Post-selection Inference Following Aggregate Level Hypothesis Testing in Large Scale Genomic Data

Ruth Heller, Department of Statistics and Operations Research, Tel-Aviv university, Tel-Aviv 6997801, Israel, and National Cancer Institute, Rockville, MD 20852, U.S.A., E-mail: ruheller@gmail.com[1]

Nilanjan Chatterjee, Department of Biostatistics, Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, U.S.A.,E-mail: nchatte2@jhu.edu

Abba Krieger, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.,E-mail: krieger@wharton.upenn.edu

Jianxin Shi, National Cancer Institute, Rockville, MD 20852, U.S.A., E-mail: jianxins@mail.nih.gov[2]

Abstract. In many genomic applications, hypotheses tests are performed by aggregating test-statistics across units within naturally defined classes for powerful identification of signals. Following class-level testing, it is naturally of interest to identify the lower level units which contain true signals. Testing the individual units within a class without taking into account the fact that the class was selected using an aggregate-level test-statistic, will produce biased inference. We develop a hypothesis testing framework that guarantees control for false positive rates conditional on the fact that the class was selected. Specifically, we develop procedures for calculating unit level p-values that allows rejection of null hypotheses controlling for two types of conditional error rates, one relating to family wise rate and the other relating to false discovery rate. We use simulation studies to illustrate validity and power of the proposed procedure in comparison to several possible alternatives. We illustrate the power of the method in a natural application involving whole-genome expression quantitative trait loci (eQTL) analysis across 17 tissue types using data from The Cancer Genome Atlas (TCGA) Project.

Keywords: Conditional p-value; False discovery rate; Multiple testing; Selective inference.

[1] This work was performed while Ruth Heller was a visiting scholar at the National Cancer Institute.

[2] Study supported by the National Cancer Institute Intramural Research Program.

1 Introduction

In large scale analysis of genomic and genetic data, it is common to perform tests for hypotheses at an aggregate level over ”classes” of multiple related units. In analysis of genome-wide association studies, for example, testing for genetic associations may be performed at a gene-, region- or pathway- level by aggregating association statistics over multiple genetic markers (Yu et al., 2009; Wu et al., 2010a,b). Similarly, for identification of susceptibility markers with pleiotropic effects, association tests for each individual genetic marker may be performed by aggregating association test-statistics over multiple related phenotypes (Bhattacharjee et al., 2012; Peterson et al., 2015). Recently, in microbiome-profiling studies, association tests for identifying taxa were performed by aggregating association test-statistics over multiple traits (Hua et al., 2016). When multiple signals are expected within classes of related units, use of suitable aggregate-level test-statistics can improve the power of discovery of the signal harboring classes, compared to one unit at a time analysis.

In using aggregate level test-statistics for large scale hypotheses testing, much research have focused on development of test-statistics that can combine evidence of signals from multiple units in the most powerful way. For any given test-statistics, standard multiple testing adjustment methods are typically applied to adjust for the number of classes, which could be potentially large, over which the analyses may be performed. An important area of research that has received less attention involves how to perform proper post-selection inference following the aggregate level hypothesis testing to isolate individual units that contain signal within the higher-level classes. For example, if a genetic marker is identified to be statistically significant for aggregate-level association in pleiotropic analysis across multiple phenotypes, it is then very natural to perform follow-up analysis to identify which underlying phenotypes contain true signals for associations.

Clearly, a naive approach that ignores the effect of class selection on the second-stage of individual units selection will produce biased inference even if multiple testing adjustment methods were used to adjust for the number of units within each selected class.

In this manuscript, we study the problem of post-selection inference following aggregate level analysis as a two-stage hypothesis testing problem. The data of interest, the unit-level test-statistics, can be organized in an $m \times n$ matrix, where m is the number of classes for which inference is of interest, and n is the number of units within each class. For example, in GWAS of different phenotypes the data matrix for analysis has rows indexed by the SNPs, and columns indexed by the phenotypes. The first goal is to select the rows (SNPs) that have signal in at least one column (phenotype), and the post-selection goal can be to identify which columns have signal within each of the chosen rows.

Approaches to various error rate controls have been proposed in the past for multi-stage or hierarchical hypothesis testing. Possibly most relevant of them in the context of applications described above is a procedure proposed by Benjamini and Bogomolov (2014) which control false positives on the average over the selected rows. The method was recently applied to pleiotropic analysis of GWAS and single tissue eQTL data (Peterson et al., 2015, 2016), a type of application that we also use for illustration in this manuscript. Other methods include a procedure proposed by Yekutieli (2008) for controlling different FDR-types on trees of hypotheses, assuming independence between the stages of the hierarchy. Barber and Ramdas (2015) suggested controlling the group FDR at an aggregate level, for several different given partitions into groups. Some other strategies have been tailored to special application fields, such as gene expression analyses (Yekutieli et al. , 2006; Li and Ghosh , 2014), electrencephalography research (Singh and Phillips , 2010), and functional magnetic resonance imaging (Schildknecht, Tabelow and Dickhaus , 2015). Conditioning on the selection event has been suggested in novel works on post-model selection (Fithian et al. , 2015; Lee et al. , 2016), as well as for spatial signals in Benjamini and Heller (2007).

For the selection of columns in the second-stage of above hypothesis testing framework, we propose developing a hypothesis testing framework that guarantees control of false positive rates conditional on the fact that the row is selected. As a researcher may often conduct different experiments for each selected row for follow-up studies, control of false positive at the level of each row may be desirable rather than controlling average rates over all selected rows as has been considered for this problem by Peterson et al. (2015). It has been noted earlier (Benjamini and Bogomolov , 2014) that development of a general procedure for controlling such error rates could be difficult. In the setting where the test-statistics can be assumed to be independent across columns, we propose general procedures that are theoretically shown to control family-wise error rates and false discovery rate at the level of each row conditional on selection. We then show through simulation studies that for a commonly used aggregation test-statistic, the proposed procedure not only provides stronger type-I error rate control but it also has a major power advantage when compared to the general procedure proposed by Benjamini and Bogomolov (2014). An application of the methods for analysis of SNP markers predictive of gene expression level across multiple tissues using data from the The Cancer Genome Atlas (TCGA) project provided further empirical validation of substantially higher power of the proposed method compared to alternatives. These results provide encouraging direction for further research in developing more powerful procedures for two-stage hypothesis testing in more general frameworks.

In section 2, we set up the proposed hypothesis testing framework more formally and define the criterion of type-I error that we propose to control. In section 3, we present a procedure for controlling the proposed type-I error criterion and provide theoretical results supporting its validity. In section 4, we conduct simulation studies to evaluate type-I error rates and power for the proposed method compared to a naive approach and to the procedure proposed by Benjamini and Bogomolov (2014). In section 5, we describe results from analysis of TCGA study. In Section 6, we conclude with a discussion about future directions.

