Accurate error control in high dimensional association testing using conditional false discovery rates

James Liley, Chris Wallace

Department of Medicine, University of Cambridge

MRC Biostatistics unit, University of Cambridge

ajl88@.cam.ac.uk, cew54@cam.ac.uk

September 27, 2018

Abstract

A common statistical procedure throughout the biomedical sciences is the high-dimensional association study, in which a large number of null hypotheses are tested against a trait of interest. These analyses can generally be strengthened by incorporating an informative covariate, often a set of p-values for tests against some related 'conditional' trait. One method to incorporate such a covariate estimates a quantity called the conditional false discovery rate (cFDR) for each observation. The cFDR is the posterior probability of association given p-value thresholds for both the principal and conditional trait, and can be used quantify evidence that the observation is associated with the principal trait. The cFDR is powerful and intuitive, but the two-dimensional nature of its construction means it is difficult to control the overall false- discovery rate (FDR), and existing attempts to do so are over-conservative. We propose a new method, based on identifying maps from the unit square to the unit interval defined by the estimated cFDR, and splitting observations so that each map is independent of the observations it is used to test. This allows theoretical FDR control at arbitrarily low or high numbers of hypotheses and more than doubles the potential improvement in power over unconditional analyses compared to the existing FDR-controlling method. We demonstrate the new method on two genome-wide association studies in autoimmune disease. This method improves the power of the cFDR

while maintaining rigorous FDR control, allowing it to be used in the discovery phase of 'omics' studies with the same confidence as the Benjamini-Hochberg procedure and improving its applicability to different data types.

1 Introduction

In the 'omics' approach to biology, a large number n of descriptive variables are considered in the analysis of a biological system, intended to provide a near-exhaustive characterisation of the system under consideration. Often, only a small proportion of the investigated variables are associated with the behaviour of the system, and we seek to identify this subset of variables, along with the magnitude and direction of their associated effect sizes. Typically, either a series of null hypotheses of non-association $H_0(i)$, $i \in \{1..n\}$ are assumed, with some of them being rejected (frequentist setting) or a series of Bernoulli-distributed indicator variables $H(i)$, $i \in \{1..n\}$ governing association are considered, seeking to infer the probabilities $\lambda_i = Pr(H(i) = 1)$ (Bayesian setting). In -omics studies using frequentist multiple hypothesis testing, control of the type-1 error rate (usually quantified by the family-wide error rate or false discovery rate (FDR)) is of critical importance, as such studies are often used as 'discovery' type procedures and follow-up of false positives is expensive.

In many areas of high-dimensional hypothesis testing, additional information may be available in the form of an external covariate, which assigns a numerical value to each hypothesis which has different (unknown) distributions amongst associations and non-associations. By attempting to infer these distributions, incorporation of such covariates into hypothesis testing can improve power to detect associations. An approach by Ferkingstad *and others* (2008) uses a Bayesian setting in which λ_i depends on the data through a frequentist p-value for the ith hypothesis and a prior probability of association modulated according to an external covariate in an unsupervised way. The (continuous) covariate observations are divided into bins, and the p-value distribution is modelled as a mixture of uniform and beta distributions, fitting separate models for observations in each bin. Priors on the difference between covariate effects in neighbouring bins are used to encourage smoothness across bins. A frequentist approach by Ignatiadis *and others* (2016) assigns a weight to each hypothesis under consideration depending on the covariate value for that hypothesis. The mapping of covariate values to weights is refitted separately for each hypothesis, leaving that hypothesis out of the data used to determine the mapping (analogous to 'leave-one-out'

cross-validation). A form of the Benjamini-Hochberg procedure can then be used on resultant weighted observations. A third approach by Zablocki *and others* (2014) parametrises the joint distribution of the covariate and p-values, allowing continuous modulation of the probability of association with changes in the covariate.

All methods enable improvement in power compared to non-covariate adjusted testing. No method is clearly dominant in every case; the general problem can take many forms according to the distribution of p-values, the form and distribution of the covariate, and the number of variables, so the best method will generally be situation-dependent. Furthermore, comparison of methods is generally difficult as multiple measures of type-1 error rates are used in covariate-adjusted analysis; Ferkingstad *and others* (2008) uses a posterior probability, Ignatiadis *and others* (2016) controls FDR, and Zablocki *and others* (2014) controls a custom measure.

A useful alternative method is the conditional false discovery rate (cFDR) (Andreassen *and others*, 2013, 2014). In most applications of cFDR, the covariate is a set of association statistics for a second study considering a related set of hypotheses; for instance association between a set of genetic variants and an epidemiologically related trait. The cFDR is defined as a posterior probability of association given p-value thresholds for both the primary and conditioned trait. Estimation of the cFDR requires only consideration of the cumulative density function (CDF) of the covariate distribution and avoids categorising (binning) covariate values, minimising parametric assumptions.

In this paper, we begin in Section 2 by introducing the usual estimator for cFDR and the current method to control the FDR of the hypothesis-testing procedure, also discussing its shortcomings. In Section 3 we introduce two new theoretical methods for controlling overall FDR; the main new idea is described in Section 3.2. These methods have the advantages of greater theoretical confidence, better performance, and ready adaptation to other cFDR estimators (although we only consider one estimator in this paper). We evaluate the performance of FDR-controlling methods on a wide range of simulated datasets in Section 4, and based on these results, we make recommendations on the best methods to use in Section 5. Finally, we demonstrate the usage of these methods on genome-wide association study (GWAS) datasets in Section 6. For the reader interested only in the application of these methods, we recommend a review of section 5.

2 Overview of existing method, shortcomings, and proposed solutions

Assume that we have results from n pairs of hypothesis tests against two outcomes in the form of a set S of generate bivariate p-values $S = (p_i, q_i)$, $i = 1..n$, considered instances of random variables (P, Q). Denote H^P as a random-valued hypothesis indicator for association in the study generating P (which we assume is the study we are primarily interested in) and assume $H^P \sim \text{Bernoulli}(\pi_1)$, and that the tri-variate random variable (P, Q, H^P) satisfies $(P, Q)|H^P \sim (1 - H^P)f_0 + H^P f_1$, where f_0 is a known PDF and f_1 has an unknown PDF. We will use the shorthand H_0^P to represent the null hypothesis event $H^P = 0$, and $H_0^P(i)$ to represent the ith null hypothesis. We assume (for the moment) association testing is done in such a way that $P \perp\!\!\!\perp Q|H_0^P$.

The cFDR is defined as a function of two values $p, q \in [0, 1]$ as

$$cFDR(p, q) = Pr(H_0^P|P \leq p, Q \leq q) \tag{1}$$

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

Quantity (1) can be geometrically considered as a Bayesian posterior (Bernoulli) of H^P given that observed p-values P, Q fall in the rectangle $[0, p] \times [0, q]$. It can be used to construct a frequentist hypothesis-testing procedure 'levering' on observations of Q, analogous to the Benjamini-Hochberg procedure in a bivariate context.

The cFDR can be estimated using a multiset X of observations (p_i, q_i) of (P, Q), where X may or may not be S. The denominator of (2) can be empirically estimated using the generally consistent empirical CDF

$$Pr(P \leq p|Q \leq q) \approx \frac{|\{i : p_i \leq p, q_i \leq q, (p_i, q_i) \in X\}|}{|\{i : q_i \leq q, (p_i, q_i) \in X\}|}$$

Under the assumption $P \perp\!\!\!\perp Q|H_0^P$, we have $Pr(P \leq p|H_0^P, Q \leq Q) = Pr(P \leq p|H_0^P) = p$. Conservatively approximating $Pr(H_0^P|Q \leq q) \approx 1$ gives the estimate

$$\widehat{cFDR}_X(p, q) = p\frac{|\{i : p_i \leq p, (p_i, q_i) \in X\}|}{|\{i : p_i \leq p, q_i \leq q, (p_i, q_i) \in X\}|} \tag{3}$$

In subsequent sections it will be important to note the explicit dependence on X, the set

of points used to estimate the cFDR. X may contain points which coincide; that is, if $X = S + (p_1, q_1)$, we would take $|\{i : p_i = p_1, q_i = q_1, (p_i, q_i) \in X\}| \geq 2$ (using the conventions '+' and '-' for the multiset X in place of $\cup$, $\setminus$). In this work, we only consider estimator (3), but other estimators are possible and most of the following results can be analogously used with other estimators.

In the existing hypothesis-testing method, for each hypothesis $H^P(i)$ with corresponding p-values p_i, q_i, he null hypothesis $H_0^P(i)$ may be rejected either using the simple criteria

$$c\widehat{FDR}_S(p_i, q_i) \leq \alpha \tag{4}$$

for some α, or using the condition (as per Andreassen *and others* (2013))

$$\exists\, p' \geq p_i : c\widehat{FDR}_{S+(p',q_i)-(p_i,q_i)}(p', q_i) \leq \alpha \tag{5}$$

that is, if the $c\widehat{FDR}$ *would* have been less than α if the observed p-value p_i had been larger. If the fixed value q' is substituted for q_i, and $H_0^P(i)$ is rejected whenever $q_i < q'$ and criterion (5) holds, this procedure is the Benjamini-Hochberg procedure on the set of p-values p_i for which the corresponding q_i are less than q', and amongst these variables, the FDR is controlled at at α.

However, criterion (5) without fixed $q_i = q'$ does not control the overall FDR at α (and nor does criterion (4)) and the true FDR may be arbitrarily larger than α (Liley and Wallace, 2015). A heuristic explanation (illustrated in figure 1) can be made by considering the case when $p_i, q_i \approx 0, 0$ whenever $H_0^P(i)$ does not hold. Suppose that for a series of hypothesis tests giving rise to pairs of p-values p_i, q_i as above, we plot the values p_i, q_i in the unit square. The value $cFDR(p, q)$ (for which $c\widehat{FDR}_S(p, q)$ is an estimate) is effectively the expected proportion of points in the rectangle $M = [0, p] \times [0, q]$ for which H_0^P holds.

Suppose $c\widehat{FDR}(p, q) \approx \alpha$ and $P, Q|(H_1^P$ or $H_1^Q) \sim (0, 0)$ (so all associations with either P or Q have $P = Q = 0$). The number of points in M for which H_0^P holds is roughly proportional to the area of M, and the number of points in M for which H_0^P does not hold is constant. The locus of all points for which $c\widehat{FDR}_S(p, q) = \alpha$ converges to a larger region L, which turns out to be a hyperbola. While L contains the same number of non-null points as M, it is expected to contain more points for which H_0^P holds as it is larger than M. The FDR of the procedure (which is asymptotically the expected proportion of points in L for

which H_0^P holds) is consequently arbitrarily higher than that in M. We include a formal exposition of that argument in appendix 1.2 using the current notation.

This difficulty in controlling overall FDR is a major shortcoming of the cFDR method, particularly as the degree by which the FDR of the procedure exceeds α increases as the degree of shared association between P, Q increases. In (Liley and Wallace, 2015) we proposed a method for controlling the overall FDR involving consideration of the region of the unit square for which $c\widehat{FDR}_{S+(p,q)}(p,q) \leq \alpha$ and a rectangle within this region (which we will later call 'method 1'), but this method is heuristic and asymptotically conservative (underestimating the true FDR), where 'asymptotically' is taken to mean as n, the number of observations of P, Q, tends to infinity.

3 False discovery rate control

We begin by reviewing the definition of FDR (Benjamini and Hochberg, 1995) in the current context and introduce relevant notation, then discuss our previous method to attempt to control FDR (Liley and Wallace, 2015), before introducing improved methods.

