may have on Ikaros binding relative to the reference alleles. I found that the alternative allele of rs1534430 (C>T at motif position 8) was predicted to have no effect on Ikaros binding, whereas the alternative allele of rs58825580 (T>G at motif position 3) was predicted to increase Ikaros binding. A hypothesis that aligns with my findings so far is that the alternative allele of rs58825580 decreases the expression of *BTN3A2* by increasing Ikaros binding, which functions as a transcriptional repressor for *BTN3A2* at this locus, and that this increases T1D risk. However, it should be noted that a G nucleotide at motif position 3 of the Ikaros binding motif represents endogenous binding and is also the consensus nucleotide for the motif at this position. This provides evidence against SNP rs58825580 (T>G) residing in an Ikaros binding site, and these results should therefore be interpreted cautiously.

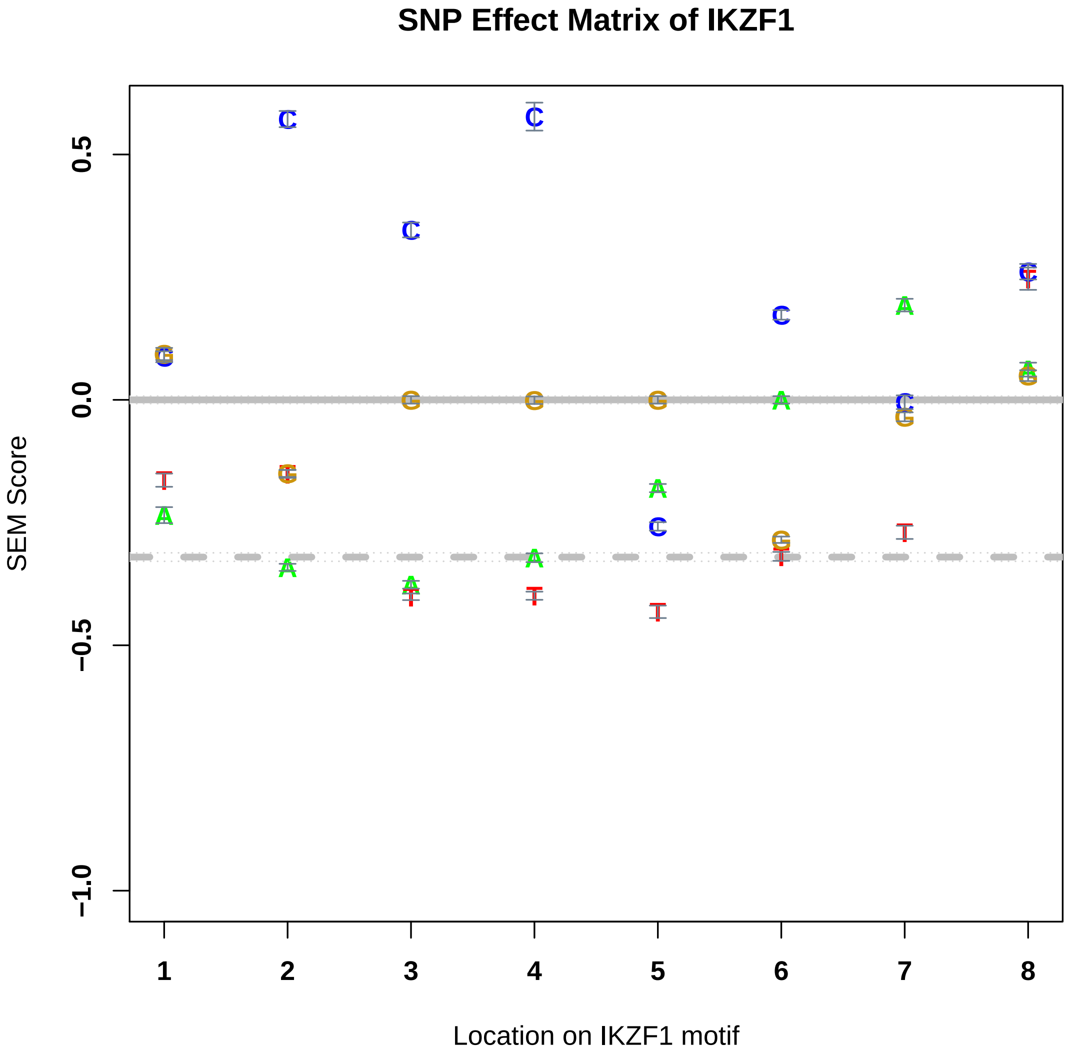


Fig. 4.13 SNP effect matrix for Ikaros binding in LCLs generated by the SEMpl method (Nishizaki et al., 2020). The solid line represents endogenous binding and the dashed grey line represents scrambled background. In the manuscript, the authors define anything above the solid grey line as predicted to increase binding on average, anything between the two lines as decreasing average binding and anything falling below the dashed grey line as ablating binding, on average.

I next sought to investigate whether the effect on T1D susceptibility for the combination of alleles at rs6944602 with alleles at either rs1534430 or rs58825580 added up to more than the sum of their individual effects (i.e. evidence of an interaction effect rather than an additive effect). An example of an additive effect would be that the reference allele of rs6944602 increases Ikaros expression and the alternative allele of rs58825580 increases Ikaros binding such that their combined effect on T1D risk is equal to the sum of their marginal effects. An example of an interaction effect would be that the reference allele of rs6944602 increases Ikaros expression and the alternative allele of rs58825580 increases Ikaros binding such that their combined effect on T1D risk is greater than the sum of their marginal effects, for example that the presence of both alternative alleles results in the gene *BTN3A2* being switched off, and that this increases the risk of T1D more than the gene just having lower expression (which is the additive consequence).

I found no evidence of an interaction effect between SNPs rs6944602 and rs1534430 or SNPs rs6944602 and rs58825580 (Table 4.3). This may be due to the lack of power to detect interaction effects and motivates the use of larger sample cohorts in SNP-SNP interaction analyses. With the caveat that these findings may be explained by a lack of statistical power, my results suggest that there is no evidence of a synergistic relationship between rs6944602 and rs1534430 or rs58825580 on T1D risk.

SNP	Correlation (WTCCC controls)	Correlation (T1DGC controls)	p-value (controls)	Correlation (WTCCC cases)	Correlation (T1DGC cases)	*p*-value (cases)
rs1534430	-0.0103	0.0110	0.9535	-0.0104	-0.0206	0.2125
rs58825580	0.0270	-0.0036	0.3606	-0.0288	0.0105	0.8092

Table 4.3 Results from SNP-SNP interaction analysis. Each SNP in the table was tested for an interaction with rs6944602 in a case-only test for interaction. Correlation is the Pearson correlation coefficient between genotypes for controls or cases in each cohort. The *p*-value columns are the *p*-values for tests of effect obtained from a fixed effect model for controls or cases. Results were obtained using the `metacor` R package.

When examining the relationship between marginal T1D GWAS *p*-values and interaction *p*-values with rs6944602 for a larger number of SNPs, I found that there was no enrichment of smaller interaction *p*-values for smaller marginal T1D GWAS *p*-values for the 58,010 SNPs included in my analysis (Fig 4.14). Since the marginal *p*-values would incorporate any interaction effect if it were present, this implies that there were no interaction effects between my tested SNPs. The lack of correlation between the interaction *p*-values ("p" in cFDR) and the marginal *p*-values ("q" in cFDR) was reflected in the results from Flexible cFDR, whereby no SNPs became newly FDR significant or newly not FDR significant. In fact, the smallest BH-adjusted interaction *p*-value was 0.8517 and this reduced to 0.6483 after applying Flexible cFDR to leverage the marginal *p*-values (for which the smallest BH-adjusted *p*-value was $6.78e - 31$).

With the caveat that power may be limiting the discovery of interaction effects in my analysis, these results imply that rs6944602 does not act synergistically with any SNPs residing in genome-wide Ikaros binding sites in T1D-relevant cell types to influence T1D risk.

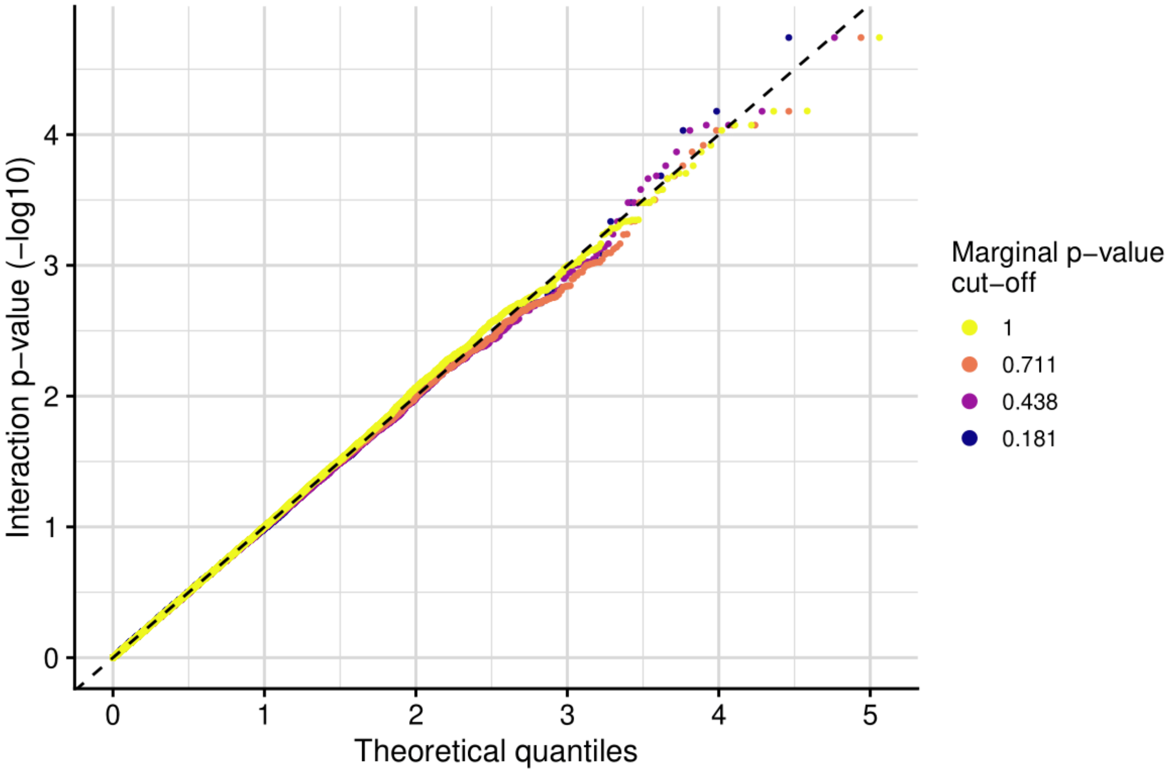


Fig. 4.14 Stratified Q-Q plot of empirical ($-log_{10}$ transformed) interaction p-values against theoretical values stratified by marginal T1D GWAS p-value. The values that were used to threshold the marginal p-values were the quantiles of the distribution (0.181 was the 0.25 quantile, 0.438 was the 0.5 quantile, 0.711 was the 0.75 quantile and 1 was the maximum value).

4.6 Discussion

Early GWAS studies found that genomic regions containing three of the five members of the Ikaros family of transcription factors associated with T1D (Barrett et al., 2009; Todd et al., 2007). However, fine-mapping was still in its infancy at the time and so the underlying causal variants were not elucidated. As fine-mapping has become more routine, there have now been several attempts made to pinpoint the specific causal variant(s) for T1D in the genomic region containing Ikaros. Alarmingly, different putative causal variants in the Ikaros gene region were identified by different fine-mapping studies when using the same, and different, GWAS data. Discrepancies between fine-mapping results across studies are unfortunately not infrequent. For example, the genomic region containing the *CASP8* gene has been fine-mapped using genotype data from the Collaborative Oncological Gene-Environment Study (COGS) Consortium at least six times (Alenazi et al., 2019; Fachal et al., 2020; Spencer et al., 2015, 2016) with little consensus reached on the likely causal variant(s) for breast cancer in the region. These examples serve as salutary reminders that fine-mapping results are method dependent and the current lack of a broad set of gold-standard true positives makes it difficult to evaluate the accuracy of different methods in the real world. Faced with such choices, other fields such as gene regulatory network inference (Hill et al., 2016) have adopted ensemble approaches, where multiple methods are applied to the same data and similar results across methods are interpreted as robust to methodological differences. Unfortunately, no consensus was reached on the likely causal variants for T1D in the Ikaros gene region by the fine-mapping attempts that I explored here.

Initiatives such as The ENCODE Project (ENCODE Project Consortium, 2012) have helped to drive the surge of publicly available data that is generated from functional genomic assays. I implemented my SNP enrichment method on the three high-quality Ikaros ChIP-seq data sets in immune-related cell types that were available at the time of the research, and found that T1D-associated variants were enriched in Ikaros binding sites in both T-cells and LCLs. It is generally recognised that Ikaros functions primarily in B-cells in leukaemia pathogenesis (Marke et al., 2018) and there is evidence that this is also the case in IMDs, such as SLE (Almlöf et al., 2017). However, T1D is a T-cell mediated disease (Roep, 2003) which may explain the significant enrichment found in T-cells. The conclusions from this SNP enrichment analysis should of course be interpreted with caution as they are based on ChIP-seq experiments using only a single or two biological replicates. A validation step is required if additional data or replicates become available.

The GoShifter methodology, and therefore my extension, uses annotation data that has been converted to a binary quantity measuring SNP-annotation overlap. Not only does this restrict the applicability of the approach to annotations that can be quantified in this way, but it discards potentially useful information relating to the annotations. For example, if ChIP-seq peaks are called using the `MACS2` software then auxiliary information on the signal value (measuring the overall enrichment for the region) and the p-value (measuring the statistical significance of the peak) are available, but are currently disregarded by the method (although they may be considered pre- or post-hoc, for example to call high-confidence peaks). Moreover, Soskic et al. (2019) found that the majority of ATAC-seq and H3K27ac ChIP-seq peaks are shared across many immune cell types and cell states, meaning that the binary SNP-annotation overlap would not distinguish between these cell types and cell states. As the granularity of functional genomic data expands, more rigorous approaches that take into account peak properties are required (such as the CHEERS approach described by Soskic et al. (2019) which models quantitative changes in read counts within peaks). In all, improvements in the granularity of the cell types and cell states within which the functional genomic assays are conducted, coupled with extensions to SNP enrichment approaches to permit quantitative data, could lead to refinements in the conclusions from SNP enrichment studies.

I found no evidence of interaction effects on T1D risk between variants in the Ikaros gene region and variants in Ikaros binding sites, but this may be due to a lack of statistical power. Extremely large sample sizes are required to detect interactions (Wang and Zhao, 2003) but since interaction analyses typically require raw genotype data, most studies are still limited by power. My SNP selection procedure may have also contributed to my negative findings. Due to the lack of publicly available data on Ikaros binding in primary immune-relevant cell types, I was required to use LCLs as the relevant cell type in my analysis. LCLs are immortal cell lines and experimental findings from studies conducted in this cell line may not directly generalise to findings in vivo. This motivates the development of functional assays performed using primary cells. I refined my results using data from Funk et al. (2020) which elucidates 7-bp regions of the genome with evidence for Ikaros binding, based on the presence of the Ikaros binding motif in genomic footprints. However, the Ikaros binding motif is redundant in the sense that it is a subset of other transcription factor binding motifs (e.g. that of the transcription factor *ZN143*) meaning that the identified regions may not even be bound by Ikaros at all. There is also redundancy in the sequence specificity of transcription factors, meaning that in general, results based on binding motifs should be interpreted cautiously.

Hemani and colleagues recently retracted their 2014 article entitled “Detection and replication of epistasis influencing transcription in humans” (Hemani et al., 2014) due to new evidence suggesting that the significant pairwise interactions that they found between eQTLs were actually due to the inflation of test statistics, owing to the presence of imperfectly tagged additive causal variants. Similar to the “haplotype effect” in which phantom epistasis can arise between two variants that are in imperfect LD with both each other and a causal variant (since a statistical interaction between the two variants can capture more of the additive variance of the causal variant than the marginal additive effects of both the variants combined) (de Los Campos et al., 2019), the authors showed that a similar phantom epistasis can occur when only one of the variants is in imperfect LD with a causal variant (Hemani et al., 2021). They found that their inflated test statistics (and therefore false positive results) were due to the use of a linear model which assumes an incorrect variance. These spurious results are therefore likely to occur in other analyses employing the same statistical test to detect interactions, which was an F-test comparing linear models with and without an interaction term (and which the authors claim is the “gold-standard statistical test to detect interactions”). It is unlikely that these spurious results would occur in analyses using alternative methods to test for interactions, such as the case-only test utilised here, however further work is warranted to support this claim.

Chapter 5

Discussion

In this thesis I have developed and examined a range of statistical methods to aid understanding of the genetic basis of complex human diseases, with application specifically to immune-mediated diseases. In this chapter I detail the key findings and contributions of the research presented in this thesis before discussing the linking themes, as well as the limitations and future directions for the disciplines of statistical genetics, and genetics, more generally.

5.1 Key findings and contributions

Genetic association studies have been a fundamental tool used to explore the genetic basis of complex human diseases for over a decade and a half, but the power to detect associations is limited by the effect size and the frequency of the genetic variation, as well as the study sample size. Researchers now have access to a wide variety of data generated from functional genomic experiments, which can be used in conjunction with genetic association data to help resolve disease aetiology. In chapter 2 I described novel methodologies based on the cFDR that leverage functional genomic data with genetic association data to increase statistical power for GWAS discovery whilst controlling the FDR. In a comprehensive simulation-based analysis, these methods were shown to exhibit greater flexibility, statistical power and improved error-rate control over several comparator methods. Application to data sets relating to asthma and T1D uncovered patterns matching those from a simulation analysis, and revealed new genetic associations that were biologically sound.

Genetic associations are now routinely fine-mapped to elucidate the underlying causal variant(s). Reporting credible sets of putative causal variants is conventional, but there is no mathematical basis for the widespread interpretations relating to the coverage probabilities of the true causal variants in these sets. In chapter 3 I found that the coverage probabilities were systematically biased and that this is due to the fact that fine-mapping data sets are not randomly sampled, but are instead sampled from a subset of those with the largest effect sizes. A practical implication of this finding is that smaller credible sets of variants can generally be derived that still achieve the desired coverage of the true causal variant, and I developed a method to derive these adjusted credible sets. Application to data sets for T1D and ankylosing spondylitis generated smaller sets of putative causal variants, thus saving time and resources in the expensive functional follow-up studies that are based on the prioritised sets of variants.

The methods developed in chapters 2 and 3 of this thesis relate to GWAS and fine-mapping, which are useful for identifying genetic determinants of complex diseases but do not necessarily resolve underlying biological mechanisms. In chapter 4 I utilised a range of statistical genetics tools to examine evidence for the role of the Ikaros family of transcription factors in T1D pathogenesis. This applied analysis highlighted many of the challenges relating to statistical approaches in genetics research more generally, including severe disparities between fine-mapping conclusions and complications arising when searching for epistatic interactions between genetic variants.

5.2 Linking themes, limitations and future directions

5.2.1 Data availability and usage

A key theme that runs through this thesis is the use of summary statistics from genetic association studies. Summary statistics have revolutionised genetics research over the past decade, facilitating data sharing through compact data storage and individual anonymity, and many genetic analyses now use summary statistics as standard input rather than raw genotype data. Publicly available curated collections of GWAS summary statistics, such as the NHGRI-EBI GWAS Catalog (Buniello et al., 2019), are a valuable tool for genetics researchers and the amalgamated data has many applications, for example facilitating the identification of causal variants or helping to streamline benchmarking analyses. Such initiatives have also been instrumental in establishing a community standard for summary statistics data representation, for example with regard to allele naming conventions that have historically obfuscated the direction of effect. The scientists behind the GWAS Catalog consistently engage with the wider scientific community to ensure that the resource remains relevant to current scientific aims (for example they have recently revealed on Twitter that they plan to make submission of effect sizes mandatory following user-feedback; https://twitter.com/GWASCatalog/status/1438449729621438474). In spite of the tremendous efforts of such initiatives, the biggest obstacle they face is the reluctance of some researchers to share their summary statistics, which is typically attributed to idleness or confidentiality concerns (Thelwall et al., 2020).

It will likely require the support of funders and peer-reviewed journals to help resolve this issue, for example to encourage researchers to share full summary statistics upon manuscript submission. Last year, the GWAS Catalog started accepting submissions for pre-published and unpublished studies, which will hopefully promote data sharing prior to publication. This will also allow for the inclusion of summary statistics from other sources, such as the UK Biobank (https://www.ukbiobank.ac.uk/) and Finngen (https://www.finngen.fi/en), which have used exceedingly large participant cohorts to reveal thousands of new genetic associations. As the GWAS Catalog expands to incorporate data from these highly-valued initiates, it will no doubt continue to be a key resource for geneticists in the coming years.

Curated collections of functional genomic data are also maturing, such as those from public research consortia such as the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) and The ENCODE Project (ENCODE Project Consortium, 2012). These initiatives have established data collection pipelines to produce and disseminate high-quality data from epigenomic assays, thereby facilitating research into the functional elements of the human genome. In contrast to genetic data, functional genomic data is very heterogeneous for example in relation to cell types, cell states and resolution. This means that whilst these public research consortia contribute enormous amounts of publicly available data relating to a wide variety of assays, targets and biosamples, it can still be challenging to find data that is relevant to a specific research question. This is what I observed in chapter 4, where I was surprised at the lack of data that was available relating to Ikaros binding in immune-relevant cell types, especially given the well-known connection between Ikaros and lymphocytic leukemia (Payne and Dovat, 2011). Increasing the availability of functional genomic data to a wider range of cell types and cell states would expand the range of scientific hypotheses that could be queried from the data, such as nominating causal cell types and cell states involved in disease pathogenesis. Due to generous funding, large-scale initiatives will likely be those responsible for generating such data, but how to do this efficiently is an open question. The data generation process should be focussed and conducted in a systematic manner, and I envisage that this could be achieved in two ways. First, the published literature could be surveyed and experiments oriented based on scientific knowledge (for example it is known that Ikaros is associated with lymphocytic leukemia which supports ChIP-seq experiments to examine Ikaros binding in T cells and B cells). Secondly, communication between the scientists behind these initiatives and the wider-scientific community would ensure that the data being collected is directly relevant to the needs of present-day researchers. Systematically increasing the amount and variety of publicly available functional genomic data will ultimately save researchers time and resources (for example because they may not be required to conduct the experiments themselves) and will also warrant fewer generalisations in the scientific literature (for example because more experimental data may be available for primary cells rather than cell lines).

One can expect that statistical methodologies that are able to make use of the increasing amounts and variety of functional genomic data will prevail over those that are unable to utilise such data. Mathematically, Flexible cFDR requires that the functional genomic data is independent at each iteration (in order to maintain type I error rate control). Large amounts of functional genomic data are likely to exhibit some correlation structure (for example ChIP-seq experiments for the same protein conducted in different cell types would produce correlated results), rendering the Flexible cFDR approach unsuitable for leveraging such data iteratively. One solution would be for researchers to choose a smaller subset of the data sets which they believe are capturing distinct disease-relevant features to include in their analysis, but this selection process is non-trivial and subjective, and this approach ultimately discards lots of potentially useful data. Another option would be to reduce the dimensionality of the functional genomic data, so that fewer iterations of Flexible cFDR are required. However, when investigating the use of dimensionality reduction techniques for the 96 annotations that are present in the baseline-LD model (version 2.2) (Gazal et al., 2017), I found that the lower-dimensional components no longer satisfied the positive stochastic monotonicity requirement of the cFDR approach, rending it unsuitable. I would therefore recommend that, as the amount of relevant functional genomic data increases, researchers should focus on the methods that have been developed for the specific purpose of utilising large amounts of data, such as FINDOR which leverages all genomics annotations present in the baseline-LD model with GWAS summary statistics to identify novel genetic associations (Kichaev et al., 2019).

5.2.2 Statistical methods in genetics research

Robust statistical methods are required to draw inferences from genetic and genomic data. Statistical geneticists face the daunting task of developing methods that exploit new biological knowledge, novel advances in statistical methodologies and the ever-evolving types of biological data. These methods are primarily developed to benefit the wider scientific community, and so it is imperative that researchers remain responsive to these new methodologies regardless of whether their development was motivated by advances in biology or statistics. In my experience, genetics researchers are often more enthusiastic about methods that are developed based on biological advances (for example to accommodate a new type of higher resolution biological data) than those based on statistical advances (for example to utilise a more efficient statistical modelling framework). This could be due to a lack of understanding for the latter and could be resolved through mandatory statistical training for geneticists. A more practical solution would be to advocate for close collaboration between the statisticians that develop the methods and the geneticists who are likely to use the methods. This could be achieved through interdisciplinary institutions whereby statisticians and geneticists interact ordinarily, interdisciplinary conferences or collaborative working groups which bring together researchers with assorted specialities. That being said, some researchers remain reluctant to step out of their comfort zone, and unwilling to dedicate time to learning about new approaches. For example, mixed-model-based approaches are emerging as an alternative methodology in GWAS and have several key advantages over traditional methods (for example they directly account for confounding and are more powerful) but are only being adopted by the GWAS community very slowly. This could be due to scepticism and stubbornness from researchers who have been using simple linear or logistic regression in GWAS for decades. The support of journal editors may be required, for example to encourage researchers to exploit novel approaches or to query why such approaches were not used, at the peer-review stage of manuscript submission.

Another important aspect relating to the uptake of statistical methodologies more generally is that they should be accessible to a wide variety of researchers. One way that this can be achieved is through well documented and open-source accompanying software, which is exemplified in chapters 2 and 3 of this thesis, whereby the methods were accompanied by open-source software and user-oriented web pages (https://annahutch.github.io/fcfdr/; https://annahutch.github.io/corrcoverage/). The web pages are suitable for a wide-variety of researchers (including those who may not be familiar with programming in R), containing easy to follow installation instructions, fully reproducible vignettes and even an interactive flowchart to enhance usability. Such approaches that encourage widespread accessibility of statistical methodologies are likely to be key driving forces for their uptake.

It is especially important for the scientists that are developing large-scale resources to remain acquainted with new methodologies, and to incorporate these into their pipelines where appropriate. This will likely improve the quality of the resource, which will in turn attract more users. In their fine-mapping pipeline, Open Targets Genetics (Carvalho-Silva et al., 2019) uses conditional regression to identify independent signals and then Bayesian fine-mapping to construct credible sets of putative causal variants, even though these approaches have been shown to be non-optimal (Asimit et al., 2019; Hutchinson et al., 2020b). The recently developed SuSiE methodology (Wang et al., 2020) provides a new statistical perspective for fine-mapping, framing it as a variable selection problem and utilising a novel "sum of single effects" model to quantify uncertainty in variables by decomposing vectors of regression coefficients (with zero elements for all non-causal variables) into sums of "single-effect" vectors (which each have only one non-zero element). It is encouraging that this statistical approach has already been widely accepted by the scientific community, for example the `coloc` software has been adapted to utilise the SuSiE approach (Wallace, 2021), and alternative prior probabilities of causality have also been investigated (Weissbrod et al., 2020). Whilst some large-scale initiatives, including the eQTL Catalog (Kerimov et al., 2021), have integrated fine-mapping using SuSiE into their pipelines, it would be beneficial for other such initiatives to follow suit.