2 Notation and Goal

Let $H_{ij}, j = 1, \ldots, n$ be the family of n null hypotheses for the ith row, and let $P_{ij}, j = 1, \ldots, n$ be their corresponding p-values, for $i = 1, \ldots, m$. The global null hypothesis for row i, $\cap_{j=1}^{n} H_{ij}$, is the hypothesis that all null hypotheses $H_{ij}, j = 1, \ldots, n$ are true. Let P_{iG} be the global null

p-value for row i, i.e., a valid p-value of the global null hypothesis $\cap_{j=1}^{n} H_{ij}$. P_{iG} is based on the p-values $p_{ij}, j = 1, \ldots, n$, or on the corresponding unit-level test-statistics. For example, it can be computed as in equation (3.1), or equation (5.1), or as suggested in Remark 3.1. Denoting the observed p-values by lower-case letters, our data matrix for analysis is:

$$\begin{array}{ccc|c} p_{11} & \cdots & p_{1n} & p_{1G} \\ \vdots & \ddots & \vdots & \vdots \\ p_{m1} & \cdots & p_{mn} & p_{mG} \end{array}$$

The first stage analysis selects rows based on the global null p-values $p_{1G}, \ldots, p_{mG}$, to answer the question of which rows are promising, i.e., show evidence that there is signal in at least one column. Let $\mathcal{S} \subseteq \{1, \ldots, m\}$ be the set of selected rows. For example, when the rows are the SNPs and the columns are different studies, the selected rows are often those that achieve genome-wide significance with FWER control, so $\mathcal{S} = \{i : p_{iG} \leq \alpha/m\}$.

The second stage analysis is based on the individual p-values $\{p_{ij} : j = 1, \ldots, n, i \in \mathcal{S}\}$, to answer questions such as which columns within selected rows contain signal.

Let V_i and R_i be the number of false and total rejections for row i, respectively. Our first error measure, which henceforth is referred to as the conditional FWER for a selected row, is the probability of at least one erroneous rejection within the row, conditional on it being selected,

$$E(I[V_i > 0] | i \in \mathcal{S}).$$

Our second error measure, which henceforth is referred to as the conditional FDR for a selected row, is the expected fraction of erroneous rejections among the rejections within the row, conditional on it being selected,

$$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}).$$

If $n = 1$, our error measures coincide with the selective type I error rate, i.e. the error rate of a test given that it was performed, as suggested by Fithian et al. (2015).

Our goal is to suggest valid multiple testing procedures for controlling the conditional FWER/FDR for the selected rows. This is done by computing conditional p-values, where for each column of a selected row the p-value is computed conditional on the event that the row was selected, holding all other p-values in the row fixed. This is shown in the next section.

3 Valid inference within a selected row

For each row $i \in \{1, \dots, m\}$, the p-values corresponding to the different columns are assumed independent. The computation of p_{iG} can be based on different aggregates of the p-values $p_{ij}, j = 1, \dots, n$, see Loughin (2004) for common choices and a review. A popular choice is Fisher's combining method, which has been shown to have excellent power properties (see, e.g., Owen (2009) and the references within). Let $f : \Re^n \to \Re$ be a combining function for testing the global null on each row. For example, Fisher's combining function is $f(p_{i1}, \dots, p_{in}) = -2\sum_{j=1}^n \log p_{ij}$. Under the global null of no signal in the row, if the p-values in the row are uniformly distribution, $-2\sum_{j=1}^n \log P_{ij}$ has a chi-squared distribution with $2n$ degrees of freedom, $-2\sum_{j=1}^n \log P_{ij} \sim \chi^2_{2n}$, so

$$p_{iG} = Pr(\chi^2_{2n} \geq -2\sum_{j=1}^{n} \log p_{ij}). \tag{3.1}$$

See Remark 3.1 and Section 5 for other combining functions.

Let t_i be the fixed row selection cut-off, so that row i is selected if $f(p_{i1}, \dots, p_{in}) \geq t_i$. The results in this section hold if t_i does not depend on $P_{i1}, \dots, P_{in}$. The independence between the threshold and the ith row p-values is clearly satisfied if the thresholds is fixed given m and n (or more generally, given the number of non-missing observations in row i, say n_i). For example, for GWAS, the threshold is often chosen so that the FWER on the m global tests is controlled at the desired nominal level α. If the Bonferroni procedure is applied for selecting the rows, then the selection threshold for Fisher's combining method is $t = \chi_{1-\alpha/m,2n}$, i.e., the $(1-\alpha/m)$ quantile of χ^2_{2n}. If the global null p-values are independent, the results in this section hold if t_i is chosen by a data-adaptive multiple testing procedure on the global null p-values, and we address this in Section 3.1.

We require that the combining method be non-increasing in each of its arguments, when the other $n-1$ values are held fixed. This is a reasonable requirement since if the individual p-values decrease, the selection of the row should become easier, so if $f(p_{i1}, \dots, p_{in}) \geq t$ and $p'_{ij} \leq p_{ij}, j = 1, \dots, n$, then $f(p'_{i1}, \dots, p'_{in}) \geq t$. For simplicity, we shall assume that the p-values are continuous, with a uniform null distribution, and therefore that f is continuous.

Since we assume we are dealing with a selected row based on the global test, we will omit the index i for row from hereon to simplify the notation. The n independent p-values for a selected row are therefore denoted by $P_1, \dots, P_n$, and their observed values are $p_1, \dots, p_n$. We observed that we rejected the global null which means that the p-values satisfy $f(p_1, \dots, p_n) \geq t$.

We let b_j satisfy

$$f(p_1, \dots, p_{j-1}, b_j, p_{j+1}, \dots, p_n) = t. \tag{3.2}$$

Clearly, $b_j \geq p_j$. If no such b_j exists, that is $f(p_1, \dots, p_{j-1}, 1, p_{j+1}, \dots, p_n) \geq t$, let $b_j = 1$.

The aim is to provide a test for each of the n hypotheses having observed that the global null was rejected. To this end, we adjust the p-values to

$$p'_j = p_j/b_j, \quad j = 1, \dots, n. \tag{3.3}$$

These are valid conditional p-values. To see this, note that P_j conditional on the row being selected and on the remaining p-values is truncated at $b_j \in (0, 1]$. So if unconditionally P_j is uniformly distributed, $P_j \sim U(0, 1)$, then conditionally on being selected and on the remaining p-values it has a uniform distribution between zero and b_j,

$$P_j \mid f(p_1,\ldots,p_{j-1},P_j,p_{j+1},\ldots,p_n) \geq t, P_1 = p_1,\ldots,P_{j-1} = p_{j-1}, P_{j+1} = p_{j+1},\ldots,P_n = p_n \sim U(0,b_j).$$

Therefore, the null distribution of P'_j, conditional on the row being selected, is uniform between zero and one.

Applying a valid FWER/FDR controlling procedure on $\{p'_j : j = 1,\ldots,n\}$ will guarantee control of the conditional FWER/FDR for a selected row. The choice of FWER/FDR controlling procedure should be guided by the observation that the n conditional p-values in the row are dependent, despite the fact that the original p-values were independent across columns. The conditional FWER for a selected row will be controlled if we use Bonferroni-Holm on the conditional p-values within each selected row, since the Bonferroni-Holm procedure guarantees FWER control for any dependency between the p-values. For conditional FDR control, we would like to use the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg , 1995) within each selected row at level α. The next theorem shows that for the dependency among the p-values induced by the conditioning step, the BH procedure is indeed valid.

Theorem 3.1. *Let $P_1,\ldots,P_n$ be independent p-values, each with a uniform null distribution. If $f(P_1,\ldots,P_n) \geq t$, then the BH procedure at level α on $P'_1,\ldots,P'_n$ controls the conditional FDR at level $\leq \frac{n_0}{n}\alpha$, where n_0 is the number of true null hypotheses.*

See Appendix A for the proof.