Pairs of p-values (p_i, q_i) correspond to points in the unit square. For a region A of the unit square, define the random variables $R(A)$, $V(A)$, $R(A) - V(A)$ as the total number of variables, number of null variables (for which H_0^P is true), and number of non-null variables for which observed p-values fall in A. For some procedure giving rise to a (possibly random-valued) rejection region A, the FDR is a functional of the procedure defined as $E\left(\frac{V(A)}{R(A)}|R(A) > 0\right) Pr\left(R(A) > 0\right)$ (Benjamini and Hochberg, 1995; Storey *and others*, 2003)

In order to control this quantity, we need to consider the chance of points falling into a given region under H_0^P. We will generally assume that the PDF f_0 of $P, Q|H_0^P$ is known (its estimation is discussed in Section 3.5), and define

$$v(A) = \iint_A f_0(x,y)dxdy = Pr\left(P, Q \in A|H_0^P\right) \tag{6}$$

which is random-valued if the region A is random-valued.

Consideration of the overall FDR associated with cFDR-based procedures requires consideration of the regions of the unit square for which (roughly) the estimated cFDR is less than some threshold. This is complicated by the dependence of the cFDR estimate (3) on

the observed points X. We denote these regions $L_X(\alpha)$, defined formally as

$$L_X(\alpha) = \{(p,q) : \exists p' \geq p : c\widehat{FD}R_{X+(p',q)}(p',q) \leq \alpha\} \tag{7}$$

that is, for each q, the set of values of p that would lead to rejection of H_0^P under rejection procedure (5) if (p,q) were added to X. Note that $\alpha \leq \alpha' \Leftrightarrow L_X(\alpha) \subseteq L_X(\alpha')$. We may thus associate each point $(p,q) \in [0,1]^2$ with the smallest region L_X which contains it, or with the corresponding 'curve' derived from replacing '$\leq \alpha$' with '$= \alpha$' in (7), although this curve may have nonzero area.

An ideal cFDR-based hypothesis testing procedure aiming to control the FDR at α would define a value α_0 so that rejecting H_0^P when $(p_i, q_i) \in L_S(\alpha_0)$ leads to overall FDR control at α, where the value α_0 may depend on observations in the set S and on α. Rejection procedure (5) is slightly different, corresponding to $(p_i, q_i) \in L_{S-(p_i,q_i)}(\alpha_0)$.

Setting $V\{L_S(\alpha_0)\} = \left|\left\{i : (p_i, q_i) \in L_S(\alpha_0), H_0^P\right\}\right|$ (the number of true nulls in $L_S(\alpha_0)$) and $R\{L_S(\alpha_0)\} = |\{i : (p_i, q_i) \in L_S(\alpha_0)\}|$ (the total number of data points in $L_S(\alpha_0)$), we thus seek to find

$$\alpha_0 : E\left(\frac{V\{L_S(\alpha_0)\}}{R\{L_S(\alpha_0)\}}\middle| R\{L_S(\alpha_0)\} > 0\right) Pr\left(R\{L_S(\alpha_0)\} > 0\right) \leq \alpha \tag{8}$$

Computation of quantity (8) is complex due to the dependence of regions L_S on values p_i,q_i. A more tractable approximation of this quantity can be made by assuming that this dependence is negligible so that regions $L_S(\alpha_0)$ take the fixed values $L(\alpha_0)$ independent of S, and assuming that

$$\begin{aligned} &E\left(\frac{V\{L_S(\alpha_0)\}}{R\{L_S(\alpha_0)\}}\middle| R\{L_S(\alpha_0)\} > 0\right) Pr\left(R\{L_S(\alpha_0)\} > 0\right) \\ &\approx E\left(\frac{V\{L(\alpha_0)\}}{R\{L(\alpha_0)\}}\middle| R\{L(\alpha_0)\} > 0\right) Pr\left(R\{L(\alpha_0)\} > 0\right) \end{aligned} \tag{9}$$

The critical difference between the two sides of (9) are that on the left, the regions L are random variables depending on the set of observed data $S = \{p_i, q_i\}$, and on the right they are assumed to be fixed whatever S is. A useful rationalisation of the right hand side is to imagine that we set regions L_S using an initial set of observations $S = \{p_i, q_i\}$ of P, Q. The right-hand side of (9) then corresponds to expectations across a *new* set of observations $S' = \{p'_i, q'_i\}$ of P, Q, using the regions L_S from the *old* observations.

For α_0 held constant, the approximation (9) is asymptotically accurate as $n \to \infty$, since the regions L converge almost surely to fixed regions independent of observations S. The problems with assumption (9) are discussed in Section 3.2.

3.1 Original method for FDR control (method 1): largest-rectangle

Our original proposal (henceforth method 1) in (Liley and Wallace, 2015) was to directly estimate α_0 by considering a rectangle $M(\alpha_0)$ with vertices $(0,0)$, $(p^*,0)$, $(0,q^*)$ and (p^*,q^*) contained within a region $L_S(\alpha_0)$ for some α_0.

Approximation (9) is used, assuming independence of $L_S(\alpha_0)$ and S. Under the assumptions on P, Q stated in the introduction (dropping arguments S, α_0 on L for brevity), this allows use of the identity (Storey *and others*, 2003):

$$E\left(\frac{V(L)}{R(L)}|R(L)>0\right) = \frac{E\{V(L)\}}{E\{R(L)\}} = Pr(H_0^P|(p,q)\in L)$$

Using this twice and recalling definition (6), we have

$$\begin{aligned}
E\left(\frac{V(L)}{R(L)}|R(L)>0\right) &= \frac{E\{V(L)\}}{E\{R(L)\}} && (10)\\
&\leq \frac{E\{V(L)\}}{E\{R(M)\}} && (11)\\
&= \frac{E\{V(L)\}}{E\{V(M)\}}\frac{E\{V(M)\}}{E\{R(M)\}}\\
&= \frac{v(L)}{v(M)}\frac{E\{V(M)\}}{E\{R(M)\}}\\
&= \frac{v(L)}{v(M)}E\left(\frac{V(M)}{R(M)}|R(M)>0\right)\\
&= \frac{v(L)}{v(M)}Pr(H_0^P|(p,q)\in M)
\end{aligned}$$

The value $\alpha^* = c\widehat{FDR}_{S+(p^*,q^*)}(p^*,q^*)$ is an estimate for $Pr(H_0^P|(p,q)\in M)$ which, as shown in appendix 1.1, is almost surely conservative as $n\to\infty$ (under reasonable conditions on f_0, f_1). The maximum value $c\widehat{FDR}_{S+(p^*,q^*)}(p^*,q^*)$ can take for $(p^*,q^*)\in L(\alpha_0)$ is approximately α_0 (since $L(\alpha_0)\approx L_S(\alpha_0)$). Finally, the RHS of approximation (9) is less than or equal to the LHS of equation (10). Thus we may choose any α_0 such that

$$\frac{v\{L(\alpha_0)\}}{v\{M(\alpha_0)\}}\alpha_0 \leq \alpha$$

It is clearly desirable to choose the rectangle $M(\alpha_0)$ for which $v\{M(\alpha_0)\}$ is maximised (effectively the largest rectangle) allowing the largest α_0 possible.

The approximation is attractive in relating an FDR bound to the cFDR cutoff α_0. However, it is conservative (and thus inconsistent) due to the approximation in inequality (11).

3.2 New method for FDR control - method 2, naive Benjamini-Hochberg

Much of the difficulty in FDR control stems from the lack of a natural ordering of points (p_i, q_i) on the unit square. This is fundamentally why a procedure rejecting H_0 when $c\widehat{FDR} < \alpha$ does not control FDR at α while the analogous one-dimensional Benjamini-Hochberg procedure does. In this section, we propose a function which maps the unit square onto the unit interval, allowing an ordering of data points corresponding to $c\widehat{FDR}$ values. With the regions $L_S(\alpha) = L(\alpha)$ assumed independent of observations S (as per (9)), define

$$f(p,q) = \min\{\alpha : (p,q) \in L(\alpha)\} = \min_{p \geq p_i}\{c\widehat{FDR}_{S+(p,q_i)}(p_i, q_i)\} \tag{12}$$

$$g(p,q) = v\left[L\{f(p,q)\}\right] \tag{13}$$

This function g maps the unit square onto the unit interval. In particular

$$g(p,q) = Pr\left\{f(P,Q) \leq f(p,q) | H_0^P\right\}$$

so $g(p,q)$ can be considered a p-value against H_0^P using $f(p,q)$ as a test statistic. This observation is at the core of the remainder of this paper. Since the values $g(p,q)$ are computed as values of the function $v(.)$ and are similar to p-values, we will generically refer to them as 'v-values'.

If we generate the v-values $g(p_1,q_1), g(p_2,q_2), ...,$ and rank them to get $g(p_{(1)}, q_{(1)})$, $g(p_{(2)}, q_{(2)})$, ..., we may then use the Benjamini-Hochberg procedure to control the FDR at α, finding the largest i such that

$$g(p_{(i)}, q_{(i)})\frac{n}{i} \leq \alpha \tag{14}$$

and rejecting H_0^P for the hypotheses corresponding to (1), (2), ...(i). This comprises our

simplest new method for (attempted) FDR control, which we will denote method 2.

The control of FDR cannot be guaranteed using Benjamini and Hochberg's theorem in method 2 due to the falsity of assumption (9) of independence of $L(\alpha)$ on values p_i, q_i. If $L = L_S$, then the v-values $g(p_i, q_i)|H_0^P$ are not necessarily uniformly distributed, since the function f depends on these points.

Using our standard estimator $c\widehat{FD}R_S$ (equation (3)), the v-values $g(p_i, q_i)|H_0^P$ may tend to be smaller than values $g(p, q)|H_0^P$ for random (p, q) not used in the construction of L_S, consequently leading to loss of FDR control. This is demonstrated in figure 2, and the effect is explained in appendix 1.3.

We now present two more methods for FDR control which aim to overcome this problem.

3.3 Method 3: fold-based FDR control

A simple way to overcome the problem of defining the function f (equation (12)) using the points (p_i, q_i) at which it is evaluated is by an analogy to N-fold cross-validation (henceforth method 3). In this method, we split variables into folds $k \in 1..N$, and for each fold k, determine regions $L_{S-(\text{fold } k)}$ leaving fold k out. The values $f_k(p_i, q_i)$ for $i \in$ fold k (where the definition of f, g in equation (12) depend on k) are then determined using the regions $L_{S-(\text{fold } k)}$ instead of L_S. We compute v-values $g_k(p_i, q_i)$ for i in fold k.

We can either control FDR strongly within each fold (method 3a), or overall (method 3b). For method 3a, v-values $g_k(p_i, q_i)$ are sorted to give $g_k(p_{(1)}, q_{(1)}) \leq g_k(p_{(2)}, q_{(2)}) \leq ... \leq g_k(p_{(n_k)}, q_{(n_k)})$ where n_k is the number of variables in fold f. We then compute

$$(i^k_{max}) = \max\left\{i : g_k(p_{(i)}, q_{(i)})\frac{n_k}{i} \leq \alpha\right\}$$

and reject H_0^P for $i \in (1), (2), ...(i^k_{max})$, repeating this procedure for every fold. Since the function f within each fold is defined independently of the points at which it is evaluated the FDR within each fold is strongly controlled at α. This controls the FDR in each fold, and the mean FDR across folds, but not the overall FDR (see appendix 1.5 for more detail)

In method 3b, variables are subdivided *only* for the computation of values $f_k(p_i, q_i)$ and v-values $g_k(p_i, q_i)$. The v-values are then pooled across all k and used in criterion (14). An advantage of method 3a is that observations $g_k(p_i, q_i)$ are independent within-fold. In methods 3b (and later method 4), values are not completely independent. If i and j are

in different folds ki and kj, v-values $g_{ki}(p_i, q_i)$, $g_{kj}(p_j, q_j)$ are generated using overlapping datasets, and will be correlated. We show later that this does not lead to a measurable effect on FDR in realistic datasets, although it theoretically could do.

