BMC Bioinformatics

SOFTWARE

Open Access

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

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

*Correspondence:
cew54@cam.ac.uk

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2] MRC Human Genetics Unit, University of Edinburgh, Edinburgh, UK
[3] The Alan Turing Institute, London, UK
[4] Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), University of Cambridge, Cambridge, UK
[5] Department of Medicine, University of Cambridge, Cambridge, UK

Abstract

Background: Genome-wide association studies (GWAS) are limited in power to detect associations that exceed the stringent genome-wide significance threshold. This limitation can be alleviated by leveraging relevant auxiliary data, such as functional genomic data. Frameworks utilising the conditional false discovery rate have been developed for this purpose, and have been shown to increase power for GWAS discovery whilst controlling the false discovery rate. However, the methods are currently only applicable for continuous auxiliary data and cannot 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.

Results: We describe an extension to the cFDR framework for binary auxiliary data, called "Binary cFDR". We demonstrate FDR control of our method using detailed simulations, and show that Binary cFDR performs better than a comparator method in terms of sensitivity and FDR control. We introduce an all-encompassing user-oriented CRAN R package (https://annahutch.github.io/fcfdr/; https://cran.r-project.org/web/packages/fcfdr/index.html) and demonstrate its utility in an application to type 1 diabetes, where we identify additional genetic associations.

Conclusions: Our all-encompassing R package, `fcfdr`, serves as a comprehensive toolkit to unite GWAS and functional genomic data in order to increase statistical power to detect genetic associations.

Keywords: GWAS, Functional genomics, Power, FDR, Multiple testing

Background

A stringent significance threshold is required to identify robust genetic associations in genome-wide association studies (GWAS) due to multiple testing constraints. Leveraging relevant auxiliary data, such as functional genomic data, has the potential to boost statistical power in order to detect associations that exceed the stringent significance threshold.

The conditional false discovery rate (cFDR) is a Bayesian FDR measure that additionally conditions on auxiliary data to call significant associations. Let $p_1, \ldots, p_m \in (0, 1]$ be a set of p values corresponding to the null hypotheses of no association between SNPs $1, \ldots, m$ and a trait of interest (denoted by H_0). Let $q_1, \ldots, q_m$ be auxiliary data

values corresponding to the same SNPs. Assume that p and q are realisations of random variables P, Q satisfying:

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

The cFDR is then 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]. That is,

$$cFDR(p,q) = Pr(H_0|P \leq p, Q \leq q). \tag{2}$$

It should be noted that, although the Bayes-optimal decision quantity $Pr(H_0|P = p, Q = q)$ [3, 4] is asymptotically more powerful for hypothesis testing, it is practically more difficult to estimate accurately in finite-sample settings [5].

The cFDR approach was originally developed to leverage GWAS p values from related traits, thereby exploiting genetic pleiotropy to increase GWAS discovery [1, 2, 6]; however, these early methods failed to control the FDR. Consequently, Liley and Wallace [5] developed an extension to the cFDR approach that transforms cFDR estimates into "v-values" which are analogous to p values and can therefore be used to control FDR (for example in the Benjamini–Hochberg procedure [7]).

Motivated by the enrichment of GWAS SNPs in particular functional genomic annotations [8], Flexible cFDR was developed to extend the usage of the cFDR approach to the accelerating field of functional genomics [9]. Several related methods exist for multiple testing in the presence of auxiliary information [4, 10–12], but Flexible cFDR has been shown to outperform these methods in terms of usability, versatility, accessibility and FDR control [9]. Nonetheless, a disadvantage of Flexible cFDR is that it cannot 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, called Binary cFDR. In a simulation-based analysis, we compare the performance of Binary cFDR to that of an existing approach, Boca and Leek's FDR regression [13], which has been shown to outperform other methods in terms of FDR control, power, applicability and consistency of results by an independent research group [14]. We introduce a cFDR toolbox in the form of an R package (https://github.com/annahutch/fcfdr) that supports various auxiliary data types and which is available on CRAN (https://cran.r-project.org/web/packages/fcfdr/index.html). Finally, 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 to uncover additional genetic associations.

Implementation

The cFDR framework

We begin by describing the standard cFDR framework. Bayes theorem and standard probability rules are used to derive:

$$\begin{aligned} cFDR(p,q) &= Pr(H_0|P \leq p, Q \leq q) \\ &= \frac{Pr(P \leq p|H_0, Q \leq q) \times Pr(H_0|Q \leq q)}{Pr(P \leq p|Q \leq q)} \\ &= \frac{Pr(P \leq p|H_0, Q \leq q) \times Pr(Q \leq q|H_0)Pr(H_0)}{Pr(P \leq p, Q \leq q)}. \end{aligned} \tag{3}$$

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):

$$\widehat{cFDR}(p,q) = \frac{p \times \widehat{Pr(Q \leq q|H_0)}}{\widehat{Pr(P \leq p, Q \leq q)}}, \tag{4}$$

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 (4) vary across approaches. For example, in the original cFDR approach they are estimated using empirical cumulative density functions [1, 5, 15] whilst in Flexible cFDR they are estimated using kernel density estimation [9].

