fcfdr: an R package to leverage continuous and binary functional genomic data in GWAS

Anna Hutchinson[1,*], James Liley[2,3], Chris Wallace[1,4,5]

1 MRC Biostatistics Unit, Cambridge Biomedical Campus, University of Cambridge, Cambridge, CB2 0SR, UK.
2 MRC Human Genetics Unit, IGMM, University of Edinburgh, Crewe Rd S, Edinburgh, UK.
3 The Alan Turing Institute, 96 Euston Rd, Somers Town, London, NW1 2DB, UK.
4 Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge, CB2 0AW, UK.
5 Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, CB2 2QQ, UK.

* anna.hutchinson@mrc-bsu.cam.ac.uk

Abstract

Summary: GWAS discovery is limited in power to detect associations that exceed the stringent genome-wide significance threshold, but this limitation can be alleviated by leveraging relevant auxiliary data. Frameworks utilising the conditional false discovery rate (cFDR) can be used to leverage continuous auxiliary data (including GWAS and functional genomic data) with GWAS test statistics and have been shown to increase power for GWAS discovery whilst controlling the FDR. Here, we describe an extension to the cFDR framework for binary auxiliary data (such as whether SNPs reside in regions of the genome with specific activity states) and introduce an all-encompassing R package to implement the cFDR approach, `fcfdr`, demonstrating its utility in an application to type 1 diabetes.

Availability and implementation: The `fcfdr` R package is freely available at: `https://github.com/annahutch/fcfdr`. Scripts and data to reproduce the analysis in this paper are freely available at: `https://annahutch.github.io/fcfdr/articles/t1d_app.html`

1 Introduction

A stringent significance threshold is required to identify robust genetic associations in GWAS due to multiple testing constraints. Leveraging relevant auxiliary data has the potential to boost statistical power to exceed the significance threshold. The conditional FDR (cFDR) is a Bayesian FDR measure that additionally conditions on auxiliary data to call significant associations. The cFDR approach was originally developed to leverage GWAS p-values from related traits, thereby exploiting genetic pleiotropy to increase GWAS discovery[1,2,3], and has been shown to increase power for GWAS discovery whilst controlling the frequentist FDR[11].

Motivated by the enrichment of GWAS SNPs in particular functional genomic annotations[14], Flexible cFDR was developed to extend the usage of the cFDR approach to the accelerating field of functional genomics[9]. However, at-present no cFDR methodology exists that permits binary auxiliary data, meaning that the 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 regions of the genome with specific activity states.

Here we present an extension to the cFDR approach that supports binary auxiliary data and we thus introduce a cFDR toolbox in the form of an R package (`https://github.com/annahutch/fcfdr`) that supports various auxiliary data types. We demonstrate the utility of our methods and software by iteratively leveraging three distinct types of relevant auxiliary data with GWAS p-values for type 1 diabetes (T1D)[12] to uncover new genetic associations.

2 The cFDR framework

Let $p_1,...,p_m \in (0,1]$ be a set of p-values corresponding to the null hypotheses of no association between the SNPs and a trait of interest (denoted by H_0). Let $q_1,...,q_m$ be auxiliary data values corresponding to the same m SNPs. Assume that p and q are realisations of random variables P,Q satisfying:

$$\begin{gathered}(P|H_0) \sim U(0,1) \\ P \perp\!\!\!\perp Q|H_0.\end{gathered} \tag{1}$$

The cFDR is defined as the probability that a random SNP is null for the trait given that the observed p-values and auxiliary data values at that SNP are less than or equal to values p and q respectively[1,2]. Bayes theorem and standard probability rules are used to derive:

$$\begin{aligned} cFDR(p,q) &= Pr(H_0|P \leq p, Q \leq q) \\ &= \frac{Pr(P \leq p|H_0, Q \leq q) \times Pr(H_0|Q \leq q)}{Pr(P \leq p|Q \leq q)} \\ &= \frac{Pr(P \leq p|H_0, Q \leq q) \times Pr(Q \leq q|H_0)Pr(H_0)}{Pr(P \leq p, Q \leq q)}. \end{aligned} \tag{2}$$