3.4 Method 4: Leave-one-out approach

Our final method (method 4) is an extreme version of method 3b. We compute different mappings of the unit square to the unit interval for each hypothesis individually. Specifically, for each hypothesis, we evaluate the regions

$$L_{S-(p_i,q_i)}(\alpha) \tag{15}$$

for all α, and set

$$f_i(p_i, q_i) = \min\left\{\alpha : (p, q) \in L_{S-(p_i,q_i)}(\alpha)\right\} = \min_{p \geq p_i}\{c\widehat{FDR}_{S-(p_i,q_i)+(p,q_i)}(p_i, q_i)\}$$

$$g_i(p_i, q_i) = v\left[L_{S-(p_i,q_i)}\{f_i(p, q)\}\right]$$

where the subscript i on f_i and g_i indicates the specificity of the function to hypothesis i. Since (p_i, q_i) is not used in the definition of the regions in (15) the set of v-values $g_i(p_i, q_i)$, $i \in 1..n$, behave like p-values against H_0^P. The v-values may be correlated, but since the pairwise correlation will generally be positive (since the influence of variables $\neq i, j$ will be the same for the generation of $g_i(p_i, q_i)$, $g_j(p_j, q_j)$) no adjustment is needed to the Benjamini-Hochberg procedure (Benjamini and Yekutieli, 2001). The effect of this leave-one-out procedure in the avoidance of the 'points on stalks' effect in figure 2 is shown in figure 3. This method corresponds more closely to rejection procedure (5) than does method 2, due to the use of the curves $L_{S-(p_i,q_i)}$ rather than curves L_S

If the random error between observations (p_i, q_i) may be correlated (for example, correlations due to linkage disequilibrium in genome-wide association studies) then method 4 is not appropriate. If observations (p_i, q_i), and (p_j, q_j) are correlated, then removing observation (p_i, q_i) does not lead to independence of $L_{S-(p_i,q_i)}$ on (p_i, q_i).

In the case where variables can be separated into 'blocks' of possibly-correlated observations (as is typically the case in genome-wide association studies) then methods 3a or 3b may be used, with variables separated into folds so as to not split any blocks of correlated

variables over different folds. If necessary, the resultant v-values used in the Benjamini-Hochberg procedure can be adjusted for possible correlation in the usual way Benjamini and Yekutieli (2001).

Algorithms for methods 1-4 are shown in appendix 1.4.

3.5 Estimation of $P, Q|H_0^P$

Estimation of values $v(L)$ as per equation (6) requires approximation of the distribution $P, Q|H_0^P$. We have $P|H_0^P \sim U(0,1)$ and generally assume $P \perp\!\!\!\perp Q|H_0^P$. To estimate of the density $Pr(Q \leq q|H_0^P)$, we suggest assuming that $Q|H_0^P$ has approximately the same distribution as $Q|P \geq \frac{1}{2}$, of the form

$$Z_Q|P \geq \frac{1}{2} \sim \begin{cases} |N(0,1)| & \text{prob} = \pi_0 \\ |N(0,\sigma^2)| & \text{prob} = 1 - \pi_0 \end{cases} \tag{16}$$

where $Z_Q = -\Phi^{-1}(Q/2)$. Parameters π_0 and σ can be estimated using an expectation-maximisation algorithm, using the values q_i for which the corresponding p_i is $\geq 1/2$. If P, Q are dependent under H_0^P, an alternative distribution must be used for computing $v(L)$. If dependence between P, Q arises, for example due to shared samples in the corresponding studies, the assumption is made that

$$Z_P, Z_Q|H_0^P \sim \pi_0 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) + (1-\pi_0)N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & \sigma^2\end{pmatrix}\right) \tag{17}$$

The value $Pr(P \leq p|H_0^P, Q \leq q)$ in equation (2) can be replaced with the estimate described in (Liley and Wallace, 2015), although this is not essential if methods 3b or 4 are used for FDR control.

Since regions L widen with lower Q (see figures 2, 3), an approximation F' of the true CDF F of $Q|H_0^P$ will lead to a conservative estimate of $v(L)$ (that is, an overestimate of the integral) if

$$F'(q) \geq F(q) \;\; \forall\, q \in (0,1) \tag{18}$$

Conservatism is usually preferable, since it ensures control of the type-1 error rate (at the cost of type-2 error). We did not specifically attempt to ensure the truth of (18) in our

simulations, but it could be ensured by forcing an overestimate of dispersion (analogous to σ) and proportion of non-null variables (analogous to π_0)

4 Assessment of performance through simulations

4.1 Simulation design

In order to assess the performance of each method above, we simulated a series of datasets and estimated the FDR of each method in each dataset. In each simulation, we generated a set of values p_i, q_i, $i \in 1..n$ from random variables P, Q. We considered an extensive range of underlying parameters governing the distributions of P, Q. Details of the simulation strategy are shown in appendix 1.6. In general, each dataset contained n_1^p associations uniquely in the study P, n_1^q association uniquely in Q, and n_1^{pq} with both. We used various alternative distributions (normal, t (3df) and Cauchy) for simulating distributions test statistics (which we converted to p-values) under the alternative to examine robustness to the normality assumption in (16).

We analysed each simulated dataset using each of the four methods to attempt to control FDR at $\alpha = 0.1$. We also performed a standard Benjamini-Hochberg procedure (which we will abbreviate as BH) on the values of P (without considering the values of Q) as a control. Methods 1-4 assume the distribution $P, Q|H_0^P$ is known, in order to generate v-values. We estimate $P, Q|H_0^P$ assuming $P|H_0^P \sim U(0,1)$, $Q|H_0^P \approx Q|P > \frac{1}{2}$, $P \perp\!\!\!\perp Q|H_0^P$, as per Section 3.5. We also computed results using the true distribution under which P, Q were simulated, and compared results from true and estimated distributions.

For each simulation, we computed the 'false discovery proportion' (FDP) and 'true detection rate' as:

$$FDP = \begin{cases} 0 & \text{if no rejections} \\ \dfrac{\text{number of falsely rejected null hypotheses}}{\text{total number of rejections}} & \text{if} \geq 1 \text{ rejection} \end{cases}$$

$$TDR = \begin{cases} 0 & \text{if no rejections} \\ \dfrac{\text{number of correctly rejected null hypotheses}}{\text{true number of associations}} & \text{if} \geq 1 \text{ rejection} \end{cases}$$

so $FDR = E(FDP)$ and (for a randomly-chosen association) Power $= E(TDR)$. We examined how FDP and TDR varied with the value of $n_1^p + n_1^{pq}$, the total number of non-

null hypotheses for the study of interest, separately considering the cases $n_1^p + n_1^{pq} = 0$ (where FDP is either 0 or 1) and $n_1^p + n_1^{pq} > 0$.

We conducted both numerical and graphical analyses of data. We used binomial tests of proportion when assessing FDR control at $n_1^p + n_1^{pq} = 0$, t-tests when assessing FDR control at $n_1^p + n_1^{pq} > 0$, and paired Wilcoxon rank-sum tests when comparing FDR/TDR values between different methods. In assessing FDR control all tests had 90% power to detect a 10% error in FDR at $p < 0.05$. In comparing FDR or TDR between methods, all Wilcoxon tests had 90% power to detect a difference between Pr(FDP/TDR in method A > FDP/TDR in method B) and Pr(FDP/TDR in method B > FDP/TDR in method B) of 5% at $p < 0.05$. P-values are not shown but 'difference' refers to $p < 1 \times 10^{-3}$ and 'no difference' refers to $p > 0.1$.

4.2 Comparison of false-discovery rate control methods

Figure 4 shows the variation in FDP of each FDR-controlling method (BH, 1-4) applied to $c\widehat{FDR}$ at varying values of $n_1^p + n_1^{pq}$, the total number of true associations.

Methods 2 and 3 both failed to control the FDR at $\alpha = 0.1$ when $n_1^p + n_1^{pq} = 0$ and when $n_1^p + n_1^{pq} > 0$. Methods 1, 3b, and 4 all controlled the FDR at $\alpha = 0.1$ (mean FDP indistinguishable from 0.1 when $n_1^p + n_1^{pq} = 0$, mean FDP < 0.1 when $n_1^p + n_1^{pq} > 0$). The FDR of method 1 was lower than than the FDR of BH, indicating that it over-controlled FDR. The FDRs of methods 3b and 4 were indistinguishable from those of BH.

Figure 5 shows a comparison of within-fold FDRs for method 3 against $n_1^p + n_1^{pq}$. Within each fold, FDR is controlled at α, and FDRs are equivalent to those of the B-H procedure.

4.3 Power by FDR-controlling method

We compared the power of each successful FDR-controlling method: that is, BH, method 1, method 3b, and method 4.

All methods were more powerful than BH. Methods 3b and 4 (new) were both more powerful than method 1 (original), but there was no difference in power between them. Figure 6 shows the results from this comparison. Power in methods 3b and 4 is visually very similar.

4.4 Estimation of null distribution

We compared the observed FDRs in which FDR-controlling methods used the true distribution of $Q|H_0^P$ with the observed FDRs using the estimated distribution of $Q|H_0^P$.

The agreement between FDRs using the true and approximate distributions was reasonable (figure 7) even when a Cauchy distribution was used to generate effect sizes. More importantly, observed FDRs for which an approximated distribution were used were universally lower than observed FDRs for which the true distribution was used, meaning that FDR control was maintained across all simulations.

5 Conclusions and recommendations

Based on our theoretical and simulated findings, we recommend the use of method 3b or 4 to control FDR at a desired level α. If observations p_i or q_i are correlated across i, method 3 is preferable, and folds should be chosen to minimise correlations between folds. Method 3a can be used to control within-fold FDR, if this is sufficient, or method 3b can be used to control FDR overall. Methods are detailed in appendix 1.4

We summarise FDR control of each method in table 1. The existing FDR control method (method 1) asymptotically controls the FDR at α (Section 3.1) but has lower power than other methods due to over-control of the FDR (Section 4.2, figure 4). Method 2 will control FDR if approximation (9) holds; however, as shown by simulation, this approximation is too strong and FDR control fails if we rely on it; FDR control methods 2 and 3a do not control overall FDR at α (figure 4, Sections 3.2, 3.3, figure 5).

Even with independence between the observations p_i and between the observations q_i, the v-values in methods 3b and 4 are slightly correlated, since the datasets used to determine L-regions for each v-value are overlapping. If we ignore this residual correlation, methods 3b and 4 are theoretically guaranteed to control FDR. It is possible to conservatively adjust for correlation in the Benjamini-Hochberg method (Benjamini and Yekutieli, 2001) but from our simulations, the correlation does not have any measurable effect on the FDR and is of negligible importance. All theoretical FDR control relies on knowing the distribution of $Q|H_0^P$; however, as we show in simulations, a simple empirical approximation of this distribution generally retains FDR control. Methods 3b (fold-based) and 4 (leave-one-out) are theoretically guaranteed to control FDR if we discount possible residual correlation

between v-values.

Methods 3a, 3b, and 4 are generalisable to other estimators of cFDR, with the subscript X in equation (3) and elsewhere more generally indicating that the cFDR estimate only depends on points in X. Method 1 relies on asymptotic conservatism of the estimator, so will work (asymptotically) if $\lim_{n\to\infty} cFDR(p,q) - \widehat{cFDR}(p,q) \geq 0$.

6 Demonstration of method on GWAS data

In this section, we demonstrate the performance of the cFDR method on a real-world GWAS dataset. In a GWAS, multiple null hypotheses are tested in a retrospectively sampled cohort of disease cases and controls, with each null hypothesis defined as independence between a genetic variant and disease. As sample sizes have grown, GWAS have rejected increasing numbers of null hypothesis suggesting the current sample sizes are still limited in power to detect all but the strongest effects. In juvenile idiopathic arthritis (JIA), fewer than 3000 cases were available for a recent GWAS (Hinks *and others*, 2013). Genetic overlap of JIA with another autoimmune disease, type 1 diabetes (T1D), has been reported (Onengut-Gumuscu *and others*, 2015), and we sought to lever information in the larger T1D GWAS ($\sim$8000 cases). The variants (corresponding to hypotheses) are physically located on one of 22 chromosomes (excluding sex chromosomes).