For Fisher's combining function, $f(p_1,\ldots,p_n) \geq t$ is equivalent to $\Pi_{j=1}^n p_j \leq e^{-t/2}$. Therefore, $b_j = \min\left(e^{-t/2}/(\Pi_{l=1,l\neq j}^n p_l), 1\right)$, and the conditional p-values are:

$$p'_j = \begin{cases} p_j & \text{if } \Pi_{l=1,l\neq j}^n p_l \leq e^{-\frac{1}{2}t}, \\ \frac{\Pi_{l=1}^n p_l}{e^{-\frac{1}{2}t}} & \text{otherwise.} \end{cases} \qquad j = 1,\ldots,n \tag{3.4}$$

Interestingly, for Fisher's selection rule it can be shown that the conditional FDR is exactly $n_0\alpha/n$.

Corollary 3.1. *Let $P_1, \ldots, P_n$ be independent p-values, each with a uniform null distribution. If $f(p_1, \ldots, p_n) = -2\sum_{j=1}^{n} \log p_j \geq t$, then the conditional FDR of the BH procedure at level α on $p'_1, \ldots, p'_n$ is equal to $\frac{n_0}{n}\alpha$.*

See Appendix B for the proof.

The magnitude of the inflation in the post-selection p-values, $p'_j - p_j$, depends on how the signal is distributed in the row. Using Fisher's combining method, there can be no inflation (i.e., no cost for selection!) if the signal is strong in at least two columns. This is a direct result of the conditional p-value computation (3.4), where $p'_j = p_j$ if $\Pi_{l=1, l\neq j}^{n} p_l \leq e^{-\frac{1}{2}t}$. Note that even though it is enough to have one strong signal for the selection of the row, at least two columns need to contain strong signal for all n conditional p-values to coincide with the original p-values. Therefore, computing the conditional p-values as suggested in (3.3) will typically lead to smaller conditional p-values whenever the signal is present in more than one column, than the use of the following probabilities, which are computed under the global null:

$$Pr_{\cap_{j=1}^{n} H_j}(P_j \leq p_j \mid f(P_1, \ldots, P_n) \geq t), j = 1, \ldots, n. \tag{3.5}$$

The computation in (3.5) uses conservatively the uniform distribution for the p-values, $P_k \sim U(0,1)$ for $k = 1, \ldots, n$, and the resulting probabilities can be substantially larger than the original p-values: for Fisher's combining method, if $t = \chi_{1-\alpha/m, 2n}$ and $p_j \leq e^{-\frac{1}{2}t}$, the value will be $\frac{m}{\alpha}p_j$.

Remark 3.1 (Other p-values combining methods)**.** *Conceptually, our approach for computing the conditional p-values can be applied to many different tests of the global null. The complexity of the conditional p-value computation in (3.3) can vary greatly between combining methods. For one-sided tests, the computation is straightforward using Stouffer's combining method, $f(p_1, \ldots, p_n) = \sum_{j=1}^{n} z_j/\sqrt{n}$, where $z_j = \Phi^{-1}(1 - p_j)$, which has a standard normal distribution under the global null. So the global p-value is $1 - \Phi(\sum_{j=1}^{n} z_j/\sqrt{n})$. If the row is selected when $\sum_{j=1}^{n} z_j/\sqrt{n} > t$, then* *the conditional p-values are:*

$$\left\{\frac{p_j}{1-\Phi(\sqrt{n}t-\sum_{l=1,l\neq j}^{n} z_l)}, j=1,\ldots,n\right\}. \qquad (3.6)$$

The computation of the conditional p-values is more complex using the test of Bhattacharjee et al. (2012) for the global null, called ASSET. Their test is based on identifying the set $S_{\max}$ that has the largest weighted Stouffer test statistic, and the significance at the aggregate row level is based on the maximal test statistic, $\sum_{j\in S_{\max}} w_j Z_j/\sqrt{|S_{\max}|}$. Bhattacharjee et al. (2012) showed that their global test can be more powerful than a test statistic that aggregates all the information in the row without subset selection.

Remark 3.2 (Estimation of the fraction of nulls in selected rows)**.** *Theorem 3.1 shows that the FDR of the level-α BH procedure applied on the conditional p-values is bounded by $\pi_0\alpha$ in a selected row, where $\pi_0 = n_0/n$. When n is large enough (say 30 or more), it may be useful to estimate π_0 and incorporate the estimate in the multiple testing procedure to gain power. The gain in power can be substantial if π_0 is far less than one, as appears to be the case in the application considered in Section 5. Schweder and Spjotvoll (1982) proposed estimating π_0 by $\frac{\#\{p-values>\lambda\}}{n(1-\lambda)}$, where $\lambda \in (0,1)$. The slightly inflated plug-in estimator $\hat{\pi}_0 = \frac{\#\{p-values>\lambda\}+1}{n(1-\lambda)}$ has been incorporated into multiple testing procedures in recent years. For independent p-values, Storey (2003) showed that the BH procedure at level $\alpha/(\hat{\pi}_0)$ controls the FDR at level α. Benjamini et al. (2006) suggested another estimator, and noted that the BH procedure which incorporates the plug-in estimator with $\lambda = 0.5$ is sensitive to deviations from the assumption of independence, and its FDR level may be inflated above the nominal level under dependency. We need to investigate which estimator/s will result in an adaptive BH procedure that controls the conditional FDR. Also, a confidence statement about π_0 within selected rows may be of interest in itself, in providing an upper bound on the fraction of nulls (or a lower bound on the fraction of columns with signal) within each selected row.*

3.1 Valid inference using data-adaptive row selection rules

So far we considered only fixed thresholds for selection. If the global null p-values are independent, more general data-adaptive thresholds can also lead to valid inference within selected rows. For example, if the rows are selected based on the discoveries from a BH procedure at level α on the global null p-values, as suggested by Peterson et al. (2015), then the conditional p-values can still be computed in a straightforward manner. Relabeling a selected row as row number one, and the remaining rows in order of their global null p-values, i.e., $p_{2G} \leq \ldots \leq p_{mG}$, let $J = \arg\max_{r=2,\ldots,m}\{p_{rG} \leq \frac{r}{m}\alpha\}$. Then for Fisher's combining function, the selection threshold for a selected row is $t(J) = \chi^2_{1-\frac{J+1}{m}\alpha,2n}$, and the conditional p-values for each selected row are computed by the formulas in (3.4) with $t = t(J)$.

For data-adaptive selection rules, we need the following requirement on a selection rule for computation of conditional p-values.

Definition 3.1 (Definition 1 in Benjamini and Bogomolov (2014)). *A selection rule is called simple if for each selected row, when the p-values not belonging to that row are fixed and the p-values in that row can change as long as the row is selected, the number of selected rows remains unchanged.*

For each $i \in \mathcal{S}$, let $\mathcal{S}^{(i)} = \mathcal{S} \setminus i$, $S^{(i)} = |\mathcal{S}^{(i)}|$, and $P^{(i)} = (P_{1G}, \ldots, P_{(i-1)G}, P_{(i+1)G}, \ldots, P_{mG})$. Row-selection by the BH procedure on the global null p-values, or by Bonferroni-Holm on the global null p-values, are simple row-selection rules. Therefore, the threshold for selection for a row i is a function of the p-values only through $S^{(i)}$. The threshold $t(S^{(i)})$ is independent of $P_{i1}, \ldots, P_{in}$ if the p-values across rows are independent. Therefore, the post-selection inference will remain valid, as formally stated in the following Theorem.