However, the $\widehat{cFDR}$ values do not directly control the FDR [15]. Instead, a method proposed by Liley and Wallace [5] can be used to generate ν-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 ν-values are therefore analogous to p values and can be used in any conventional error-controlling multiple testing procedure. The derivation of ν-values also allows for the method to be applied iteratively to incorporate additional layers of auxiliary data.

Extension for binary covariate data

We introduce an extension to the cFDR framework that permits binary covariate data, and call our method "Binary cFDR".

As before, let $p_1, \ldots, p_m \in (0, 1]$ be a set of p values corresponding to the null hypotheses of no association between the SNP and the trait of interest. Now, let $q_1, \ldots, q_m \in \{0, 1\}$ be a set of binary covariates for the same m SNPs. Denote the null (no association) and alternative (association) hypotheses as H_0 and H_1 respectively and assume that p and q are realisations of random variables P, Q satisfying property (1). We follow the standard methodology introduced by Liley and Wallace [5] to derive a ν-value, ν_i, for each (p_i, q_i) pair.

Since all q are binary, the support of P, Q is two lines $(0, 1) \times \{0, 1\}$. We consider rejection regions of the form $L(p_0, p_1) = (P \leq p_0, Q = 0) \cup (P \leq p_1, Q = 1)$, where p_0 and p_1 are to be determined.

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

$$\begin{aligned} &Pr(\nu_i < \alpha|H_0) = \alpha \\ &Pr(\nu_i < \alpha|H_1) \text{ is maximal.} \end{aligned} \tag{5}$$

That is, the ν-values behave like typical p values in that they are uniform under the null, but are as small as possible under the alternative hypothesis. Appendix A.1 in [5]

(and also [16] and [17], 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 $f_0(p,q) = f(P=p, Q=q|H_0)$ and $f_1(p,q) = f(P=p, Q=q|H_1)$. If $f_1(p,q)$ is non-increasing in p, then such optimal rejection regions are of the type $L(p_0,p_1)$ defined above (we describe behaviour in other cases in Additional File 1). That is, p_0 and p_1 will satisfy the property

$$\frac{f_0(p_0,0)}{f_1(p_0,0)} = \frac{f_0(p_1,1)}{f_1(p_1,1)}. \tag{6}$$

Let

$$f(p,q) = f(P=p, Q=q) = \pi_0 f_0(p,q) + (1-\pi_0) f_1(p,q), \tag{7}$$

where $\pi_0 = Pr(H_0)$. Then equation (6) implies that

$$\frac{f(p_1,1)}{f_0(p_1,1)} = \frac{f(p_0,0)}{f_0(p_0,0)}. \tag{8}$$

To solve equation (6) for p_0 and p_1, we approximate

$$\frac{f_0(p_i,q_i)}{f(p_i,q_i)} = \frac{Pr(P=p_i, Q=q_i|H_0)}{Pr(P=p_i, Q=q_i)} \tag{9}$$

$$\approx \frac{Pr(P\le p_i, Q=q_i|H_0)}{Pr(P\le p_i, Q=q_i)} \tag{10}$$

$$= \frac{Pr(P\le p_i|Q=q_i, H_0)Pr(Q=q_i|H_0)}{Pr(P\le p_i|Q=q_i)Pr(Q=q_i)} \tag{11}$$

$$\approx \frac{p_i \times \widehat{Pr(Q=q_i|H_0)}}{|j : p_j \le p_i, q_j = q_i|/m} \tag{12}$$

where $\widehat{Pr(Q=q_i|H_0)} = \frac{|j : q_j = q_i, p_j > 1/2|}{|j : p_j > 1/2|}$ and m is the number of SNPs. Approximation (10) is discussed in Additional File 1. If $q_i = 0$ then we set $p_0 = p_i$ and use approximation (12) to solve equation (6) for p_1. If $q_i = 1$, then we set $p_1 = p_i$ and solve for p_0. In practise, we 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. Specifically, we leave out each chromosome in turn and use the remaining SNPs to estimate the values for the held out SNPs.

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

$$\int_{L(p_0,p_1)} df_0 = Pr((P,Q) \in L(p_0,p_1)|H_0) \tag{13}$$

$$= Pr((P\le p_0, Q=0) \cup (P\le p_1, Q=1)|H_0) \tag{14}$$

$$= Pr(P \leq p_0, Q = 0|H_0) + Pr(P \leq p_1, Q = 1|H_0) \quad (15)$$

$$= Pr(P \leq p_0|Q = 0, H_0)Pr(Q = 0|H_0) + Pr(P \leq p_1|Q = 1, H_0)Pr(Q = 1|H_0) \quad (16)$$

$$= p_0 \times (1 - q_0) + p_1 \times q_0 \quad (17)$$

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

The ν-value, ν_i, can be interpreted as the probability that a randomly-chosen (p, q) pair has a more extreme cFDR value than cFDR(p_i, q_i) under H_0. That is, a quantity analogous to a p value. This means that, as in the original cFDR approach [5], the Binary cFDR method can be applied iteratively to incorporate additional layers of auxiliary data, whereby the ν-values from the previous iteration are used as the principal trait p values in the current iteration. The derivation of ν-values analogous to p values also means that they can be readily FDR controlled using any FDR controlling procedure that allows for slightly dependent p values (as in GWAS), such as the Benjamini–Hochberg procedure [5].

fcfdr R package

We have created a CRAN R package, `fcfdr`, that implements the Flexible cFDR and Binary cFDR approaches (https://cran.r-project.org/web/packages/fcfdr/index.html). Our recently updated package supports a wide range of auxiliary data types and is particularly suited to leveraging functional genomic data with GWAS test statistics, as explored below and also in several fully reproducible vignettes that are available on the package web-page (https://annahutch.github.io/fcfdr/).