We considered JIA as the 'principal' phenotype under investigation (that is, P/p) and T1D as the 'conditional' phenotype (Q/q). SNPs were excluded on the basis of quality summaries calculated on 12,888 control samples genotyped on the ImmunoChip: call rate less than 99%, minor allele frequency less than 0.02, or deviation from Hardy-Weinberg equilibrium ($|Z|> 5$). We excluded the MHC region (chr6, 24.5- 34.8 Mb, build GChr37) as it is known to harbour strong effects disproportionate to the rest of the genome (The Wellcome Trust Case Control Consortium, 2007). Since controls were shared between the two studies, we assumed a distribution of $P, Q|H_0^P$ of the form (17). We used an FDR threshold of 1.6×10^{-5}, which if used in an unconditional analysis with the B-H procedure would reject H_0^P whenever $P < 5 \times 10^{-8}$, the standard GWAS significance threshold. We emphasise that using thresholds on p-values and thresholds on FDRs control different measures of type-1 error, with the former typical in GWAS. We used method 3b, with 22 folds corresponding to different chromosomes to remove correlation effects from linkage disequilibrium.

Data was of sufficient quality in 99781 variants. Using the unconditional analysis, H_0^P

could be rejected for 305 variants. Using the cFDR, H_0^P could be rejected for 395 variants (figure 8), including variants in three regions for which no null hypothesis was rejected in unconditional analyses using the JIA p-value alone. One region had variants which reached significance by p-value, but no variants reaching significance by cFDR. One of the three additional regions (rs7808122 on chromosome 7) neighboured a variant which was found associated in denser (imputation-based) genotyping of the same samples (Hinks *and others*, 2013), while another (rs79893749 on chromosome 3) only reached "suggestive association" in that study (p=1×10^{-7}) (Hinks *and others*, 2013) and could be considered for follow-up replication studies. The third additional variant (rs74956615 on chromosome 19) neighboured a variant not in our dataset (not passing quality control criteria for both diseases) for which the JIA p-value reached significance ($p < 5 \times 10^{-8}$) in the original study. A Manhattan plot of p- and v- values is shown in Supplementary figure 1.

This section is intended to be demonstrative rather than exploratory and we do not claim that putative cFDR-based discoveries are novel. Since this demonstration only involves a single dataset, any judgement about the appropriateness of cFDR to a particular application should be made using the results of the simulations above, rather than the performance on this dataset alone.

We also do not endorse the universal use of any FDR threshold (including the one used above) in GWAS; a study-appropriate threshold should be determined *a priori* by consideration of the relative cost of type 1 and 2 errors. In the absence of formal replication or other follow-up procedures (which are standard in GWAS), the associations in this section should not be considered as confirmed.

7 Discussion

We present an improvement to the conditional false discovery rate method. Our new methods essentially involve computing an analogy of the p-value (broadly 'the probability under H_0^P of an observed cFDR value being less than that observed') and then using the Benjamini-Hochberg procedure. Our new methods for FDR control (method 3b/4) are markedly stronger than our previous proposed approach (method 1). Method 1 is complicated, lacks theoretical guarantee, and over-controls FDR, allowing only a small increase in power over an analysis with no informative covariate. As well as having stronger theoretical justification, the improvement in power over BH with method 3b or 4 is two or more times as

large as that of method 1 across all tested parameters. Methods 3b/4 enable the cFDR to be used definitively in the discovery phase of -omics studies with confident theoretical control of a standard measure of type-1 error rate (FDR), the setting for which it was originally proposed.

The standard cFDR approach is minimally parametric, minimising the distributional assumptions on P,Q. The only parametrisation required is in approximating the distribution of $Q|H_0^P$, which generally leads to a conservative procedure; that is, leading to an FDR controlled below the requisite level α (see Section 4.4). The approximation of $Q|H_0^P$ is additionally only used in computation of integrals reducing dependence on its pointwise accuracy.

The avoidance of asymptotic assumptions in the justification of methods 3 and 4 allows cFDR methods to be used with fewer hypotheses (lower n) than it has previously been used while still asserting FDR control, opening a wider range of applications. Our simulations extended to $n = 1000$, far fewer than the number of hypotheses encountered in the GWAS setting (where $n \approx 10^5 - 10^7$). Small n does mean that the distribution of $Q|H_0^P$ is less well-approximated, but in our usual method the fitted distribution has only two degrees of freedom (π_0, σ) and the error in approximation is still small.

Control of FDR at α using cFDR and using unconditional analysis on p-values (BH) will give differing results, and some associations discovered by p-value may be 'lost' when using cFDR. This effect is inevitable (seen in figure 8). However, since the cFDR (as opposed to p-value) uses a flexible rejection region 'learned' from the data, the loss/gain of associations will generally fall in favour of the cFDR when the hypotheses H_0^P and H_0^Q tend to co-occur.

As sample sizes tend to infinity, the order of v-values will tend to a different order than the order of p-values. Furthermore, the usefulness of cFDR depends on the 'two groups model' Efron *and others* (2008) in which many variables are not associated. In alternative models of association such as the 'universal pleiotropy hypothesis' (Paaby and Rockman, 2013), common in random-effects models (Yang *and others*, 2011), use of cFDR only changes the order in which associations are discovered, since all associations are eventually discovered as sample sizes tend to infinity.

Our method improves the power of cFDR analyses and improves confidence in the FDR associated with the rejection procedure. This can enable more efficient use of data, and more information to be gained from the same datasets. Our method contributes to a set

of tools for high-dimensional statistical analysis and has wide application across a range of fields in biomedicine and elsewhere.

8 Software

Software is available as an R package at `https://github.com/jamesliley/cfdr`. All data and a pipeline to deterministically generate all plots and results in this paper can be found at `https://github.com/jamesliley/cfdr_control_pipeline`

9 Supplementary Material

1. Description of the behaviour of cFDR on a rectangular region
2. Demonstration of the failure of the rejection procedure $c\widehat{FDR} < \alpha$ to control FDR at α
3. Explanation of the 'points on stalks' effect regarding FDR control method 2
4. Algorithms for each FDR control method
5. Discussion of FDR control in method 3a
6. Our protocol for simulations
7. Supplementary figure: Manhattan plots for JIA, T1D.

Acknowledgments

CW is funded by the Wellcome Trust (WT107881) and the MRC (MC_UU_00002/4). JL is funded under by WT107881 and for part of this work was on the Wellcome Trust PhD programme in Mathematical Genomics and Medicine, funded by the NIHR Cambridge BRC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: None declared.

References

Andreassen, Ole A., McEvoy, Linda K, Thompson, Wesley K, Wang, Yunpeng, Reppe, Sjur, Schork, Andrew J, Zuber, Verena, Barrett-Connor, Elizabeth, Gautvik, Kaare and Aukrust, Pal. (2014, April). Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes. *Hypertension* **63**.

Andreassen, Ole A., Thompson, Wesley K, Schork, Andrew J, Ripke, Steven, Mattingsdal, Morten, Kelsoe, John R, Kendler, Kenneth S, O'Donovan, Michael C, Rujescu, Dan, Werge, Thomas *and others*. (2013, April). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genetics* **9(4)**.

Benjamini, Yoav and Hochberg, Yosef. (1995). 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.

Benjamini, Yoav and Yekutieli, Daniel. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Efron, Bradley *and others*. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**(1), 1–22.

Ferkingstad, Egil, Frigessi, Arnoldo, Rue, Havard, Thorleifsson, Gudmar and Kong, Augustine. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Annals of Applied Statistics* **2**(2), 714–735.

Hinks, Anne, Cobb, Joanna, Marion, Miranda C, Prahalad, Sampath, Sudman, Marc, Bowes, John, Martin, Paul, Comeau, Mary E, Sajuthi, Satria, Andrews, Robert, Brown, Milton, Chen, Wei-Min, Concannon, Patrick, Deloukas, Panos, Edkins, Sarah, Eyre, Stephen, Gaffney, Patrick M, Guthery, Stephen L, Guthridge, Joel M, Hunt, Sarah E, James, Judith A, Keddache, Mehdi, Moser, Kathy L, Nigrovic, Peter A, Onengut-Gumuscu, Suna, Onslow, Mitchell L, Rose, Carlos D, Rich, Stephen S, Steel, Kathryn J A, Wakeland, Edward K, Wallace, Carol A, Wedderburn, Lucy R, Woo, Patricia, Bohnsack, John F, Haas, Johannes Peter, Glass, David N, Langefeld,

Carl D, Thomson, Wendy *and others*. (2013, 06). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet* **45**(6), 664–669.

Ignatiadis, Nikolaos, Klaus, Bernd, Zaugg, Judith B and Huber, Wolfgang. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods* **13**(7), 577.

Liley, James and Wallace, Chris. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLOS Genetics*.

Onengut-Gumuscu, Suna, Chen, Wei-Min, Burren, Oliver, Cooper, Nick J, Quinlan, Aaron R, Mychaleckyj, Josyf C, Farber, Emily, Bonnie, Jessica K, Szpak, Michal, Schofield, Ellen *and others*. (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.

Paaby, Annalise B and Rockman, Matthew V. (2013, February). The many faces of pleiotropy. *Trends in Genetics* **29(2)**, 66–73.

Storey, John D *and others*. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31**(6), 2013–2035.

The Wellcome Trust Case Control Consortium. (2007, June). Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature* **447**(7145), 661–678.

Yang, Jian, Manolio, Teri A, Pasquale, Louis R, Boerwinkle, Eric, Caporaso, Neil, Cunningham, Julie M, de Andrade, Mariza, Feenstra, Bjarke, Feingold, Eleanor, Hayes, M Geoffrey *and others*. (2011, June). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43(6)**, 519–525.

Zablocki, Rong W, Schork, Andrew J, Levine, Richard A, Andreassen, Ole A, Dale, Anders M and Thompson, Wesley K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30**(15), 2098–2104.

General notes for notation

1. P, p refer to study under investigation, Q, q to study for conditioning on. P and Q are also used to refer to these two studies as opposed to the random variables corresponding to p-values.

2. i and j index hypotheses and variables corresponding to hypotheses, (i) generally meaning the ith smallest

3. S refers to the set $(p_1, q_1), (p_2, q_2), ..., (p_n, q_n)$

4. k indexes folds/subdivisions

5. n is the total number of hypotheses

6. N is the number of folds for cross-validation/resampling

7. P, Q are random variables, p, q are observations

8. H^P is an indicator variable for association in study P. It can be considered in a Bayesian sense as a Bernoulli-distributed random variable, or a frequentist sense as a null hypothesis eg $H^P = 0$.

9. v-values refer to values $v(L)$ used for testing in some sense. The exact definition depends on the method used for FDR control.

Table 1: Theoretical and simulated FDR control. Columns 2-5 show successively stronger assumptions needed for theoretical FDR control. Method 1 controls FDR only asymptotically. All methods assume $Q|H_0^P$ is known.

Method	Theor. controlled	Theor. given negligible correlation between v-values	Theor. given approximation (9) holds ($L_S \perp\!\!\!\perp S$)	Controlled in simulation	Over-controls FDR in sim. and theory
1				✓	✓
2			✓		
3a within-fold	✓	✓	✓	✓	
3a overall					
3b		✓	✓	✓	
4		✓	✓	✓	

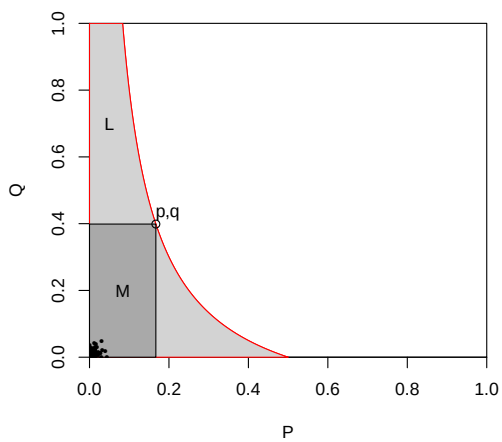

Figure 1: Illustration of failure of FDR control in region with $c\widehat{FDR}_S^1 < \alpha$. Non-null variables are shown as black dots. Regions M and L are indicated.