The authors of the SuSiE approach state in their manuscript that the resultant credible sets are generally anti-conservative in terms of coverage probabilities (their 95% credible sets "typically had coverage slightly below 0.95, and in most cases above 0.90"). It should be noted that the authors of the original Bayesian fine-mapping approach (Maller et al., 2012) did not interpret the resultant credible sets in terms of coverage probabilities, but rather in terms of accounting for a proportion of the total posterior probability. Since it is now conventional in the literature to interpret credible sets in terms of coverage probabilities, I was surprised at the lack of research that had been conducted into examining the empirical coverage of these credible sets, especially given their widespread use over the past decade. Whilst van de Bunt et al. (2015) showed (in a supplemental figure) that the coverage probabilities for credible sets varied according to power, to the best of my knowledge, the research detailed in chapter 3 of this thesis was the first systematic examination into the accuracy of the inferences from the approach. My findings, that credible sets from conventional Bayesian fine-mapping are generally over-conservative (in terms of coverage probabilities) and that fewer variants are required in these sets, has widespread implications relating to published sets of variants and also the follow-up analyses that are based on these sets of variants. Indeed, if this research was conducted earlier, then considerable time and resources could have been saved in the follow-up analyses based on the smaller sets of variants. Therefore, it is important that statistical methodologies are scrutinised by the wider scientific community, preferably before they gain widespread use.

Due to technological and computational advances, there is now a plethora of statistical methodologies available that aim to achieve the same goal. Researchers are therefore faced with a daunting, and often overwhelming, task of choosing one of these methods for their analysis. This is most apparent in fine-mapping, where there have been many approaches developed in the past few years that accommodate both new biological knowledge (such as widespread allelic heterogeneity) and new data types (such as functional genomic data). The advantage of having a collection of relevant methods is that researchers are able to select the one that best suits their data requirements, computational restrictions and scientific assumptions. But this can lead to problems whereby researchers publish results from the method that gives them the "best" results. Rather alarmingly, I found that there were huge discrepancies in fine-mapping results in a simple comparative analysis for the Ikaros gene region. Not only did different fine-mapping methods give vastly different results when using different genetic association data, but these discrepancies remained when using different fine-mapping methods with the same genetic association data. This suggests that the data from genetic association studies plays at least some part in these apparent discrepancies.

A formal benchmarking analysis is required to evaluate the accuracy of fine-mapping methods relative to each other, and to also investigate how the input data impacts the results. Ideally, this analysis should be conducted by independent researchers, but this research may appear uninteresting or dull, and may not provide attractive publication or funding opportunities. It will therefore require journal editors and funders to realise the gains in this type of research in order to achieve this goal. Indeed, discrepancies in fine-mapping results ultimately leads to mistrust of the methods and results in longer lists of potentially causal variants, thus costing researchers time and resources. An additional challenge for such fine-mapping benchmarking analyses is the lack of a "gold-standard" data set of fully validated causal variants. A solution may be to use simulated data, with the caveat that findings from simulated data may not reflect true findings from real fine-mapping analyses. Establishing a consensus approach for fine-mapping will help to prevent discrepancies in results across studies and could also initiate a public database of credible variants generated from the consensus approach (CAUSALdb (http://mulinlab.org/causaldb/index.html) lists credible variants, but these are currently derived from three arbitrary fine-mapping methods: PAINTOR, CAVIARBF and FINEMAP). As I stated in Hutchinson et al. (2020a), the accurate identification of causal variants will in turn enable the creation of more accurate polygenic risk scores, which can help to disentangle the causal pathways to disease, for example, through Mendelian randomisation analysis. They will also assist considerably with the primary motivation of fine-mapping studies: efficient design of functional studies, leading to understanding the biological mechanisms which underlie disease risk and thus, ultimately, intervene upon these mechanisms.

5.2.3 Considerations relating to the future of genetics research

We are now situated in the "post-GWAS era", and focus has shifted from identifying genetic associations to investigating the biological mechanisms underpinning these associations. In chapter 4 I used genetic and functional genomic data to examine evidence for the role of a specific biological factor (Ikaros) in the pathogenesis of a specific disease (T1D). In hindsight this approach was inefficient, and as the amount of genetic and functional genomic data increases, it becomes increasingly impractical to expend time and resources studying such specific biological mechanisms in this way. For example, for immune-mediated diseases that share an immune component, a better approach may be to investigate immune mechanisms more generally and then to interpret these in the context of specific diseases. That is, merit could be gained by focussing on the broader picture and developing more holistic approaches that characterise some feature of the immune response, that can then be interpreted in relation to many immune-mediated diseases. Identifying genetic architectures that are shared across immune-mediated diseases is one such approach. Burren et al. (2020) exploited the pervasive sharing of genetic architectures across multiple immune-mediated diseases to extract disease-relevant components underlying genetic risk for immune-mediated disease. The authors were able to interpret several of the components in terms of disease biology, for example one component distinguished between autoantibody seronegative to seropositive diseases whilst another described eosinophilic involvement in disease. A follow-up analysis could be to identify the driver variants and genes for each of these components, to ultimately link genetic variation to the biological processes that are involved in disease pathogenesis. This analysis would likely be extremely involved, and would require an intricate balance of a wide range of data types (for example eQTLs and pQTLs) and tools (for example gene network inference and pathway enrichment analyses). The challenge of interpretation in terms of biological mechanisms is likely to be the most difficult yet most rewarding.

In contrast to monogenic diseases and some cancers, we are still far from translating complex disease genetics to routine clinical practise. The ideal outcome would be to "solve" a complex trait (for example in terms of aetiology) and to use these observations to aid diagnosis and treatment options, as has been done in the past for monogenic diseases such as cystic fibrosis. Yet with increasing genetic associations arising with increasing GWAS sample sizes and functional studies highlighting the exquisite complexity of disease biology, it is unclear whether we will ever be able to "solve" the aetiology of any (or all) complex disease(s). Instead, the genetics community should focus on more realistic goals in terms of translating research findings to clinical practise. For example, it is well-known that genetic risk prediction tools do not perform well at the individual level, but integrating genetic data with external risk factors to stratify patients for targeted lifestyle changes, and potential interventions, may be more attainable. In particular, Genomic plc's "integrated risk tool" (Riveros-Mckay et al., 2021) integrates polygenic risk scores with established risk factors for coronary artery disease, and has been shown to exhibit superior performance compared with other published polygenic risk scores. In my opinion, this tool is one of the most promising advances towards translating genetics research into routine clinical practise in recent years, but several challenges remain. Firstly, as usual, the polygenic risk scores that are used in the tool are primarily trained on European samples, and therefore show weaker performance in other ancestry groups. The predictive power of such polygenic risk scores will increase as more genetic data becomes available, especially in individuals of non-European ancestries. Secondly, I suspect that clinicians would be sceptical to use such a tool due to a lack of understanding regarding the derivation of the polygenic risk scores, which may hinder the utility of the tool in practise. A close collaboration between researchers and clinicians is required if we expect to see such breakthroughs for genetics in routine clinical practise.

5.2.4 Closing remarks

In all, our understanding of complex disease genetics is still very limited. Instrumental to expanding our understanding is increasing the amount and variety of publicly available genetic and functional genomic data. Novel statistical methodologies will be required, and systematic benchmarking of these will be essential in order to fully understand the opportunities and obstacles faced by both existing and novel methodologies. Statistical geneticists must continue to adapt to the evolving types, resolutions and quantities of biological data whilst also incorporating relevant advances in statistical methods, technologies and computation, and this will require systematic collaboration. Thus, dissecting disease aetiology has now become as much a statistical question as it is a biological one, and it is only through robust statistical methods that we will begin to gain a comprehensive understanding of complex disease pathogenesis to ultimately translate research into equitable clinical care.

References

D. Adapa. A Brief Review on Immune Mediated Diseases. *Journal of Clinical & Cellular Immunology*, 02, Jan. 2011. doi: 10.4172/2155-9899.S11-001.

C. Albanesi, H. R. Fairchild, S. Madonna, et al. IL-4 and IL-13 Negatively Regulate TNF-α- and IFN-γ-Induced β-Defensin Expression through STAT-6, Suppressor of Cytokine Signaling (SOCS)-1, and SOCS-3. *The Journal of Immunology*, 179(2):984–992, July 2007. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.179.2.984.

B. Alberts, A. Johnson, J. Lewis, et al. Innate Immunity. *Molecular Biology of the Cell. 4th edition*, 2002.

A. A. Alenazi, A. Cox, M. Juarez, et al. Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies. *Genetic Epidemiology*, 43(6):690–703, Sept. 2019. ISSN 1098-2272. doi: 10.1002/gepi.22213.

K. Alishahi, A. R. Ehyaei, and A. Shojaie. A Generalized Benjamini-Hochberg Procedure for Multivariate Hypothesis Testing. *arXiv:1606.02386 [stat]*, June 2016.

J. C. Almlöf, A. Alexsson, J. Imgenberg-Kreuz, et al. Novel risk genes for systemic lupus erythematosus predicted by random forest classification. *Scientific Reports*, 7(1):6236, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-06516-1.

C. A. Anderson, G. Boucher, C. W. Lees, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics*, 43(3):246–252, Mar. 2011. ISSN 1546-1718. doi: 10.1038/ng.764.

O. A. Andreassen, W. K. Thompson, A. J. Schork, et al. Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLOS Genetics*, 9(4):e1003455, Apr. 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003455.

O. A. Andreassen, McEvoy Linda K., Thompson Wesley K., et al. Identifying Common Genetic Variants in Blood Pressure Due to Polygenic Pleiotropy With Associated Phenotypes. *Hypertension*, 63(4):819–826, Apr. 2014a. doi: 10.1161/HYPERTENSIONAHA.113.02077.

O. A. Andreassen, W. K. Thompson, and A. M. Dale. Boosting the Power of Schizophrenia Genetics by Leveraging New Statistical Tools. *Schizophrenia Bulletin*, 40(1):13–17, Jan. 2014b. ISSN 0586-7614. doi: 10.1093/schbul/sbt168.

O. A. Andreassen, V. Zuber, W. K. Thompson, et al. Shared common variants in prostate cancer and blood lipids. *International Journal of Epidemiology*, 43(4):1205–1214, Aug. 2014c. ISSN 0300-5771. doi: 10.1093/ije/dyu090.

O. A. Andreassen, H. F. Harbo, Y. Wang, et al. Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: Differential involvement of immune-related gene loci. *Molecular Psychiatry*, 20(2): 207–214, Feb. 2015. ISSN 1476-5578. doi: 10.1038/mp.2013.195.

J. L. Asimit, D. B. Rainbow, M. D. Fortune, et al. Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nature Communications*, 10 (1):3216, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11271-0.

J. C. Barrett, S. Hansoul, D. L. Nicolae, et al. Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nature genetics*, 40(8):955–962, Aug. 2008. ISSN 1061-4036. doi: 10.1038/NG.175.

J. C. Barrett, D. G. Clayton, P. Concannon, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 41(6):703–707, June 2009. ISSN 1546-1718. doi: 10.1038/ng.381.

P. Basu, T. T. Cai, K. Das, and W. Sun. Weighted False Discovery Rate Control in Large-Scale Multiple Testing. *Journal of the American Statistical Association*, 113(523):1172–1183, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1336443.

W. Bateson. *Mendel's Principles of Heredity*. Cambridge University Press, 1909.

A. Bell, M. Fairbrother, and K. Jones. Fixed and random effects models: Making an informed choice. *Quality & Quantity*, 53(2):1051–1074, Mar. 2019. ISSN 1573-7845. doi: 10.1007/s11135-018-0802-x.

Y. Benjamini. Comment: Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1): 23–28, Feb. 2008. ISSN 0883-4237, 2168-8745. doi: 10.1214/07-STS236B.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246.

Y. Benjamini and Y. Hochberg. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000. ISSN 1076-9986. doi: 10.2307/1165312.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, Aug. 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013699998.

Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive Linear Step-up Procedures That Control the False Discovery Rate. *Biometrika*, 93(3):491–507, 2006. ISSN 0006-3444.

C. Benner, C. C. A. Spencer, A. S. Havulinna, et al. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics (Oxford, England)*, 32(10):1493–1501, May 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw018.

T. Berisa and J. K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285, Jan. 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv546.

B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10):1045–1048, Oct. 2010. ISSN 1087-0156. doi: 10.1038/nbt1010-1045.

S. M. Boca and J. T. Leek. A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035, Dec. 2018. ISSN 2167-8359. doi: 10.7717/peerj.6035.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

K. Borodulin, H. Tolonen, P. Jousilahti, et al. Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology*, 47(3):696–696i, June 2018. ISSN 1464-3685. doi: 10.1093/ije/dyx239.

S. Bottardi, L. Mavoungou, H. Pak, et al. The IKAROS Interaction with a Complex Including Chromatin Remodeling and Transcription Elongation Activities Is Required for Hematopoiesis. *PLOS Genetics*, 10(12): e1004827, Dec. 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004827.

L. Bottolo and S. Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, Sept. 2010. ISSN 1936-0975, 1931-6690. doi: 10.1214/10-BA523.

L. Bottolo, M. Chadeau-Hyam, D. I. Hastie, et al. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS genetics*, 9(8):e1003657, 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003657.

C. Bourges, A. F. Groff, O. S. Burren, et al. Resolving mechanisms of immune-mediated disease in primary CD4 T cells. *EMBO Molecular Medicine*, 12(5):e12112, May 2020. ISSN 1757-4676. doi: 10.15252/emmm.202012112.

R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551, May 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0914005107.

A. P. Boyle, E. L. Hong, M. Hariharan, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9):1790–1797, Jan. 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.137323.112.

M. S. Breen, C. Kemena, P. K. Vlasov, et al. Epistasis as the primary factor in molecular evolution. *Nature*, 490 (7421):535–538, Oct. 2012. ISSN 1476-4687. doi: 10.1038/nature11510.

D. A. Brewerton, F. D. Hart, A. Nicholls, et al. Ankylosing spondylitis and HL-A 27. *Lancet (London, England)*, 1(7809):904–907, Apr. 1973. ISSN 0140-6736. doi: 10.1016/s0140-6736(73)91360-3.

H. A. Bruns, U. Schindler, and M. H. Kaplan. Expression of a constitutively active Stat6 in vivo alters lymphocyte homeostasis with distinct effects in T and B cells. *Journal of Immunology*, 170(7):3478–3487, Apr. 2003. ISSN 0022-1767. doi: 10.4049/jimmunol.170.7.3478.

A. Buniello, J. A. L. MacArthur, M. Cerezo, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Research*, 47(D1):D1005–D1012, Aug. 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1120.

O. S. Burren, H. Guo, and C. Wallace. VSEAMS: A pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics*, 30(23):3342–3348, Dec. 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu571.

O. S. Burren, G. Reales, L. Wong, et al. Genetic feature engineering enables characterisation of shared risk factors in immune-mediated diseases. *Genome Medicine*, 12(1):106, Nov. 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00797-4.

W. S. Bush and J. H. Moore. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, 8 (12):e1002822, Dec. 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002822.

M. F. P. Caffrey and D. C. O. James. Human Lymphocyte Antigen Association in Ankylosing Spondylitis. *Nature*, 242(5393):121–121, Mar. 1973. ISSN 1476-4687. doi: 10.1038/242121a0.

T. T. Cai, W. Sun, and W. Wang. Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):187–234, 2019. ISSN 1467-9868. doi: 10.1111/rssb.12304.

X. Cai, Y. Qiao, C. Diao, et al. Association between polymorphisms of the IKZF3 gene and systemic lupus erythematosus in a Chinese Han population. *PloS One*, 9(10):e108661, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0108661.

B. Calabrese. Linkage Disequilibrium. In S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 763–765. Academic Press, Oxford, Jan. 2019. ISBN 978-0-12-811432-2. doi: 10.1016/B978-0-12-809633-8.20234-3.

E. Cano-Gamez and G. Trynka. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00424.

D. Carvalho-Silva, A. Pierleoni, M. Pignatelli, et al. Open Targets Platform: New developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065, Jan. 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1133.

C. Chai, Z. Xie, and E. Grotewold. SELEX (Systematic Evolution of Ligands by EXponential Enrichment), as a powerful tool for deciphering the protein-DNA interaction space. *Methods in Molecular Biology (Clifton, N.J.)*, 754:249–258, 2011. ISSN 1940-6029. doi: 10.1007/978-1-61779-154-3_14.

C. C. Chang, C. C. Chow, L. C. Tellier, et al. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), Dec. 2015. doi: 10.1186/s13742-015-0047-8.

L. Chen, Q. Niu, Z. Huang, et al. IKZF1 polymorphisms are associated with susceptibility, cytokine levels, and clinical features in systemic lupus erythematosus. *Medicine*, 99(41):e22607, Oct. 2020. ISSN 0025-7974. doi: 10.1097/MD.0000000000022607.

W. Chen, B. R. Larrabee, I. G. Ovsyannikova, et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, 200(3):719–736, July 2015. ISSN 1943-2631. doi: 10.1534/genetics.115.176107.

J. Chèneby, Z. Ménétrier, M. Mestdagh, et al. ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Research*, 48(D1): D180–D188, Jan. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz945.

C.-Y. Cheng, C.-H. Chu, H.-W. Hsu, et al. An improved ChIP-seq peak detection system for simultaneously identifying post-translational modified transcription factors by combinatorial fusion, using SUMOylation as an example. *BMC genomics*, 15 Suppl 1:S1, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-S1-S1.

J. Chiou, R. J. Geusz, M.-L. Okino, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*, 594(7863):398–402, June 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03552-w.

F. Ciccarelli, M. D. Martinis, and L. Ginaldi. An Update on Autoinflammatory Diseases. *Current Medicinal Chemistry*, 21(3):261–269, Jan. 2013. ISSN 0929-8673. doi: 10.2174/09298673113206660303.

M. Claussnitzer, S. N. Dankel, K.-H. Kim, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. https://www.nejm.org/doi/10.1056/NEJMoa1502214, Sept. 2015.

J. D. Cooper, D. J. Smyth, A. M. Smiles, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, 40(12):1399–1401, Dec. 2008. ISSN 1546-1718. doi: 10.1038/ng.249.

N. J. Cooper, C. Wallace, O. Burren, et al. Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions. *bioRxiv*, Apr. 2017. doi: 10.1101/120022.

H. J. Cordell. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, Oct. 2002. ISSN 0964-6906. doi: 10.1093/hmg/11.20.2463.

H. J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6): 392–404, June 2009. ISSN 1471-0064. doi: 10.1038/nrg2579.

O. Corradin and P. C. Scacheri. Enhancer variants: Evaluating functions in common disease. *Genome Medicine*, 6(10), Oct. 2014. ISSN 1756-994X. doi: 10.1186/s13073-014-0085-3.

A. Cortes and M. A. Brown. Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, 13(1):101, 2011. ISSN 1478-6354. doi: 10.1186/ar3204.

F. Costantino, M. Breban, and H.-J. Garchon. Genetics and Functional Genomics of Spondyloarthritis. *Frontiers in Immunology*, 9, 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.02933.

M. Costanzo, A. Baryshnikova, J. Bellay, et al. The Genetic Landscape of a Cell. *Science*, 327(5964):425–431, Jan. 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1180823.

R. Cowper-Sal·lari, X. Zhang, J. B. Wright, et al. Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*, 44(11):1191–1198, Nov. 2012. ISSN 1546-1718. doi: 10.1038/ng.2416.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931, Dec. 2010. doi: 10.1073/pnas.1016071107.

A. G. Cudworth and J. C. Woodrow. HL-A System and Diabetes Mellitus. *Diabetes*, 24(4):345–349, Apr. 1975. ISSN 0012-1797, 1939-327X. doi: 10.2337/diab.24.4.345.

D. S. Cunninghame Graham, D. L. Morris, T. R. Bhangale, et al. Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with Systemic Lupus Erythematosus. *PLoS Genetics*, 7(10):e1002341, Oct. 2011. ISSN 1553-7390. doi: 10.1371/journal.pgen.1002341.

D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad. The Functional Consequences of Variation in Transcription Factor Binding. *PLOS Genetics*, 10(3):e1004226, Mar. 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004226.

T. Dadaev, E. J. Saunders, P. J. Newcombe, et al. Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nature Communications*, 9(1):2256, June 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04109-8.

G. Darnell, D. Duong, B. Han, and E. Eskin. Incorporating prior information into association studies. *Bioinformatics*, 28(12):i147–i153, June 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts235.

S. Das, L. Forer, S. Schönherr, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, Oct. 2016. ISSN 1546-1718. doi: 10.1038/ng.3656.

L. de la Torre-Ubieta, J. L. Stein, H. Won, et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell*, 172(1):289–304.e18, Jan. 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2017.12.014.

G. de Los Campos, D. A. Sorensen, and M. A. Toro. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). *G3 (Bethesda, Md.)*, 9(5):1429–1436, May 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400101.

F. Demenais, P. Margaritte-Jeannin, K. C. Barnes, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*, 50(1):42–53, Jan. 2018. ISSN 1546-1718. doi: 10.1038/s41588-017-0014-7.

D. Demontis, R. K. Walters, J. Martin, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*, 51(1):63–75, Jan. 2019. ISSN 1546-1718. doi: 10.1038/s41588-018-0269-7.

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–244, Mar. 2014. doi: doi.org/10.1038/ng.2897.

S. Dovat, E. Montecino-Rodriguez, V. Schuman, et al. Transgenic Expression of Helios in B Lineage Cells Alters B Cell Properties and Promotes Lymphomagenesis. *The Journal of Immunology*, 175(6):3508–3515, Sept. 2005. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.175.6.3508.

L. Du and C. Zhang. Single-index modulated multiple testing. *Annals of Statistics*, 42(4):1262–1311, Aug. 2014. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1222.

W. Du, Y.-W. Shen, W.-H. Lee, et al. Foxp3+ Treg expanded from patients with established diabetes reduce Helios expression while retaining normal function compared to healthy individuals. *PloS One*, 8(2):e56209, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0056209.

M. Duhamel, I. Arrouss, H. Merle-Béral, and A. Rebollo. The Aiolos transcription factor is up-regulated in chronic lymphocytic leukemia. *Blood*, 111(6):3225–3228, Mar. 2008. ISSN 0006-4971. doi: 10.1182/blood-2007-09-113191.

C. W. Dunnett. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272):1096–1121, Dec. 1955. ISSN 0162-1459. doi: 10.1080/01621459.1955.10501294.

B. Efron. Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, 99(465):96–104, Mar. 2004. ISSN 0162-1459. doi: 10.1198/016214504000000089.

B. Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, Aug. 2007. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053606000001460.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, Dec. 2001. ISSN 0162-1459. doi: 10.1198/016214501753382129.

D. Ellinghaus, L. Jostins, S. L. Spain, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nature Genetics*, 48(5):510–518, May 2016. ISSN 1546-1718. doi: 10.1038/ng.3528.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

D. J. Erle and D. Sheppard. The cell biology of asthma. *The Journal of Cell Biology*, 205(5):621–631, June 2014. ISSN 0021-9525. doi: 10.1083/jcb.201401050.

E. Eskin. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research*, 18(4):653–660, Jan. 2008. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.072785.107.

S. Eyre, J. Bowes, D. Diogo, et al. High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature genetics*, 44(12):1336–1340, Dec. 2012. ISSN 1061-4036. doi: 10.1038/ng.2462.

L. Fachal, H. Aschard, J. Beesley, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics*, 52(1):56–73, Jan. 2020. ISSN 1546-1718. doi: 10.1038/s41588-019-0537-1.

K. K.-H. Farh, A. Marson, J. Zhu, et al. Genetic and Epigenetic Fine-Mapping of Causal Autoimmune Disease Variants. *Nature*, 518(7539):337–343, Feb. 2015. ISSN 0028-0836. doi: 10.1038/nature13835.

E. Ferkingstad, A. Frigessi, H. Rue, et al. Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2(2):714–735, June 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS158.

M. M. A. Fernando, C. R. Stevens, E. C. Walsh, et al. Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. *PLoS Genetics*, 4(4):e1000024, Apr. 2008. ISSN 1553-7390. doi: 10.1371/journal.pgen.1000024.

M. A. Ferreira, R. Jansen, G. Willemsen, et al. Gene-based analysis of regulatory variants identifies four putative novel asthma risk genes related to nucleotide synthesis and signaling. *The Journal of allergy and clinical immunology*, 139(4):1148–1157, Apr. 2017. ISSN 0091-6749. doi: 10.1016/j.jaci.2016.07.017.

M. A. R. Ferreira, M. C. Matheson, C. S. Tang, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *The Journal of Allergy and Clinical Immunology*, 133 (6):1564–1571, June 2014. ISSN 1097-6825. doi: 10.1016/j.jaci.2013.10.030.

R. C. Ferreira, X. Castro Dopico, J. J. Oliveira, et al. Chronic Immune Activation in Systemic Lupus Erythematosus and the Autoimmune PTPN22 Trp620 Risk Allele Drive the Expansion of FOXP3+ Regulatory T Cells and PD-1 Expression. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.02606.

FinnGen. {FinnGen} Documentation of R5 release. https://finngen.gitbook.io/documentation/, 2021.

H. K. Finucane, B. Bulik-Sullivan, A. Gusev, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, Nov. 2015. ISSN 1546-1718. doi: 10.1038/ng.3404.

M. D. Fortune and C. Wallace. simGWAS: A fast method for simulation of large scale case–control GWAS summary statistics. *Bioinformatics*, 35(11):1901–1906, June 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty898.

L. G. Fritsche, W. Igl, J. N. C. Bailey, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, 48(2):134–143, Feb. 2016. ISSN 1546-1718. doi: 10.1038/ng.3448.

C. C. Funk, A. M. Casella, S. Jung, et al. Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data across 27 Tissue Types. *Cell Reports*, 32(7):108029, Aug. 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2020.108029.

T. S. Furey. ChIP-seq and Beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*, 13(12):840–852, Dec. 2012. ISSN 1471-0056. doi: 10.1038/nrg3306.

A. R. García, A. Paterou, M. Lee, et al. HLA class II mediates type 1 diabetes risk by anti-insulin repertoire selection, Sept. 2021.