Results

Simulation based analysis

We evaluated the performance of Binary cFDR as implemented in the `fcfdr` R package using a simulation-based analysis. In each simulation, we applied Binary cFDR iteratively 5 times to represent leveraging multi-dimensional binary covariates. We additionally compared our results to those when using a comparator method, Boca and Leek's FDR regression (BL) [13], which has been shown to outperform other methods by an independent research group [14].

We expect that leveraging irrelevant data should not change our conclusions about a study. Figure 1A shows that the sensitivity and specificity remain stable across iterations and that the FDR was controlled at a pre-defined level when using Binary cFDR to leverage independent binary auxiliary data with arbitrary GWAS p values. In contrast, when leveraging relevant data we hope that the sensitivity improves whilst the specificity remains high, which is what we observed for Binary cFDR in Fig. 1B.

It is known that the cFDR approach should not be used to iterate over correlated auxiliary data that is capturing the same functional mark, as SNPs with a modest p but extreme q will incorrectly attain greater significance with each iteration (for a more detailed explanation see [9]). Our final set of simulations involved iterating over

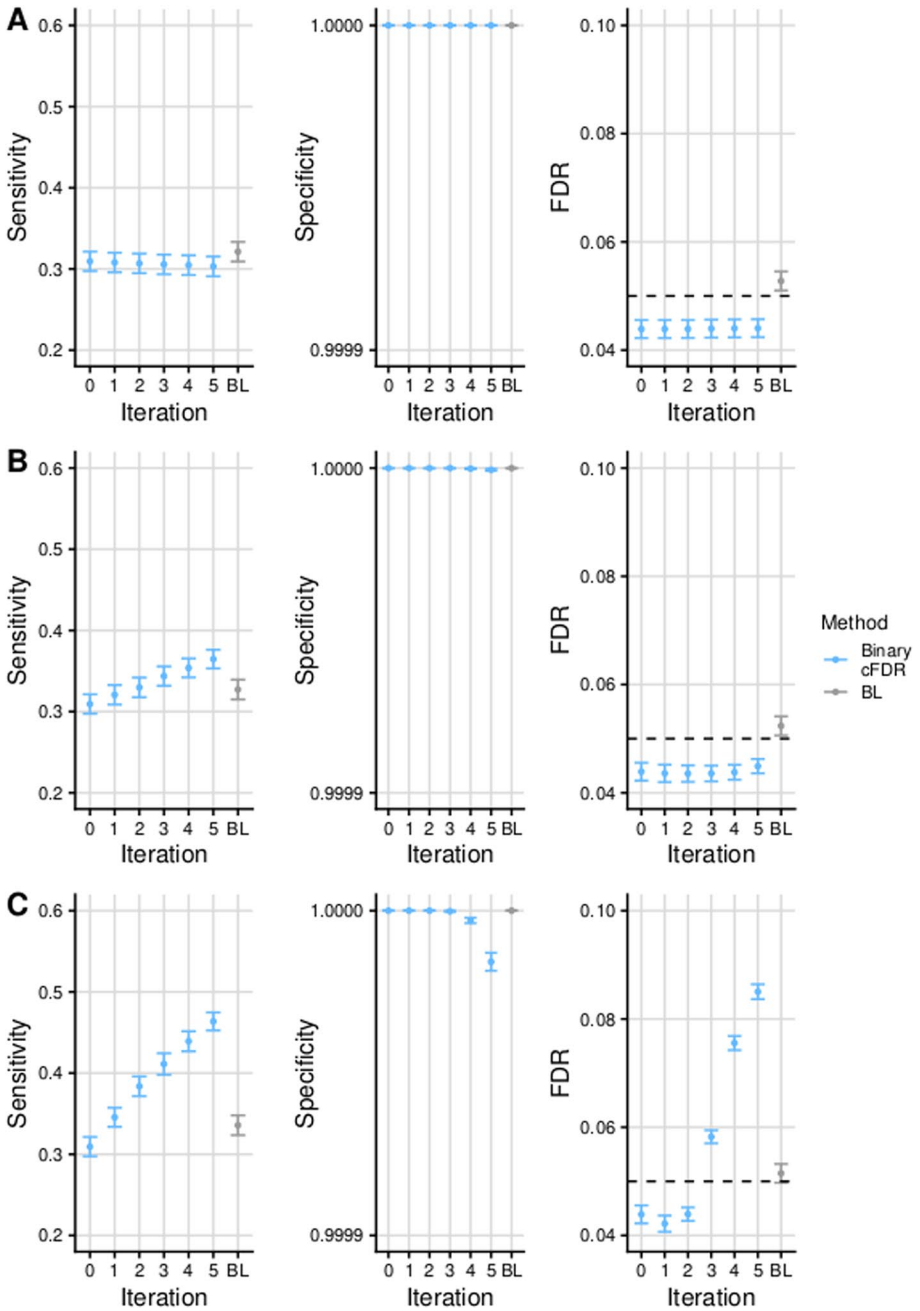

Fig. 1 Simulation results for Binary cFDR and BL. Mean +/− standard error for the sensitivity, specificity and FDR of FDR values (derived from the Benjamini–Hochberg procedure) from Binary cFDR when iterating over independent (**A**; "simulation A") and dependent (**B**; "simulation B" and **C**; "simulation C") 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

correlated auxiliary data values (mean Pearson correlation coefficient = 0.3) that capture the same "functional mark" (80% of functional SNPs were expected to have an auxiliary data value of 1 in each iteration). The lack of FDR control in these sets of simulations (Fig. 1C) serves as a salutary reminder that care should be taken not to repeatedly iterate over functional data that is capturing the same genomic feature.