Theorem 3.2. *Assume that rows are selected by a simple selection rule such that row i is selected if $f(p_{i1}, \ldots, p_{in}) \geq t(S^{(i)})$. If the p-values across rows are independent, then if we compute the conditional p-values for selected row i as in (3.3) with $t = t(S^{(i)})$:*

1. *$E(I[V_i > 0] | i \in \mathcal{S}, P^{(i)}) \leq \alpha$ if we use the Bonferroni-Holm procedure on the conditional p-values.*

2. *$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}, P^{(i)}) \leq \alpha$ if we use the BH procedure on the conditional p-values.*

The proof of item 1 follows from the fact that the Bonferroni-Holm procedure is valid for any type of dependency, and the proof of item 2 follows from the proof of Theorem 3.1, using the fact that the conditional p-values in (3.3) are valid p-values due to the independence of the rows and the selection rule being simple.

3.2 Relation to the approach of Benjamini and Bogomolov (2014)

Benjamini and Bogomolov (2014) considered selective inference on families of hypotheses, which are defined by rows in our paper. They showed that applying a Bonferroni procedure in each selected row may result in a highly inflated conditional FWER when the selection is based on within-row p-values. Moreover, they noted that the goal of conditional control for any combination of selection rule and testing procedure and for any configuration of true null hypotheses is difficult to achieve. Indeed, our procedures for conditional control rely on the ability to compute the conditional p-values, which requires well-defined selection rules and independence across the p-values within the rows. Benjamini and Bogomolov (2014) considered a different error measure addressing selective inference, which can be controlled under more general conditions. Specifically, they suggested to control an expected average error measure over the selected rows. While they considered a general class of error rates, we focus on two special cases. The control of the FWER on the average,

$$E\left(\frac{\sum_{i \in \mathcal{S}} I[V_i > 0]}{\max\{|\mathcal{S}|, 1\}}\right),$$

and the control of the FDR on the average,

$$E\left(\frac{\sum_{i\in\mathcal{S}} V_i/\max\{R_i,1\}}{\max\{|\mathcal{S}|,1\}}\right).$$

Benjamini and Bogomolov (2014) showed that by applying an FWER/FDR controlling procedure within each row at level $|\mathcal{S}|\alpha/m$, if the set $\mathcal{S}$ is selected by a simple selection rule, then the FWER/FDR on the average is controlled at level α. However, the conditional FWER/FDR for a selected row may exceed α in some of the rows as our simulations show.

Controlling the FWER/FDR on the average is useful when we require control over false positives on the combined set of discoveries within selected rows. However, when we require control over false positives within each selected row separately, the conditional FWER/FDR is more appropriate.

If the selection threshold is fixed, then a valid FWER/FDR-controlling procedure on the conditional p-values guarantees control of FWER/FDR on the average over the selected rows if the p-values across rows are independent, as formally stated in the next lemma.

Let C_i be the (unobserved) random variable whose expectation is the desired error rate, so $C_i = I[V_i > 0]$ for FWER control, and $C_i = V_i/\max(R_i, 1)$ for FDR control.

Lemma 3.1. *Assume that row i is selected if $f(p_{i1},\dots,p_{in}) \geq t$ where t is a fixed threshold. If the p-values across rows are independent, then $E(C_i|i\in\mathcal{S}) \leq \alpha$ implies $E\left(\frac{\sum_{i\in\mathcal{S}} C_i}{\max(|\mathcal{S}|,1)}\right) \leq \alpha$.*

If the selection is made by a simple selection rule (definition 3.1) on the global null p-values, then Theorem 3.2 states that the level α Bonferroni-Holm and BH procedures on the conditional p-values are valid FWER and FDR controlling procedures, respectively. A valid conditional FWER/FDR-controlling procedure also controls the FWER/FDR on the average over the selected rows if the p-values across rows are independent, as stated in the next Lemma.

Lemma 3.2. *Assume that rows are selected by a simple selection rule such that row i is selected if $f(p_{i1},\dots,p_{in}) \geq t(S^{(i)})$. If the p-values across rows are independent, then $E(C_i|i\in\mathcal{S}, P^{(i)}) \leq \alpha$*

implies $E\left(\frac{\sum_{i\in\mathcal{S}} C_i}{\max(|\mathcal{S}|,1)}\right) \leq \alpha$.

See Supplementary Section S1 for the proofs.

4 Simulations

In order to assess the performance of different post-selection approaches, we carry out simulation studies. We have two specific aims in this simulation: (1) to examine the possible inflation in the number of false positives when failing to account for the selection (i.e., when using the original p-values in selected rows); and (2) to compare the power of methods that properly account for selection, i.e., our novel approach that uses conditional p-values, and the approach of Benjamini and Bogomolov (2014).

Our data generation scenario is as follows. We sample the unit-level test statistics, Z_{ij}, $i = 1,\ldots,m, j = 1,\ldots,n$, independently from the normal distribution as follows. If the null hypothesis H_{ij} is true, the test statistic has a standard normal distribution; if the null hypothesis H_{ij} is false, the test statistic has a normal distribution with mean $\mu \in \{0.5,\ldots,7\}$ and variance of one. The p-value is $p_{ij} = 1 - \Phi(z_{ij})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. In each of n columns, $m = 1000$ rows are examined, and $m_1 = 10$ rows contain signal in n_1 of the n columns. So we generated $n_1 \times m_1$ entries in the $n \times m$ matrix of p-values that contained signal we aim to discover.

The selection of rows is done by applying the Bonferroni procedure at level $\alpha = 0.05$ on the global null p-values, computed using equation (3.1). So row i is selected if $p_{iG} \leq \alpha/m$.

We consider the following post-selection analyses using the BH procedure. The BH procedure on each selected row $i \in \mathcal{S}$: at level α on the original p-values (BH-naive); at level $\frac{|\mathcal{S}|}{m}\alpha$ on the original p-values, as suggested by Benjamini and Bogomolov (2014) (BH-BB); at level α on the conditional p-values in (3.4) (BH-cond). We also considered using the Bonferroni-Holm procedure, and these results turned out to be qualitatively the same as the results using the BH procedure, see details in the Supplementary Material (SM).

The results on the average and conditional FDR control show that the control of false positives was tightest using the conditional approach, since it controlled all the error measures (as expected) at the nominal 0.05 level. The approach of Benjamini and Bogomolov (2014) controlled at the nominal 0.05 level the average FDR/FWER (as expected), as well as the conditional FDR/FWER for a correctly selected row (i.e., a row that contains signal), but did not control the conditional FDR/FWER for an incorrectly selected row (i.e., a row that contains no signal). The naive approach exceeded the nominal 0.05 level in all error measures considered (as expected, since this procedure does not account for selection). Figure 1 shows the actual level of each error measures for the three post selection analyses procedures. We see that BH-naive can have a high inflation of false positives. The inflation for an incorrectly selected row was so high that it was not plotted in row 3. The fact that the inflation can be substantial clearly demonstrates that accounting for selection is necessary even when the selection criterion is very stringent. The fourth row of Table S1 in the SM shows the exact levels of the error measures for BH-naive for the setting in the last row of Figure 1 when $\mu = 4$: on a correctly selected row (i.e., a row that contains signal), the conditional FDR was 0.065; on an incorrectly selected row the conditional FDR level was 0.996; and the average FDR was 0.074.