G. Garg, J. R. Tyler, J. H. M. Yang, et al. Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+CD25+ regulatory T cell function. *Journal of Immunology (Baltimore, Md.: 1950)*, 188(9):4644–4653, May 2012. ISSN 1550-6606. doi: 10.4049/jimmunol.1100272.

S. Gazal, H. K. Finucane, N. A. Furlotte, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, 49(10):1421–1427, Oct. 2017. ISSN 1546-1718. doi: 10.1038/ng.3954.

Y. Ge, T. K. Paisie, J. R. B. Newman, et al. UBASH3A Mediates Risk for Type 1 Diabetes Through Inhibition of T-Cell Receptor-Induced NF-κB Signaling. *Diabetes*, 66(7):2033–2043, July 2017. ISSN 1939-327X. doi: 10.2337/db16-1023.

C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00347.

C. R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 93 (3):509–524, Sept. 2006. ISSN 1464-3510, 0006-3444. doi: 10.1093/biomet/93.3.509.

K. Georgopoulos, D. D. Moore, and B. Derfler. Ikaros, an early lymphoid-specific transcription factor and a putative mediator for T cell commitment. *Science (New York, N.Y.)*, 258(5083):808–812, Oct. 1992. ISSN 0036-8075. doi: 10.1126/science.1439790.

J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014. ISSN 1097-0258. doi: 10.1002/sim.6082.

A. S. Gokhale, A. Gangaplara, M. Lopez-Occasio, et al. Selective deletion of Eos (Ikzf4) in T-regulatory cells leads to loss of suppressive function and development of systemic autoimmunity. *Journal of Autoimmunity*, 105:102300, Dec. 2019. ISSN 1095-9157. doi: 10.1016/j.jaut.2019.06.011.

P. Gormley, V. Anttila, B. S. Winsvold, et al. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics*, 48(8):856–866, Aug. 2016. ISSN 1546-1718. doi: 10.1038/ng.3598.

C. E. Grant, T. L. Bailey, and W. S. Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, Apr. 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr064.

J. M. Greene. Locating three-dimensional roots by a bisection method. *Journal of Computational Physics*, 98(2): 194–198, Feb. 1992. ISSN 0021-9991. doi: 10.1016/0021-9991(92)90137-N.

E. Grundberg, K. S. Small, Å. K. Hedman, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, Oct. 2012. ISSN 1546-1718. doi: 10.1038/ng.2394.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. ISSN 1546-1718. doi: 10.1038/ng.2653.

R. Gupta, J. Hadaya, A. Trehan, et al. A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell*, 170(3):522–533.e15, 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.06.049.

A. Gusev, S. H. Lee, G. Trynka, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American Journal of Human Genetics*, 95(5):535–552, Nov. 2014. ISSN 0002-9297. doi: 10.1016/j.ajhg.2014.10.004.

E. G. Gusmao, C. Dieterich, M. Zenke, and I. G. Costa. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151, Nov. 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu519.

E. G. Gusmao, M. Allhoff, M. Zenke, and I. G. Costa. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, 13(4):303–309, Apr. 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3772.

K. Hahm, B. S. Cobb, A. S. McCarty, et al. Helios, a T cell-restricted Ikaros family member that quantitatively associates with Ikaros at centromeric heterochromatin. *Genes & Development*, 12(6):782–796, Mar. 1998. ISSN 0890-9369, 1549-5477.

H. Hakonarson, H.-Q. Qu, J. P. Bradfield, et al. A Novel Susceptibility Locus for Type 1 Diabetes on Chr12q13 Identified by a Genome-Wide Association Study. *Diabetes*, 57(4):1143–1146, Apr. 2008. ISSN 0012-1797, 1939-327X. doi: 10.2337/db07-1305.

J.-W. Han, H.-F. Zheng, Y. Cui, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genetics*, 41(11):1234–1237, Nov. 2009. ISSN 1546-1718. doi: 10.1038/ng.472.

Y. Han, Q. Jia, P. S. Jahani, et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nature Communications*, 11(1):1776, Apr. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15649-3.

A. Hanson and M. A. Brown. Genetics and the causes of ankylosing spondylitis. *Rheumatic diseases clinics of North America*, 43(3):401–414, Aug. 2017. ISSN 0889-857X. doi: 10.1016/j.rdc.2017.04.006.

K. Hao, Y. Bossé, D. C. Nickle, et al. Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *PLOS Genetics*, 8(11):e1003029, Nov. 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003029.

J. Harrow, A. Frankish, J. M. Gonzalez, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, Sept. 2012. ISSN 1088-9051. doi: 10.1101/gr.135350.111.

M. Hedl, S. Zheng, and C. Abraham. The IL18RAP Region Disease Polymorphism Decreases IL-18RAP/IL-18R1/IL-1R1 Expression and Signaling through Innate Receptor–Initiated Pathways. *The Journal of Immunology*, 192(12):5924–5932, June 2014. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1302727.

M. Heinig, E. Petretto, C. Wallace, et al. A trans -acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467(7314):460–464, Sept. 2010. ISSN 1476-4687. doi: 10.1038/nature09386.

G. Hemani, K. Shakhbazov, H.-J. Westra, et al. RETRACTED ARTICLE: Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–253, Apr. 2014. ISSN 1476-4687. doi: 10.1038/nature13005.

G. Hemani, J. E. Powell, H. Wang, et al. Phantom epistasis between unlinked loci. *Nature*, 596(7871):E1–E3, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03765-z.

E. W. Hewitt. The MHC class I antigen presentation pathway: Strategies for viral immune evasion. *Immunology*, 110(2):163–169, Oct. 2003. ISSN 0019-2805. doi: 10.1046/j.1365-2567.2003.01738.x.

S. M. Hill, L. M. Heiser, T. Cokelaer, et al. Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, Apr. 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3773.

T. Hirota, A. Takahashi, M. Kubo, et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. *Nature genetics*, 44(11):1222–1226, Nov. 2012. ISSN 1061-4036. doi: 10.1038/ng.2438.

D. Hnisz, B. J. Abraham, T. I. Lee, et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, 155 (4):934–947, Nov. 2013. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2013.09.053.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979. ISSN 0303-6898.

F. Hormozdiari, E. Kostem, E. Y. Kang, et al. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, Oct. 2014. ISSN 1943-2631. doi: 10.1534/genetics.114.167908.

L. Hou and H. Zhao. A review of post-GWAS prioritization approaches. *Frontiers in Genetics*, 4, 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00280.

M. D. Howell, P. Gao, B. E. Kim, et al. The signal transducer and activator of transcription 6 gene (STAT6) increases the propensity of patients with atopic dermatitis toward disseminated viral skin infections. *The Journal of Allergy and Clinical Immunology*, 128(5):1006–1014, Nov. 2011. ISSN 1097-6825. doi: 10.1016/j.jaci.2011.06.003.

B. Howie, J. Marchini, and M. Stephens. Genotype Imputation with Thousands of Genomes. *G3 Genes|Genomes|Genetics*, 1(6):457–470, Nov. 2011. ISSN 2160-1836. doi: 10.1534/g3.111.001198.

J. X. Hu, H. Zhao, and H. H. Zhou. False Discovery Rate Control With Groups. *Journal of the American Statistical Association*, 105(491):1215–1227, Sept. 2010. ISSN 0162-1459. doi: 10.1198/jasa.2010.tm09329.

S.-j. Hu, L.-l. Wen, X. Hu, et al. IKZF1: A critical role in the pathogenesis of systemic lupus erythematosus? *Modern Rheumatology*, 23(2):205–209, Mar. 2013. ISSN 1439-7609. doi: 10.1007/s10165-012-0706-x.

H. Huang, M. Fang, L. Jostins, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662):173–178, July 2017. ISSN 1476-4687. doi: 10.1038/nature22969.

M. L. A. Hujoel, S. Gazal, F. Hormozdiari, et al. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. *The American Journal of Human Genetics*, 104(4):611–624, Apr. 2019. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2019.02.008.

A. Hutchinson, J. Asimit, and C. Wallace. Fine-mapping genetic associations. *Human Molecular Genetics*, 29 (R1):R81–R88, Sept. 2020a. ISSN 0964-6906. doi: 10.1093/hmg/ddaa148.

A. Hutchinson, H. Watson, and C. Wallace. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Computational Biology*, 16(4):e1007829, Apr. 2020b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007829.

A. Hutchinson, G. Reales, T. Willis, and C. Wallace. Leveraging auxiliary data from arbitrary distributions to boost GWAS discovery with Flexible cFDR. *bioRxiv*, Apr. 2021. doi: 10.1101/2020.12.04.411710.

C. Igartua, R. A. Myers, R. A. Mathias, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nature Communications*, 6:5965, Jan. 2015. ISSN 2041-1723. doi: 10.1038/ncomms6965.

N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, July 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3885.

S. Ikegawa. A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going. *Genomics & Informatics*, 10(4):220–225, Dec. 2012. ISSN 1598-866X. doi: 10.5808/GI.2012.10.4.220.

J. R. J. Inshaw, A. J. Cutler, D. J. M. Crouch, 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. *Diabetes Care*, 43(1):169–177, Jan. 2020. ISSN 1935-5548. doi: 10.2337/dc19-0803.

International Genetics of Ankylosing Spondylitis Consortium (IGAS). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature Genetics*, 45(7):730–738, July 2013. ISSN 1546-1718. doi: 10.1038/ng.2667.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. doi: 10.1038/35057062.

V. Iotchkova, G. R. Ritchie, M. Geihs, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature genetics*, 51(2):343–353, Feb. 2019. ISSN 1061-4036. doi: 10.1038/s41588-018-0322-6.

B. M. Javierre, O. S. Burren, S. P. Wilder, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, Nov. 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.09.037.

L. Jostins, S. Ripke, R. K. Weersma, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, Nov. 2012. ISSN 1476-4687. doi: 10.1038/nature11582.

M. H. Kaplan, S. Sehra, H.-C. Chang, et al. Constitutively active STAT6 predisposes toward a lymphoproliferative disorder. *Blood*, 110(13):4367–4369, Dec. 2007. ISSN 0006-4971. doi: 10.1182/blood-2007-06-098244.

K. L. Keene, A. R. Quinlan, X. Hou, et al. Evidence for two independent associations with type 1 diabetes at the 12q13 locus. *Genes and Immunity*, 13(1):66–70, Jan. 2012. ISSN 1476-5470. doi: 10.1038/gene.2011.56.

C. Kelley, T. Ikeda, J. Koipally, et al. Helios, a novel dimerization partner of Ikaros expressed in the earliest hematopoietic progenitors. *Current Biology*, 8(9):508–515, 1998. doi: 10.1016/s0960-9822(98)70202-7.

N. Kerimov, J. D. Hayhurst, K. Peikova, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics*, 53(9):1290–1299, Sept. 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00924-w.

E. Khurana, Y. Fu, V. Colonna, et al. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*, 342(6154), Oct. 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1235587.

G. Kichaev, W.-Y. Yang, S. Lindstrom, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genetics*, 10(10):e1004722, Oct. 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004722.

G. Kichaev, G. Bhatia, P.-R. Loh, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *American Journal of Human Genetics*, 104(1):65–75, Mar. 2019. ISSN 1537-6605. doi: 10.1016/j.ajhg.2018.11.008.

M. Kircher, D. M. Witten, P. Jain, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Mar. 2014. ISSN 1546-1718. doi: 10.1038/ng.2892.

R. J. Klein, C. Zeiss, E. Y. Chew, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, N.Y.)*, 308(5720):385–389, Apr. 2005. ISSN 0036-8075. doi: 10.1126/science.1109557.

K. Korthauer, P. K. Kimes, C. Duvallet, et al. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):118, June 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1716-1.

R. M. Kuhn, D. Haussler, and W. J. Kent. The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 14(2):144, Mar. 2013. doi: 10.1093/bib/bbs038.

I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, Jan. 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx1106.

D. A. Kuperman, X. Huang, L. L. Koth, et al. Direct effects of interleukin-13 on epithelial cells cause airway hyperreactivity and mucus overproduction in asthma. *Nature Medicine*, 8(8):885–889, Aug. 2002. ISSN 1078-8956. doi: 10.1038/nm734.

S. J. Kuritz, J. R. Landis, and G. G. Koch. A General Overview of Mantel-Haenszel Methods: Applications and Recent Developments. *Annual Review of Public Health*, 9(1):123–160, 1988. doi: 10.1146/annurev.pu.09.050188.001011.

J. R. Landis, E. R. Heyman, and G. G. Koch. Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. *International Statistical Review / Revue Internationale de Statistique*, 46(3):237–254, 1978. ISSN 0306-7734. doi: 10.2307/1402373.

N. LaPierre, K. Taraszka, H. Huang, et al. Identifying causal variants by fine mapping across multiple studies. *PLOS Genetics*, 17(9):e1009733, Sept. 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009733.

T. Lappalainen. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Research*, 25(10):1427–1431, Oct. 2015. ISSN 1088-9051. doi: 10.1101/gr.190983.115.

Y. Lee, F. Luca, R. Pique-Regi, and X. Wen. Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. *bioRxiv*, May 2018. doi: 10.1101/316471.

Y. L. Lee, J. J.-Y. Yen, L.-C. Hsu, et al. Association of STAT6 genetic variants with childhood atopic dermatitis in Taiwanese population. *Journal of Dermatological Science*, 79(3):222–228, Sept. 2015. ISSN 1873-569X. doi: 10.1016/j.jdermsci.2015.05.006.

J. T. Leek, L. Jager, S. M. Boca, and T. Konopka. Swfdr: Science-wise false discovery rate and proportion of true null hypotheses estimation. Bioconductor version: Release (3.12), 2021.

L. Lei and W. Fithian. AdaPT: An interactive procedure for multiple testing with side information. *arXiv:1609.06035 [stat]*, July 2018.

E. Leinoe, M. Kjaersgaard, E. Zetterberg, et al. Highly impaired platelet ultrastructure in two families with novel IKZF5 variants. *Platelets*, 32(4):492–497, May 2021. ISSN 0953-7104. doi: 10.1080/09537104.2020.1764921.

J. Lempainen, T. Härkönen, A. Laine, et al. Associations of polymorphisms in non-HLA loci with autoantibodies at the diagnosis of type 1 diabetes: INS and IKZF4 associate with insulin autoantibodies. *Pediatric Diabetes*, 14(7):490–496, Nov. 2013. ISSN 1399-5448. doi: 10.1111/pedi.12046.

C. Lentaigne, D. Greene, S. Sivapalaratnam, et al. Germline mutations in the transcription factor IKZF5 cause thrombocytopenia. *Blood*, 134(23):2070–2081, Dec. 2019. ISSN 0006-4971. doi: 10.1182/blood.2019000782.

R. C. Lewontin. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67, Jan. 1964. ISSN 0016-6731.

A. Li and R. F. Barber. Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *arXiv:1606.07926 [stat]*, Sept. 2017.

H. Li, B. Handsaker, A. Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352.

Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, Sept. 2011. ISSN 1932-6157. doi: 10.1214/11-AOAS466.

L. Liang, N. Morar, A. L. Dixon, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, 23(4):716–726, Apr. 2013. ISSN 1088-9051. doi: 10.1101/gr.142521.112.

J. Liley and C. Wallace. A Pleiotropy-Informed Bayesian False Discovery Rate Adapted to a Shared Control Design Finds New Disease Associations From GWAS Summary Statistics. *PLOS Genetics*, 11(2):e1004926, Feb. 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004926.

J. Liley and C. Wallace. Accurate error control in high-dimensional association testing using conditional false discovery rates. *Biometrical Journal*, 2021. ISSN 1521-4036. doi: 10.1002/bimj.201900254.

D. Y. Lin. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787, Mar. 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti053.

P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, Mar. 2015. ISSN 1546-1718. doi: 10.1038/ng.3190.

P.-R. Loh, P. Danecek, P. F. Palamara, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, Nov. 2016. ISSN 1546-1718. doi: 10.1038/ng.3679.

Q. Lu, Y. Hu, J. Sun, et al. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. *Scientific Reports*, 5(1):10576, May 2015. ISSN 2045-2322. doi: 10.1038/srep10576.

Q. Lu, R. L. Powles, Q. Wang, et al. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS genetics*, 12(4):e1005947, Apr. 2016a. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005947.

Q. Lu, X. Yao, Y. Hu, and H. Zhao. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, 32(4):542–548, Feb. 2016b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv610.

Q. Lu, R. L. Powles, S. Abdallah, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLOS Genetics*, 13(7):e1006933, July 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006933.

C. Lyon de Ana, K. Arakcheeva, P. Agnihotri, et al. Lack of Ikaros deregulates inflammatory gene programs in T cells. *Journal of immunology (Baltimore, Md. : 1950)*, 202(4):1112–1123, Feb. 2019. ISSN 0022-1767. doi: 10.4049/jimmunol.1801270.

X. Ma, P. Wang, G. Xu, et al. Integrative genomics analysis of various omics data and networks identify risk genes and variants vulnerable to childhood-onset asthma. *BMC Medical Genomics*, 13(1):123, Aug. 2020. ISSN 1755-8794. doi: 10.1186/s12920-020-00768-z.

M. J. Machiela and S. J. Chanock. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557, Nov. 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv402.

J. B. Maller, G. McVean, J. Byrnes, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294–1301, Dec. 2012. ISSN 1546-1718. doi: 10.1038/ng.2435.

R. Marke, F. N. van Leeuwen, and B. Scheijen. The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. *Haematologica*, 103(4):565–574, Apr. 2018. ISSN 1592-8721. doi: 10.3324/haematol.2017.185603.

D. Marnetto, F. Mantica, I. Molineris, et al. Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *The American Journal of Human Genetics*, 102(2):207–218, Feb. 2018. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.12.014.

R. Martin and S. Tokdar. A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–439, July 2012. ISSN 1465-4644. doi: 10.1093/biostatistics/kxr039.

G. Martinelli, I. Iacobucci, C. T. Storlazzi, et al. IKZF1 (Ikaros) deletions in BCR-ABL1-positive acute lymphoblastic leukemia are associated with short disease-free survival and high rate of cumulative incidence of relapse: A GIMEMA AL WP report. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 27(31):5202–5207, Nov. 2009. ISSN 1527-7755. doi: 10.1200/JCO.2008.21.6408.

J. Mbatchou, L. Barnard, J. Backman, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):1097–1103, July 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00870-7.

S. McCarthy, S. Das, W. Kretzschmar, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279–1283, Oct. 2016. ISSN 1061-4036. doi: 10.1038/ng.3643.

R. Moore, L. Georgatou-Politou, J. Liley, et al. Genome-wide scale analyses identify novel BMI genotype-environment interactions using a conditional false discovery rate. *bioRxiv*, Jan. 2020. doi: 10.1101/2020.01.22.908038.

B. Morgan, L. Sun, N. Avitahl, et al. Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *The EMBO Journal*, 16(8):2004–2013, Apr. 1997. ISSN 0261-4189. doi: 10.1093/emboj/16.8.2004.

A. P. Morris. Transethnic Meta-Analysis of Genomewide Association Studies. *Genetic Epidemiology*, 35(8): 809–822, Dec. 2011. ISSN 0741-0395. doi: 10.1002/gepi.20630.

C. G. Mullighan, C. B. Miller, I. Radtke, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature*, 453(7191):110–114, May 2008. ISSN 1476-4687. doi: 10.1038/nature06866.

M. Naderi, M. Hashemi, and S. Amininia. Association of TAP1 and TAP2 Gene Polymorphisms with Susceptibility to Pulmonary Tuberculosis. *Iranian Journal of Allergy, Asthma and Immunology*, pages 62–68, Jan. 2016. ISSN 1735-5249.

S. Neph, M. S. Kuehn, A. P. Reynolds, et al. BEDOPS: High-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, July 2012a. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts277.

S. Neph, J. Vierstra, A. B. Stergachis, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, Sept. 2012b. ISSN 1476-4687. doi: 10.1038/nature11212.

P. J. Newcombe, D. V. Conti, and S. Richardson. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genetic Epidemiology*, 40(3):188–201, Apr. 2016. ISSN 1098-2272. doi: 10.1002/gepi.21953.

M. A. Newton. On a Nonparametric Recursive Estimator of the Mixing Distribution. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 64(2):306–322, 2002. ISSN 0581-572X.

A. Nichogiannopoulou, M. Trevisan, C. Friedrich, and K. Georgopoulos. Ikaros in hemopoietic lineage determination and homeostasis. *Seminars in Immunology*, 10(2):119–125, Apr. 1998. ISSN 1044-5323. doi: 10.1006/smim.1998.0113.

J. Nishino, Y. Kochi, D. Shigemizu, et al. Empirical Bayes Estimation of Semi-parametric Hierarchical Mixture Models for Unbiased Characterization of Polygenic Disease Architectures. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00115.

S. S. Nishizaki, N. Ng, S. Dong, et al. Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, 36(2):364–372, Jan. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz612.

H. Nückel, U. H. Frey, L. Sellmann, et al. The IKZF3 (Aiolos) transcription factor is highly upregulated and inversely correlated with clinical progression in chronic lymphocytic leukaemia. *British Journal of Haematology*, 144(2):268–270, 2009. ISSN 1365-2141. doi: 10.1111/j.1365-2141.2008.07442.x.

Y. Ohmori and T. A. Hamilton. Interleukin-4/STAT6 represses STAT1 and NF-kappa B-dependent transcription through distinct mechanisms. *The Journal of Biological Chemistry*, 275(48):38095–38103, Dec. 2000. ISSN 0021-9258. doi: 10.1074/jbc.M006227200.

Y. Ohnishi, T. Tanaka, K. Ozaki, et al. A high-throughput SNP typing system for genome-wide association studies. *Journal of Human Genetics*, 46(8):471–477, 2001. ISSN 1434-5161. doi: 10.1007/s100380170047.

S. Onengut-Gumuscu, W.-M. Chen, O. Burren, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics*, 47(4):381–386, Apr. 2015. ISSN 1546-1718. doi: 10.1038/ng.3245.

K. Ozaki, Y. Ohnishi, A. Iida, et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–654, Dec. 2002. ISSN 1546-1718. doi: 10.1038/ng1047.

O. A. Panagiotou, J. P. A. Ioannidis, and f. t. G.-W. S. Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1):273–286, Feb. 2012. ISSN 0300-5771. doi: 10.1093/ije/dyr178.

E. Papaemmanuil, F. J. Hosking, J. Vijayakrishnan, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nature Genetics*, 41(9):1006–1010, Sept. 2009. ISSN 1546-1718. doi: 10.1038/ng.430.

K. J. Payne and S. Dovat. Ikaros and Tumor Suppression in Acute Lymphoblastic Leukemia. *Critical reviews in oncogenesis*, 16(1-2):3–12, 2011. ISSN 0893-9675.

J. Perdomo, M. Holmes, B. Chong, and M. Crossley. Eos and Pegasus, Two Members of the Ikaros Family of Proteins with Distinct DNA Binding Activities*. *Journal of Biological Chemistry*, 275(49):38347–38354, Dec. 2000. ISSN 0021-9258. doi: 10.1074/jbc.M005457200.

J. K. Pickrell. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *American Journal of Human Genetics*, 94(4):559–573, Apr. 2014. ISSN 0002-9297. doi: 10.1016/j.ajhg.2014.03.004.

J. Piper, M. C. Elze, P. Cauchy, et al. Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, 41(21):e201, Nov. 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt850.

P. Platz, B. Jakobsen, N. Morling, et al. HLA-D AND -DR antigens in genetic analysis of Insulin Dependent Diabetes Mellitus. *Diabetologia*, 21:108–15, Sept. 1981. doi: 10.1007/BF00251276.

M. D. Powell, K. A. Read, B. K. Sreekumar, and K. J. Oestreich. Ikaros Zinc Finger Transcription Factors: Regulators of Cytokine Signaling Pathways and CD4+ T Helper Cell Differentiation. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.01299.

C. Power and J. Elliott. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, 35(1):34–41, Feb. 2006. ISSN 0300-5771. doi: 10.1093/ije/dyi183.

S. Qu, Y. Du, S. Chang, et al. Common variants near IKZF1 are associated with primary Sjögren's syndrome in Han Chinese. *PloS One*, 12(5):e0177320, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177320.

A. R. Quinlan and I. M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar. 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.

F. J. Quintana, H. Jin, E. J. Burns, et al. Aiolos promotes TH17 differentiation by directly silencing Il2 expression. *Nature Immunology*, 13(8):770–777, July 2012. ISSN 1529-2916. doi: 10.1038/ni.2363.

M. A. Reimers, C. Craver, M. Dozmorov, et al. The Coherence Problem: Finding Meaning in GWAS Complexity. *Behavior Genetics*, 49(2):187–195, Mar. 2019. ISSN 1573-3297. doi: 10.1007/s10519-018-9935-x.

R. S. Reshma and D. N. Das. Chapter 9 - Molecular markers and its application in animal breeding. In S. Mondal and R. L. Singh, editors, *Advances in Animal Genomics*, pages 123–140. Academic Press, Jan. 2021. ISBN 978-0-12-820595-2. doi: 10.1016/B978-0-12-820595-2.00009-6.

S. S. Rich, B. Akolkar, P. Concannon, et al. Overview of the Type I Diabetes Genetics Consortium. *Genes & Immunity*, 10(1):S1–S4, Dec. 2009. ISSN 1476-5470. doi: 10.1038/gene.2009.84.

G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek. Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–296, Mar. 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2832.

F. Rivellese, S. Manou-Stathopoulou, D. Mauro, et al. Effects of targeting the transcription factors Ikaros and Aiolos on B cell activation and differentiation in systemic lupus erythematosus. *Lupus Science & Medicine*, 8 (1):e000445, Mar. 2021. ISSN 2053-8790. doi: 10.1136/lupus-2020-000445.

F. Riveros-Mckay, M. E. Weale, R. Moore, et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circulation: Genomic and Precision Medicine*, 14(2):e003304, Apr. 2021. doi: 10.1161/CIRCGEN.120.003304.

C. C. Robertson, J. R. J. Inshaw, S. Onengut-Gumuscu, et al. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nature Genetics*, pages 1–10, June 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00880-5.

K. Roeder and L. Wasserman. Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Statistical Science*, 24(4):398–413, Nov. 2009. ISSN 0883-4237, 2168-8745. doi: 10.1214/09-STS289.

K. Roeder, B. Devlin, and L. Wasserman. Improving power in genome-wide association studies: Weights tip the scale. *Genetic Epidemiology*, 31(7):741–747, 2007. ISSN 1098-2272. doi: 10.1002/gepi.20237.

B. O. Roep. The role of T-cells in the pathogenesis of Type 1 diabetes: From cause to cure. *Diabetologia*, 46(3): 305–321, Mar. 2003. ISSN 0012-186X. doi: 10.1007/s00125-003-1089-5.

D. Rubin, S. Dudoit, and M. van der Laan. A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5:Article19, 2006. ISSN 1544-6115. doi: 10.2202/1544-6115.1148.