To construct a conservative estimator of the cFDR, approximate $Pr(P \leq p|H_0, Q \leq q) \approx p$ (from property 1; note that if property 1 holds and P is correctly calibrated then this approximation is an equality) and $Pr(H_0) \approx 1$ (since associations are rare in GWAS):

$$\widehat{cFDR}(p,q) = \frac{p \times \widehat{Pr(Q \leq q|H_0)}}{\widehat{Pr(P \leq p, Q \leq q)}}, \tag{3}$$

where $\widehat{}$ is used to denote that these are estimates under the assumption $H_0 \perp\!\!\!\perp Q|P$. The methods used to estimate the cumulative densities in equation (3) vary across approaches. In the original cFDR approach they are estimated using empirical cumulative density functions[1,10,11] whilst in Flexible cFDR they are estimated using kernel density estimation[9].

However, the $\widehat{cFDR}$ values do not directly control the FDR[10]. Instead, a method proposed by Liley and Wallace[11] can be used to generate v-values, which are essentially the probability of a newly-sampled realisation (p,q) of P,Q attaining an as extreme or more extreme $\widehat{cFDR}$ value than that observed, given H_0. The v-values are therefore analogous to p-values and can be used in any conventional error-controlling multiple testing procedure that allows for slightly dependent p-values (e.g. the Benjamini-Hochberg procedure). The derivation of v-values also allows for the method to be applied iteratively to incorporate additional layers of auxiliary data.

Since binary auxiliary data can only take two values, we introduce an alternative methodology called "Binary cFDR" which is based on finding optimal rejection regions to derive v-values (see Supplementary Methods for full details on the Binary cFDR methodology). We show in a simulation-based analysis that applying Binary cFDR iteratively over informative auxiliary data increases power whilst controlling the frequentist FDR (Supplementary Results, Supplementary Fig. 2).

3 R package and T1D application

We present an R package that implements both Flexible cFDR and Binary cFDR, named `fcfdr` (`https://github.com/annahutch/fcfdr`), and demonstrate its utility in an application to T1D which is fully reproducible (see `https://annahutch.github.io/fcfdr/articles/t1d_app.html`).

We used p-values from an Immunochip study of T1D[12] as our primary data set. In the first iteration we used Flexible cFDR to leverage Immunochip p-values for a genetically related trait, rheumatoid arthritis (RA)[6] (Fig. 1A). In the second iteration we used Binary cFDR to leverage data measuring SNP overlap with regulatory factor binding sites[5,8,7] (Fig. 1B) and in the third iteration we used Flexible cFDR to leverage average enhancer-associated H3K27ac fold change values derived from ChIP-seq experiments conducted in T1D-relevant cell types[4] (Fig. 1C) (see Supplementary Methods for full details on the data).

Our implementation of cFDR identified 101 SNPs as newly genome-wide significant ($FDR \leq 3.3e-06$ which corresponds to $p \leq 5e-08$; Supplementary Methods). These SNPs had relatively small p-values for RA (median $p = 0.007$ compared with median $p = 0.422$ in full data set), were more likely to be found in regulatory factor binding sites (mean binary value was 0.406 compared to 0.234 in full data set) and had larger H3K27ac fold change values in T1D-relevant cell types (median fold change value was 1.44 compared with 0.576 in full data set). Similarly, 45 SNPs were identified as newly not significant (i.e. they were significant in the original GWAS data set but became not significant after applying cFDR). These SNPs had relatively high p-values for RA (median $p = 0.620$), were less likely to be found in regulatory factor binding sites (mean binary value was 0.044) and had smaller H3K27ac fold change values in T1D-relevant cell types (median fold change value was 0.431).