The results on the power show that the conditional approach is best when (1) many correctly selected rows are expected to contain signal in more than two columns; and (2) the number of rows selected out of the large number of rows examined is expected to be small (Figure 1). Note that in this simulation the average power is the same as the probability of discovering a true signal (i.e., rejecting a single non-null hypothesis), since we generated the signal (i.e., non-null test statistics) from the same distribution, and the same number of signals (i.e., non-nulls) in each of the m_1 rows. Although the naive procedure should not be used due to its unacceptable inflation of false positives, we plot its power so it can serve as a benchmark for the power loss due to the necessary adjustment for selection. Examining the power of BH-BB and BH-cond, we see that the conditional approach is more powerful than the approach of Benjamini and Bogomolov (2014) when $(n, n_1) \in \{(21, 7), (15, 5), (10, 4)\}$. The gain in power can be very large. For example, when $\mu = 3$ the power difference between our approach and that of Benjamini and Bogomolov (2014) was greater than 40% for $(n_1, n) = (7, 21)$ and $(n_1, n) = (5, 15)$ and about 30% for $(n_1, n) = (4, 10)$. Our approach was slightly less powerful than the approach of Benjamini and Bogomolov (2014) for $(n, n_1) = (10, 2)$. Additional results are provided in Supplemental Table S1.

5 Cross-tissue eQTL analysis in The Cancer Genome Atlas (TCGA) Project

Expression quantitative trait loci (eQTLs) are genomic regions with genetic variants that influence the expression level of genes. Identifying eQTLs is important for understanding biological mechanisms that controls various normal physiological processes and their aberrations that can lead to complex diseases. Because gene regulation is tissue specific, eQTL analysis is most informative using relevant tissue samples, which suffers frequently from the lack of statistical power because of the small sample size for the tissue. It is, however, observed that some eQTL SNPs are predictive of gene expression levels across multiple tissues and identification of such eQTLs could be facilitated by aggregated analysis across tissue types. A number of studies (Rivas et al. (2015) and Li et al., (2016), among others) have reported results based on such cross-tissue eQTL analysis using the data from the Genotype-Tissue Expression (GTEx) project. In this section, we illustrate the post-selection procedure in an eQTL analysis using 17 tumor tissues in The Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/). We first performed an aggregated eQTL analysis across 17 tissue types to identify eQTL SNPs influencing the gene expression in at least one tissue type. For significant eQTLs, we performed post-selection inference to identify tissue types with the eQTL effect. We downloaded genotype and total gene expression data based on

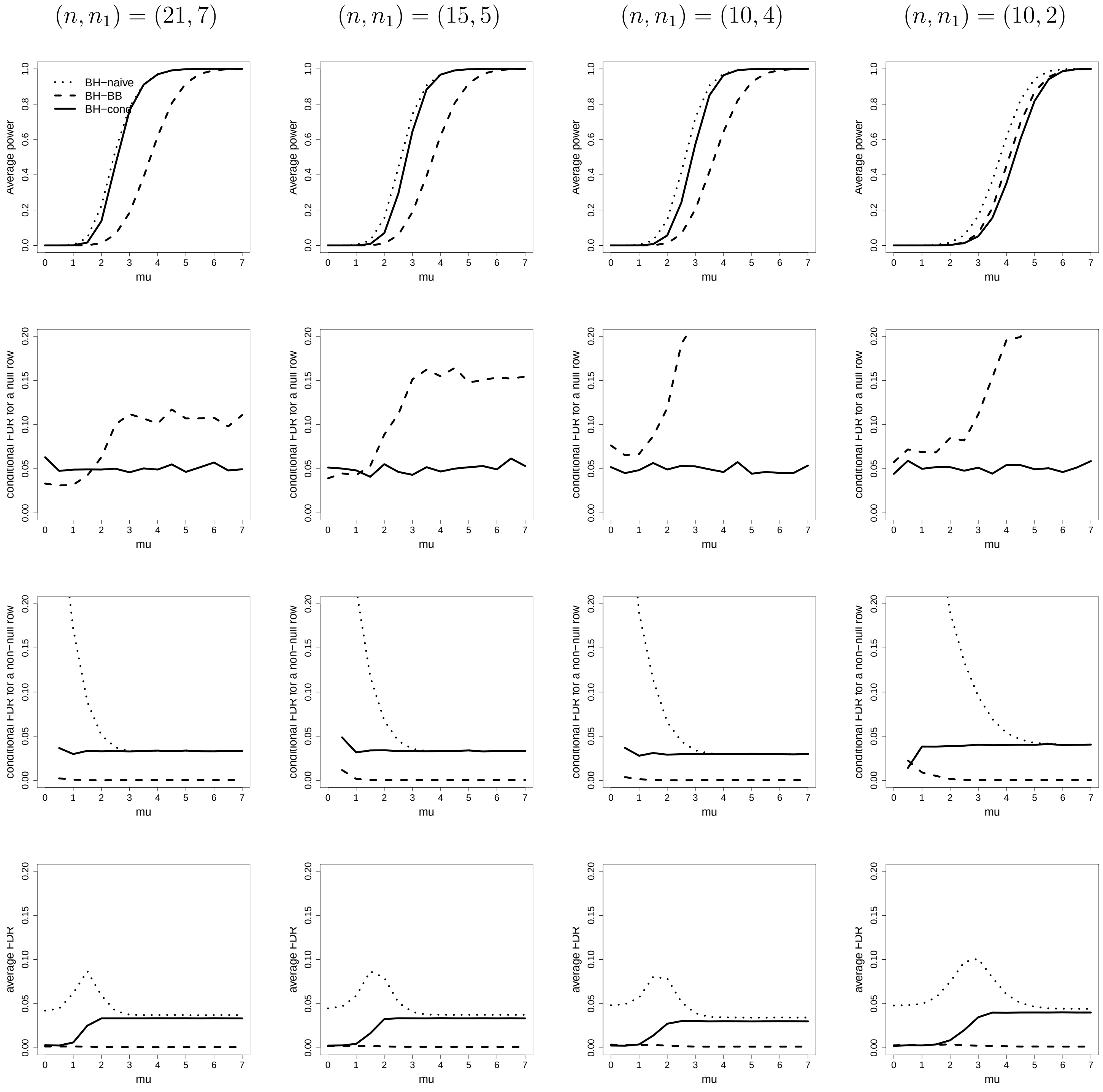

Figure 1: From left to right, variations in the total number of studies (n), and the number of studies with signal (n_1) within the $m_1 = 10$ rows that contained signal, out of a total of $m = 1000$ rows. From top to bottom: (1) average power, i.e., the expected fraction of discoveries among the true signals; (2) conditional FDR for a null row, i.e., a row that was selected despite the fact that all columns are null (the naive procedure does not appear because its value was above 0.80); (3) conditional FDR for a non-null row, i.e., a row that was correctly selected since it has signal in at least one column; and (4) average FDR. The power and error rates as a function of μ (the signal strength in the non-null studies) of the three post selection analyses using the BH procedure: at level 0.05 on the original p-values (dotted line); at level $|\mathcal{S}|0.05/m$ on the original p-values (dashed line); at level 0.05 on the conditional p-values in (3.4) (solid line). Estimated using 50000 datasets.