D. J. Schaid, W. Chen, and N. B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature reviews. Genetics*, 19(8):491–504, Aug. 2018. ISSN 1471-0056. doi: 10.1038/s41576-018-0016-z.

L. Schlosstein, P. I. Terasaki, R. Bluestone, and C. M. Pearson. High association of an HL-A antigen, W27, with ankylosing spondylitis. *The New England Journal of Medicine*, 288(14):704–706, Apr. 1973. ISSN 0028-4793. doi: 10.1056/NEJM197304052881403.

E. M. Schmidt, J. Zhang, W. Zhou, et al. GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, 31(16):2601–2606, Aug. 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv201.

C. Schmitt, C. Tonnelle, A. Dalloul, et al. Aiolos and Ikaros: Regulators of lymphocyte development, homeostasis and lymphoproliferation. *Apoptosis*, 7(3):277–284, June 2002. ISSN 1573-675X. doi: 10.1023/A:1015372322419.

A. J. Schork, W. K. Thompson, P. Pham, et al. All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics*, 9(4): e1003449, Apr. 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003449.

J. G. Scott, R. C. Kelly, M. A. Smith, et al. False Discovery Rate Regression: An Application to Neural Synchrony Detection in Primary Visual Cortex. *Journal of the American Statistical Association*, 110(510): 459–471, Apr. 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.990973.

F. Seidl, R. Linder, and I. M. Ehrenreich. Quantitative Trait Variation, Molecular Basis of. In R. M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 388–394. Academic Press, Oxford, Jan. 2016. ISBN 978-0-12-800426-5. doi: 10.1016/B978-0-12-800049-6.00059-7.

H. Shim, D. I. Chasman, J. D. Smith, et al. A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLOS ONE*, 10(4):e0120758, Apr. 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0120758.

N. Shrine, M. A. Portelli, C. John, et al. Moderate-to-severe asthma in individuals of European ancestry: A genome-wide association study. *The Lancet Respiratory Medicine*, 7(1):20–34, Jan. 2019. ISSN 2213-2600, 2213-2619. doi: 10.1016/S2213-2600(18)30389-8.

B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.

D. P. Singal and M. A. Blajchman. Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes*, 22(6):429–432, June 1973. ISSN 0012-1797. doi: 10.2337/diab.22.6.429.

S. Sivakumaran, F. Agakov, E. Theodoratou, et al. Abundant pleiotropy in human complex diseases and traits. *American Journal of Human Genetics*, 89(5):607–618, Nov. 2011. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.10.004.

M. Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, 9(6):477–485, June 2008. ISSN 1471-0056. doi: 10.1038/nrg2361.

D. J. Smyth, V. Plagnol, N. M. Walker, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *The New England Journal of Medicine*, 359(26):2767–2777, Dec. 2008. ISSN 1533-4406. doi: 10.1056/NEJMoa0807917.

B. Sorić. Statistical "Discoveries" and Effect-Size Estimation. *Journal of the American Statistical Association*, 84(406):608–610, June 1989. ISSN 0162-1459. doi: 10.1080/01621459.1989.10478811.

B. Soskic, E. Cano-Gamez, D. J. Smyth, et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics*, 51(10):1486–1493, Oct. 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0493-9.

S. L. Spain and J. C. Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1): R111–R119, Oct. 2015. ISSN 0964-6906. doi: 10.1093/hmg/ddv260.

D. Speed, J. Holmes, and D. J. Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, Apr. 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0600-y.

A. V. Spencer, A. Cox, W.-Y. Lin, et al. Novel Bayes Factors That Capture Expert Uncertainty in Prior Density Specification in Genetic Association Studies. *Genetic Epidemiology*, 39(4):239–248, 2015. ISSN 1098-2272. doi: 10.1002/gepi.21891.

A. V. Spencer, A. Cox, W.-Y. Lin, et al. Incorporating Functional Genomic Information in Genetic Association Studies Using an Empirical Bayes Approach. *Genetic Epidemiology*, 40(3):176–187, 2016. ISSN 1098-2272. doi: 10.1002/gepi.21956.

P. Sriaroon, Y. Chang, B. Ujhazi, et al. Familial Immune Thrombocytopenia Associated With a Novel Variant in IKZF1. *Frontiers in Pediatrics*, 7:139, 2019. ISSN 2296-2360. doi: 10.3389/fped.2019.00139.

E. A. Stahl, S. Raychaudhuri, E. F. Remmers, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, 42(6):508–514, June 2010. ISSN 1546-1718. doi: 10.1038/ng.582.

G. Stelzer, N. Rosen, I. Plaschkes, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54(1):1.30.1–1.30.33, 2016. ISSN 1934-340X. doi: 10.1002/cpbi.5.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00346.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, Aug. 2003. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1530509100.

G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10(9):2997–3011, May 1982. ISSN 0305-1048. doi: 10.1093/nar/10.9.2997.

H. G. Stunnenberg, International Human Epigenome Consortium, and M. Hirst. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, Nov. 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2016.11.007.

C. Sudlow, J. Gallacher, N. Allen, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), Mar. 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.

L. Sun, R. V. Craiu, A. D. Paterson, and S. B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, Sept. 2006. ISSN 0741-0395. doi: 10.1002/gepi.20164.

G. Sveinbjornsson, A. Albrechtsen, F. Zink, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics*, 48(3):314–317, Mar. 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3507.

A. Svejgaard, P. Plartz, and L. P. Ryder. HLA and Disease 1982 - A Survey. *Immunological Reviews*, 70(1): 193–218, 1983. ISSN 1600-065X. doi: 10.1111/j.1600-065X.1983.tb00715.x.

A. D.-E. Swafford, J. M. Howson, L. J. Davison, et al. An Allele of IKZF1 (Ikaros) Conferring Susceptibility to Childhood Acute Lymphoblastic Leukemia Protects Against Type 1 Diabetes. *Diabetes*, 60(3):1041–1044, Mar. 2011. ISSN 0012-1797. doi: 10.2337/db10-0446.

A. Syreeni, N. Sandholm, C. Sidore, et al. Genome-wide search for genes affecting the age at diagnosis of type 1 diabetes. *Journal of Internal Medicine*, 289(5):662–674, May 2021. ISSN 1365-2796. doi: 10.1111/joim.13187.

K. Takeda, M. Kamanaka, T. Tanaka, et al. Impaired IL-13-mediated functions of macrophages in STAT6-deficient mice. *The Journal of Immunology*, 157(8):3220–3222, Oct. 1996a. ISSN 0022-1767, 1550-6606.

K. Takeda, T. Tanaka, W. Shi, et al. Essential role of Stat6 in IL-4 signalling. *Nature*, 380(6575):627–630, Apr. 1996b. ISSN 0028-0836. doi: 10.1038/380627a0.

D. Taliun, D. N. Harris, M. D. Kessler, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, Feb. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03205-y.

L.-S. Tam, J. Gu, and D. Yu. Pathogenesis of ankylosing spondylitis. *Nature Reviews Rheumatology*, 6(7): 399–405, July 2010. ISSN 1759-4804. doi: 10.1038/nrrheum.2010.79.

M. Tang. Understanding p value, multiple comparisons, FDR and q value, Jan. 2019.

M. D. Teare, A. M. Dunning, F. Durocher, et al. Sampling distribution of summary linkage disequilibrium measures. *Annals of Human Genetics*, 66(3):223–233, 2002. ISSN 1469-1809. doi: 10.1046/j.1469-1809.2002.00108.x.

J. E. D. Thaventhiran, H. Lango Allen, O. S. Burren, et al. Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature*, 583(7814):90–95, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2265-1.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, Oct. 2015. ISSN 1476-4687. doi: 10.1038/nature15393.

The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics*, 47(12): 1415–1425, Dec. 2015. ISSN 1546-1718. doi: 10.1038/ng.3437.

The EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*, 47(12):1449–1456, Dec. 2015. doi: 10.1038/ng.3424.

The International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, 2003. ISSN 1476-4687. doi: 10.1038/nature02168.

The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571): 82–90, Oct. 2015. doi: 10.1038/nature14962.

M. Thelwall, M. Munafò, A. Mas-Bleda, et al. Is useful research data usually shared? An investigation of genome-wide association study summary statistics. *PLOS ONE*, 15(2):e0229578, Feb. 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0229578.

M. Thomsen, P. Platz, O. O. Andersen, et al. MLC typing in juvenile diabetes mellitus and idiopathic Addison's disease. *Transplantation Reviews*, 22:125–147, 1975. ISSN 0082-5948. doi: 10.1111/j.1600-065x.1975.tb01555.x.

J. A. Todd, N. M. Walker, J. D. Cooper, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, 39(7):857–864, July 2007. ISSN 1546-1718. doi: 10.1038/ng2068.

Y. Tomer, L. M. Dolan, G. Kahaly, et al. Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes. *Journal of Autoimmunity*, 60:32–39, June 2015. ISSN 1095-9157. doi: 10.1016/j.jaut.2015.03.006.

C. Tonnelle, B. Calmels, C. Maroc, et al. Ikaros gene expression and leukemia. *Leukemia and Lymphoma*, 43(1): 29–35, 2002. doi: 10.1080/10428190210186.

H. Touzet and J.-S. Varré. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for Molecular Biology*, 2(1):15, Dec. 2007. ISSN 1748-7188. doi: 10.1186/1748-7188-2-15.

L. R. Treviño, W. Yang, D. French, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nature Genetics*, 41(9):1001–1005, Sept. 2009. ISSN 1546-1718. doi: 10.1038/ng.432.

G. Trynka, H.-J. Westra, K. Slowikowski, et al. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics*, 97(1):139–152, July 2015. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2015.05.016.

E. Turkheimer. Still Missing. *Research in Human Development*, 8(3-4):227–241, July 2011. ISSN 1542-7609. doi: 10.1080/15427609.2011.625321.

E. Turkheimer. Genome Wide Association Studies of Behavior are Social Science. In K. S. Plaisance and T. A. Reydon, editors, *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science, pages 43–64. Springer Netherlands, Dordrecht, 2012. ISBN 978-94-007-1951-4. doi: 10.1007/978-94-007-1951-4_3.

S. D. TURNER and W. S. BUSH. MULTIVARIATE ANALYSIS OF REGULATORY SNPS: EMPOWERING PERSONAL GENOMICS BY CONSIDERING CIS-EPISTASIS AND HETEROGENEITY. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 276–287, 2011.

M. Valta, A. M. Gazali, T. Viisanen, et al. Type 1 diabetes linked PTPN22 gene polymorphism is associated with the frequency of circulating regulatory T cells. *European Journal of Immunology*, 50(4):581–588, 2020. ISSN 1521-4141. doi: 10.1002/eji.201948378.

M. van de Bunt, A. Cortes, IGAS Consortium, et al. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS genetics*, 11(9):e1005535, 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005535.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Statistics and Computing. Springer-Verlag, New York, fourth edition, 2002. ISBN 978-0-387-95457-8. doi: 10.1007/978-0-387-21706-2.

A. Vibhuti, K. Gupta, H. Subramanian, et al. Distinct and Shared Roles of β-Arrestin-1 and β-Arrestin-2 on the Regulation of C3a Receptor Signaling in Human Mast Cells. *PLoS ONE*, 6(5), May 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019585.

J. Vierstra, J. Lazar, R. Sandstrom, et al. Global reference mapping of human transcription factor footprints. *Nature*, 583(7818):729–736, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2528-x.

D. Villar, C. Berthelot, S. Aldridge, et al. Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3): 554–566, Jan. 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.01.006.

A. Villarreal-Martínez, H. Gallardo-Blanco, R. Cerda-Flores, et al. Candidate gene polymorphisms and risk of psoriasis: A pilot study. *Experimental and Therapeutic Medicine*, 11(4):1217–1222, Apr. 2016. ISSN 1792-0981. doi: 10.3892/etm.2016.3066.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five Years of GWAS Discovery. *American Journal of Human Genetics*, 90(1):7–24, Jan. 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.11.029.

P. M. Visscher, N. R. Wray, Q. Zhang, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2017.06.005.

T. J. Vyse and D. S. C. Graham. Trans-Ancestral Fine-Mapping and Epigenetic Annotation as Tools to Delineate Functionally Relevant Risk Alleles at *IKZF1* and *IKZF3* in Systemic Lupus Erythematosus. *International Journal of Molecular Sciences*, 21(21):8383, Nov. 2020. ISSN 14220067. doi: 10.3390/ijms21218383.

J. Wakefield. Bayes factors for genome-wide association studies: Comparison with P-values. *Genetic Epidemiology*, 33(1):79–86, 2009. ISSN 1098-2272. doi: 10.1002/gepi.20359.

C. Wallace. A more accurate method for colocalisation analysis allowing for multiple causal variants, May 2021.

C. Wallace, D. J. Smyth, M. Maisuria-Armer, et al. The imprinted DLK1 - MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nature Genetics*, 42(1):68–71, Jan. 2010. ISSN 1546-1718. doi: 10.1038/ng.493.

C. Wallace, A. J. Cutler, N. Pontikos, et al. Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLOS Genetics*, 11(6):e1005272, June 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005272.

K. Walters, A. Cox, and H. Yaacob. Using GWAS top hits to inform priors in Bayesian fine-mapping association studies. *Genetic Epidemiology*, 43(6):675–689, 2019. ISSN 1098-2272. doi: 10.1002/gepi.22212.

K. Walters, A. Cox, and H. Yaacob. The utility of the Laplace effect size prior distribution in Bayesian fine-mapping studies. *Genetic Epidemiology*, 45(4):386–401, 2021. ISSN 1098-2272. doi: 10.1002/gepi.22375.

G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a), July 2020. ISSN 1467-9868. doi: 10.1111/rssb.12388.

J. H. Wang, A. Nichogiannopoulou, L. Wu, et al. Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. *Immunity*, 5(6):537–549, Dec. 1996. ISSN 1074-7613. doi: 10.1016/s1074-7613(00)80269-1.

J. H. Wang, N. Avitahl, A. Cariappa, et al. Aiolos regulates B cell activation and maturation to effector state. *Immunity*, 9(4):543–553, Oct. 1998. ISSN 1074-7613. doi: 10.1016/s1074-7613(00)80637-8.

S. Wang and H. Zhao. Sample Size Needed to Detect Gene-Gene Interactions using Association Designs. *American Journal of Epidemiology*, 158(9):899–914, Nov. 2003. ISSN 0002-9262. doi: 10.1093/aje/kwg233.

L. Wasserman and K. Roeder. Weighted Hypothesis Testing. *arXiv:math/0604172*, Apr. 2006.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, Apr. 2004. ISSN 1471-0064. doi: 10.1038/nrg1315.

M. T. Weirauch, A. Cote, R. Norel, et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, Feb. 2013. ISSN 1546-1696. doi: 10.1038/nbt.2486.

O. Weissbrod, F. Hormozdiari, C. Benner, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics*, 52(12):1355–1363, Dec. 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-00735-5.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007. ISSN 1476-4687. doi: 10.1038/nature05911.

X. Wen. A Unified View of False Discovery Rate Control: Reconciliation of Bayesian and Frequentist Approaches. *arXiv:1803.05284 [stat]*, Mar. 2018.

X. Wen, Y. Lee, F. Luca, and R. Pique-Regi. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *American Journal of Human Genetics*, 98(6):1114–1129, June 2016. ISSN 0002-9297. doi: 10.1016/j.ajhg.2016.03.029.

H.-J. Westra, M. J. Peters, T. Esko, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, Oct. 2013. ISSN 1546-1718. doi: 10.1038/ng.2756.

J. L. Wiemels, K. M. Walsh, A. J. de Smith, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nature Communications*, 9(1):286, Jan. 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02596-9.

J. E. Wigginton, D. J. Cutler, and G. R. Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, 76(5):887–893, May 2005. ISSN 0002-9297. doi: 10.1086/429864.

S. Winandy, P. Wu, and K. Georgopoulos. A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. *Cell*, 83(2):289–299, Oct. 1995. ISSN 0092-8674. doi: 10.1016/0092-8674(95)90170-1.

T. Yoshida, S. Y.-M. Ng, J. C. Zuniga-Pflucker, and K. Georgopoulos. Early hematopoietic lineage restrictions directed by Ikaros. *Nature Immunology*, 7(4):382–391, Apr. 2006. ISSN 1529-2908. doi: 10.1038/ni1314.

R. Yurko, M. G'Sell, K. Roeder, and B. Devlin. A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proceedings of the National Academy of Sciences*, 117(26):15028–15035, June 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1918862117.

T. Zeller, P. Wild, S. Szymczak, et al. Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. *PLOS ONE*, 5(5):e10693, May 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010693.

M. J. Zhang, F. Xia, and J. Zou. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, 10(1):3433, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11247-0.

Y. Zhang, T. Liu, C. A. Meyer, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, Sept. 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

W. Zhou, J. B. Nielsen, L. G. Fritsche, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, Sept. 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0184-y.

O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, Jan. 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1119675109.

Appendix A

Appendix to chapter 2

A.1 fcfdr vignettes

A.1.1 Introductory vignette

The `fcfdr` R package implements the cFDR framework and is applicable for a wide variety of auxiliary covariates. This is in contrast to earlier empirical cFDR methods (Liley and Wallace 2021; https://github.com/jamesliley/cfdr) that only support auxiliary p-values from related traits. A direct utility of `fcfdr` is to leverage relevant functional genomic data with GWAS p-values to increase power for GWAS discovery. The method generates "v-values" which can be interpreted as GWAS p-values that have been re-weighted according to the auxiliary data values. Since the v-values are analogous to p-values, they can be used directly in any error-rate controlling procedure.

The `fcfdr` R package contains two key functions:

1. `flexible_cfdr`: Implements cFDR leveraging **continuous** auxiliary covariates.
2. `binary_cfdr`: Implements cFDR leveraging **binary** auxiliary covariates.

Both functions require two parameters to be specified:

- `p`: GWAS p-values for the trait of interest (vector of per-SNP p-values)
- `q`: Auxiliary data values (vector of per-SNP auxiliary data values).

The `flexible_cfdr` function also requires the indices of an independent subset of SNPs (`indep_index` parameter; see LDAK vignette) whilst the `binary_cfdr` function requires group indices for each SNP (`group` parameter; typically the chromosome or LD block index for each SNP).

For further details, including instructions to generate `indep_index`, examples of auxiliary data to leverage and instructions to apply cFDR iteratively, please see the Extra Information vignette.

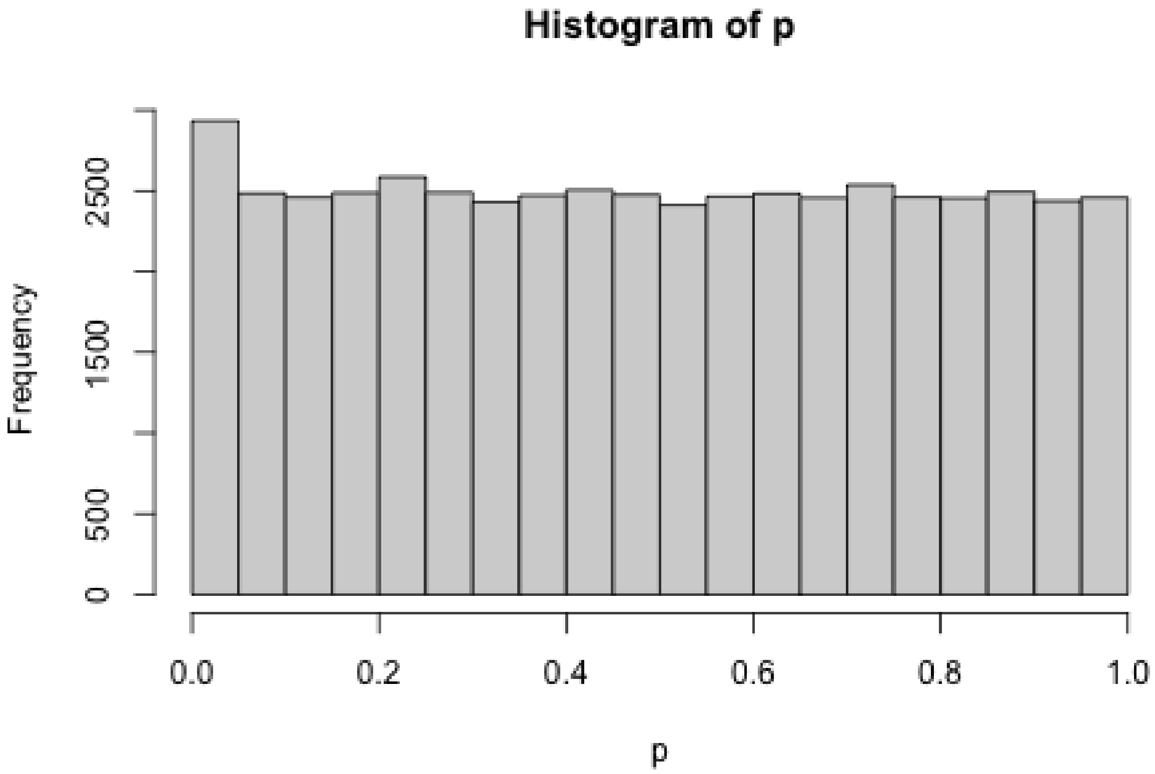
Simple walk-through example

Firstly, load the Flexible cFDR R package:

```
library(fcfdr)
```

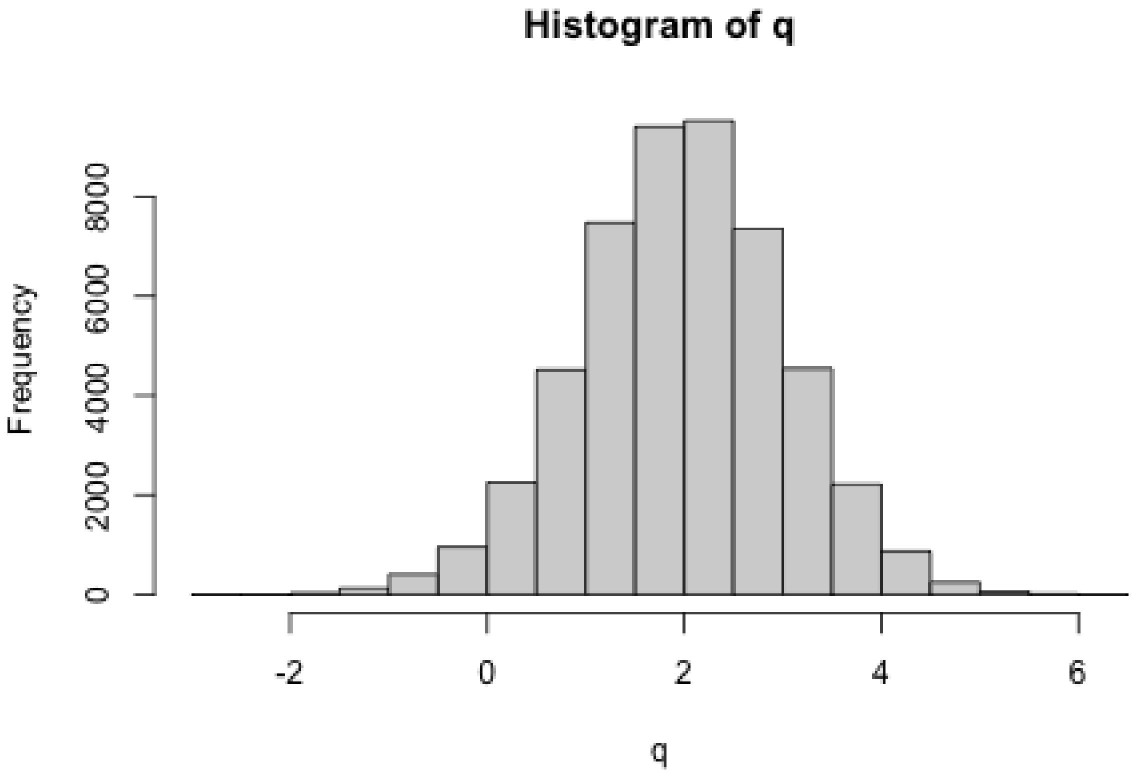
Next, simulate p-values for 50,000 genetic variants, including 500 associated variants.

```
set.seed(1)
n = 50000
n1p = 500 # associated variants
zp = c(rnorm(n1p, sd=5), rnorm(n-n1p, sd=1)) # z-scores
p = 2*pnorm(-abs(zp)) # convert to p-values
hist(p)
```



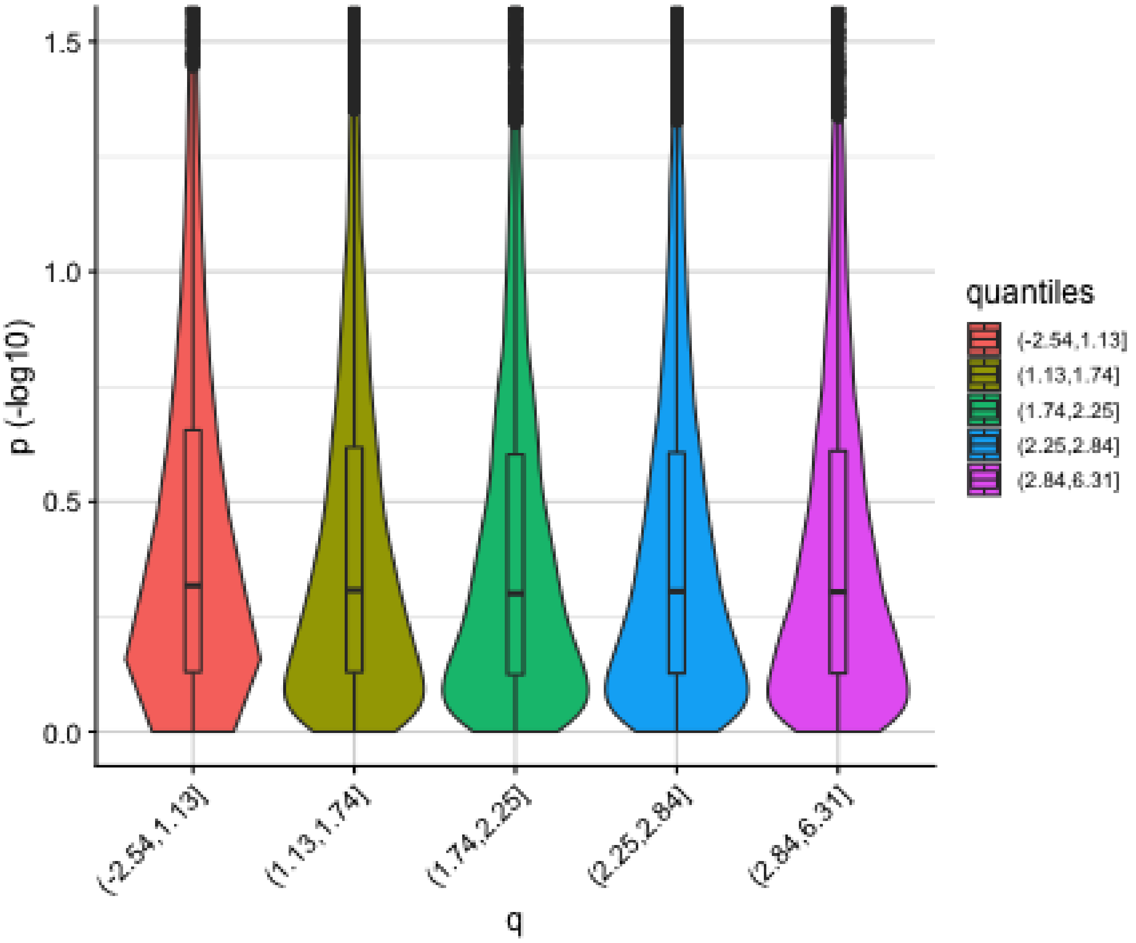
We simulate relevant auxiliary data from a mixture normal distribution (q). The associated SNPs (with indices 1-500) are sampled from $N(-0.5, 0.5^2)$ and the non-associated SNPs (with indices 500-50000) are sampled from $N(2, 1)$.

```
mixture_comp1 <- function(x) rnorm(x, mean = -0.5, sd = 0.5)
mixture_comp2 <- function(x) rnorm(x, mean = 2, sd = 1)
n = length(p)
z = runif(n)

q <- c(mixture_comp1(n1p), mixture_comp2(n-n1p))
hist(q)
```