The original GWAS identified 38 significant genomic regions (based on our definition of genomic regions, see Supplementary Methods). All of these were found to be significant in the cFDR analysis, which additionally identified 4 genomic regions that were newly significant (with lead variants:

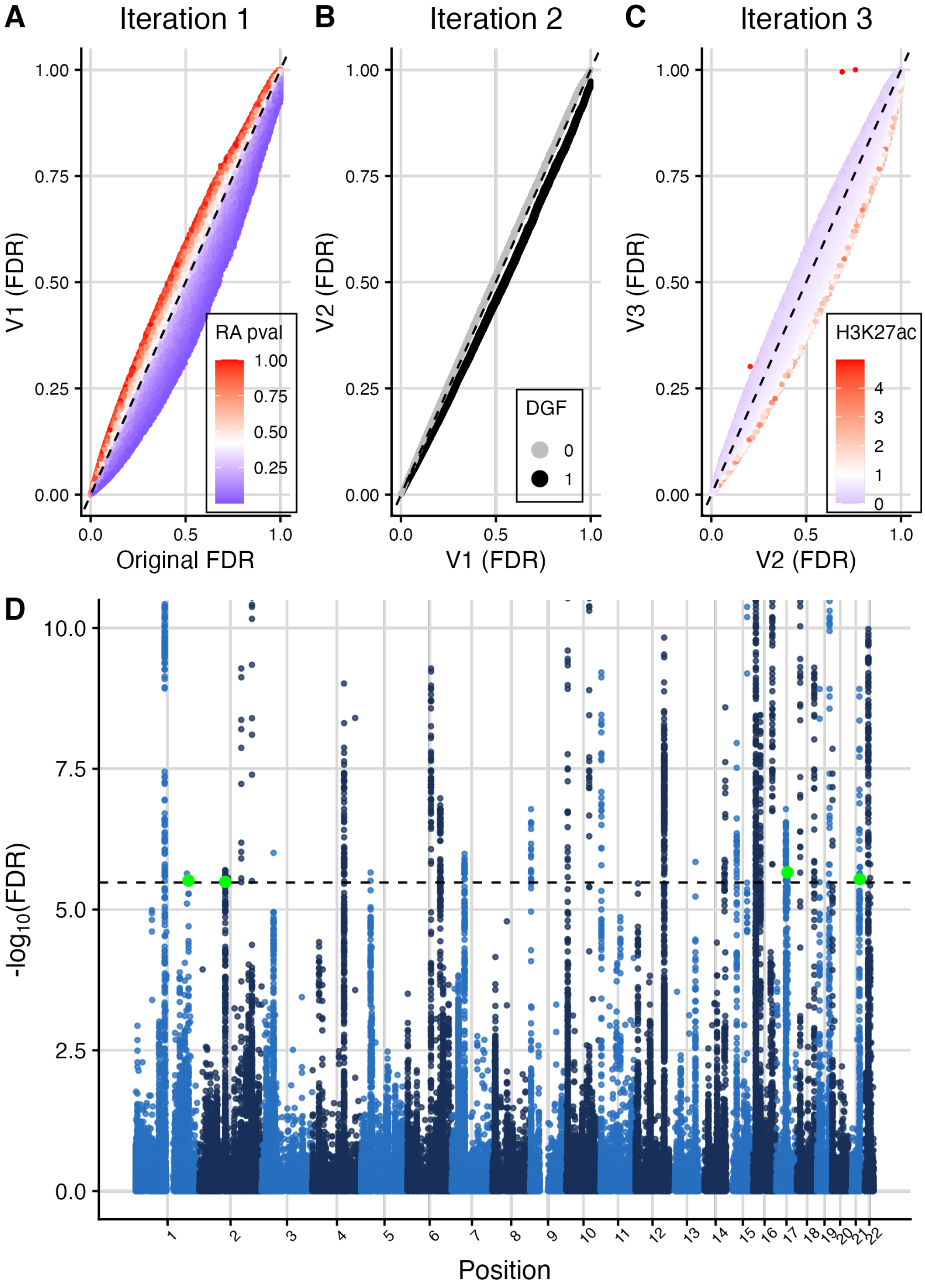

Figure 1: Summary of cFDR results for T1D application. "FDR values" were obtained from the raw p-values and v-values from each iteration of the cFDR approach using the Benjamini-Hochberg procedure. Top panel shows FDR values before and after (A) iteration 1 (B) iteration 2 and (C) iteration 3 of the method, coloured by the value of the auxiliary data (p-values for RA in iteration 1, DGF annotation values in iteration 2 and average H3K27ac fold change values relative to expected background counts in iteration 3). (D) Manhattan plot of ($-log_{10}$ transformed) FDR values. Green points indicate the four lead variants that were newly FDR significant after cFDR. Black dashed line at FDR significance threshold ($FDR = 3.3e-06$; which was the maximum FDR value amongst SNPs with raw p-values $\leq 5e-08$ - see Supplementary Methods). y-axis has been truncated in panel (D) to aid visualisation.