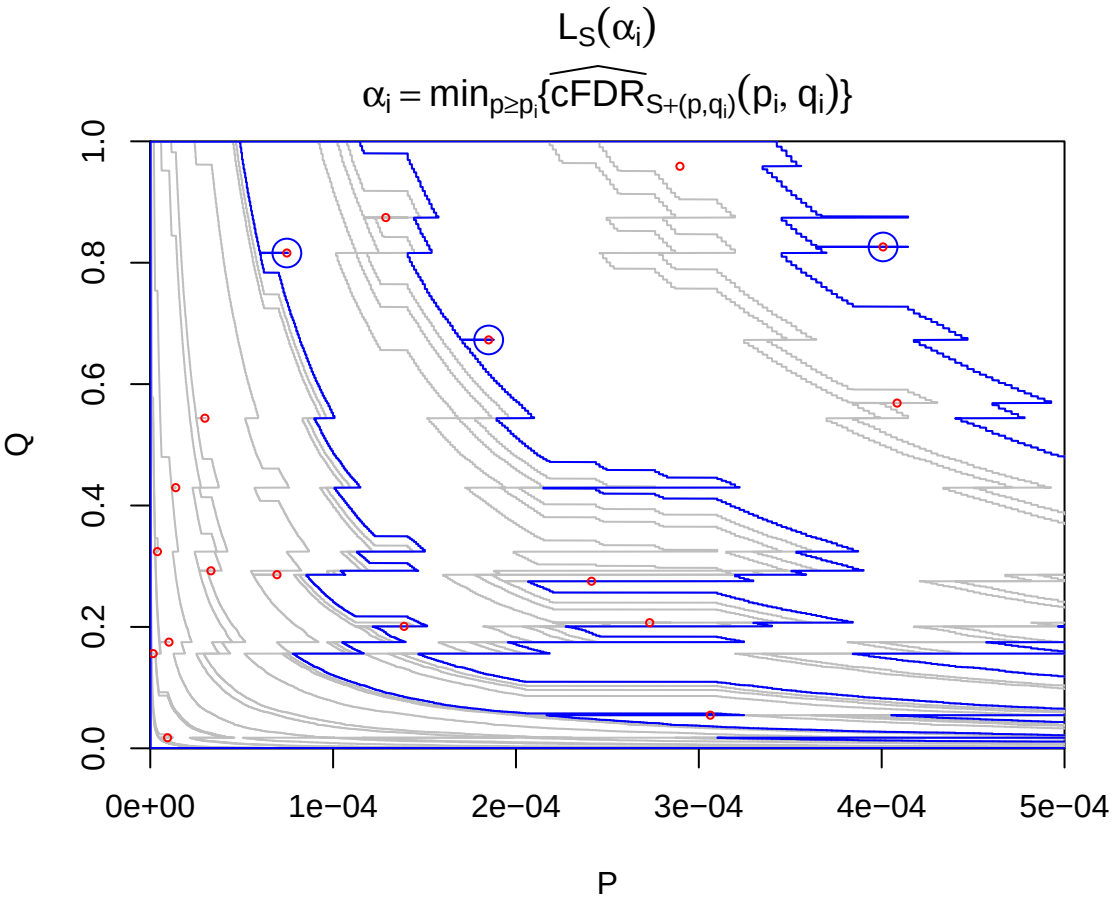

Figure 2: Demonstration of rightmost borders of regions $L_S(\alpha)$ where $\alpha = f(p_i, q_i)$. Grey/red/blue curves show the rightmost border of regions L_S, and points are shown in red. For each blue-ringed point p_i, q_i, the curve $L_S(f(p_i, q_i))$ is shown in blue. Points may lie on 'stalks' emanating from the remainder of the curve L (blue curves), indicating that $g(p_i, q_i)$ may be substantially lower than v-values $g(p, q)$ for (p, q) in a neighbourhood of (p_i, q_i).

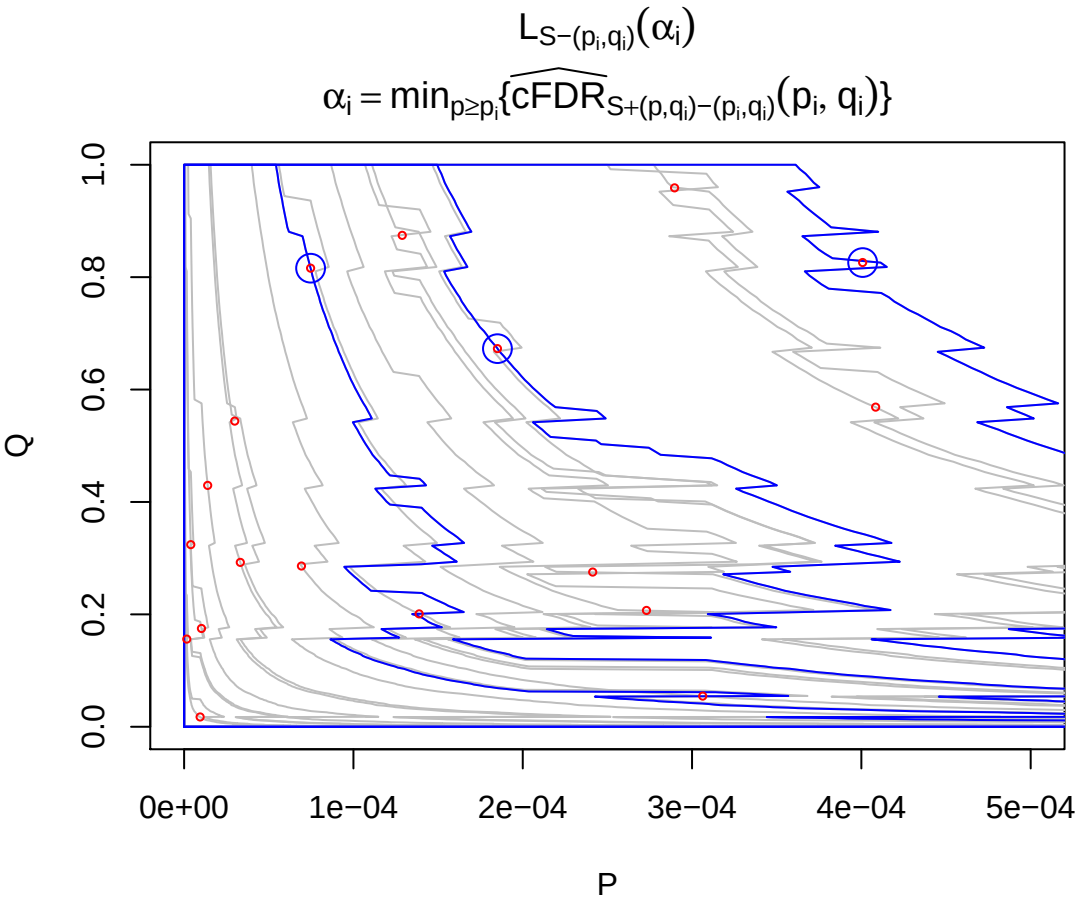

Figure 3: Demonstration of rightmost borders of regions $L_{S-(p_i,q_i)}(f(p_i,q_i))$, on the same data as used for figure 2. Grey/red curves show the rightmost border of regions L_S, and points are shown in red. For each ringed point p_i, q_i, the corresponding curve $L_S(f(p_i,q_i))$ is shown in blue. Ringed points were not used in the corresponding curves. Points which lay on 'stalks' in figure 2 no longer do.

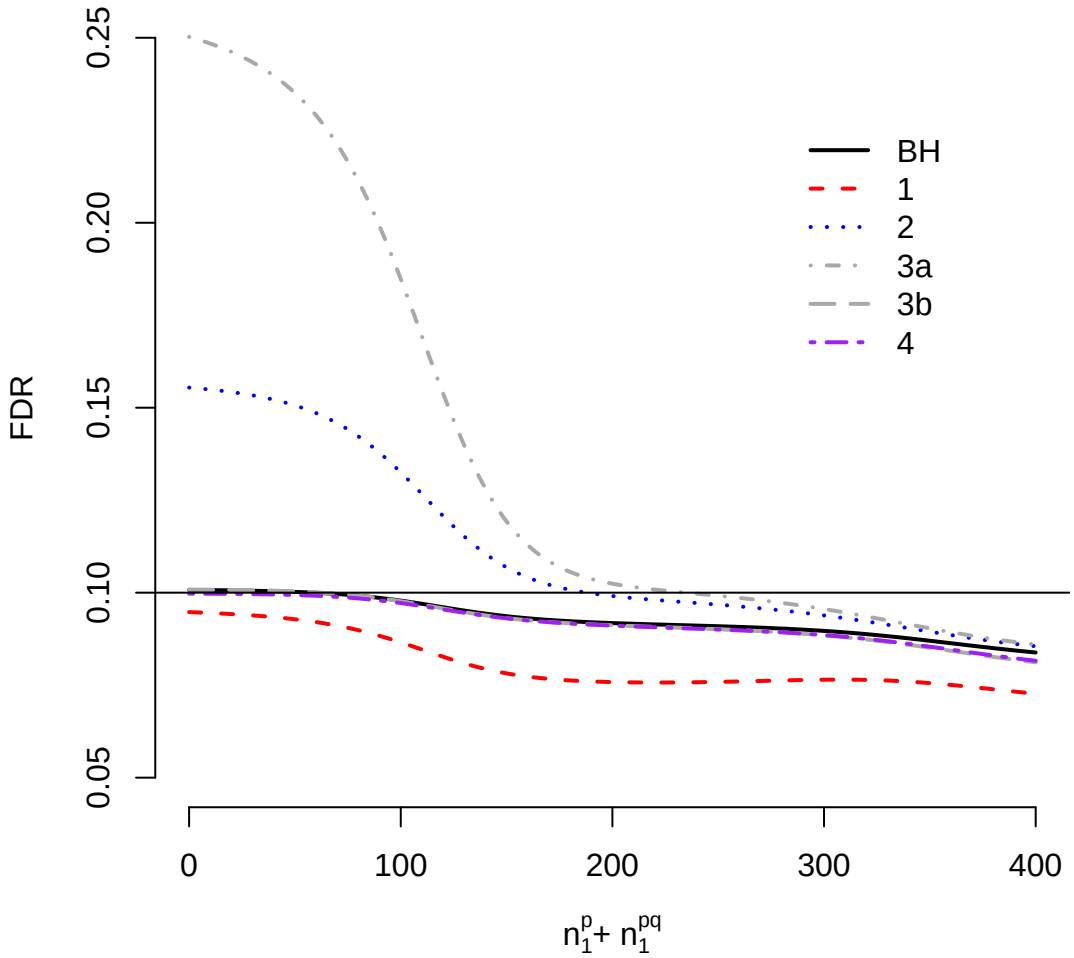

Figure 4: Performance of FDR-control methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration). The horizontal line shows α, the desired FDR control level. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