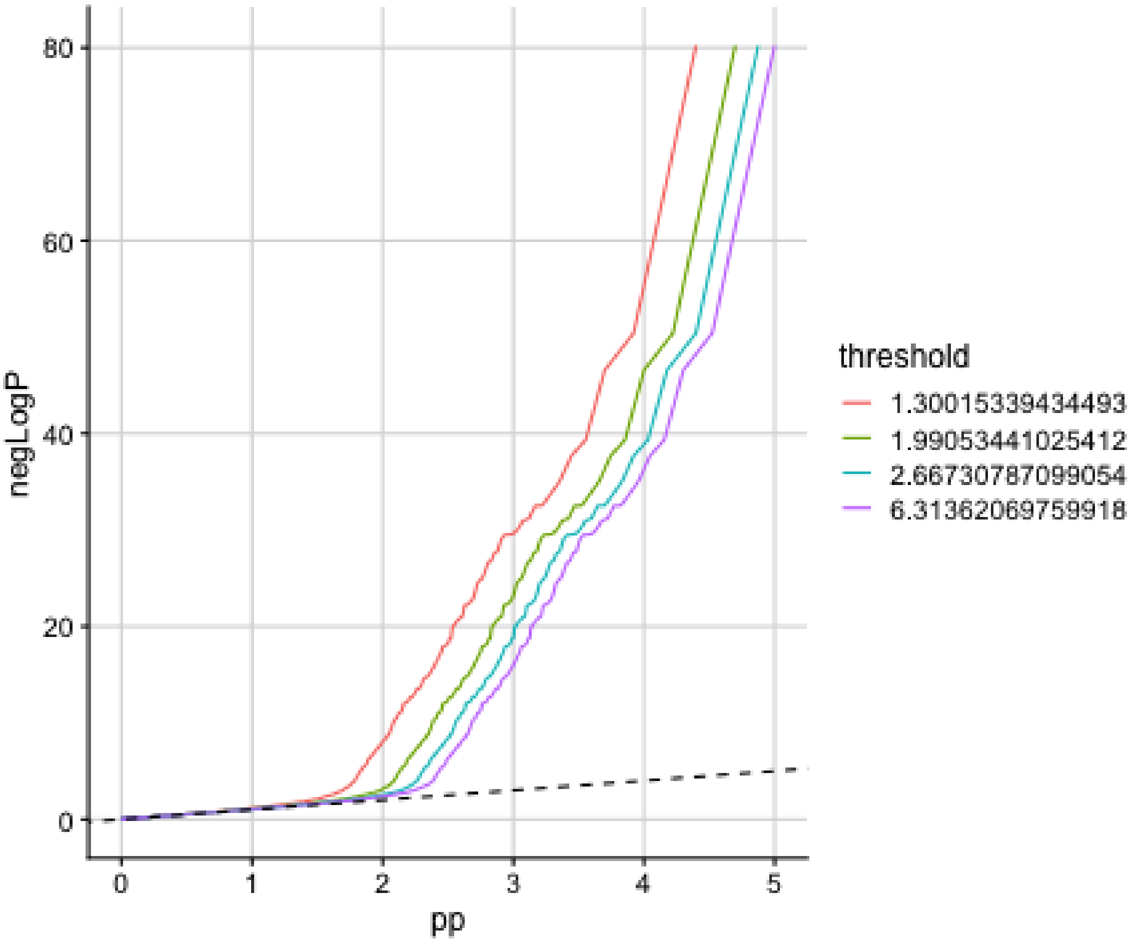
We can use the `corr_plot` function to visualise the relationship between p and q. We observe that low p-values (i.e. high $-log10(p)$) are enriched for low q values.

```
corr_plot(p, q)
```



This is also clear from the stratified QQ plot.

```
stratified_qqplot(data_frame = data.frame(p, q), prin_value_label = "p",
cond_value_label = "q", thresholds = quantile(q)[-1])
```

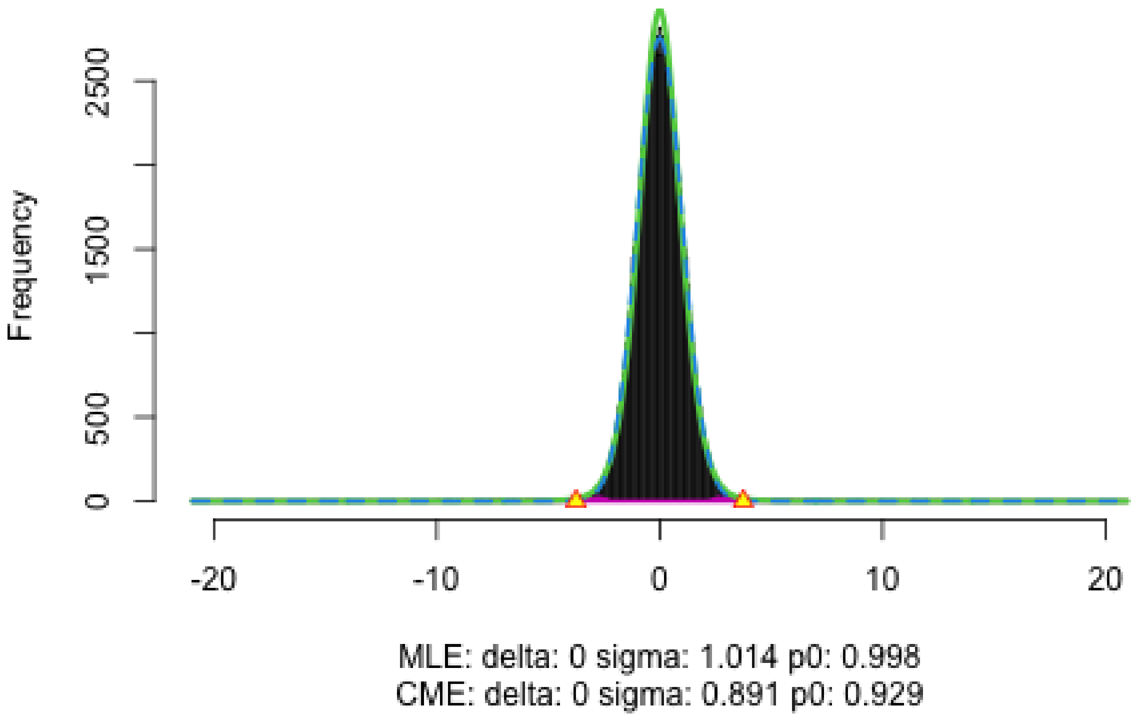


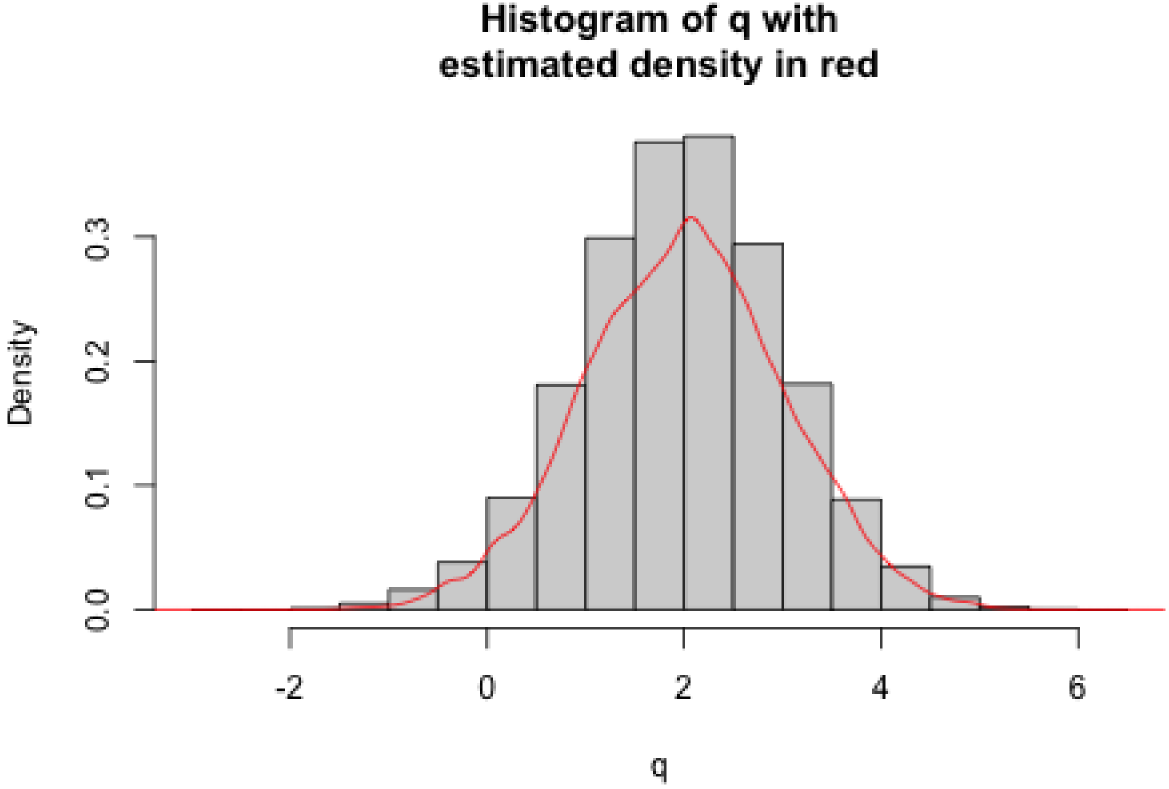
We are now ready to use the `flexible_cfdr` function to derive the v-values. Note that for the purpose of the vignette, we do not specify an independent subset of SNPs, however for real analyses this parameter should be specified appropriately to avoid biased bandwidth estimations when fitting the KDE. Subsets of independent SNPs can be readily found using PLINK or LDAK - see the vignette for deriving LDAK weights here.

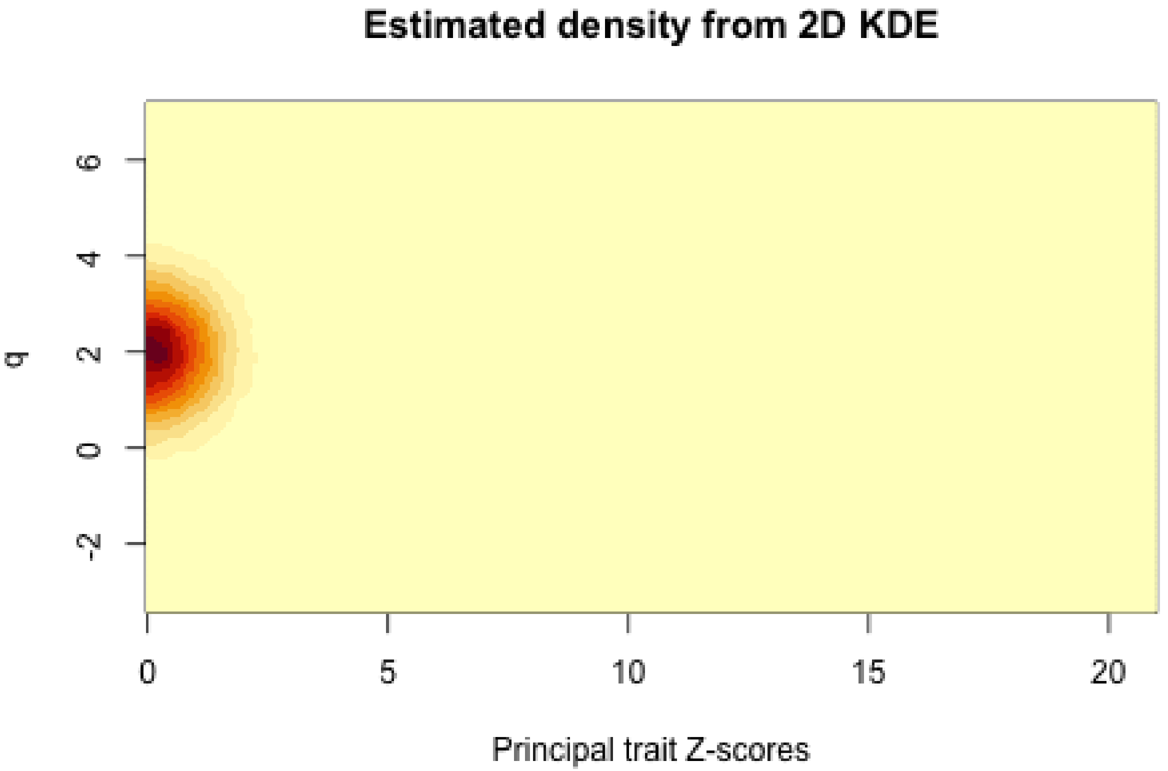
By default, `flexible_cfdr` prints some useful plots so that users can evaluate the accuracy of the KDE that is estimated in the method.

(This line of code will take approximately 4 minutes to run.)

```
res <- flexible_cfdr(p, q, indep_index = seq(1, n, 1))
```


Frequency
0
500
1500
2500
-20
-10
0
10
20
MLE: delta: 0 sigma: 1.014 p0: 0.998
CME: delta: 0 sigma: 0.891 p0: 0.929


Histogram of q with
estimated density in red
Density
0.0
0.1
0.2
0.3
-2
0
2
4
6
q



The output from the function is a list of length two. The first element is a data.frame containing the p-values (input parameter `p`), the auxiliary data values (input parameter `q`) and the generated v-values. The second element contains auxiliary data, such as how many data-points were left-censored in the method and/or spline corrected. In this example, we can see that a value of $q = -0.169$ was used for left censoring (any auxiliary data values smaller than this were set to this value) which results in 1108/50000 (2% of) data points being left censored. We also see that 109 (0.2% of) data points were spline corrected.

```
str(res)
#> List of 2
#>  $ :'data.frame':    50000 obs. of  3 variables:
#>   ..$ p: num [1:50000] 1.73e-03 3.59e-01 2.94e-05 1.51e-15 9.94e-02 ...
#>   ..$ q: num [1:50000] -0.325 -0.297 -0.458 -0.383 -0.578 ...
#>   ..$ v: num [1:50000] 2.07e-04 4.94e-02 3.48e-06 1.70e-16 1.22e-02 ...
#>  $ :'data.frame':    1 obs. of  3 variables:
#>   ..$ q_low      : num -0.169
#>   ..$ left_cens : int 1108
#>   ..$ splinecorr: int 109

p = res[[1]]$p
```

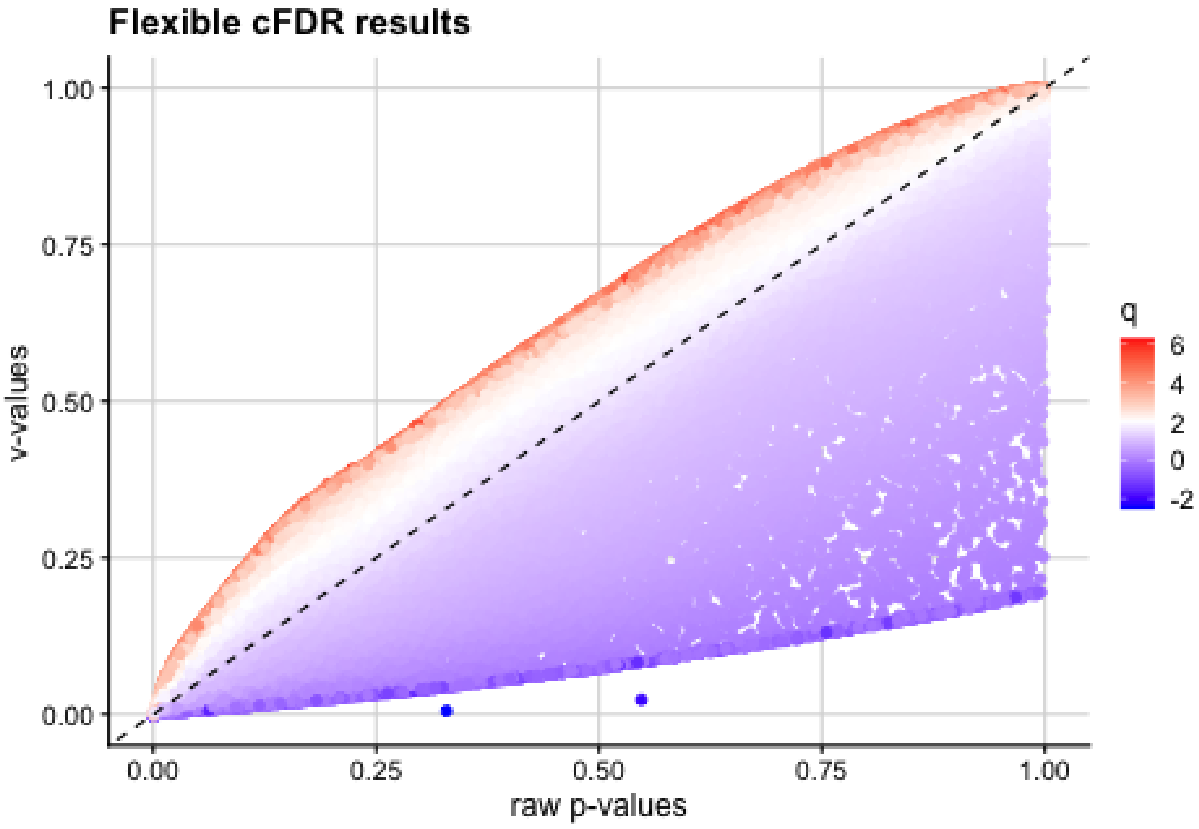
```
q = res[[1]]$q
v = res[[1]]$v
```

Note that the cFDR framework requires that low p-values are enriched for low q values, so that if the correlation between p and q is negative then the function intrinsically flips the sign of q, meaning that the q values reported in the data.frame output may be $q := -q$.

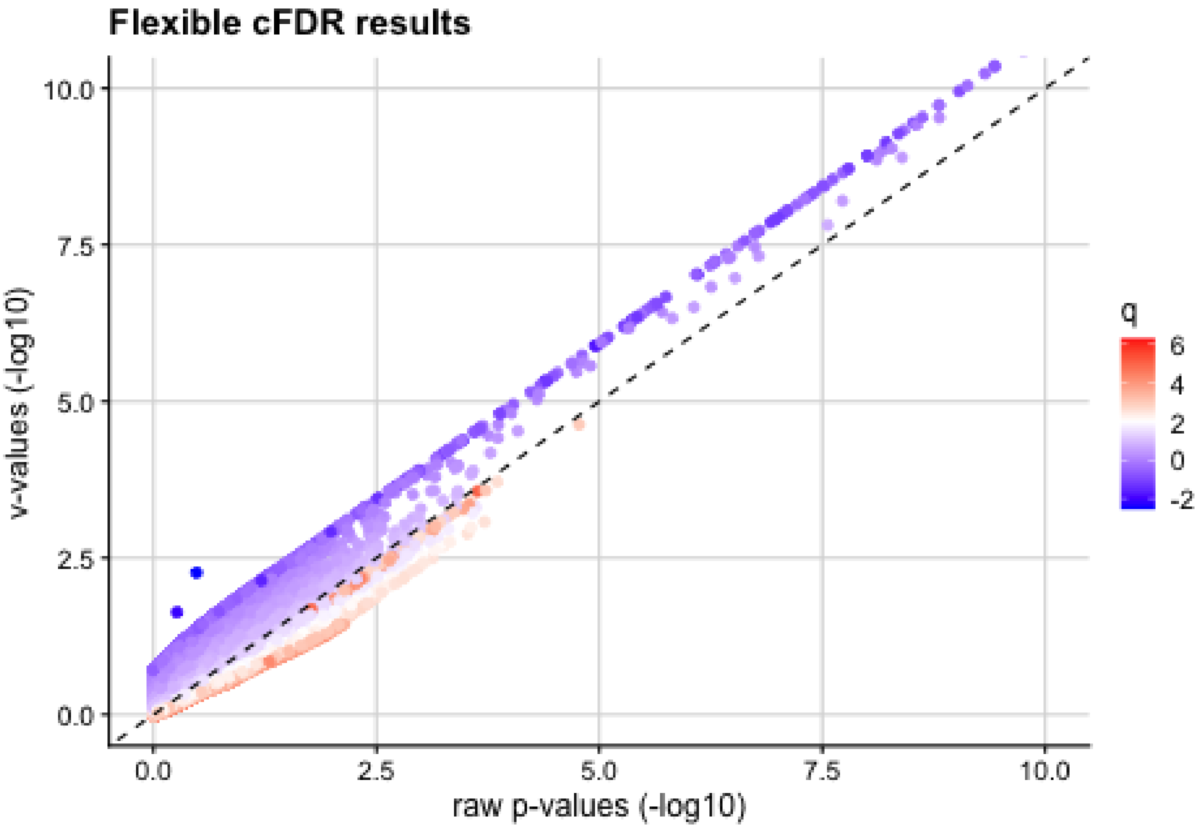
We can then visualise the results using the `pv_plot` and `log10pv_plot` functions.

We observe that v-values for SNPs with high q values are increased, whilst those for SNPs with low q values are decreased.

```
pv_plot(p = p, q = q, v = v)
```



```
log10pv_plot(p = p, q = q, v = v,
             axis_lim = c(0, 10)) # zoom in to interesting region
```



Finally, we run the Benjamini-Hochberg procedure on the v-values and control the FDR at 0.05. This means that we are willing to accept up to 5% of the associations to be false positives.

```
hit = which(p.adjust(v, method = "BH") <= 0.05)
```

For comparison, we do the same to the raw p-values:

```
hit_p = which(p.adjust(p, method = "BH") <= 0.05)
```

‘True’ associations are those with indices 1-500, so the proportions of false discoveries are

```
# cFDR
1 - (length(intersect(hit,c(1:500)))/length(hit))
#> [1] 0.04363636

# p-value
1 - (length(intersect(hit_p,c(1:500)))/length(hit_p))
#> [1] 0.02262443
```

Altogether, the cFDR method has found 47 new associations that are true whilst controlling the FDR.

```r
# number of extra true associations identified by flexible cFDR
length(which(hit[!hit %in% hit_p] <= 500))
#> [1] 47
```

A.1.2 T1D application

In this vignette, we walk through an example to illustrate how the `fcfdr` R package can be used to leverage various functional genomic data with GWAS *p*-values for type 1 diabetes (T1D) to find new genetic associations. This vignette will take about 30 minutes to complete.

The data required for this example is available to download within the `fcfdr` R package and includes:

1. GWAS *p*-values for T1D (Onengut-Gumuscu et al. 2015) downloaded from the GWAS Catalog
2. GWAS *p*-values for Rheumatoid Arthritis (RA) (Eyre et al. 2012) downloaded from the GWAS Catalog
3. Binary measure of SNP overlap with transcription factor binding site (TFBS), derived from merging all DNaseI digital genomic footprinting (DGF) regions from the narrow-peak classifications across 57 cell types (see https://www.nature.com/articles/nature11247; https://doi.org/10.1016/j.ajhg.2014.10.004). SNP annotations were downloaded for all 1000 Genomes phase 3 SNPs from the LDSC data repository and the binary `DGF_ENCODE` annotation was extracted for all T1D SNPs in our analysis.
4. Fold-enrichment ratio of H3K27ac ChIP-seq counts relative to expected background counts in naive CD4+ T helper cells (https://www.nature.com/articles/nbt1010-1045). Downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E043-H3K27ac.fc.signal.bigwig.

First, we download the data:

```
set.seed(1)
library(fcfdr)
data(T1D_df, package = "fcfdr")
head(T1D_df)
#>                     rsid chrom       pos other_allele effect_allele         p
#> 1      1kg_10_59614461    10  59944455            C             A 0.8245356
#> 2      1kg_14_80445068    14  81375315            G             A 0.3503642
#> 3      1kg_14_87503642    14  88433889            A             G 0.1048239
#> 4      1kg_2_207220639     2 207512394            G             C 0.9821215
#> 5      1kg_5_173458951     5 173526345            C             G 0.4549943
#> 6 ccc-2-102079765-A-G     2 102713333            ?             ? 0.4395651
#>   DGF_ENCODE Th_H3K27ac   RA_p ldak_weight       maf
#> 1          0    0.59652 0.1553    0.900318 0.1413714
#> 2          0    0.66973 0.6732    0.000000 0.1735724
#> 3          0    0.42419 0.5976    0.000000 0.3747790
#> 4          0    0.00000 0.8109    0.000000 0.4781095
#> 5          0    0.00000 0.7110    0.829307 0.3555777
#> 6          0    0.89836 0.3508    0.816463 0.4957808
```

In this application we leverage GWAS p-values for RA (q1), binary SNP overlap with TFBS (q2) and H3K27ac counts in naive CD4+ T helper cells (q3) with GWAS p-values for T1D (orig_p) to generate adjusted p-values (called v-values).