When bench-marking the performance of Binary cFDR against that of BL, we found that BL was consistently less powerful than Binary cFDR when leveraging dependent auxiliary data (Fig. 1B, C). In contrast, BL was more powerful than Binary cFDR when leveraging independent auxiliary data, but this was at the cost of a marginal loss of FDR control (Fig. 1A). In fact, the FDR control of BL was similar across all simulations, even when using correlated auxiliary data in Fig. 1C.

Application to type 1 diabetes

We demonstrate the utility of `fcfdr` in an application to type 1 diabetes which is fully reproducible (https://annahutch.github.io/fcfdr/articles/t1d_app.html). Using p values from an Immunochip study of type 1 diabetes [18] as our primary data set, we iteratively leveraged p values from an Immunochip study of a related immune-mediated trait (rheumatoid arthritis; RA), binary data measuring SNP overlap with regulatory factor binding sites and enhancer-associated H3K27ac ChIP-seq data in cell types relevant to type 1 diabetes (Fig. 2).

Our method identified 101 SNPs as newly FDR significant ($FDR \leq 3.3 \times 10^{-6}$ which corresponds to $p \leq 5 \times 10^{-8}$; see 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

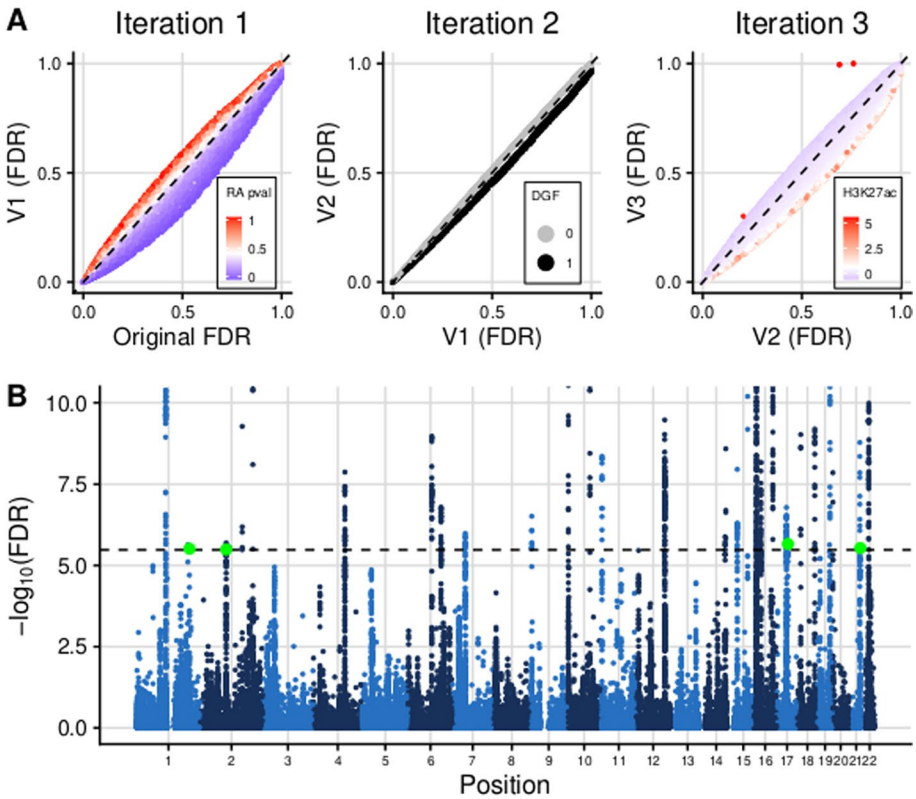

Fig. 2 Summary of cFDR results from type 1 diabetes application. **A** FDR values (derived from the Benjamini–Hochberg procedure) before and after each iteration of cFDR, coloured by the auxiliary data values. **B** Manhattan plot of ($-log_{10}$) FDR values (y-axis truncated to aid visualisation). Green points indicate the four lead variants that were newly FDR significant after cFDR. Black dashed line at FDR significance threshold ($FDR = 3.3 \times 10^{-6}$)

likely to be found in regulatory factor binding sites (40.6% in binding sites compared to 23.4% in full data set) and had larger H3K27ac fold change values in relevant cell types (median value was 1.44 compared with 0.576 in full data set). In contrast, 45 SNPs that were significant in the original GWAS data set became not significant after applying cFDR, and these had relatively high p values for RA (median $p = 0.620$), were less likely to be found in regulatory factor binding sites (4.4% in binding sites) and had smaller H3K27ac fold change values (median value was 0.431).