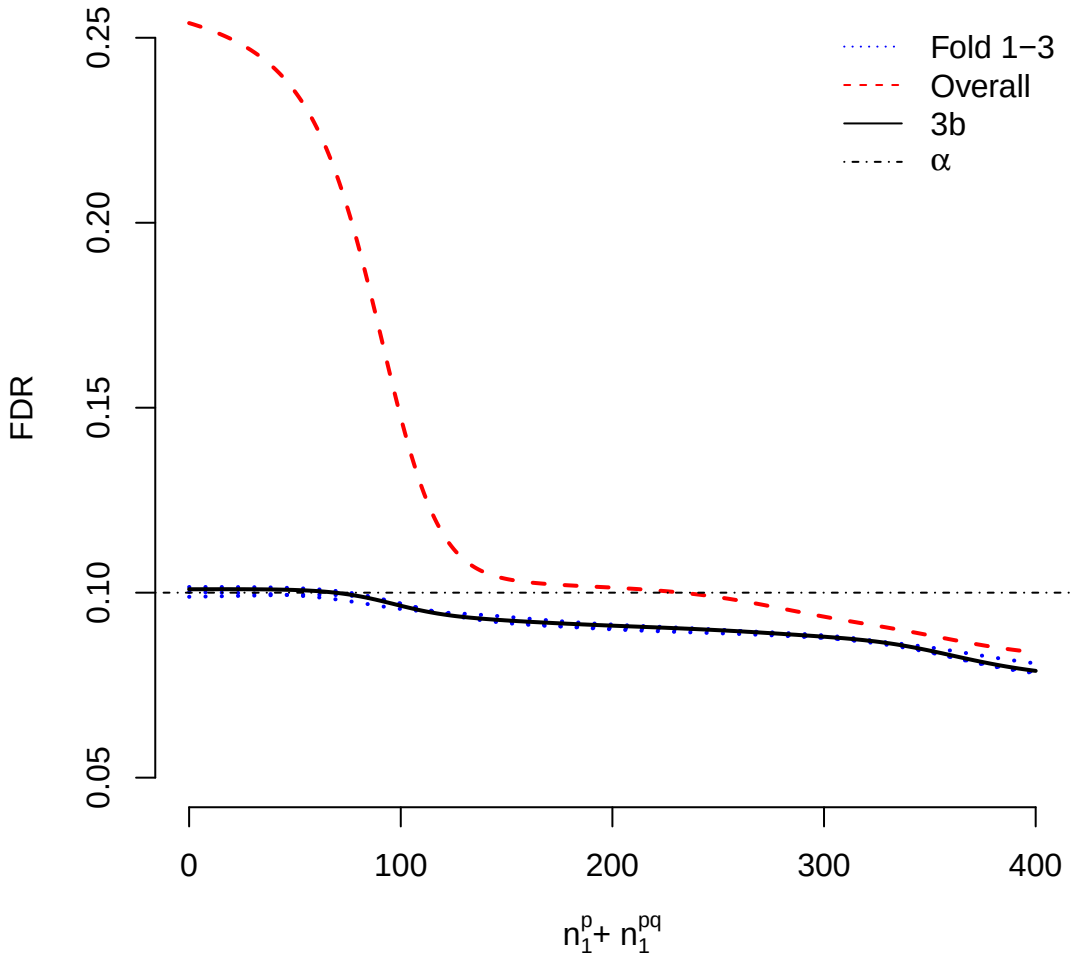

Figure 5: FDR control in methods 3a and 3b overall and within each fold against $n_1^p + n_1^{pq}$, the total number of variables associated with P, the primary study under consideration. The horizontal line shows α, the desired FDR control level. Curves show moving weighted averages as above.

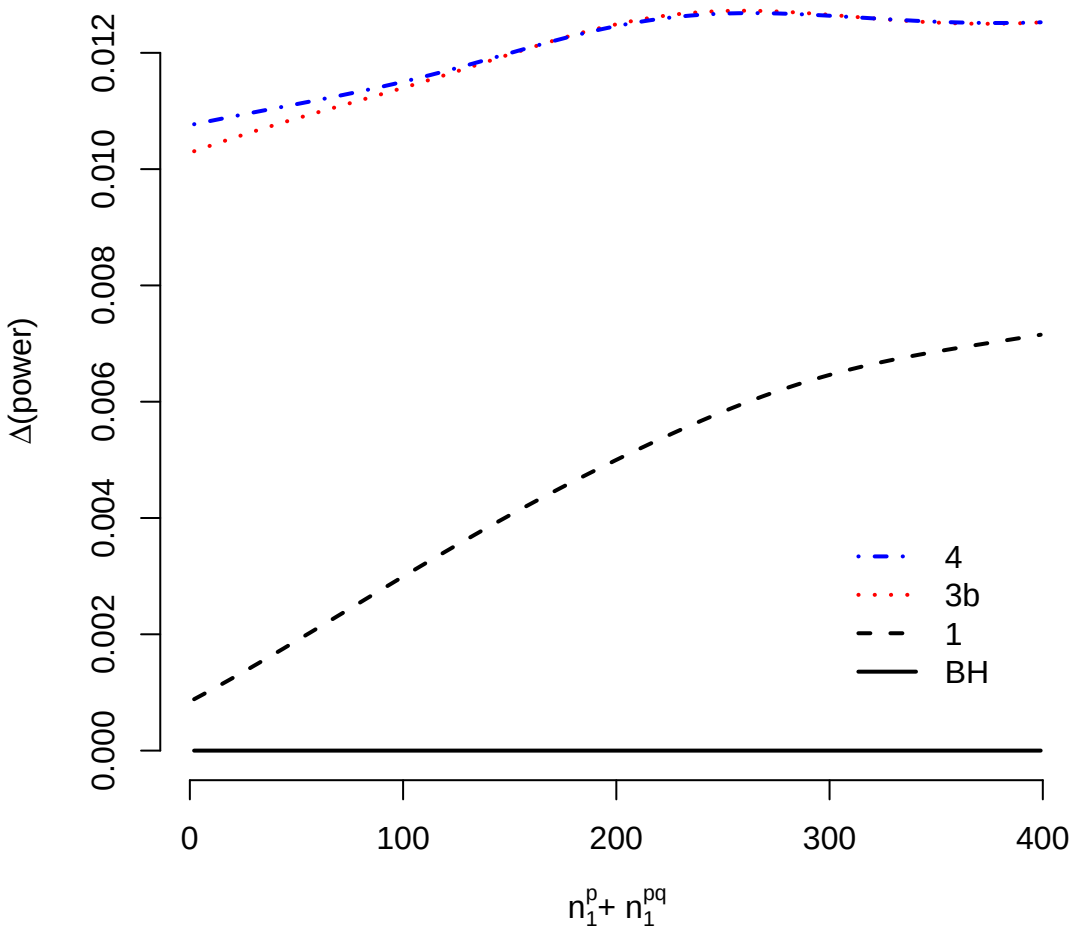

Figure 6: Power of FDR-controlling methods 1, 3b, 4, and BH (unconditional analysis of p-values). Power is highest in method 4 and 3b, then (substantially ahead of) method 1 and BH. Curves show moving weighted averages as above.

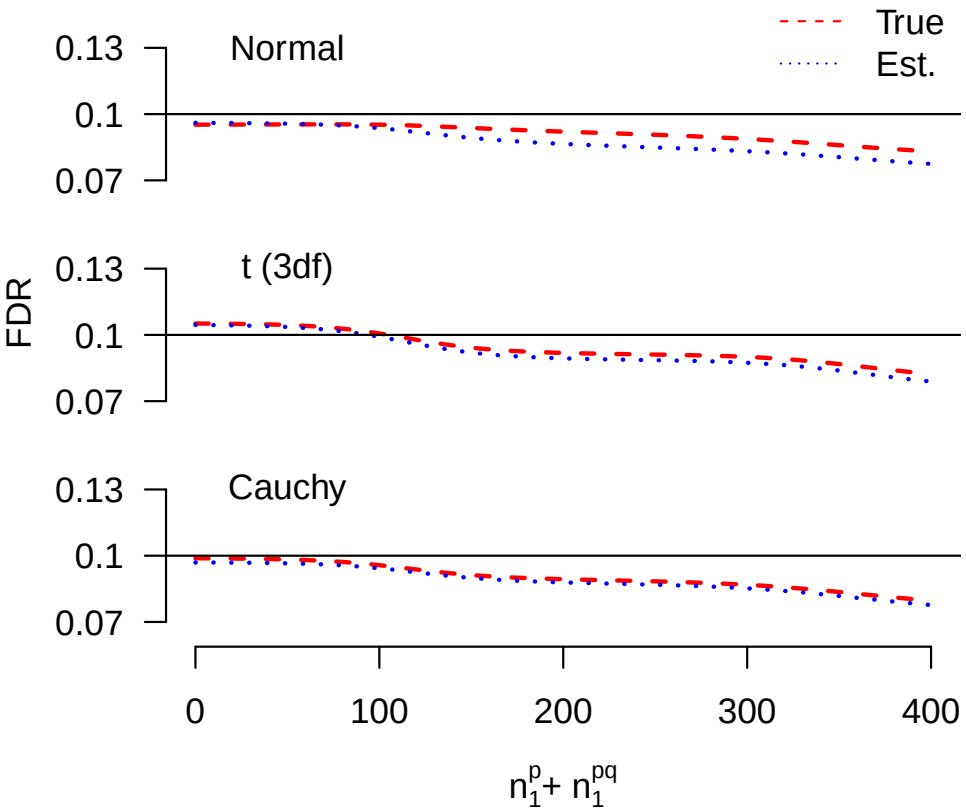

Figure 7: FDR control with method 4, using either the true null distribution $Pr(H_0^P|Q \leq q)$ or a mixture-normal estimate from the observed distribution of $Q|P \geq \frac{1}{2}$, across the three types of alternative distribution for association statistics. Approximation is reasonable even when effect sizes are Cauchy-distributed. Gaussian-weighted moving averages are used as above. Black lines show the FDR control level, $\alpha = 0.1$.

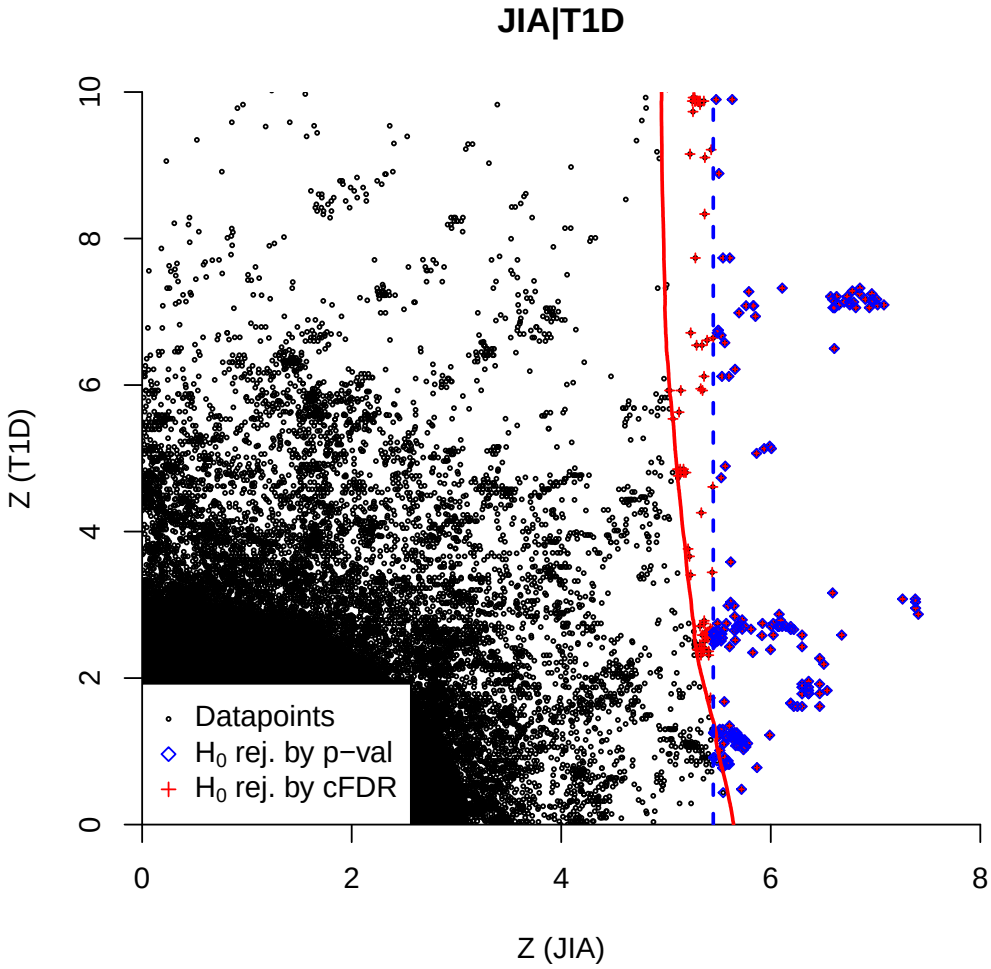

Figure 8: Z scores from GWAS-like study (ImmunoChip) on JIA and T1D. Red points (crosses) show rejections using cFDR, blue points (diamonds) using p-value alone, controlling overall FDR at $\alpha = 1.6 \times 10^{-5}$. Red (solid) and blue (dashed) lines indicate approximate rejection regions for cFDR and p-values respectively

1 Appendix

1.1 FDR of a rectangular region

Define all variables as per Section 1 and 2.