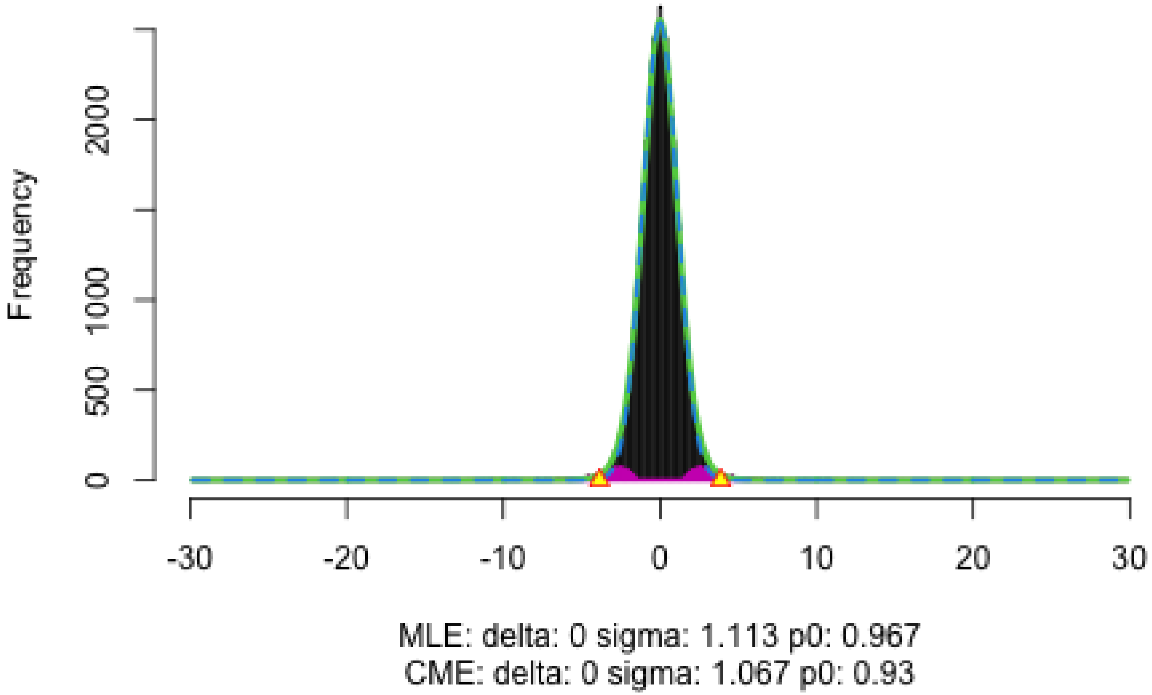
```
orig_p <- T1D_df$p
chr <- T1D_df$chrom
MAF <- T1D_df$maf
q1 <- T1D_df$RA_p
q2 <- T1D_df$DGF_ENCODE
q3 <- log(T1D_df$Th_H3K27ac+1) # deal with long tail
```

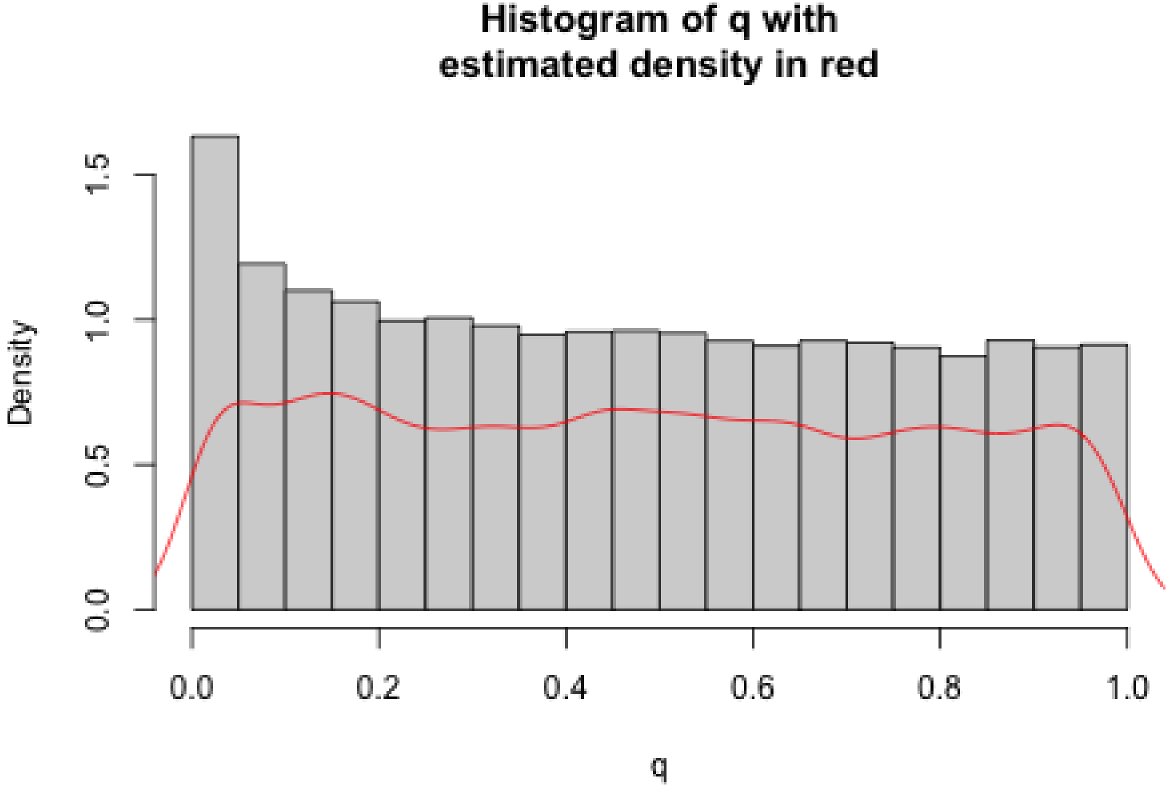
The data frame also contains a column of LDAK weights for each SNP (http://dougspeed.com/calculate-weightings/). An LDAK weight of zero means that the signal is (almost) perfectly captured by neighbouring SNPs and so we use the subset of SNPs with non-zero LDAK weights as our independent subset of SNPs.

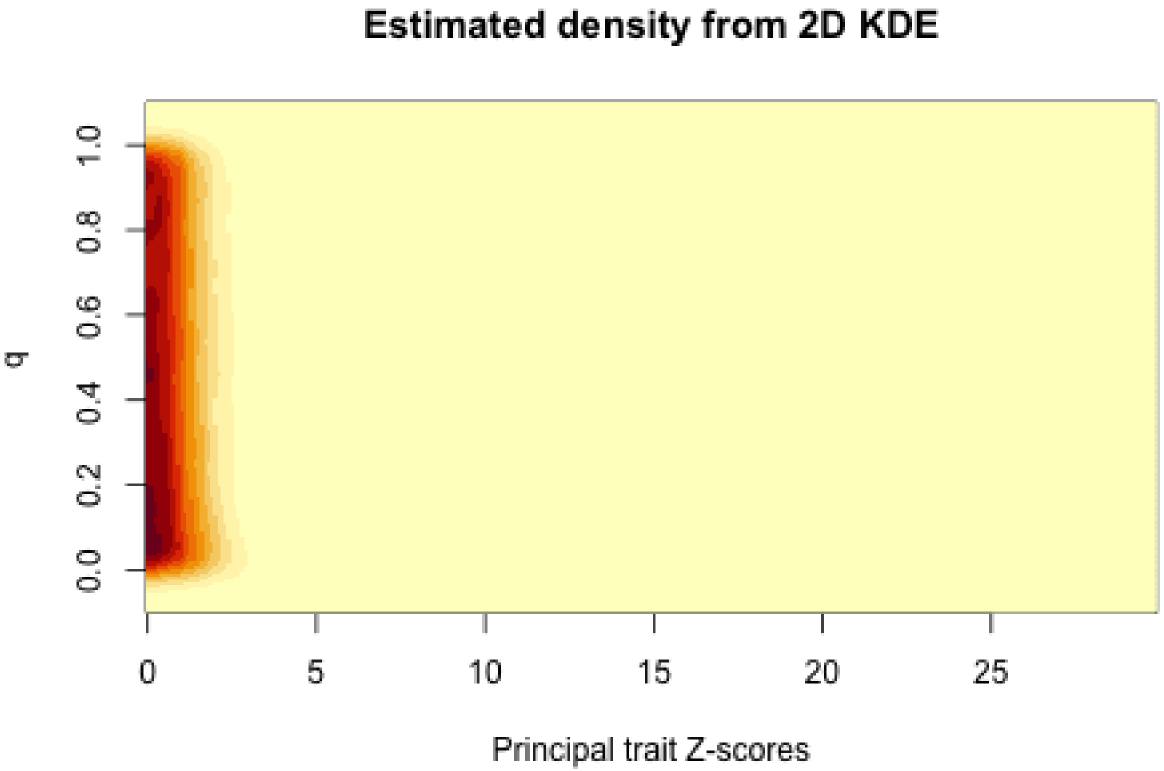
```
ind_snps <- which(T1D_df$ldak_weight != 0)
```

We are now ready to use the `fcfdr` R package to generate v-values. Firstly, we generate v-values leveraging GWAS p-values for RA. We supply MAF values to prevent a bias of the KDE fit towards the behaviour of rarer SNPs (the function intrinsically down-samples the independent subset of SNPs to match the MAF distribution in this subset to that in the whole set of SNPs).

```
iter1_res <- flexible_cfdr(p = orig_p,
                           q = q1,
                           indep_index = ind_snps,
                           maf = MAF)
#> Warning: glm.fit: fitted rates numerically 0 occurred
#> Warning from locfdr:
#> Warning: glm.fit: fitted rates numerically 0 occurred
```


Frequency
0
500
1000
2000
-30
-20
-10
0
10
20
30
MLE: delta: 0 sigma: 1.113 p0: 0.967
CME: delta: 0 sigma: 1.067 p0: 0.93


Histogram of q with
estimated density in red
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
q



Ah ha! We've encountered a warning suggesting that the estimated density from `locfdr` may be inaccurate. Let's do as suggested and examine the fit to the data. This means examining the first plot returned by the function. We can see that the fit to the data from the `locfdr::locfdr` function looks fine so we can ignore the warning in this instance. (Note that an alternative approach would be to use the `fcfdr::parameters_in_locfdr` function to extract the parameter values used intrinsically in `locfdr` and examine the effect of changing these within the `locfdr::locfdr` function).

```
v1 <- iter1_res[[1]]$v
```

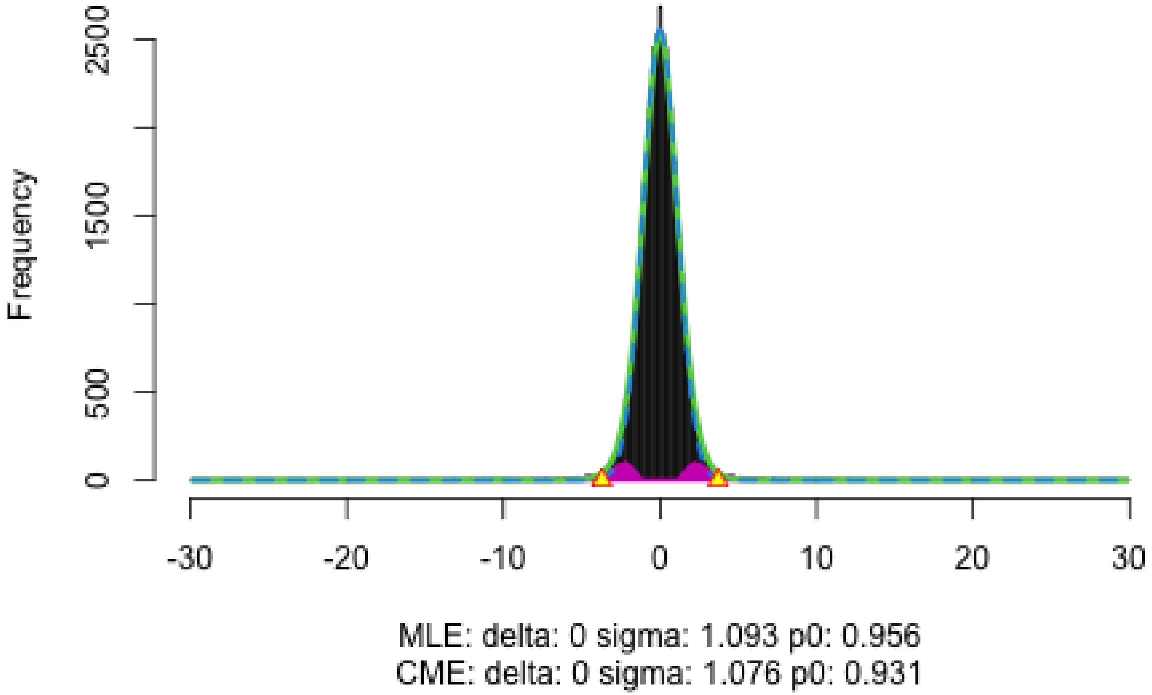
The resultant v-values for this first iteration (`v1`) are then used in the next iteration to leverage binary data on SNP overlap with TFBS. Note that the binary cFDR function implements a leave-one-out procedure and therefore requires a group index for each SNP. This will generally be the chromosome on which that SNP resides but can also be indices relating to LD blocks, for example.

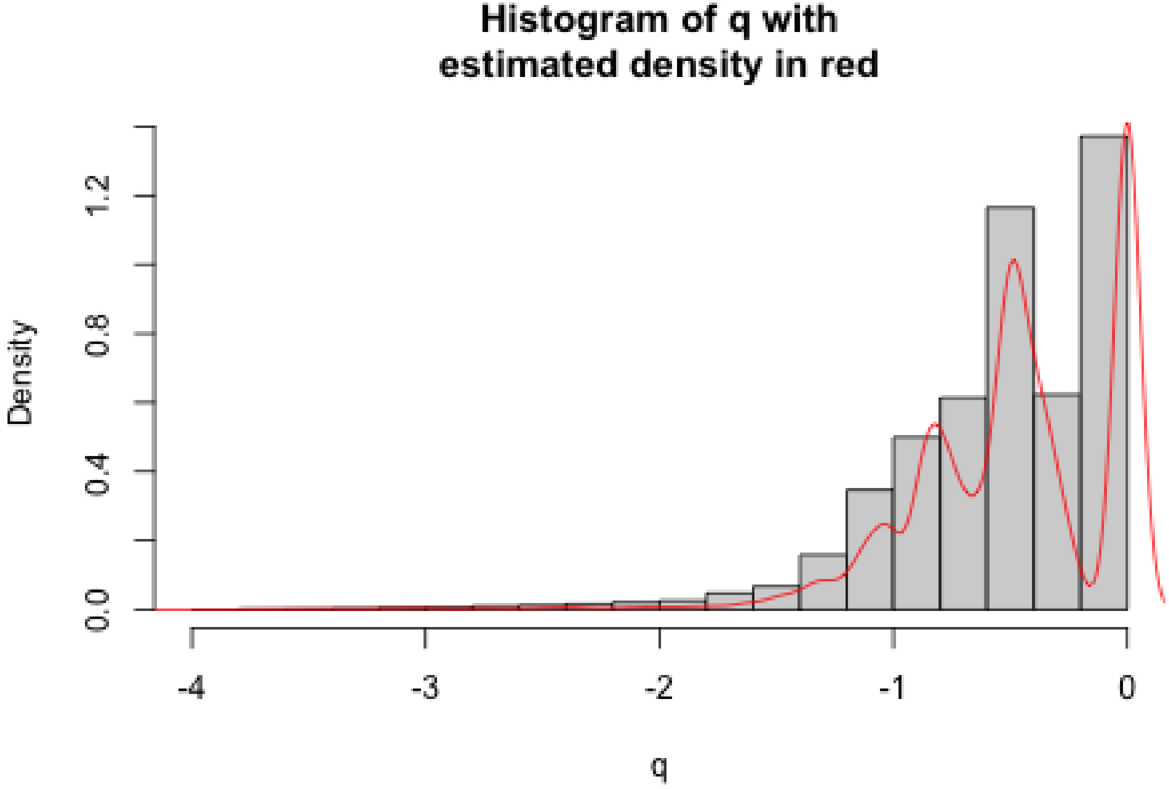
```
iter2_res <- binary_cfdr(p = v1,
                         q = q2,
                         group = chr)
```

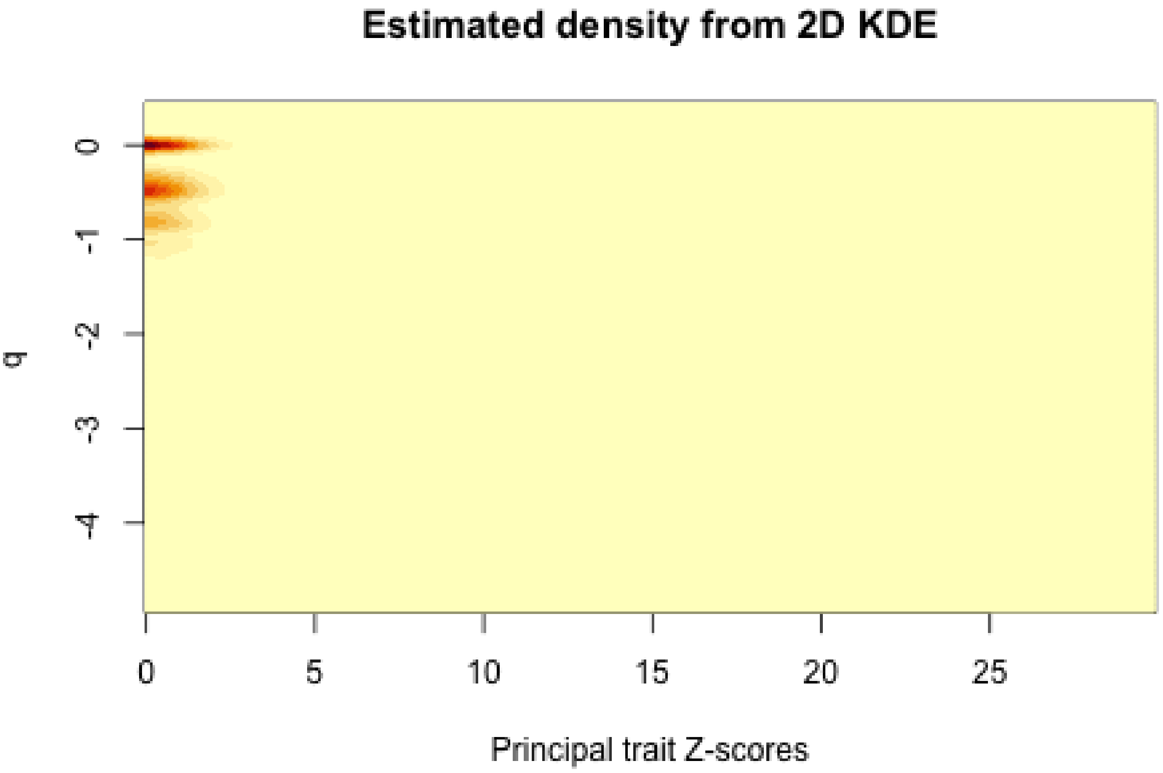
```
v2 <- iter2_res$v
```

The resultant *v*-values for this second iteration (v2) are then used in the next iteration to leverage H3K27ac counts.

```
iter3_res <- flexible_cfdr(p = v2,
                           q = q3,
                           indep_index = ind_snps,
                           maf = MAF)
```


Frequency
0
500
1500
2500
-30
-20
-10
0
10
20
30
MLE: delta: 0 sigma: 1.093 p0: 0.956
CME: delta: 0 sigma: 1.076 p0: 0.931


Histogram of q with
estimated density in red
Density
0.0
0.4
0.8
1.2
-4
-3
-2
-1
0
q



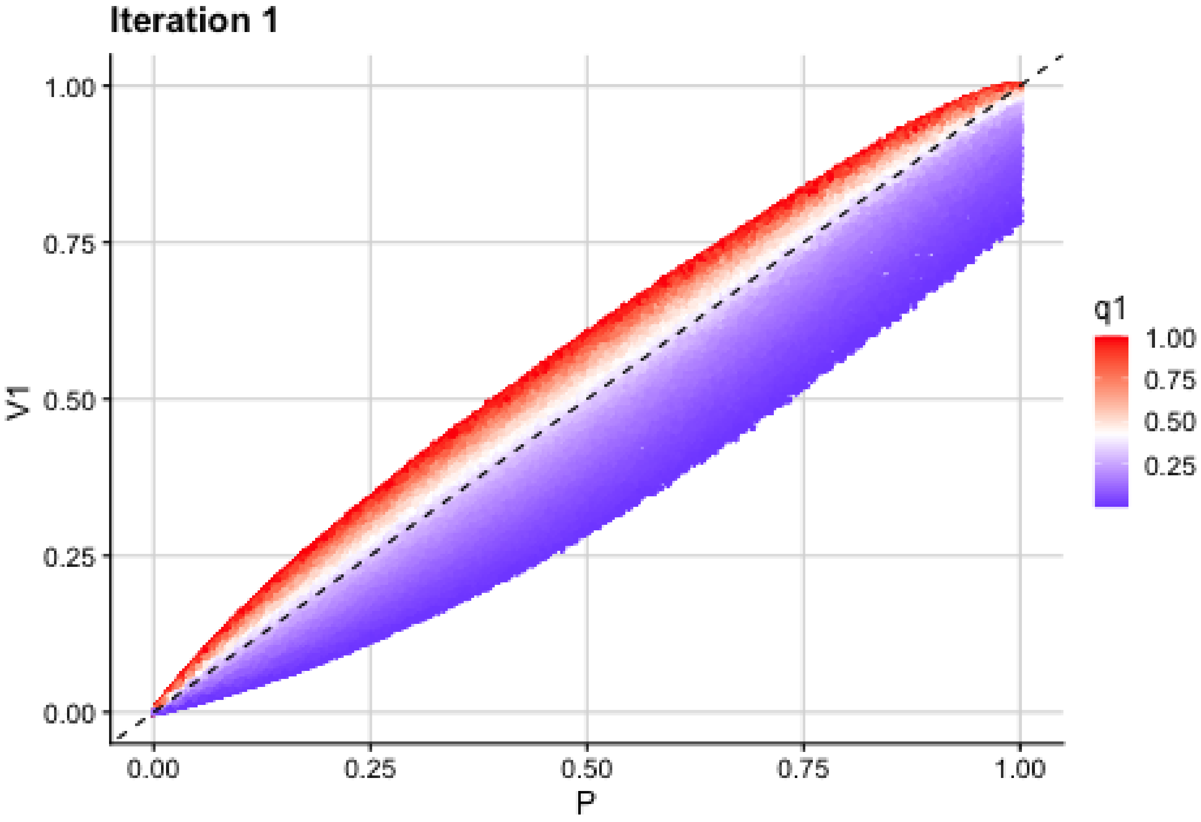
```
v3 <- iter3_res[[1]]$v
```

We then create a final data frame for the `fcfdr` results. Note that the sign is flipped for $q2$ and $q3$. This is because these are negatively correlated with `p` and the flexible cFDR software automatically flips the sign of `q` to ensure that low `p` are enriched for low `q`.

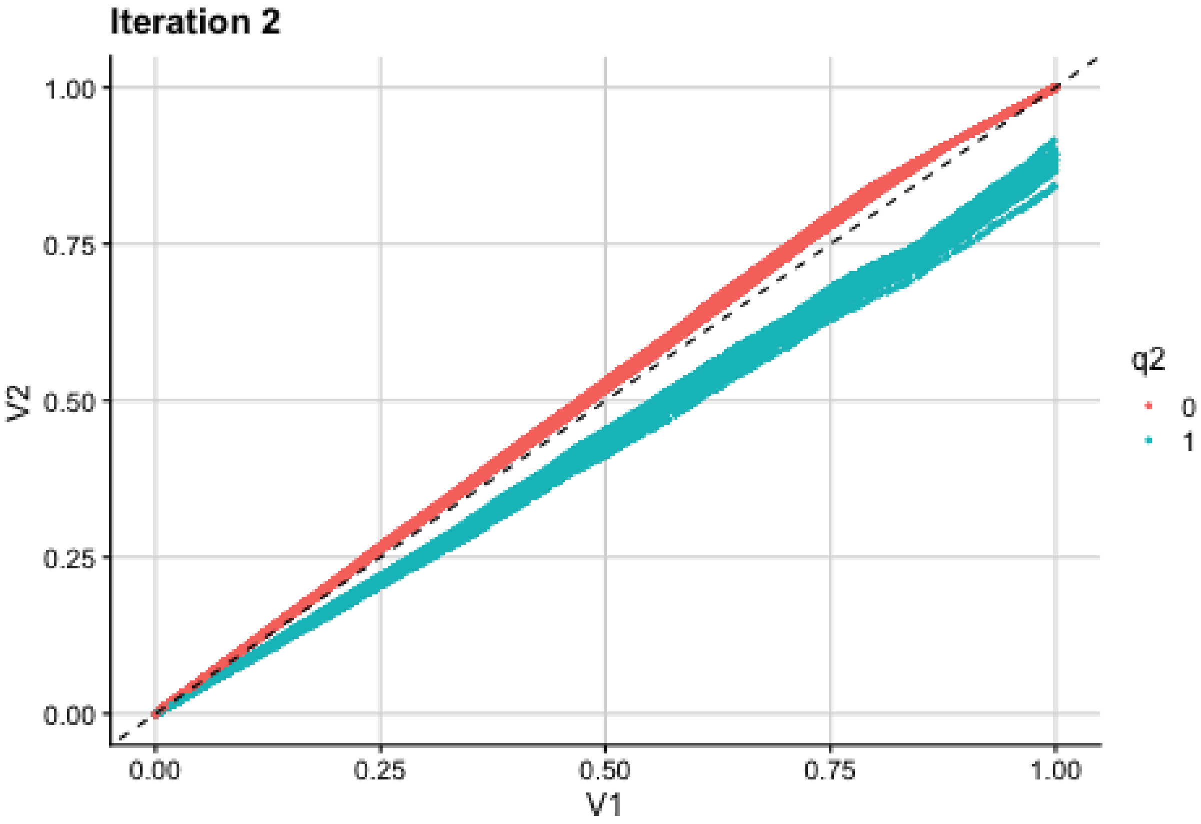
```
res <- data.frame(orig_p, q1 = iter1_res[[1]]$q,
                  q2 = as.factor(iter2_res$q),
                  q3 = iter3_res[[1]]$q,
                  v1, v2, v3)

head(res)
#>      orig_p     q1 q2         q3        v1        v2        v3
#> 1 0.8245356 0.1553  0 -0.4678263 0.7555302 0.7925922 0.8245212
#> 2 0.3503642 0.6732  0 -0.5126619 0.4173777 0.4450330 0.4760522
#> 3 0.1048239 0.5976  0 -0.3536032 0.1381179 0.1487443 0.1769102
#> 4 0.9821215 0.8109  0  0.0000000 0.9929203 0.9945451 0.9975206
#> 5 0.4549943 0.7110  0  0.0000000 0.5240799 0.5461018 0.6166161
#> 6 0.4395651 0.3508  0 -0.6409904 0.4447297 0.4671311 0.4751084
```

We inspect the resultant v-values after each iteration by plotting these against the values used as p in that iteration.