RNA sequencing from the TCGA website. The data quality control (QC) for genetic data and the normalization for the RNA-seq data are described in Supplementary Materials. After QC, 4,476 subjects with European ancestry and 19,285 genes were included for analysis. The sample size for each tissue type is reported in Supplementary Table S2. For the purpose of illustration, we performed cis-eQTL analysis (i.e., analyzing SNPs less than 1,000,000 base pairs from the target gene) using matrixeQTL (Shabalin, 2012), adjusting for sex, age and the top five principal component scores to eliminate the potential confounding due to population stratification. In total, we analyzed $m = 7,732,750$ SNP-gene pairs to identify cis-eQTLs. Because samples are independent across tissue types in TCGA, the p-values are independent across tissues.

For selection of eQTL SNPs (i.e., the rows), we combined p-values across the different tissue types (columns) by a Fisher style test suggested by Pearson, as described in Owen (2009): it runs the Fisher combining method for left-sided alternatives, and separately for right-sided alternatives, and takes the maximum of the two resulting statistics. The global null p-value is therefore

$$p_{iG} = 2Pr\left[\chi^2_{2n} \geq \max\left\{-2\sum_{j=1}^{n}\log p^L_{ij}, -2\sum_{j=1}^{n}\log(1-p^L_{ij})\right\}\right], \tag{5.1}$$

where p^L_{ij} is the p-value when testing the left sided alternative for feature i in study j. This test has a strong preference for common directionality (i.e., it will have greater power than a test based on Fisher's combining method on two-sided p-values when the direction of the signal is consistent across tissues), while not requiring us to know the common direction. We identified 19,690 significant SNP-gene pairs using the global null p-value threshold $0.05/7,732,750 = 6.47 \times 10^{-9}$ based on the Bonferroni correction for FWER control at the 0.05 level. For each of the 19,690 SNP-gene pair, we proceeded to post-selection inference to identify relevant tissue types using our conditional approach.

Table 1 demonstrates the post-selection analysis for three SNP-gene pairs, which differ in the number of identified tissue with signal in the post-selection inference. For the pair rs13066873-LARS2, the conditional p-values are identical to the original p-values, i.e., there is no cost for selection, since this pair would have been selected regardless of the realized p-value in a single tissue, when conditioning on all the other tissue p-values for this pair. From the BH adjusted p-values provided in column 10, we see that with the conditional approach 15 tissues are discovered which have adjusted p-values at most 0.05, but with the approach of Benjamini and Bogomolov (2014) only the tissue LUAD is discovered, i.e., the single tissue with BH-adjusted p-value at most $0.05 \times 19,690/7,732,750 = 0.00013$. For the pair rs1437891-ASNSD1 in columns 5-7, with the conditional approach six tissues are discovered, and these tissues have conditional p-values larger than the original p-values. With the approach of Benjamini and Bogomolov (2014) no discoveries are made. For the pair rs7977641-GALNT9 in columns 2-4, no discoveries are made in either method.

Using the BH procedure on selected rows, the median number of tissue discoveries was 6, with an inter-quartile range (IQR) of [4,8]. For comparison, we also applied the approach of Benjamini and Bogomolov (2014), which made far fewer discoveries: the median number of discoveries was 1, with IQR [0,2]. The conditional approach results in many more discoveries than the approach of Benjamini and Bogomolov (2014) for two reasons. First, because the number of selected SNP-expression pairs is far smaller than the number originally examined, 19,690/7,732,750=0.0025, and the approach of Benjamini and Bogomolov (2014) is more conservative the smaller this ratio is. Second, because for many SNP-expression pairs there are at least two tissues which are highly significant, thus the conditional p-values coincide with the original p-values, and there was essentially no cost for selection within these pairs.

An R implementation of the conditional p-value computation after selection by thresholding the global null p-values computed using (5.1) or (3.1) is available upon request from the first author (and will soon become available as a Bioconductor package).

Table 1: The original two-sided p-values, conditional two-sided p-values, and BH-adjusted conditional two-sided p-values for each tissue, for three eQTL SNPs that differ in the number of post-selection discoveries: rs10896016-CTSW identified no tissues (columns 2–4); rs1437891-ASNSD1 identified 6 tissues (columns 5–7); rs13066873-LARS2 identified 15 tissues (columns 8-10). Significant discoveries at the 0.05 FDR nominal level are in bold for the BH-adjusted conditional p-values. The global null p-value for each SNP-gene pair are provided in the last row.

	rs10896016-CTSW p-values			rs1437891-ASNSD1 p-values			rs13066873-LARS2 p-values		
	original	conditional	BH-adjusted	original	conditional	BH-adjusted	original	conditional	BH-adjusted
BLCA	0.0126	0.2951	0.3859	0.4552	0.4552	0.6449	0.0020	0.0020	**0.0048**
BRCA	0.7327	0.7327	0.8304	0.0003	0.0080	**0.0228**	0.0003	0.0003	**0.0015**
COAD	0.2660	0.2951	0.3859	0.0023	0.0023	**0.0228**	0.0010	0.0010	**0.0036**
GBM	0.3609	0.2951	0.3859	0.9023	0.9023	0.9023	0.0072	0.0072	**0.0135**
HNSC	0.9225	0.9225	0.9801	0.5471	0.5471	0.6643	0.5439	0.5439	0.5439
KIRC	0.0074	0.2951	0.3859	0.0000	0.0080	**0.0228**	0.0136	0.0136	**0.0178**
KIRP	0.9958	0.9958	0.9958	0.5197	0.5197	0.6643	0.0083	0.0083	**0.0142**
LAML	0.0235	0.2951	0.3859	0.7783	0.7783	0.8269	0.0034	0.0034	**0.0073**
LGG	0.1396	0.2951	0.3859	0.0000	0.0080	**0.0228**	0.0011	0.0011	**0.0036**
LIHC	0.0157	0.2951	0.3859	0.3441	0.3441	0.6449	0.0101	0.0101	**0.0143**
LUAD	0.0000	0.2951	0.3859	0.0008	0.0080	**0.0228**	0.0000	0.0000	**0.0000**
LUSC	0.1291	0.2951	0.3859	0.3034	0.3034	0.6448	0.0407	0.0407	**0.0483**
OV	0.0666	0.2951	0.3859	0.1626	0.1626	0.3948	0.0096	0.0096	**0.0143**
PAAD	0.2567	0.2567	0.3859	0.6417	0.6417	0.7272	0.0426	0.0426	**0.0483**
PRAD	0.1409	0.2951	0.3859	0.0050	0.0080	**0.0228**	0.0641	0.0641	0.0681
SKCM	0.0158	0.2951	0.3859	0.4150	0.4150	0.6449	0.0002	0.0002	**0.0015**
UCEC	0.5923	0.5923	0.7192	0.4291	0.4291	0.6449	0.0017	0.0017	**0.0047**
Global null p-value	3×10^{-9}			2×10^{-10}			$< 10^{-20}$		

5.1 Valid inference within the selected most significant SNP-expression pair in a gene

For a target gene, there might be multiple SNPs in the cis region that achieve the genome-wide significance. Most likely, these SNPs are in strong linkage disequilibrium (LD) and represent one eQTL. To avoid reporting redundant eQTLs, one natural step is to choose the SNP with the smallest global p-value and perform post-selection inference to identify relevant tissue types. However, the post-selection inference may suffer from high false positive rate if this second selection is not appropriately accounted for. Related simulation results are provided in Supplemental Figure S2. Thus, we propose a simple modification to our post-selection inference method to account for the second selection. We denote the recommended procedure for FDR control by BH-cond-MT, where MT stands for the additional Multiple Tests (of the SNPs that passed the first selection threshold at the aggregate level in the gene) that we need to correct for, after computing the conditional p-values.