rs1052553, rs3024505, rs6518350 and rs13415583). Three of these SNPs had small p-values for RA (rs1052553: RA $p = 0.007$; rs6518350: RA $p = 0.06161$ and rs13415583: RA $p = 1.913e-06$ whereas rs3024505 had RA $p = 0.6008$) and two of these SNPs had high H3K27ac fold change values (rs3024505 had 87.4th percentile and rs6518350 had 72.7th percentile of H3K27ac fold change values). Two of the lead variants overlapped regulatory factor binding sites (rs1052553 and rs3024505). When using a larger Immunochip study of T1D for validation ($16,159$ T1D cases compared to $6,670$)[13], we found that three out of the four lead variants were present and that these had smaller p-values in the validation GWAS data set than the discovery GWAS data set: rs1052553 had $p = 1.649e-15$, rs3024505 had $p = 9.127e-14$, rs13415583 had $p = 4.764e-09$ in the validation data set[13] compared to $p = 8.156e-08$, $p = 6.394e-08$ and $p = 1.062e-07$ respectively in the discovery data set[12].

4 Conclusion

We have described a novel implementation of the cFDR approach that supports binary auxiliary data and have introduced an all-encompassing R package, `fcfdr`, that can be used to implement the cFDR approach for a wide variety of auxiliary data types. We have demonstrated the versatility of this tool in an application to T1D where we uncovered new genetic associations.

Funding

AH is funded by the Engineering and Physical Sciences Research Council (EPSRC) `https://epsrc.ukri.org/` (EP/R511870/1) and GlaxoSmithKline (GSK) `https://www.gsk.com/`. CW is funded by the Wellcome Trust `https://wellcome.ac.uk/` (WT107881, WT220788), the Medical Research Council (MRC) `https://mrc.ukri.org/` (MC UU 00002/4) and supported by the NIHR Cambridge BRC `https://cambridgebrc.nihr.ac.uk/` (BRC-1215-20014). JL is partially supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Health" theme within that grant and The Alan Turing Institute, and partially supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), the devolved administrations, and leading medical research charities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Andreassen, O.A. et al (2013). Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLOS Genetics*, **9**(4), e1003455.

[2] Andreassen, O.A. et al (2014). Identifying Common Genetic Variants in Blood Pressure Due to Polygenic Pleiotropy With Associated Phenotypes. *Hypertension*, **63**(4), 819–826.

[3] Andreassen, O.A. et al (2015). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: Differential involvement of immune-related gene loci. *Molecular Psychiatry*, **20**(2), 207–214.

[4] Bernstein, B.E. et al (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10), 1045–1048.

[5] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

[6] Eyre, S. et al (2012). High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature genetics*, **44**(12), 1336–1340.

[7] Gazal, S. et al (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, **49**(10), 1421–1427.

[8] Gusev, A. et al (2014). Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American Journal of Human Genetics*, **95**(5), 535–552.

[9] Hutchinson, A., Reales, G., Willis, T. and Wallace, C. (2021). Leveraging auxiliary data from arbitrary distributions to boost GWAS discovery with Flexible cFDR. *PLOS Genetics*, **17**(10), e1009853.

[10] Liley, J. and Wallace, C. (2015). 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.

[11] Liley, J. and Wallace, C. (2021). Accurate error control in high-dimensional association testing using conditional false discovery rates. *Biometrical Journal*.

[12] Onengut-Gumuscu, S. et al (2015). 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.

[13] Robertson, C.C. et al (2021). Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nature Genetics*, pages 1–10.

[14] Schork, A.J. et al (2013). 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.