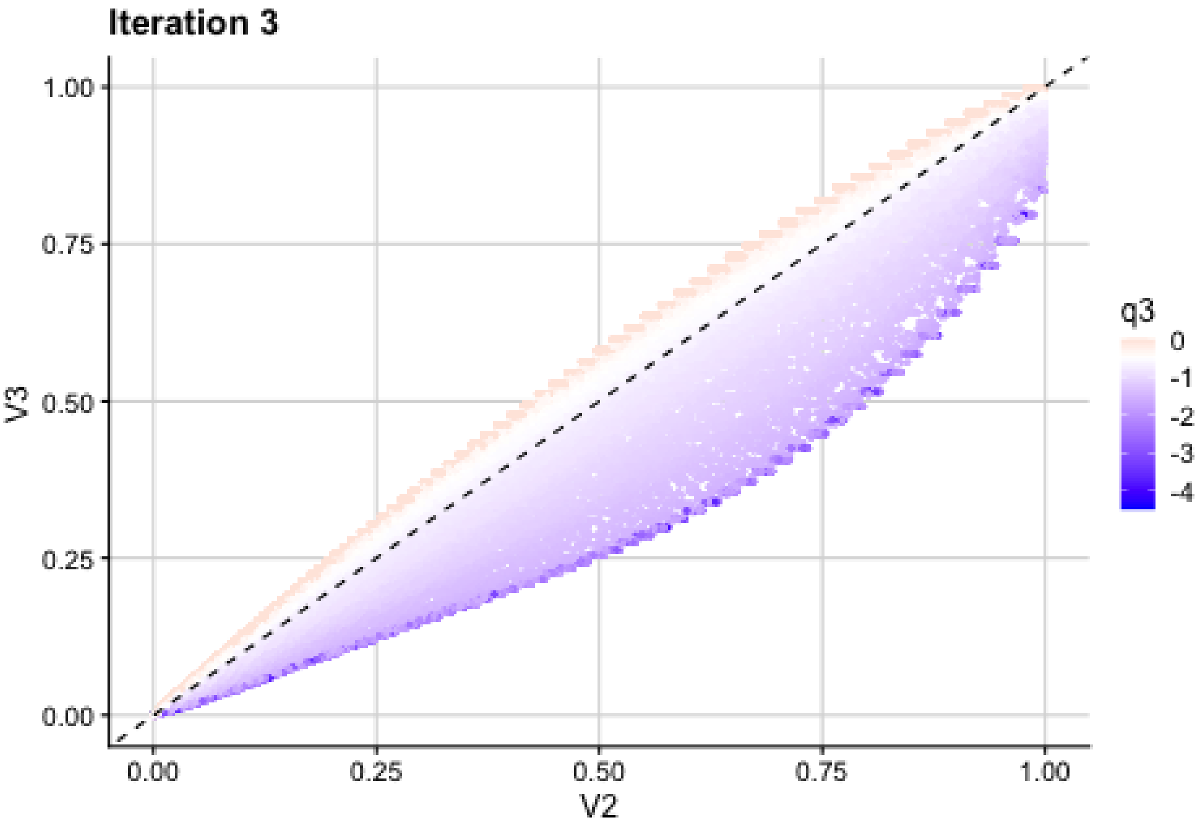
```
library(ggplot2)
library(cowplot)

mid1 <- median(res$q1)

ggplot(res, aes(x = orig_p, y = v1, col = q1)) +
  geom_point(cex = 0.5) + theme_cowplot(12) +
  background_grid(major = "xy", minor = "none") +
  geom_abline(intercept = 0, slope = 1,  linetype="dashed") +
  xlab("P") + ylab("V1") + ggtitle(paste0("Iteration 1")) +
  scale_color_gradient2(midpoint = mid1, low = "blue",
                        mid = "white", high = "red", space = "Lab")
```



```
ggplot(res, aes(x = v1, y = v2, col = q2)) +
  geom_point(cex = 0.5) + theme_cowplot(12) +
  background_grid(major = "xy", minor = "none") +
  geom_abline(intercept = 0, slope = 1,  linetype="dashed") +
  xlab("V1") + ylab("V2") + ggtitle(paste0("Iteration 2"))
```

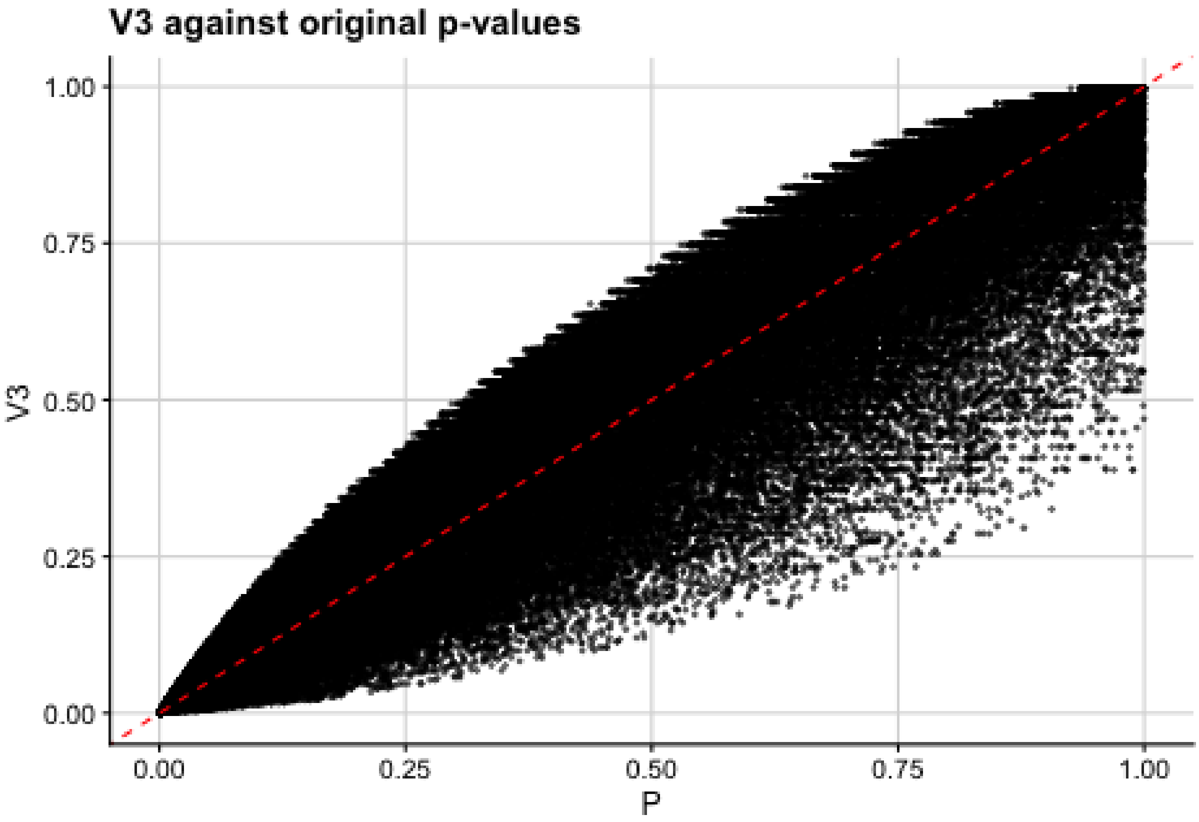


```
mid3 <- median(res$q3)

ggplot(res, aes(x = v2, y = v3, col = q3)) +
  geom_point(cex = 0.5) + theme_cowplot(12) +
  background_grid(major = "xy", minor = "none") +
  geom_abline(intercept = 0, slope = 1,  linetype="dashed") +
  xlab("V2") + ylab("V3") + ggtitle(paste0("Iteration 3")) +
  scale_color_gradient2(midpoint = mid3, low = "blue",
                        mid = "white", high = "red", space = "Lab")
```

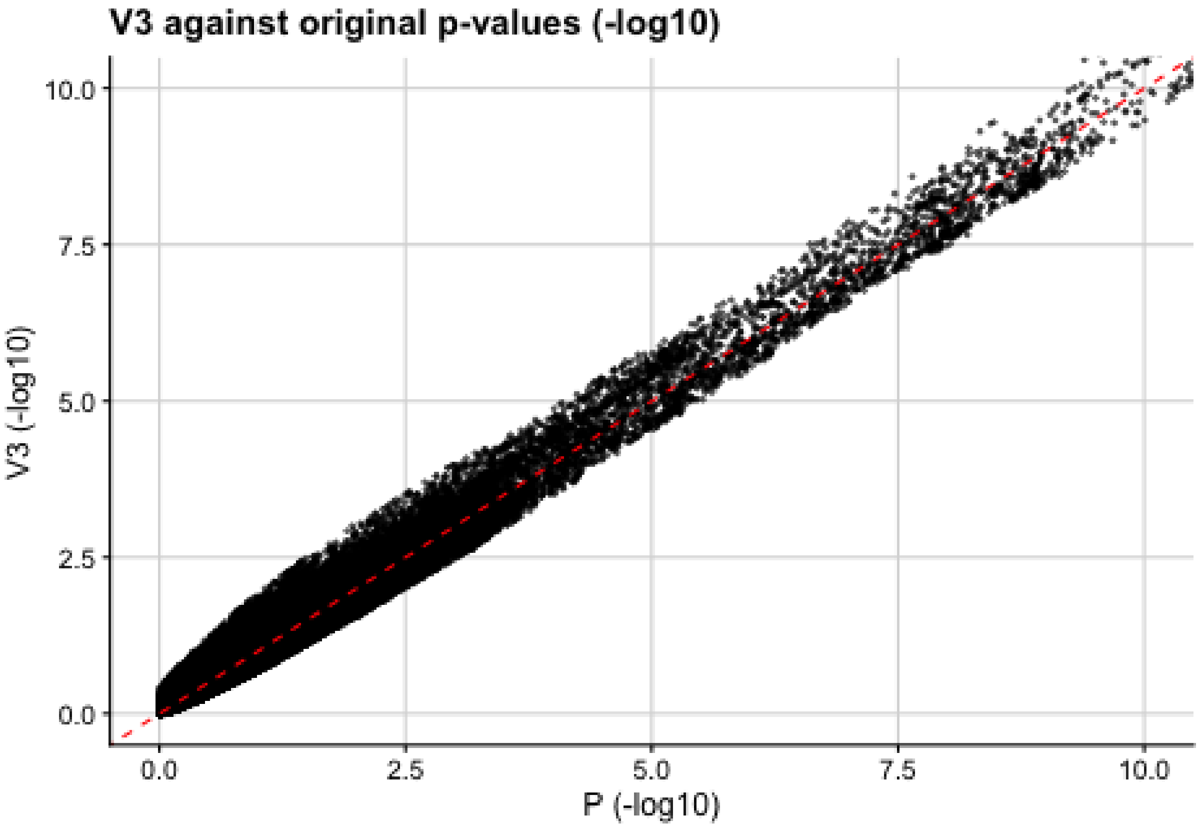


We can also plot the original *p*-values for T1D against the final adjusted *v*-values.

```
mid1 <- median(res$q1)

ggplot(res, aes(x = orig_p, y = v3)) +
  geom_point(cex = 0.5, alpha = 0.5) +
  theme_cowplot(12) + background_grid(major = "xy",
                                      minor = "none") +
  geom_abline(intercept = 0, slope = 1,  linetype="dashed",
              col = "red") + xlab("P") + ylab("V3") +
  ggtitle(paste0("V3 against original p-values"))
```



```
ggplot(res, aes(x = -log10(orig_p), y = -log10(v3))) +
  geom_point(cex = 0.5, alpha = 0.5) + theme_cowplot(12) +
  background_grid(major = "xy", minor = "none") +
  geom_abline(intercept = 0, slope = 1,  linetype="dashed",
              col = "red") + xlab("P (-log10)") +
  ylab("V3 (-log10)") +
  ggtitle(paste0("V3 against original p-values (-log10)")) +
  coord_cartesian(ylim = c(0,10), xlim = c(0,10))
```



Since the outputted v-values are analogous to p-values, they can be used directly in any error-rate controlling procedure. Here, we use the BH method to derive FDR-adjusted v-values and find that our implementation of cFDR identifies newly FDR significant SNPs that have relatively small GWAS p-values for rheumatoid arthritis, are more likely to be found in genomic regions where transcription factors may bind and have relatively high H3K27ac counts in a T1D relevant cell type.

```
fdr_thr <- 5*10^-6
p_fdr <- p.adjust(orig_p, method = "BH")
v3_fdr <- p.adjust(v3, method = "BH")

length(which(v3_fdr <= fdr_thr & p_fdr > fdr_thr))
#> [1] 59

median(T1D_df$RA_p[which(v3_fdr < fdr_thr & p_fdr > fdr_thr)])
#> [1] 0.001625
median(T1D_df$RA_p)
#> [1] 0.4228
```

```r
mean(T1D_df$DGF_ENCODE[which(v3_fdr < fdr_thr & p_fdr > fdr_thr)])
#> [1] 0.3389831
mean(T1D_df$DGF_ENCODE)
#> [1] 0.2347241

median(T1D_df$Th_H3K27ac[which(v3_fdr < fdr_thr & p_fdr > fdr_thr)])
#> [1] 1.73539
median(T1D_df$Th_H3K27ac)
#> [1] 0.63626
```

Appendix B

Appendix to chapter 3

B.1 Auxiliary simulation results

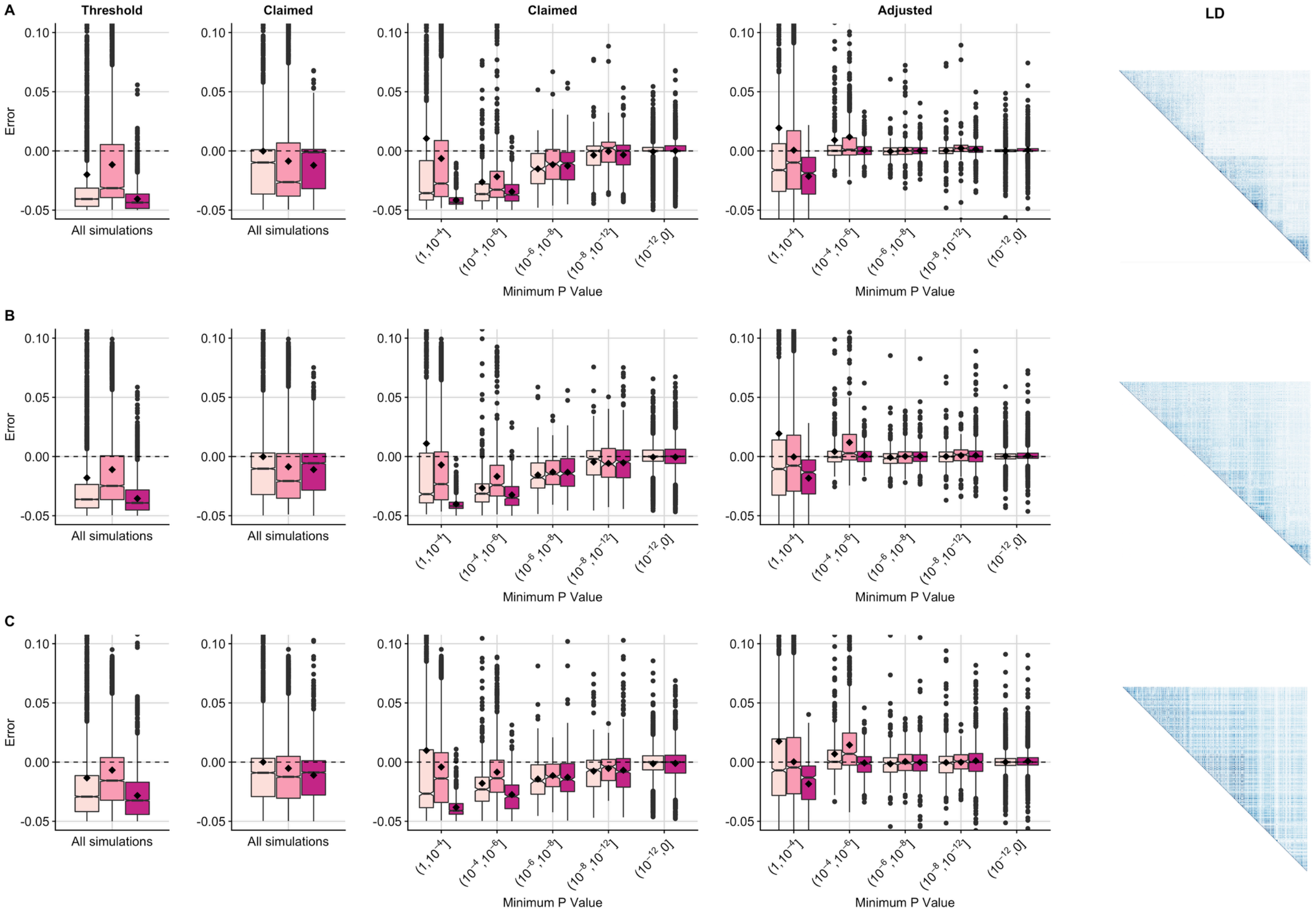


Fig. B.1 Error of conditional coverage estimates for 95% credible sets. Error is calculated as estimated conditional coverage − empirical conditional coverage, where empirical conditional coverage is the proportion of 5000 replicate credible sets that contain the causal variant. Box plots showing error in conditional coverage estimates for 5000 (A) low (B) medium and (C) high LD simulations. Conditional coverage estimates are the threshold (0.95) (left), the claimed coverage (the sum of the posterior probabilities of the variants in the credible set) averaged over all simulations (left-middle) or for simulations binned by minimum p-value in the region (right-middle) and the adjusted coverage estimate (right) binned by minimum p-value in the region. Black diamond shows mean error. Graphical display of SNP correlation matrix for each region shown, which was generated using the `corrplot` R package. The colour legend is as in Fig. 3.3, Fig. 3.4 and Fig. 3.6.

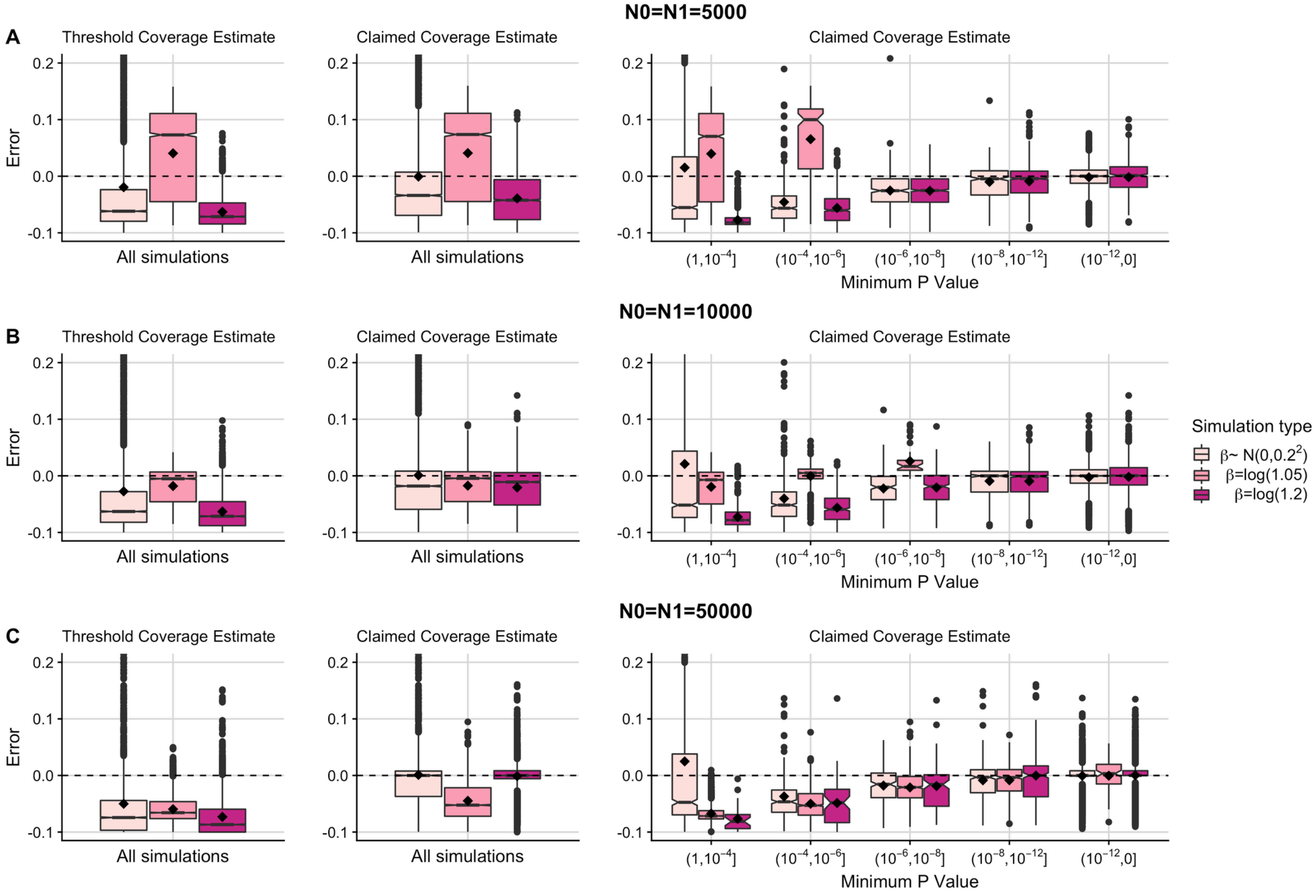


Fig. B.2 Error of conditional coverage estimates for 90% credible sets split up by sample size. Error is calculated as estimated conditional coverage − empirical conditional coverage, where empirical conditional coverage is the proportion of 5000 replicate credible sets that contain the causal variant. Box plots showing error in conditional coverage estimates for 5000 simulations with $N0$ (number of controls) $= N1$ (number of cases) $=$ (A) 5000 (B) 10000 and (C) 50000. Overall error in (left) threshold and (middle) claimed coverage estimates averaged across all 5000 simulations. Right hand plots show error in claimed coverage estimates for different p-value bins.

B.2 Selecting the prior causal variance for the adjusted coverage estimate

My adjustment method is influenced by the choice of the prior variance at the causal variant (W) through the calculation of the ABFs for each variant, which are proportional to the posterior probabilities of causality (PPs). Recall that the ABF for SNP i is the relative likelihood of a Z-score given the alternative hypothesis, $Z_i \sim N(0, \frac{V_i+W}{V_i})$, and the null hypothesis, $Z_i \sim N(0,1)$, where V_i is the variance of the estimated effect size at SNP i. So that,

$$ABF_i = \frac{P(Z_i|Z_i \sim N(0, \frac{V_i+W}{V_i}))}{P(Z_i|Z_i \sim N(0,1))}. \tag{B.1}$$

Low values of W (e.g. $W \leq 0.1^2$) decrease the variance of the Z-scores under the alternative hypothesis towards 1. This means that the distribution of Z-scores under the alternative hypothesis is similar to that under the null hypothesis. Higher values of W (e.g. $W \geq 0.2^2$) increase the variance of the Z-scores under the alternative hypothesis, meaning that there is now more evidence of low Z-scores being sampled from the null distribution compared to the alternative distribution, and resulting in the corresponding ABFs decreasing towards 0.

I investigated the empirical effect that changing W had on my adjustment method by calculating the ABFs for variants in a single GWAS simulation using different values of W, ranging from $W = 0.05^2$ to $W = 0.5^2$, and using the logarithm of the ABFs for visualisation.

For smaller absolute Z-scores, there was more variability in the corresponding ABFs for various values of W, whereby the ABF was smaller for larger values of W, reflecting more evidence for the null (Fig. B.2A). This difference was less pronounced for higher Z-scores as there was more evidence for the alternative hypothesis. Moreover, the median ABF value decreased as W increased (Fig. B.2B), further supporting the idea that larger values of W result in more evidence that smaller Z-scores are sampled from the null hypothesis.

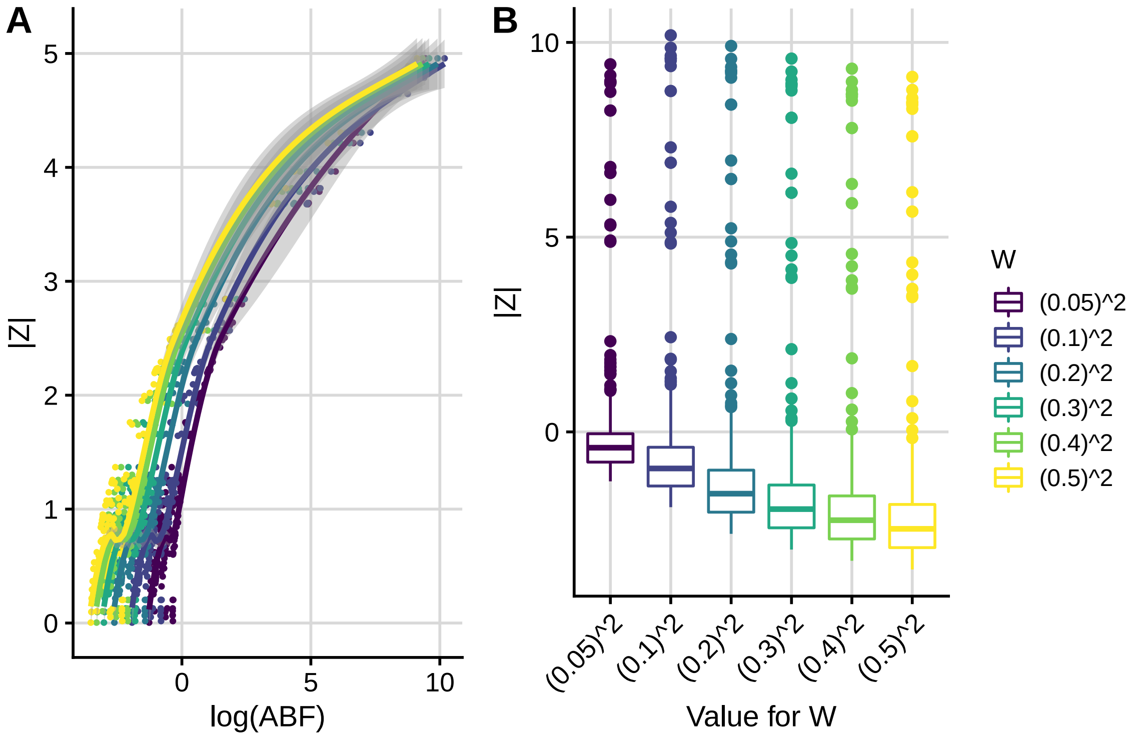


Fig. B.3 (A) Absolute Z-score plotted against log(ABF) coloured by the value of W used in the calculation of the ABF. (B) Boxplots of log(ABF) for different values of W.