Procedure 5.1. *BH-cond-MT post-selection procedure for conditional FDR control at level α:*

1. *Select all SNP-expression pairs that reach the genome-wide significance threshold $h(\alpha, m)$ (e.g., $h(\alpha, m) = \alpha/m$ if the Bonferroni correction is used), $\mathcal{S} = \{i : p_{iG} \leq h(\alpha, m)\}$.*

2. *For each gene k with at least one selected SNP, select the most significant SNP in gene k, $i_k = \arg\min_{\{i:i\in\mathcal{S},i \text{ in gene } k\}} p_{iG}$. Let $R_k = |\{i : i \in \mathcal{S}, i \text{ in gene } k\}|$ be the number of SNPs selected in gene k.*

3. *Compute the conditional p-values as in (3.3) for row i_k.*

4. *At level α/R_k on the conditional p-values in row i_k, apply the BH procedure.*

For conditional FWER control, apply Bonferroni-Holm instead of BH in step 4 of Procedure 5.1. In Supplemental Figure S3 we show that the procedure controls the FDR for dependencies across SNPs similar to those that arise in GWAS datasets within each study due to LD, and in Appendix C we formally prove that the conditional FDR is controlled when the rows are independent.

The impact of the second selection is illustrated in Table 2 for gene KIAA0141. Four SNPs (rs351260, rs164515, rs164084, rs164075) were identified for KIAA0141. Out of the four SNPs, rs351260 had the strongest overall association and was selected for post-selection inference to identify relevant tissues. Without accounting for the second selection, 14 tissue types were determined as significant based on the BH-adjusted conditional p-value < 0.05. After accounting for the second selection, 9 tissue types were counted as significant based on the new threshold BH-adjusted conditional p-value$< 0.05/4 = 0.0125$.

Finally, we compared the performance of four methods based on the whole eQTL analysis: BH-naive, BH-cond and BH-cond-MT and BH-BB (Table 3). Note that BH-naive and BH-cond do not account for second selection and thus their error rates are likely to be inflated. Consistent with simulation studies, BH-BB detected the smallest number of signals because of lower power. For example, among the 2235 selected SNPs, at least two tissue discoveries were made in 1857 of

Table 2: The original p-values for the four SNPs for gene KIAA0141 that passed the genome-wide selection threshold (columns 3 to 6), as well as the BH adjusted conditional p-values for the most significant SNP (column 6). In bold the adjusted p-values $\leq \frac{0.05}{4}$, i.e., the significant discoveres using Procedure 5.1 with $\alpha = 0.05$. Underlined are the adjusted p-values in $(0.05/4, 0.05]$, i.e., discoveries using BH-cond that are not discoveries using Procedure 5.1 with $\alpha = 0.05$. The conditional p-values were identical to the unconditional p-values in all four SNPs (since for each tissue j, the SNP would have been selected regardless of the value of the jth p-value). The global null test-statistic for each SNP is provided in the last row (the corresponding p-value was effectively zero).

tissue	p-value	p-value	p-value	p-value	BH adjusted p-value
type	rs164515	rs164084	rs164075	rs351260	rs351260
BLCA	0.0194	0.1321	0.1509	0.0028	**0.0059**
BRCA	0.0110	0.0211	0.0185	0.0003	**0.0008**
COAD	0.3479	0.3008	0.3008	0.0367	0.0445
GBM	0.0869	0.5607	0.5911	0.0792	0.0897
HNSC	0.3926	0.2164	0.2029	0.5692	0.5692
KIRC	0.0000	0.0000	0.0000	0.0000	**0.0000**
KIRP	0.0125	0.0337	0.0305	0.0300	0.0403
LAML	0.1334	0.0654	0.0901	0.0057	**0.0107**
LGG	0.0001	0.0768	0.0890	0.0000	**0.0000**
LIHC	0.0401	0.1214	0.0787	0.0308	0.0403
LUAD	0.0033	0.0142	0.0102	0.0004	**0.0009**
LUSC	0.0000	0.0000	0.0000	0.0000	**0.0000**
OV	0.0005	0.1115	0.0788	0.0003	**0.0008**
PAAD	0.1789	0.0942	0.1404	0.4318	0.4588
PRAD	0.0026	0.0029	0.0021	0.0001	**0.0003**
SKCM	0.0595	0.8552	0.6756	0.0083	0.0142
UCEC	0.0314	0.0358	0.0225	0.0216	0.0334
global null test-statistic	343.7	361.6	346.0	396.8	

the genes by BH-cond-MT and only in 678 of the genes by BH-BB. For the BH procedure based on conditional p-values, accounting for second selection noticeably reduced the significant findings. This was expected, since the number of SNPs discovered per gene is typically greater than one: out of the 19690 pairs that reached genome-wide significance, there were 2235 unique genes, and the number of SNPs per gene varied between 1 and 76, with a median number of 5.

6 Discussion

Results from both simulation studies and data analysis highlight the potential of the proposed method for valid and powerful hypothesis testing for detection of signals at the level of the finer units following selection of broader units using aggregate level test-statistics. Although the method is not as general as its existing competitor (Benjamini and Bogomolov , 2014), it can handle an important class of scenarios that involve independence of the primary test-statistics across columns,

Table 3: By each post-selection method, the number of genes with at least x discoveries of tissues, for $x = 0, 1, \ldots, 14$. The first row represents the genes with at least 0 discoveries, i.e., the number of selected genes. The theoretically valid methods, which control their respective error rates, are BH-BB (column 5), as well as the methods that adjust for the selection of a single row per gene but use the conditional p-values, BH-cond-MT (column 4), described in Procedure 5.1.

minimum #				
of discoveries	BH-naive	BH-cond	BH-cond-MT	BH-BB
0	2235	2235	2235	2235
1	2235	2003	1936	1309
2	2200	1980	1857	678
3	2105	1938	1653	352
4	1891	1808	1352	160
5	1642	1615	983	59
6	1348	1333	677	7
7	1081	1081	447	2
8	832	832	266	0
9	604	604	132	0
10	375	375	50	0
11	195	195	15	0
12	84	84	8	0
13	21	21	0	0
14	1	1	0	0

a practical context of which is demonstrated through the application involving cross-tissue eQTL analysis in the rich TCGA dataset. The superior power of the proposed procedure over that of Benjamini and Bogomolov (2014) in this particular setting implies that a general error-controlling method may not be very powerful for specific applications. Thus, substantial scope for future research exist for development of other powerful procedures tailored towards specific important application settings following the general principles we lay out.

If the columns are dependent, it is an open question how to compute valid conditional p-values. In this work we relied on the fact that the null distribution for a unit-level test statistic is known when we condition on the selection event and on all the other p-values in its row. When the columns are dependent, the null distribution after conditioning on the selection event and on all other p-values in the row may still depend on unknown parameters. The approach of Benjamini and Bogomolov (2014) remains valid in this case, since it is not sensitive to dependence across columns, as long as the within row multiple testing procedure controls the desired error rate for the dependence. In applications where the dependency across columns is approximately known, it may be possible to compute the conditional p-values and carry on the post-selection inference as we suggest in this paper. We plan to investigate the usefulness of this approach for specific applications in future work.