The original GWAS identified 38 significant genomic regions (based on our definition of genomic regions; see Methods). All of these were found to be significant in the cFDR analysis, which additionally identified 4 genomic regions with index SNPs that became newly significant (Table 1). When using a larger Immunochip study of type 1 diabetes for validation [19] (see Methods) 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 validation $p = 1.65 \times 10^{-15}$, rs3024505 validation $p = 9.13 \times 10^{-14}$ and rs13415583 validation $p = 4.76 \times 10^{-9}$.

When using BL to leverage the same auxiliary data, only 46 SNPs were identified as newly FDR significant, and these had larger p values for RA (median $p = 0.1942$), were less likely to be found in regulatory factor binding sites (32.6% in binding sites) and had similar H3K27ac fold change values (median value was 1.48) compared with the 101 SNPs identified as newly significant in the cFDR analysis. At the locus level, BL only identified 1 newly significant index SNP, rs3024505, which was also identified by cFDR. No SNPs that were significant in the original GWAS data set became not significant after applying BL.

Conclusions

We have described Binary cFDR, a novel implementation of the cFDR approach that supports binary auxiliary data. Binary cFDR controls the FDR and increases sensitivity where appropriate, and outperforms an existing method in terms of sensitivity and FDR control. Binary cFDR is implemented in an all-encompassing CRAN 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 our software in an application to type 1 diabetes, whereby we incorporated both binary and continuous auxiliary data simultaneously to uncover additional genetic associations that were replicated in a larger study.

Methods

Simulation analysis

Simulating GWAS results (p)

Following Hutchinson et al. (2021) [9], we first simulated GWAS p values for the arbitrary "principal trait". We collected haplotype data for 3781 individuals from the UK10K project (REL-2012-06-02) [20] 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). We split the haplotype data into 24 LD blocks representing approximately independent genomic regions defined by the LD detect method [21]. We 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.

Table 1 Table of newly significant index SNPs from type 1 diabetes application

rsID	Position	Ref/Alt	OR	SE	*p* value	*v*-value	RA *p* value	DGF	H3K27ac percentile	Gene
rs1052553	chr17:44073889	A/G	0.889	0.022	8.16×10^{-8}	3.10×10^{-8}	6.76×10^{-3}	1	2.2th	*STH*
rs3024505	chr1:206939904	G/A	0.864	0.027	6.39×10^{-8}	4.51×10^{-8}	0.601	1	87.4th	*IL19*
rs6518350	chr21:45621817	A/G	0.880	0.024	9.64×10^{-8}	4.26×10^{-8}	0.062	0	72.7th	*ICOSLG*
rs13415583	chr2:100764087	T/G	0.904	0.019	1.06×10^{-7}	4.81×10^{-8}	1.91×10^{-6}	0	14.4th	*AFF3*

For each of the four newly significant SNPs from the cFDR analysis, we list the rsID, genomic position (hg19; Position), reference and alternative alleles (Ref/Alt), odds ratio (OR), standard error (SE) and *p* value reported in the primary GWAS data set [18], *v*-value from the cFDR analysis, GWAS *p* value for rheumatoid arthritis [27] (RA *p* value), binary indicator of SNP overlap with regulatory factor binding sites (DGF), percentile of mean H3K27ac fold change value across asthma relevant cell types (H3K27ac percentile) and the closest protein-coding gene

We used the `simGWAS` R package (https://github.com/chr1swallace/simGWAS) [22] to simulate Z-scores for SNPs within each block. The `simulate_z_scores` function in thee `simGWAS` R package requires input for (i) the number of cases and controls (ii) the causal variants (iii) the log odds ratios at the causal variants and (iv) haplotype frequencies. For our simulation analysis, we selected 5000 cases and 5000 control samples, and within each block we 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, $N(0, 0.2^2)$ [23]. For the haplotype frequency parameter, we supplied a `data.frame` of haplotypes using the UK10K data, with a column of computed frequencies for each haplotype. We 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.

Simulating auxiliary data (q)

We considered three use-cases of Binary cFDR (simulations A-C) defined by dependence on the principal trait p value (p_i) and correlations between realisations of q. In simulation A we leveraged binary auxiliary data that was independent of p_i: $q_i \sim$ Bernoulli(0.05). In simulations B and C we leveraged binary 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. We then sampled q_i from different distributions for functional and non-functional SNPs. Specifically, in simulation B we sampled:

$$q_i \sim \begin{cases} \text{Bernoulli}(0.05), \text{ if SNP i is non-functional} \\ \text{Bernoulli}(0.4), \text{ if SNP i is functional.} \end{cases} \tag{18}$$

Our method will likely be used to leverage functional genomic data iteratively, and so we also evaluated the impact of repeatedly iterating over auxiliary data that captured the same functional mark. Thus, in simulation C we iterated over realisations of q that were highly correlated:

$$q_i \sim \begin{cases} \text{Bernoulli}(0.05), \text{ if SNP i is non-functional} \\ \text{Bernoulli}(0.8), \text{ if SNP i is functional.} \end{cases} \tag{19}$$

Note that the auxiliary data is correlated in simulation C because the functional SNPs are the same across iterations in each simulation.