Let X be the rectangle with corners $(0,0)$, $(p_i,0)$, $(0,q_i)$, (p_i,q_i), and X' the rectangle with corners $(0,0)$, $(p_i,0)$, $(0,1)$, $(p_i,1)$. As above, assume $P,Q \sim \pi_0 f_0 + \pi_1 f_1$ and that P and Q are conditionally independent given H_0^P, so $\int_X df_0 = p_i f_0(q_i)$ for some function f_0. As $n \to \infty$, we have

$$\begin{aligned}
c\widehat{FD}R(p_i,q_i) &\to_{A.S} Pr(P \le p_i | H_0^P, Q \le q_i) \frac{Pr(Q \le q_i)}{Pr(P \le p_i, Q \le q_i)} \\
&= p_i \frac{\pi_0 \int_{X'} df_0 + \pi_1 \int_{X'} df_1}{\pi_0 \int_X df_0 + \pi_1 \int_X df_1} \\
&= \frac{\pi_0 p_i f_0(q_i) + \pi_1 \int_{X'} df_1}{\pi_0 p_i f_0(q_i) + \pi_1 \int_X df_1}
\end{aligned}$$

and

$$\begin{aligned}
\frac{V(X)}{R(X)} &\to_{A.S} \frac{n\pi_0 \int_X df_0}{n\left(\pi_0 \int_X df_0 + \pi_1 \int_X df_1\right)} \\
&= \frac{\pi_0 p_i f_0(q_i)}{\pi_0 p_i f_0(q_i)\pi_0 + \int_X df_1}
\end{aligned}$$

so for large enough n the value $c\widehat{FD}R(p_i,q_i)$ almost surely overestimates $V(X)/R(X)$ for the rejection region X.

1.2 FDR of the rejection procedure $c\widehat{FD}R \le \alpha$

If the null hypothesis $H^P = 0$ is rejected whenever $c\widehat{FD}R(p_i,q_i) < \alpha$ or using rejection criterion (5), the FDR is not controlled at α. Consider an extreme case in which $P,Q|H_0^P \sim U(0,1) \times U(0,1)$ and $P,Q|H^P = 1 \equiv 0$, so all non-null variables have $P = Q = 0$ and $\int_X df_0 = \int_X 1 dX = \text{area}(X)$. Then the region for which $c\widehat{FD}R(x,y) < \alpha$ (and the corresponding region for procedure (5)) converges pointwise almost surely to the region L defined by

$$\begin{aligned}
(p_i,q_i): \quad & p_i \frac{\pi_0 q_i + \pi_1}{\pi_0 p_i q_i + \pi_1} < \alpha \\
& \Leftrightarrow q_i < \frac{\pi_1}{\pi_0(1-\alpha)} \left(\frac{\alpha}{p_i} - 1\right)
\end{aligned}$$

and, by integrating over this region, we have

$$
\begin{aligned}
V(L) &\to_{A.S} n\frac{\pi_1\alpha}{(1-\alpha)\pi_0}\log\left(\frac{1-\alpha\pi_0}{\pi_1}\right)\\
R(L) &\to_{A.S} n\left(\pi_1+\frac{\pi_1\alpha}{(1-\alpha)\pi_0}\log\left(\frac{1-\alpha\pi_0}{\pi_1}\right)\right)\\
\frac{V(L)}{R(L)} &\to_{A.S} \frac{\alpha\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right)}{\pi_0(1-\alpha)+\alpha\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right)}
\end{aligned}
$$

which is not generally $< \alpha$ (and exceed α by an arbitrarily large factor). Thus neither the procedure in which $H^P = 0$ is rejected whenever $c\widehat{FDR} < \alpha$ nor criterion (5) even asymptotically control the FDR at α, or at $f(\alpha)$ for any function f of α only.

1.3 Cause of 'points on stalks' effect

The performance of FDR-controlling method 2 on $c\widehat{FDR}$ is poor (figure 4) in that FDR is not controlled. This is because the values $g(p_i, q_i)|H_0^P$ (equation (13)) are not uniformly distributed, since the function g depends on p_i, q_i. Indeed, the failure of the FDR control indicates that the values $g(p_i, q_i)$ are in a sense smaller than they should be.

This is briefly noted in figure 2, and the resolution of the problem using FDR-controlling method 4 is demonstrated in figure 3. We note that in FDR-controlling method 2, the regions L_S corresponding to each point tend to contain 'stalks' on which the relevant points fall. In this section, we examine why this effect occurs.

The effect is most easily understood when considering a p-value pair $(p_0, q_0) \in S$ for which no other pair $(p, q) \in S$ has $p < p_0, q < q_0$ (see figure 1). Denote by α_0 the value of α corresponding to the smallest region $L_S(\alpha)$ containing (p_0, q_0). Suppose that $k \gg 1$ pairs p, q have $q < q_0$, and that points are distributed such that $c\widehat{FDR}_{S+(p,q_0)}(p, q_0) > c\widehat{FDR}_{S+(p_0,q_0)}(p_0, q_0)$ for $p > p_0$, so

$$
\begin{aligned}
\alpha_0 &= c\widehat{FDR}_{S+(p_0,q_0)}(p_0, q_0)\\
&= p_0\frac{|\{i : q_i \le q_0, (p_i, q_i) \in S + (p_0, q_0)\}|}{|\{i : p_i \le p_0, q_i \le q_0, (p_i, q_i) \in S + (p_0, q_0)\}|}\\
&= p_0\frac{k+2}{2}
\end{aligned}
$$

since there are k pairs (p_i, q_i) with $q_i < q_0$, and the pair (p_0, q_0) appears in $S + (p_0, q_0)$ twice.

Consider the locus of points (p, q) at the rightmost border of $L_S(\alpha_0)$. For a point on this

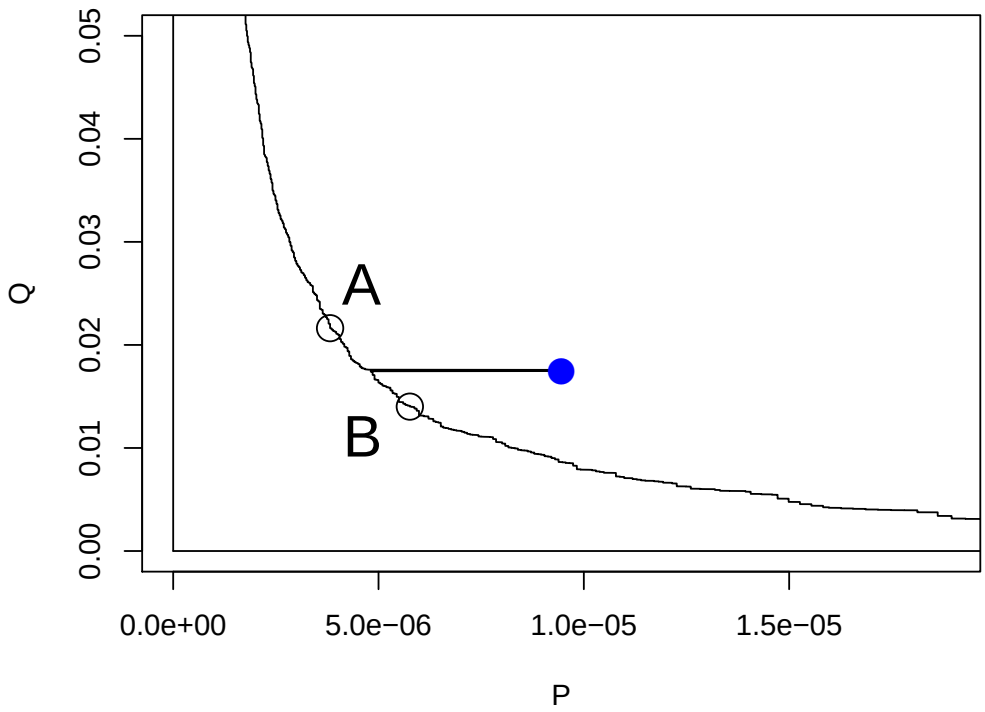

Figure 1: The blue point shows a point (p_0, q_0) with no other points to the lower-left. Points on the rightmost boundary of $L_S(\alpha_0)$ (the smallest region $L_S(\alpha)$ containing (p_0, q_0)) are shown in black. Points A and B show points (p, q) on the boundary of interest with $q = q_0 \pm \epsilon$.

locus (p', q') with $q' = q_0 + \epsilon$ (labelled in figure 1 as 'A'), the corresponding p' value would be such that $c\widehat{FDR}_{S+(p',q')}(p', q') = \alpha_0 = p_0 \frac{k+2}{2}$. For sufficiently small ϵ, the number of points $(p_i, q_i) \in S + (p', q')$ with $q_i \leq q'$ will be $k + 2$ ((p_0, q_0), (p, q) itself, and the k points with $q < q_0$). Since there are no points below and to the left of (p_0, q_0), the only point in $S + (p', q')$ with $p \leq p', q \leq q'$ is (p', q') itself for sufficiently small ϵ. Thus

$$\begin{aligned}
& c\widehat{FDR}_{S+(p,q_0+\epsilon)}(p, _0 + \epsilon) = \alpha_0 \\
\Leftrightarrow\ & p\frac{k+2}{1} = \alpha_0 \\
\Leftrightarrow\ & p\frac{k+2}{1} = p_0\frac{k+2}{2} \\
\Leftrightarrow\ & p = \frac{1}{2}p_0
\end{aligned} \tag{19}$$

so we expect, for small ϵ, the curve forming the rightmost border of $L_S(\alpha_0)$ to contain the points $\frac{1}{2}p_0, q_0 + \epsilon$. For a point (p', q') with $q' = q_0 - \epsilon$ (labelled 'B' on figure 1), there are $k+1$ points in $S + (p', q')$ with $q \leq q'$, and in a similar way we expect

$$p' = \frac{k+2}{2(k+1)} p_0 \tag{20}$$

For points $p', q_0 \pm \epsilon$ with larger ϵ, the $k+2$ in equation (19) is replaced with a larger value, and the $k+1$ in equation (20) with a smaller one. Since k is assumed to be large, a change in ϵ leads to only a small change in p'. Thus, as seen in figure 1, the rightmost boundary of L_S locally takes the form of a smooth curve, with the point (p_0, q_0) on a narrow 'stalk'. In this way, the point (p_0, q_0) has a large effect on the shape of the curve $L_S(\alpha_0)$, effectively making its area smaller than it 'should' be, leading to a lower-than expected value for $g(p_0, q_0)|H_0^P$.

1.4 Algorithms for each method

Assume we are given a set of pairs of p-values (p_i, q_i), $i \in 1..n$. We wish to test the null hypothesis H_0^P for each i.

As above, define

$$L_S(\alpha) = \{(p, q) : \exists p' \geq p | \widehat{cFDR}_{S+(p',q)}(p', q) \leq \alpha\} \tag{21}$$

The estimator $\widehat{cFDR}$ may take the form in (3) or an alternative estimator. The dependence of the estimator on the observations (p_i, q_i) should be such that $\widehat{cFDR}_X$ is taken to be independent of observations not in X. If method 1 is to be used, the estimator must be asymptotically conservative ($\lim_{n\to\infty} cFDR(p, q) - \widehat{cFDR}(p, q) \geq 0$) in order for the overall FDR to be asymptotically controlled.