I hypothesised that reducing W would improve the μ estimate in lower powered scenarios, specifically those with improbably extreme maximum Z-scores, as these would be treated more sceptically. I found that reducing W from 0.2^2 to 0.1^2 made little difference to the accuracy of the μ estimate (Fig. B.4).

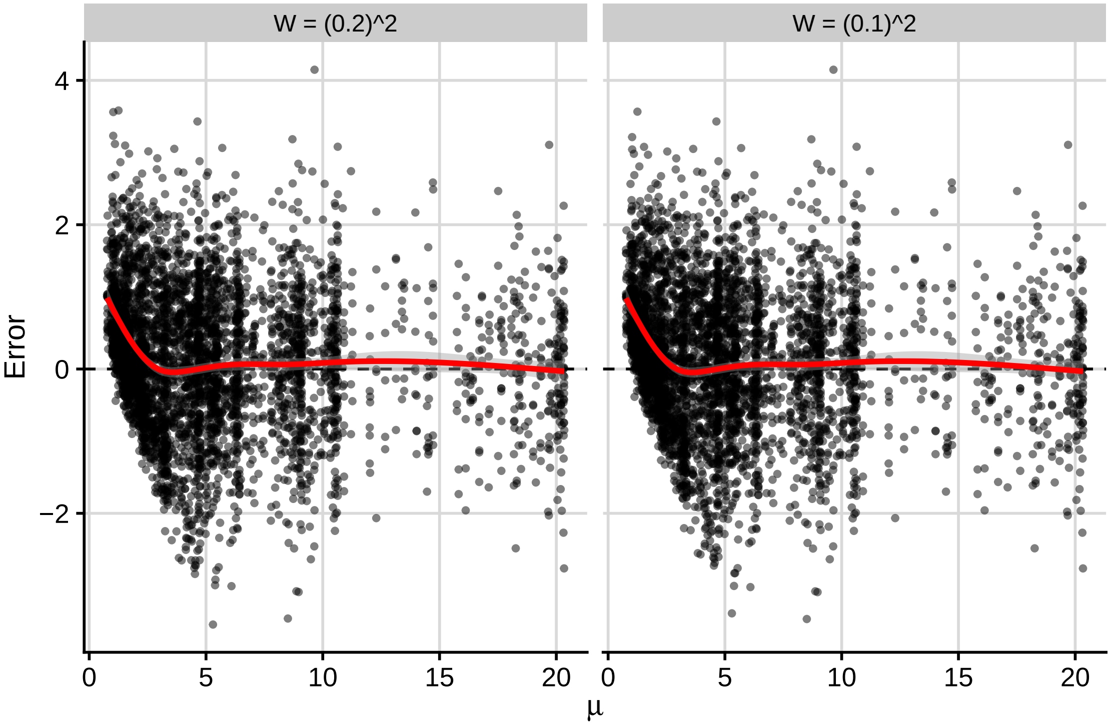


Fig. B.4 Error of μ estimates calculated as $\hat{\mu}_X - \mu$ where $\hat{\mu} = \sum_{i=1}^{p} |Z_i| \times PP_i$. The x-axis is the joint Z score at the causal variant. Line is fitted using a GAM as the smoothing function (`ggplot2::geom_smooth`). Faceted by the value of W used to calculate the ABFs in the derivation of the posterior probabilities (PP in $\hat{\mu}$).

Since the value for W did not affect the estimate of μ, I hypothesised that the value for W would not impact the adjusted coverage estimates. Accordingly, I found that the value used for W had minimal impact on the adjusted coverage estimate (Fig. B.5).

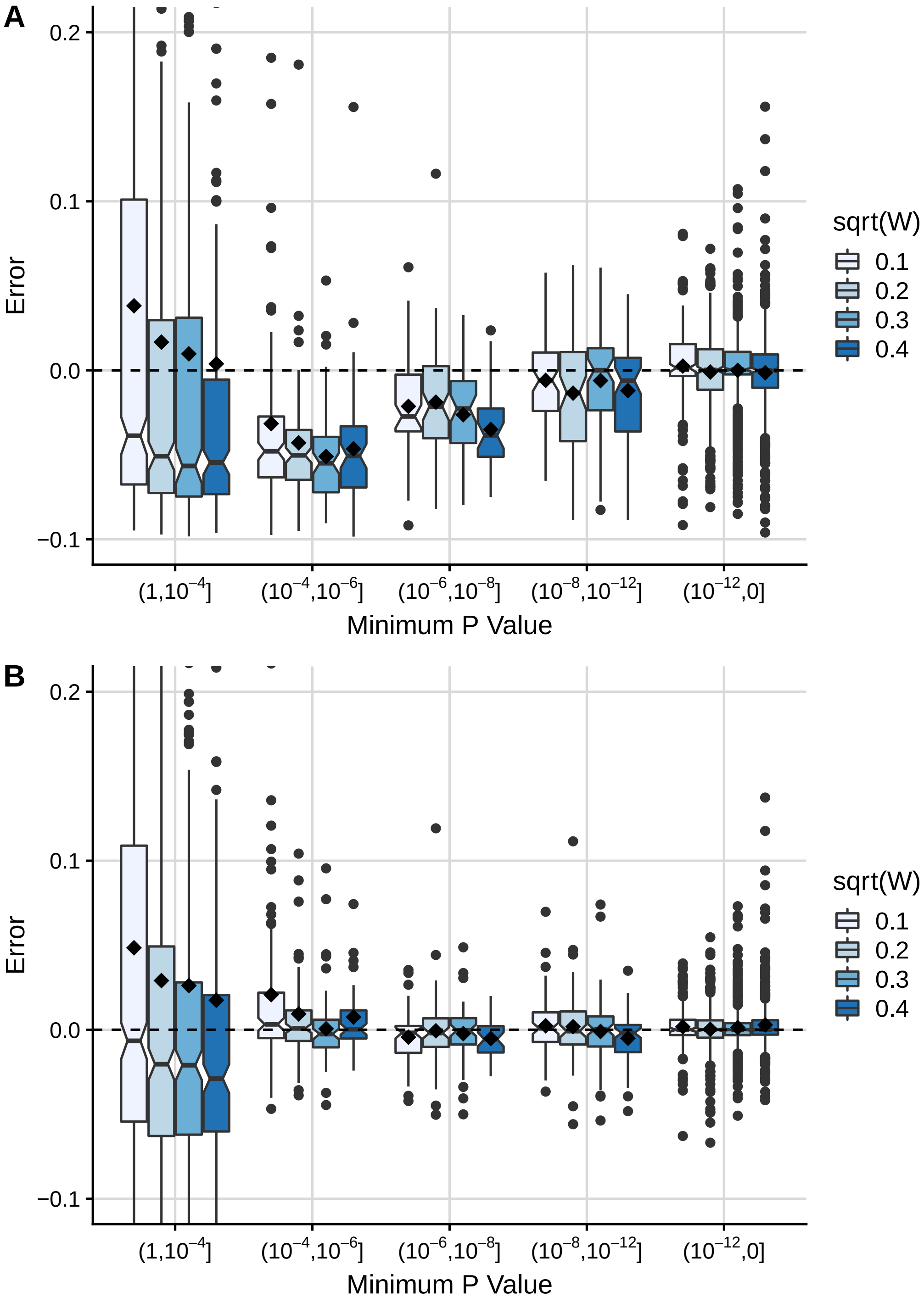


Fig. B.5 Error is calculated as estimated conditional coverage–empirical conditional coverage, where empirical conditional coverage is the proportion of 5000 replicate credible sets that contain the causal variant. Box plots showing error in conditional coverage estimates for 5000 medium LD simulations. Conditional coverage estimates are (A) the claimed coverage and (B) the adjusted coverage estimate for simulations binned by minimum p-value in the region. Black diamond shows mean error.

To conclude, higher values for W provide more evidence of low Z-scores coming from the null hypothesis (shrinking their ABFs and therefore their posterior probabilities). Higher values for W had little effect on higher Z-scores and varying W had little effect on estimating μ. Since my focus in this instance was for genetic association studies, where effect sizes are usually large, I chose to set $W = 0.2^2$ in the adjusted coverage estimate method.

B.3 corrcoverage vignettes

B.3.1 Corrected coverage vignette

This guide will show users how the `corrcoverage` package can be used for statistical single causal variant fine-mapping, including obtaining an accurate coverage estimate for the causal variant in a credible set (corrected coverage estimate). This package is specific to credible sets obtained using the Bayesian approach to fine-mapping, described by Maller et al. here, which utilities approximate Bayes factors for genetic association studies described by Wakefield here.

The `corrcoverage` R package requires only GWAS summary statistics to provide users with a corrected coverage estimate that the true causal variant is contained within the credible set, for example calculating that a 90% credible set actually has 99% probability of containing the true causal variant. In this vignette, we walk-through how the `corrcoverage` R package can be used to perform single causal variant Bayesian fine-mapping and to derive a corrected coverage estimate.

Firstly, load the library.

```
set.seed(2)
library(corrcoverage)
```

---

1. Simulate GWAS summary statistics

This package is intended for use on summary statistics obtained from GWAS, such as observed Z-scores. For the purpose of this vignette, we will simulate artificial haplotypes and GWAS data using the simGWAS package. Please refer to the walkthrough guide here from which the following is taken.

```r
library(simGWAS)
```

```r
# simulate reference haplotypes

nsnps <- 200
nhaps <- 1000
lag <- 6 # genotypes are correlated between neighbouring variants
maf <- runif(nsnps+lag,0.05,0.5) # common SNPs

laghaps <- do.call("cbind", lapply(maf, function(f) rbinom(nhaps,1,f)))
haps <- laghaps[,1:nsnps]
for(j in 1:lag)
    haps <- haps + laghaps[,(1:nsnps)+j]
haps <- round(haps/matrix(apply(haps,2,max),nhaps,nsnps,byrow=TRUE))
snps <- colnames(haps) <- paste0("s",1:nsnps)
freq <- as.data.frame(haps+1)
freq$Probability <- 1/nrow(freq)
sum(freq$Probability)
```

```
## [1] 1
```

```r
MAF <- colMeans(freq[,snps]-1)

# SNP correlation matrix

LD <- cor2(haps)
```

We specify the causal variant (CV) and it’s effect on disease, as an odds ratio. One causal variant is chosen since the Bayesian approach to fine-mapping relies on the assumption of one causal variant per region, which has been typed in the study.

```r
CV <- sample(snps[which(colMeans(haps)>0.1)],1)
iCV <- sub("s", "", CV) # index of cv
OR <- 1.1
```

Then, we simulate marginal Z-scores. Here, we consider a (relatively small) GWAS with 5000 cases and 5000 controls.

```r
z0 <- simulated_z_score(N0=5000, # number of controls
              N1=5000, # number of cases
```

```
                  snps=snps, # column names in freq of SNPs
                  W=CV, # causal variants, subset of snps
                  gamma.W=log(OR), # log odds ratios
                  freq=freq # reference haplotypes
                  )
length(z0)
```

```
## [1] 200
```

```
z0[1:5]
```

```
## [1] -1.889831 -2.534196 -3.139531 -3.753833 -1.821278
```

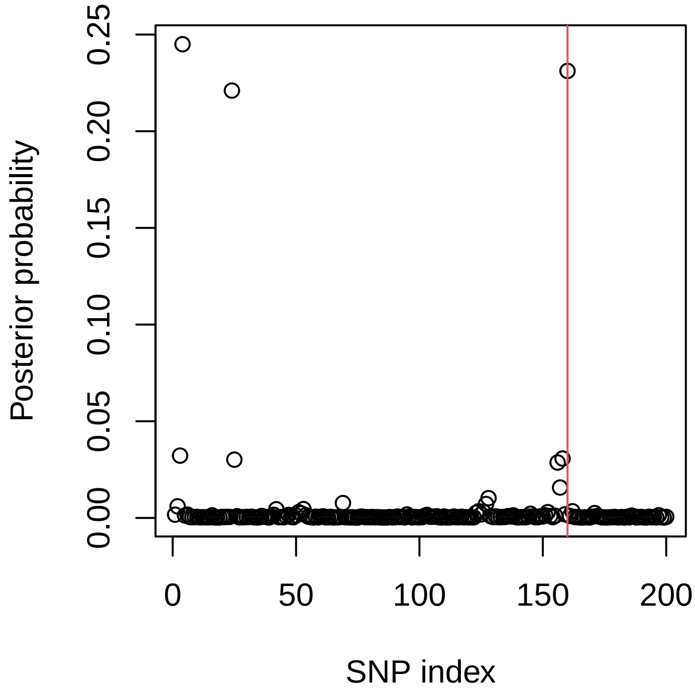
`z0` is a vector of marginal Z scores for the SNPs in our genomic region.

---

2. Convert Z scores to posterior probabilities

The `ppfunc` function is used to convert the marginal Z-scores to posterior probabilities of causality. The `ppfunc` function also requires a value for V, the variance of the estimated effect size, which can be calculated using the `Var.data.cc` function. The prior standard deviation of the effect size, W, is an optional parameter with a default value of 0.2, which is shown to be a robust choice through our analyses.

```
N0 <- 5000 # number of controls
N1 <- 5000 # number of cases

varbeta <- Var.data.cc(f = MAF, N = N1+N0, s = N1/(N0+N1)) # variance of
                                                           # estimated effect
                                                           # size

postprobs <- ppfunc(z = z0, V = varbeta)

plot(postprobs, xlab = "SNP index", ylab = "Posterior probability")
abline(v = iCV, col = 2)
```



The posterior probability plot shows the location of the causal variant (red line).

---

3. Derive credible set

The `credset` function in the `corrcoverage` package is used to obtain a 90% credible set using the Bayesian approach for fine-mapping. In brief, the function follows the standard steps from Maller et al.'s approach:

1. Order variants into descending order of posterior probabilities (PPs).
2. Cumulatively sum the PPs for these variants until the specified threshold (90%) is exceeded.
3. Group these variants to form the credible set.

The `credset` function reports the variants in the credible set, the claimed coverage[1] (the sum of the posterior probabilities of the variants in the set) and the number of variants in the credible set (nvar). If the `CV` parameter is supplied by the user (e.g. in simulation studies where the CV is known), then the output also includes a binary indicator of whether the causal variant was contained in the credible set.

[1] Researchers commonly use the size of the credible set as an estimate of the coverage probability of the causal variant in the set. It is believed to be a slightly better indicator of coverage than the threshold (claimed coverage > threshold by virtue of the method).

Let's use the `credset` function to obtain a 90% credible set.

```
thresh <- 0.9

credibleset <- credset(pp = postprobs, thr = thresh)

credibleset
```

```
## $credset
##  [1]   4 160  24   3 158  25 156 157 128  69 127   2  53  42 124 162 126 152
## [19]  51 123 171  52 145
##
## $claimed.cov
## [1] 0.9007093
##
## $nvar
## [1] 23
```

```
# is the true CV contained within the credible set?

thresh <- 0.9

credibleset <- credset(pp = postprobs, CV = iCV, thr = thresh)

credibleset
```

```
## $credset
##  [1]   4 160  24   3 158  25 156 157 128  69 127   2  53  42 124 162 126 152
## [19]  51 123 171  52 145
##
## $claimed.cov
## [1] 0.9007093
##
## $covered
## [1] 1
##
## $nvar
## [1] 23
```

The 90% credible set has a claimed coverage of ~90% and contains the true causal variant. In the literature, authors would typically report that they have found a 90% credible set which they are at least 90% confident contains the true causal variant.

Note that if the variants are named in the `pp` vector, then the `$credset` output would contain variant names rather than their indices.

---

4. Obtain corrected coverage estimate

Suppose we are suspicious of this coverage estimate (90% in our example) and wish to find a more accurate estimate using this fancy new package. The `corrcov` function can be used, which requires the following parameters to be specified:

- `z` (vector of marginal Z-scores)
- `f` (vector of minor allele frequencies)
- `N0`, `N1` (number of controls and cases respectively)
- `Sigma` (SNP correlation matrix)
- `thr` (threshold used to derive the credible set)

Note that if the estimated effect size coefficients (`bhat`) and their standard errors (`sqrt(V)`) are known instead of Z-scores and minor allele frequencies, then the `corrcov_bhat` function can be used analogously.

```
corrected_cov_estimate <- corrcov(z = z0, f = MAF, N0, N1, Sigma = LD,
                                  thr = thresh)
```

```
##  Corrected.Coverage Claimed.Coverage Threshold
##           0.9832236        0.9007093       0.9
```

In our example, the 90% credible set may have nearer to 98% coverage of the causal variant in the credible set.

---

5. Evaluate accuracy of estimate

For the purpose of this vignette and to assess the accuracy of this corrected coverage estimate, we simulate more credible sets from the same system and calculate what proportion of these contain the true causal variant.

```
z0.tmp <- simulated_z_score(N0=5000, # number of controls
                            N1=5000, # number of cases
                            snps=snps, # column names in freq
                            W=CV, # causal variants, subset of snps
                            gamma.W=log(OR), # log odds ratios
                            freq=freq, # reference haplotypes
                            nrep = 5000 # 5000 simulations
)

pps <- ppfunc.mat(zstar = z0.tmp, V = varbeta) # find pps
cs <- apply(pps, 1, function(x) credset(x, CV = iCV, thr = thresh)$cov)
true.cov.est <- mean(cs)
true.cov.est
```

```
## [1] 0.981
```

The estimated empirical coverage is found to be approximately 98%, showing that our corrected coverage estimate was indeed accurate - far more so than the standard claimed coverage of ~90%.

```
##   Empirical.Coverage Corrected.Coverage Claimed.Coverage Threshold
##                0.981          0.9832236        0.9007093       0.9
```

So what?

This vignette has shown readers how to use the `corrcoverage` R package to obtain a more accurate coverage estimate of the causal variant in a credible set. Obtaining accurate coverage estimates will allow researchers to report more specific findings from their fine-mapping analysis, and we hope that our correction will be used as an extra step in the single causal variant fine-mapping pipeline. Finally, reporting these corrected coverage estimates will allow for more efficient allocation and expenditure of resources in the laborious follow-up wet lab analyses of the variants in the credible set.

This method can be extended to find a corrected credible set, please see https://annahutch.github.io/corrcoverage/articles/New-Credible-Set.html.

B.3.2 Generating a new credible set

This vignette will show users how the `corrcoverage` R package can be used to obtain a new credible set of variants that contains the true causal variant with some specified desired coverage value whilst containing as few variants as possible.

---

1. Simulate GWAS summary statistics

As in the corrected coverage vignette, let's begin by simulating some GWAS data using the `simGWAS` package.

```r
set.seed(18)
library(corrcoverage)
library(simGWAS)

#  Simulate reference haplotypes
nsnps <- 200
nhaps <- 1000
lag <- 30  # genotypes are correlated between neighbouring variants
maf <- runif(nsnps + lag, 0.05, 0.5)  # common SNPs
laghaps <- do.call("cbind", lapply(maf, function(f) rbinom(nhaps, 1, f)))
haps <- laghaps[, 1:nsnps]
for (j in 1:lag) haps <- haps + laghaps[, (1:nsnps) + j]
haps <- round(haps/matrix(apply(haps, 2, max), nhaps, nsnps, byrow = TRUE))
snps <- colnames(haps) <- paste0("s", 1:nsnps)
freq <- as.data.frame(haps + 1)
freq$Probability <- 1/nrow(freq)
sum(freq$Probability)
```

```
## [1] 1
```

```r
MAF <- colMeans(freq[, snps] - 1)  # minor allele frequencies
CV <- sample(snps[which(colMeans(haps) > 0.1)], 1)
iCV <- sub("s", "", CV)  # index of cv
LD <- cor2(haps) # correlation between SNPs
```

```
OR <- 1.1 # odds ratios
N0 <- 10000 # number of controls
N1 <- 10000 # number of cases

z0 <- simulated_z_score(N0 = N0, # number of controls
                        N1 = N1, # number of cases
                        snps = snps, # column names in freq
                        W = CV, # causal variants, subset of snps
                        gamma.W = log(OR), # log odds ratios
                        freq = freq) # reference haplotypes
```

To calculate V, the prior variance for the estimated effect size, we use `Var.data.cc`.

```
# variance of estimated effect size
varbeta <- Var.data.cc(f = MAF, N = N1+N0, s = N1/(N0+N1))
```

We can then use the `ppfunc` function to calculate the posterior probabilities of causality for each variant.

```
postprobs <- ppfunc(z = z0, V = varbeta)
```

We use the `est_mu` function to obtain an estimate of the true effect at the causal variant.

```
muhat <- est_mu(z0, MAF, N0, N1)
muhat
```

```
## [1] 4.970273
```

2. Derive corrected coverage estimate

The `corrected_cov` function is used to find the corrected coverage of a credible set with specified threshold, say 0.9.

Note that this function is similar to using `corrcov` as explained in the "Corrected Coverage" vignette; which would require Z-scores, minor allele frequencies and sample sizes. Here, we have already calculated some of the intermediaries calculated in the `corrcov` function (muhat, varbeta etc.) so we can use `corrected_cov` instead.

```
thr = 0.9
corrcov <- corrected_cov(pp0 = postprobs, mu = muhat, V = varbeta,
                         Sigma = LD, thr = thr, nrep = 1000)
cs <- credset(pp = postprobs, thr = thr)
data.frame(claimed.cov = cs$claimed.cov, corr.cov =  corrcov, nvar = cs$nvar)
```

```
##   claimed.cov corr.cov nvar
## 1   0.9307395 0.967282    9
```

Using the Bayesian approach for statistical fine-mapping we obtain a 90% credible set consisting of 9 variants. The claimed coverage of this credible set is ~0.93, yet the corrected coverage estimate is ~0.97, suggesting that we can afford to be ‘more confident’ that we have captured the causal variant in our credible set.

---

3. Evaluate accuracy of estimate

Again, for the purpose of this vignette we can investigate how accurate this estimate is by simulating many credible sets from the same system, and finding the proportion of these that contain the true causal variant.

```
z0.tmp <- simulated_z_score(N0 = N0, # number of controls
                            N1 = N1, # number of cases
                            snps = snps, # column names in freq
                            W = CV, # causal variants, subset of snps
                            gamma.W = log(OR), # log odds ratios
                            freq = freq, # reference haplotypes
                            nrep = 5000)

pps <- ppfunc.mat(zstar = z0.tmp, V = varbeta)  # find pps
cs.cov <- apply(pps, 1, function(x) credset(x, CV = iCV, thr = thr)$cov)
true.cov.est <- mean(cs.cov)
data.frame(claimed.cov = cs$claimed.cov, corr.cov =  corrcov,
           true.cov = true.cov.est, nvar = cs$nvar)
```

```
##   claimed.cov corr.cov true.cov nvar
## 1   0.9307395 0.967282   0.9696    9
```

We find that our corrected coverage value is very close to the empirical coverage of the credible set.

4. Obtain a corrected credible set

Our results suggest that we may be able to remove some variants from the credible set, whilst still achieving the desired coverage of 90%.

The `corrected_cs` function uses GWAS summary statistics and some user-defined parameters to find the smallest credible set such that the coverage estimate is within some accuracy of the desired coverage.

The function requires the following parameters to be specified:

- `z` (vector of marginal Z-scores)
- `f` (vector of minor allele frequencies)
- `N0`, `N1` (number of controls and cases respectively)
- `Sigma` (SNP correlation matrix)
- `desired.cov` (desired coverage of the credible set)

In addition, there are a number of optional parameters:

- `lower` (the lower value to try for the threshold)
- `upper` (the upper value to try for the threshold)
- `acc` (the accuracy required for the coverage of the resultant credible set)
- `max.iter` (maximum number of iterations)

The function uses the bisection root finding method to converge to the smallest threshold such that the corrected coverage is larger than the desired coverage. The `lower` and `upper` parameters define the boundaries of the threshold values for the root finding method, with default values of 0 and 1 respectively. The `acc` parameter has a default value set to 0.005, meaning that the algorithm will keep attempting to find a corrected credible set (until the number of iterations reaches `max.iter`) that has coverage within 0.005 of the desired coverage value (e.g. between 0.895 and 0.905 for a 90% credible set).

The function reports the threshold values tested and their corresponding corrected coverage value at each iteration. The maximum number of iterations for the bisecting root finding algorithm is an optional parameter, with default value 20. The functions stops when either the number of iterations reaches the maximum, or the corrected coverage is within some accuracy of the desired coverage.

```
res <- corrected_cs(z = z0, f = MAF, N0, N1,
                    Sigma = LD, lower = 0.5, upper = 1, desired.cov = 0.9)
```

```
## [1] "thr:  0.75 , cov:  0.917720814111386"
## [1] "thr:  0.625 , cov:  0.878894070913302"
## [1] "thr:  0.6875 , cov:  0.896797621472936"
## [1] "thr:  0.71875 , cov:  0.907034158326262"
## [1] "thr:  0.703125 , cov:  0.901702620560411"
```

```
res
```

```
## $credset
## [1] "s54" "s58" "s57" "s67"
##
## $req.thr
## [1] 0.703125
##
## $corr.cov
## [1] 0.9017026
##
## $size
## [1] 0.7343958
```

The first threshold value tested is the midpoint of the `lower` (0.5) and `upper` (1) parameter values, which here is 0.75. The coverage of this credible set is too high (> 0.9) and so a smaller threshold value is tested, the midpoint of the `lower` parameter value (0.5) and the previously tried threshold value (0.75). This credible set has a coverage that is too low, and so a higher threshold value is tested, and so on.

In this example we see that a much smaller threshold value is required to obtain a credible set with 90% corrected coverage of the causal variant, containing only 4 variants. In the standard Bayesian approach, the threshold value of 90% leads to over-coverage.

Finally, we can compare this coverage estimate of the corrected credible set to an empirical estimate of the true coverage.

```
new.cs.sims <- apply(pps, 1, function(x)
                    credset(x, CV = iCV, thr = res$req.thr)$cov)
true.cov.est2 <- mean(new.cs.sims)
```

Original 90% credible set:

```
df1 <- data.frame(claimed.cov = round(cs$claimed.cov, 3),
                  corr.cov =  round(corrcov, 3),
                  true.cov = round(true.cov.est, 3), nvar = cs$nvar)
print(df1, row.names = FALSE)
```

```
##  claimed.cov corr.cov true.cov nvar
##        0.931    0.967     0.97    9
```

New 90% credible set:

```
df2 <- data.frame(claimed.cov = round(res$size, 3),
                  corr.cov = round(res$corr.cov, 3),
                  true.cov = round(true.cov.est2, 3),
                  nvar = length(res$credset))
print(df2, row.names = FALSE)
```

```
##  claimed.cov corr.cov true.cov nvar
##        0.734    0.902    0.897    4
```

This vignette has shown how the `corrcoverage` R package can be used to improve the resolution of a credible set from Bayesian genetic fine-mapping, without the use of any additional data.