Implementing Binary cFDR and BL

We used the `fcfdr::binary_cfdr` function to implement Binary cFDR in our simulation analysis. To avoid overfitting we used a leave-one-out procedure, whereby the LD block [21] was used as the group variable. In each simulation for each simulation scenario, we applied Binary cFDR iteratively 5 times to represent leveraging multi-dimensional covariates.

To implement BL, we used the `lm_qvalue` function in the `swfdr` Bioconductor R package (version 1.16.0) [24], using a covariate matrix that consisted of five columns for the auxiliary data values to derive adjusted p values.

Evaluating sensitivity, specificity and FDR control

To quantify the results from our simulations, we used the Benjamini–Hochberg procedure to derive FDR-adjusted ν-values from Flexible cFDR, which we call "FDR values" for conciseness (that is, we used the `stats::p.adjust` R function with `method="BH"`). We then calculated proxies for the sensitivity (true positive rate) and the specificity (true negative rate) at an FDR threshold of $\alpha = 5 \times 10^{-6}$, which roughly corresponds to the genome-wide significance p value threshold of 5×10^{-8} (the maximum FDR value amongst SNPs with raw p value $\leq \times 10^{-8}$ was 5.4×10^{-6}). We defined a subset of "truly associated SNPs" as any SNPs with $r^2 \geq 0.8$ with any of the causal variants. Similarly, we 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 are 3 non-overlapping sets of SNPs: "truly associated", "truly not-associated" and neither of these). We 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.

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

Application to type 1 diabetes

GWAS data

We downloaded full harmonised GWAS summary statistics for type 1 diabetes [18] from the NHGRI-EBI GWAS Catalog [25] (study GCST005536 accessed on 08/10/21) and used these as the principal trait p values. This data was for 6670 European type 1 diabetes cases and 12,262 European controls. We used the LDAK software (https://dougspeed.com/ldak/) to obtain LDAK weights for each SNP, and defined our independent SNP set (used to fit the KDE in Flexible cFDR) as the set of SNPs given a non-zero LDAK weight (an LDAK weight of 0 means that its signal is (almost) perfectly captured by neighbouring SNPs). We used MAFs estimated from the CEU sub-population samples in the 1000 Genomes Project Phase 3 data set [26], and for any SNPs with missing MAF we randomly sampled a value from the empirical distribution of non-missing MAFs.

To define independent loci for our locus-level results, we first calculated LD between each pair of SNPs using haplotype data from the 503 individuals of European ancestry in the 1000 Genomes Project Phase 3 data set [26]. We then used PLINK's LD-clumping algorithm with a 5-Mb window and an r^2 threshold of 0.01. This conservative clumping approach sorts SNPs into ascending order of p value and then moves down the list, sequentially removing SNPs within a 5-Mb window and with $r^2 > 0.01$. The SNP with the smallest p value in the data set in each LD clump was called the "lead variant".

Validation GWAS data set

We downloaded full harmonised GWAS summary statistics for type 1 diabetes [19] from the NHGRI-EBI GWAS Catalog [25] (study GCST90013445 accessed on 08/10/21) and

used this as our validation GWAS data set. The samples in the discovery GWAS data set [18] were a subset of those in the validation data set, and so we said that a discovery validated if it's corresponding p value was smaller in [19] than [18]. The validation data set was for 16,159 European type 1 diabetes cases and 25,386 European controls.

Auxiliary data

We downloaded full harmonised GWAS summary statistics for rheumatoid arthritis (RA) [27] from the NHGRI-EBI GWAS Catalog [25] (study GCST005569 accessed on 08/10/21). We mapped each SNP in the type 1 diabetes GWAS data set to its corresponding p value for RA using genomic coordinates and rsIDs. We removed 6044 SNPs from the analysis which did not have a corresponding p value for RA.

We downloaded SNP-level annotations for all 1000 Genomes SNPs from the baseline-LD model (version 2.2) described in [28]. We extracted values for the binary annotation "DGF_ENCODE" which quantifies sites of transcription 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 [29, 30]. DGF regions (corresponding to DGF annotation values of 1) are expected to precisely map sites where regulatory factors bind to the genome [31]. We matched each SNP in the type 1 diabetes GWAS data set to its binary DGF annotation using genomic coordinates. We removed 2811 SNPs from the analysis that did not have a corresponding DGF annotation value.