The algorithms to control the FDR at α are as follows

1.4.1 Method 1

1. For a given region $L_S(\alpha_0)$, considered as a function of α', define $M(L_S(\alpha_0))$ as the rectangle M contained within $L_S(\alpha_0)$ for which $v(M)$ is maximised

2. Find the largest value α_0 such that

$$\frac{v(L_S(\alpha_0))}{v(M(L_S(\alpha_0)))}\alpha_0 \leq \alpha$$

3. Reject the null hypothesis H_0^P for all $i : (p_i, q_i) \in L_S(\alpha_0)$

1.4.2 Method 2

1. For each (p_i, q_i) set

$$\begin{aligned} f(p_i, q_i) &= \min\left(\alpha' : (p_i, q_i) \in L_S(\alpha')\right) \\ g(p_i, q_i) &= v\left(L_S(f(p_i, q_i))\right) \qquad \text{(v-values)} \end{aligned}$$

2. Order v-values $g(p_i, q_i)$ to get $g(p_{(1)}, q_{(1)}) \leq g(p_{(2)}, q_{(2)}) \leq ... \leq g(p_{(n)}, q_{(n)})$

3. Find (i_{max}) for which

$$(i_{max}) = \max\left\{i : g(p_{(i)}, q_{(i)})\frac{n}{i} \leq \alpha\right\}$$

4. Reject the null H_0^P for all $(i) \leq (i_{max})$

1.4.3 Method 3

Method 3a: To control FDR in each fold

1. Divide $1..n$ into N folds. Denote by n_k the size of fold k and (p_i^k, q_i^k) the ith sample in fold k.

2. For each fold k:

   (a) Compute the values

$$\begin{aligned} f_k(p_i^k, q_i^k) &= \min\left(\alpha' : (p_i^k, q_i^k) \in L_{S-\text{fold } k}(\alpha')\right) \\ g_k(p_i^k, q_i^k) &= v\left(L_{S-\text{fold } k}(f(p_i, q_i))\right) \qquad \text{(v-values)} \end{aligned}$$

   (b) Rank the v-values $g_k(p_i^k, q_i^k)$ to give $g_k(p_{(1)}^k, q_{(1)}^k) \leq g_k(p_{(2)}^k, q_{(2)}^k) \leq ... \leq g_k(p_{(n)}^k, q_{(n)}^k)$

(c) Find (i^k_{max}) such that

$$(i^k_{max}) = \max\left\{i : g_k(p^k_{(i)}, q^k_{(i)})\frac{n_k}{i} \leq \alpha\right\}$$

(d) Reject H^P_0 for all the (i) in fold k such that $(i) \leq (i^k_{max})$

Method 3b: To control FDR overall

1. Divide $1..n$ into N folds. Denote by n_k the size of fold k and (p^k_i, q^k_i) the ith sample in fold k.

2. For each fold k:

   (a) Compute the values

$$f_k(p^k_i, q^k_i) = \min\left(\alpha' : (p^k_i, q^k_i) \in L_{S-\text{fold } k}(\alpha')\right)$$
$$g_k(p^k_i, q^k_i) = v\left(L_{S-\text{fold } k}(f(p_i, q_i))\right) \qquad \text{(v-values)}$$

3. Collect the v-values $g_k(p^k_i, q^k_i)$ for all k. There is only one such value for each i; denote it by g_i. Arrange the values g_i in order so $g_{(1)} \leq g_{(2)} \leq ... \leq g_{(n)}$

4. Find (i_{max}) such that

$$(i_{max}) = \max\left\{i : g_{(i)}\frac{n}{i} \leq \alpha\right\}$$

5. Reject H^P_0 for all $(i) \leq (i^k_{max})$

1.4.4 Method 4

1. For each (p_i, q_i) set

$$f_i(p_i, q_i) = \min\left(\alpha' : (p_i, q_i) \in L_{S-(p_i,q_i)}(\alpha')\right)$$
$$g_i(p_i, q_i) = v\left(L_{S-(p_i,q_i)}(f_i(p_i, q_i))\right) \qquad \text{(v-values)}$$

2. Order v-values $g_i(p_i, q_i)$ to get $g_{(1)}(p_{(1)}, q_{(1)}) \leq g_{(2)}(p_{(2)}, q_{(2)}) \leq ... \leq g_{(n)}(p_{(n)}, q_{(n)})$

3. Find (i_{max}) for which

$$(i_{max}) = \max\left\{i : g_{(i)}(p_{(i)}, q_{(i)})\frac{n}{i} \leq \alpha\right\}$$

4. Reject the null H_0^P for all $(i) \leq (i_{max})$

1.5 FDR control in method 3a

Method 3a controls FDR within each fold, but not overall. Control within each fold can, in a sense, be considered as strong a metric of type-1 error rate as FDR, in that it controls the mean FDR across folds, but not the 'right' mean.

Denoting by V_k, R_k the number of false discoveries and total discoveries in fold k, this method guarantees that

$$E\left(\frac{V_k}{R_k}\middle| R_k > 0\right) Pr(R_k > 0) = E\left(\left(\frac{V_k}{R_k}\right)'\right) \leq \alpha$$

where $(Z)' = 0$ if Z is indeterminate, and Z otherwise. Thus

$$E\left(\frac{1}{N}\sum_k \left(\frac{V_k}{R_k}\right)'\right) \leq \alpha \tag{22}$$

However, the FDR of the procedure as a whole is

$$E\left(\left(\frac{\sum_k V_k}{\sum_k R_k}\right)'\right) \tag{23}$$

The two quantities (22) and (23) are not equal. If (fortuitously) all $R_k > 0$ and the orders of $R_1, R_2, ...R_N$ and $\frac{V_1}{R_1}, \frac{V_2}{R_2}, ..., \frac{V_N}{R_N}$ are the same, then the LHS of (22) exceeds the LHS of (23) by Chebyshev's inequality, although the opposite is true if the orders are reversed (Set $x_k = V_k$, $y_k = \frac{V_k}{R_k}$, so $x_i \leq x_j \Leftrightarrow y_i \leq y_j$. Then $\left(\sum_f (V_k/R_k)\right)(\sum_k R_k) = (\sum x_k)(\sum y_k) \leq N \sum x_k y_k = N \sum V_k$ and the inequality follows; reversed if $x_i \leq x_j \Leftrightarrow y_i \geq y_j$). More usefully, if $\frac{V_k}{R_k}$ converges in probability to a positive limit as $n \to \infty$ with k finite, then (22) and (23) also converge to that limit.

The discrepancy between the left-hand sides of (22) and (23) is clearest when $R_k = 0$ for some k. In the extreme case in which H_0^P holds for all i (so $V_k = R_k$) with large N, if $E\left(\left(\frac{V_k}{R_k}\right)'\right) = \alpha$, we have

$$E\left(\left(\frac{\sum_k V_k}{\sum_k R_k}\right)'\right) = Pr\left(R_1 > 0 \cup R_2 > 0 ... R_n > 0\right) \to 1$$

For this reason, we recommend that the number of folds be low if method 3a is used.

1.6 Simulation protocol

The distributions and variables we considered are shown in table 1. In general, our simulations took the form of: a fixed number n_1^{pq} of variables associated in both P and Q, a fixed number of variables n_1^p associated only with p, a fixed number of variables n_1^q associated only with Q and the remainder of variables associated with neither P nor Q. We will refer to these four classes of variables as C_1, C_2, C_3, C_4. We designated that within each class of variables, P and Q were iid.

Table 1: Variables used in simulations

Variable	Description	Point values	Sampling distribution for other values
n	Total number of variables	10^3, 10^4	$10^{U(3,4)}$ (rounded)
n_1^{pq}	Number of variables assoc. with P, Q	0, 10, 200	$U(0, 200)$
n_1^p	Number of variables associated with P	0,10,200	$U(0, 200)$ (rounded)
n_1^q	Number of variables associated with Q	0,10,200	$U(0, 200)$ (rounded)
s_p	Scale for distribution of P (see below)	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
s_q	Scale for distribution of P_j	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
d	Form of distributions	Normal, t (3df), Cauchy (eq. prob.)	Normal, t (3df), Cauchy (eq. prob.)

For variables in C_1, C_2, we set the distribution of P (determined by d, s_p) by first simulating Z scores:

d=1: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim N(0,1)$

d=2: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim t(\mathrm{df} = 3, \mathrm{ncp} = 0)$

d=3: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim \mathrm{Cauchy}(\mathrm{location} = 0, \mathrm{scale} = 1)$

where $-\Phi^{-1}\left(\frac{P}{2}\right)$ can be considered a Z-score corresponding to P, and s_p a scaling factor for the distribution. We set the distribution of Q in C_1, C_3 similarly, with s_q in place of s_p. The values p_i, q_i for $i \in C_4$ were sampled from $U(0, 1)$.

Within each class $C_1 - C_4$ we simulated independent P, Q, although P, Q are still dependent when not conditioning on class. We also considered s_p to be constant across C_1 and C_2, s_q to be constant across C_1 and C_3, and d to be constant in each simulation. We

deliberately oversampled parameter sets with $n_1^p + n_1^{pq} = 0$ given the importance of this part of the parameter space.

We only used the value $\alpha = 0.1$. The reason for only using one relatively high value of α was to assert FDR control at $n_1^p + n_1^{pq} = 0$ with reasonable confidence. Since FDP is Bernoulli in this case, to find a given relative change in FDR at α needs $O\left(\frac{1}{\alpha}\right)$ simulation runs. For instance, to determine that the FDR of a method was 10% too high at $\alpha = 1e-4$ with type-1 and 2 error rates of 5 and 10% respectively would require $\approx 10^7$ simulation runs, which was beyond our computational resources.

2 Supplementary figure: Manhattan plots for JIA/T1D

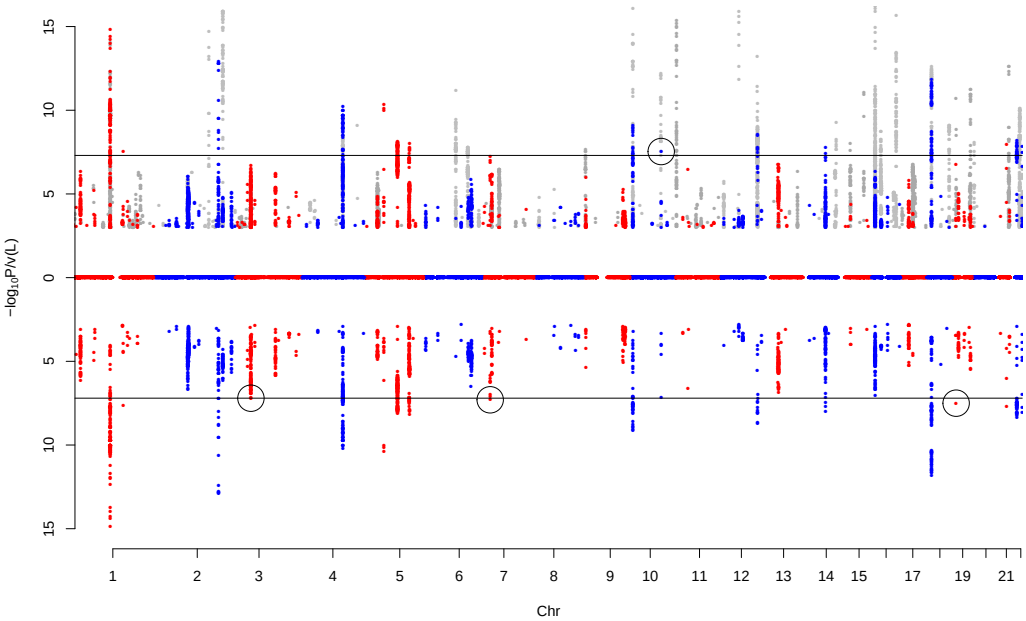

Figure 1: Back-to-back Manhattan plots for JIA/T1D analysis. The top plot shows $-\log_{10}(p)$ values for JIA, with $-\log_{10}(p)$ values for T1D overlaid behind in grey. The bottom plot (pointing downwards) shows $-\log_{10}(v)$ values for the cFDR analysis of JIA—T1D. Only p/v values $< 10^{-4}$ are shown. Black lines correspond to FDR thresholds. Ringed points indicate SNPs designated as associated in only one of the two analyses (p-value/v-value).