We downloaded consolidated fold-enrichment ratios of H3K27ac ChIP-seq counts relative to expected background counts from NIH Roadmap Epigenomics Mapping Consortium [32] in nine primary tissues and cells relevant for type 1 diabetes (CD3, CD4+ CD25int CD127+ Tmem, CD4+ CD25+ CD127- Treg, CD4+ CD25- Th, CD4+ CD25- CD45RA+, CD4 memory, CD4 naive, CD8 memory, CD8 naive). Specifically, we downloaded the `bigWig` files, converted these to `wig` files and then to `bed` files, and then mapped each SNP in the type 1 diabetes GWAS data set to its corresponding genomic region in the `bed` files and recorded the H3K27ac fold change values in each cell type using the `bedtools intersect` utility. For SNPs on the boundary of a genomic region (and therefore mapping to two regions) we randomly selected one of the regions. We observed that the fold change values across relevant cell types were highly correlated ($r > 0.65$) and therefore averaged values across cell types to avoid iterating over highly correlated auxiliary data that is likely capturing the same functional mark. We transformed the averaged fold change values ($q := log(q + 1)$) to deal with long tails.

Implementation

We used the `fcfdr::flexible_cfdr` and `fcfdr::binary_cfdr` functions to leverage the auxiliary data with type 1 diabetes GWAS p values iteratively. We used the chromosome for which each SNP resided for the `group` parameter in `fcfdr::binary_cfdr`, and we used the estimated MAF values for the optional `maf` parameter in the `fcfdr::flexible_cfdr` function. We 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. We used an FDR threshold of $FDR \leq 3.3 \times 10^{-6}$ to call significant SNPs, which corresponded to the genome-wide significance threshold $p \leq 5 \times 10^{-8}$ (it was the maximum FDR value amongst SNPs with raw p values $\leq 5 \times 10^{-8}$ in the discovery GWAS data set). The full data and code to replicate the analysis are available from https://annahutch.github.io/fcfdr/articles/t1d_app.html.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04838-0.

Additional file 1. Further details on Binary cFDR methodology. Further details on Binary cFDR methodology, including an overview and comments on the key assumptions.

Additional file 2. Supplementary results from T1D application. Supplementary results from T1D application quantifying the relationship between the "principal p-values" (p) and the auxiliary data (q) in each iteration of the T1D application.

Acknowledgements
This work has been previously published in Anna Hutchinson's PhD thesis.

Author contributions
AH lead the simulation-based analysis and the application to type 1 diabetes, was a major contributor in writing the manuscript and created the software. JL and CW conceived the binary cFDR methodology and contributed to the software. CW supervised the project and designed the simulated-based analysis. All authors read and approved the final manuscript.

Funding
AH is funded by the Engineering and Physical Sciences Research Council (EPSRC; EP/R511870/1), GlaxoSmithKline (GSK) and the Medical Research Council (MRC; MC UU 00002/4). CW is funded by the Wellcome Trust (WT220788), the Medical Research Council (MRC; MC UU 00002/4) and supported by the NIHR Cambridge BRC (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.

Availability of data and materials
The datasets analysed during the current study are available in the github repository: https://github.com/annahutch/fcfdr. ***Availability and requirements***: Project name: fcfdr Project home page: https://annahutch.github.io/fcfdr/ Operating system(s): Tested on Linux, MacOS and Windows Programming language: R Other requirements: R ($\geq$ 3.5.0) License: MIT license Any restrictions to use by non-academics: None.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 11 February 2022 Accepted: 13 July 2022
Published online: 30 July 2022

References

1. Andreassen OA, Thompson WK, Schork AJ, Ripke S, Mattingsdal M, Kelsoe JR, Kendler KS, O'Donovan MC, Rujescu D, Werge T, Sklar P, Consortium (PGC) TPG, Groups BDaSW, Roddey JC, Chen C-H, McEvoy L, Desikan RS, Djurovic S, Dale AM. Improved detection of common variants associated with Schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. PLOS Genet.2013;9(4):1003455. https://doi.org/10.1371/journal.pgen.1003455.
2. Andreassen OA, McEvoy LK, Thompson WK, Wang Y, Reppe S, Schork AJ, Zuber V, Barrett-Connor E, Gautvik K, Aukrust P, Karlsen TH, Djurovic S, Desikan RS, Dale AM. Identifying common genetic variants in blood pressure due

to polygenic pleiotropy with associated phenotypes. Hypertension. 2014;63(4):819–26. https://doi.org/10.1161/HYPERTENSIONAHA.113.02077.
3. Tony Cai T, Sun W, Wang W. Covariate-assisted ranking and screening for large-scale two-sample inference. J R Stat Soc Ser B Stat Methodol. 2019;81(2):187–234. https://doi.org/10.1111/rssb.12304.
4. Lei L, Fithian W. AdaPT: an interactive procedure for multiple testing with side information. J R Stat Soc Ser B Stat Methodol. 2018;80(4):649–79. https://doi.org/10.1111/rssb.12274.
5. Liley J, Wallace C. Accurate error control in high-dimensional association testing using conditional false discovery rates. Biom J. 2021. https://doi.org/10.1002/bimj.201900254.
6. Andreassen OA, Harbo HF, Wang Y, Thompson WK, Schork AJ, Mattingsdal M, Zuber V, Bettella F, Ripke S, Kelsoe JR, Kendler KS, O'Donovan MC, Sklar P, McEvoy LK, Desikan RS, Lie BA, Djurovic S, Dale AM. Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. Mol Psychiatry. 2015;20(2):207–14. https://doi.org/10.1038/mp.2013.195.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
8. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Tobacco and Genetics Consortium, Bipolar Disorder Psychiatric Genomics Consortium, Schizophrenia Psychiatric Genomics Consortium, Schork NJ, Andreassen OA, Dale AM. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. 2013;9(4):1003449. https://doi.org/10.1371/journal.pgen.1003449.
9. Hutchinson A, Reales G, Willis T, Wallace C. Leveraging auxiliary data from arbitrary distributions to boost GWAS discovery with Flexible cFDR. PLOS Genet. 2021;17(10):1009853. https://doi.org/10.1371/journal.pgen.1009853.
10. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016;13(7):577–80. https://doi.org/10.1038/nmeth.3885.
11. Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics. 2016;32(4):542–8. https://doi.org/10.1093/bioinformatics/btv610.
12. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. Leveraging polygenic functional enrichment to improve GWAS power. A J Hum Genet. 2019;104(1):65–75. https://doi.org/10.1016/j.ajhg.2018.11.008.
13. Boca SM, Leek JT. A direct approach to estimating false discovery rates conditional on covariates. PeerJ. 2018;6:6035. https://doi.org/10.7717/peerj.6035.
14. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, Shukla C, Alm EJ, Hicks SC. A practical guide to methods controlling false discoveries in computational biology. Genome Biol. 2019;20(1):118. https://doi.org/10.1186/s13059-019-1716-1.
15. Liley J, Wallace C. A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. PLOS Genet. 2015;11(2):1004926. https://doi.org/10.1371/journal.pgen.1004926.
16. Du L, Zhang C. Single-index modulated multiple testing. Ann Stat. 2014;42(4):1262–311. https://doi.org/10.1214/14-AOS1222.
17. Alishahi K, Ehyaei AR, Shojaie A. A generalized Benjamini–Hochberg procedure for multivariate hypothesis testing. arXiv:1606.02386 [stat];2016.
18. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, Achuthan P, Guo H, Fortune MD, Stevens H, Walker NM, Ward LD, Kundaje A, Kellis M, Daly MJ, Barrett JC, Cooper JD, Deloukas P, Type 1 Diabetes Genetics Consortium, Todd JA, Wallace C, Concannon P, Rich SS. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47(4), 381–86. https://doi.org/10.1038/ng.3245.
19. Robertson CC, Inshaw JRJ, Onengut-Gumuscu S, Chen W-M, Santa Cruz DF, Yang H, Cutler AJ, Crouch DJM, Farber E, Bridges SL, Edberg JC, Kimberly RP, Buckner JH, Deloukas P, Divers J, Dabelea D, Lawrence JM, Marcovina S, Shah AS, Greenbaum CJ, Atkinson MA, Gregersen PK, Oksenberg JR, Pociot F, Rewers MJ, Steck AK, Dunger DB, Wicker LS, Concannon P, Todd JA, Rich SS. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. Nat Genet. 2021. https://doi.org/10.1038/s41588-021-00880-5.
20. The UK10K Consortium: The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82–90. https://doi.org/10.1038/nature14962.
21. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32(2):283–5. https://doi.org/10.1093/bioinformatics/btv546.
22. Fortune MD, Wallace C. simGWAS: a fast method for simulation of large scale case–control GWAS summary statistics. Bioinformatics. 2019;35(11):1901–6. https://doi.org/10.1093/bioinformatics/bty898.
23. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78. https://doi.org/10.1038/nature05911.
24. Leek JT, Jager L, Boca SM, Konopka T. Swfdr: science-wise false discovery rate and proportion of true null hypotheses estimation. Bioconductor version: Release. 2021(3.12). https://doi.org/10.18129/B9.bioc.swfdr.
25. ...Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. Nucleic Acids Res. 2019;47(D1):1005–12. https://doi.org/10.1093/nar/gky1120.
26. The 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
27. ...Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, Zhernakova A, Stahl E, Viatte S, McAllister K, Amos CI, Padyukov L, Toes REM, Huizinga TWJ, Wijmenga C, Trynka G, Franke L, Westra H-J, Alfredsson L, Hu X, Sandor C, de Bakker PIW, Davila S, Khor CC, Heng KK, Andrews R, Edkins S, Hunt SE, Langford C, Symmons D, Concannon P, Onengut-Gumuscu S, Rich SS, Deloukas P, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Ärlsetig L, Martin J, Rantapää-Dahlqvist

S, Plenge R, Raychaudhuri S, Klareskog L, Gregersen PK, Worthington J. High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44(12):1336–40. https://doi.org/10.1038/ng.2462.
28. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, Price AL. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017;49(10):1421–7. https://doi.org/10.1038/ng.3954.
29. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
30. ...Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Kähler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014;95(5):535–52. https://doi.org/10.1016/j.ajhg.2014.10.004.
31. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489(7414):83–90. https://doi.org/10.1038/nature11212.
32. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28(10):1045–8. https://doi.org/10.1038/nbt1010-1045.

Publisher's Note