Other post-selection analyses may be of interest. For example, estimation of the fraction of columns containing signal within each selected row. Such estimates can be useful in separating the selected rows where there is signal in most columns, from the rows driven by very few (one or two) columns only that contain signal. Another example is the estimation of a linear combination of the effect sizes. A conditional approach can be useful for these post-selection estimation problems.

A Proof of Theorem 3.1

Proof. Without loss of generality, relabel the columns so that the first n_0 columns have a true null hypothesis. We make use of the representation of FDR from Benjamini and Yekutieli (2001), so the conditional FDR is

$$n_0 \sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1 \text{ and } R = k \mid f(P_1, \ldots, P_n) \geq t\right), \tag{A.1}$$

where $I = 1$ if H_1 (the first hypothesis), which we assume to be null, is rejected, and R is the number of rejected hypotheses. We condition on $p_2, \ldots, p_n$ in which case b_1 is fixed. It is sufficient to show that

$$\sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1 \text{ and } R = k \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) \leq \frac{\alpha}{n}. \tag{A.2}$$

for all $p_2, \ldots, p_n$.

The only random quantity is P_1, the p-value for H_1. It follows that

$$P_1 \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n \sim U(0, b_1).$$

We then need to determine the joint behavior of I and R as P_1 varies from zero to b_1 (or P_1' varies from zero to one).

As p_1 varies, so will $b_2, \ldots, b_n$: as p_1 gets larger $b_2, \ldots, b_n$ will be non-increasing. This implies that $p_2', \ldots, p_n'$ are non-decreasing. Hence R as a function of p_1 is non-increasing.

What is critical is that $I = 1$ if and only if $p_1' \leq \frac{R(p_1)\alpha}{n}$, or

$$p_1 \leq \frac{R(p_1)\alpha}{n} b_1.$$

The left-hand side increases from zero to b_1. The right-hand side is a step-function where the steps decrease. If we let

$$x = \sup_{p_1} \left\{ \frac{R(p_1)\alpha}{n} b_1 - p_1 \geq 0 \right\},$$

then $I = 1$ if $p_1 \in [0, x]$. Let $h = R(x)$. Then h is the smallest value that R attains while I remains one. So

$$\begin{aligned}
&\sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1 \text{ and } R = k \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) \\
&\leq \frac{1}{h} \sum_{k=1}^{n} Pr\left(I = 1 \text{ and } R = k \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) \\
&= \frac{1}{h} Pr\left(I = 1 \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) \\
&= \frac{1}{h} Pr\left(P_1 \leq x \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) \\
&= \frac{1}{h} \frac{x}{b_1} \leq \frac{\alpha}{n}
\end{aligned}$$

where the last inequality follows from the definition of x. □

B Proof of Corollary 3.1

Proof. Without loss of generality order the p-values such that $p_2 \leq \ldots \leq p_n$, and let $J = \arg\max_{2\leq k\leq n}\{p_k \leq \alpha k/n\}$. Note that $R \leq J$, since the number of $p_2, \ldots, p_m$ that does not exceed $\alpha k/n$ is at most $k-2$ for $k > J$. Since $p'_j \geq p_j$ there are at most $k-2$ of the p'_js that do not exceed $\alpha k/n$. Even if $p'_1 \leq \alpha k/n$, the kth largest from among $p'_1, \ldots, p'_n$ exceeds $\alpha k/n$. It follows that if $p'_1 > \alpha J/n$ then $I = 0$.

If $p'_1 \leq \alpha J/n$ and $b_1 = 1$, for $r \leq J$: p'_r is either equal to p_r and therefore $p'_r \leq \alpha J/n$, or $p'_r = \frac{\Pi_{l=1}^{n} p_l}{e^{-\frac{1}{2}t}} = p_1\left(\frac{\Pi_{l=2}^{n} p_l}{e^{-\frac{1}{2}t}}\right) \leq p_1 \leq \alpha J/n$. If $p'_1 \leq \alpha J/n$ and $b_1 < 1$, for $r \leq J$: since $p'_1 = \frac{\Pi_{l=1}^{n} p_l}{e^{-\frac{1}{2}t}} \leq \alpha J/n$, then $p'_r \leq \alpha J/n$. Therefore, if $p'_1 \leq \alpha J/n$ then $p'_r \leq \alpha J/n$ for $r \leq J$, i.e., $R = J$.

Therefore,

$$\sum_{k=1}^{n} \frac{1}{k} Pr\left(I = 1 \text{ and } R = k \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right)$$
$$= \frac{1}{J} Pr\left(P'_1 \leq \alpha J/n \mid f(P_1, p_2, \ldots, p_n) \geq t, P_2 = p_2, \ldots, P_n = p_n\right) = \frac{\alpha}{n}.$$

□

C Proof of conditional FDR control for Procedure 5.1

Procedure 5.1 controls the conditional FDR when the rows are independent, as formally stated in the following theorem.

Theorem C.1. *For i in gene k, let $P_{i1}, \ldots, P_{in}$ be independent p-values, which are also independent of P_{jG}, j in gene $k, j \neq i$. If R_k of the SNPs in gene k have global null p-values at most $h(\alpha, m)$, then the BH procedure at level α/R_k on $P'_{i_k 1}, \ldots, P'_{i_k n}$ controls the conditional FDR at level $\leq \alpha$.*

Proof. Relabel the rows so that the first m_k rows are the SNPs for gene k, and the first R_k rows are the selected SNPs for gene k. Denote the selection event that $R_k \geq 1$ rows for gene k are selected by $\mathcal{A}$, i.e.,

$$\mathcal{A} = \{P_{1G} \leq h(\alpha, m), \ldots, P_{R_k G} \leq h(\alpha, m), P_{(R_k+1)G} > h(\alpha, m), \ldots, P_{m_k G} > h(\alpha, m)\}.$$

Within each row i, rearrange the columns so that the first n_{0i} columns have a true null hypothesis for SNP i. Let $\phi_{ij} = 1$ if the jth column for SNP i is rejected. The (unobserved) number of false rejections for gene k is $V = \sum_{r=1}^{R_k} \sum_{j=1}^{n_{0r}} I[\phi_{rj} = 1 \text{ and } \min_{i=1,\ldots,R_k} P_{iG} = P_{rG}]$, and the (observed) number of rejections is $R = \sum_{r=1}^{R_k} \sum_{j=1}^{n} I[\phi_{rj} = 1 \text{ and } \min_{i=1,\ldots,R_k} P_{iG} = P_{rG}]$, The conditional FDR is therefore

$$\begin{aligned}
&\sum_{r=1}^{R_k}\sum_{j=1}^{n_{0r}}\sum_{k=1}^{n} \frac{1}{k} Pr\left(\phi_{rj} = 1, \min_{i=1,\ldots,m_k} P_{iG} = P_{rG}, R = k \mid \mathcal{A}\right) \\
&= \sum_{r=1}^{R_k} n_{0r} \sum_{k=1}^{n} \frac{1}{k} Pr\left(\phi_{r1} = 1, P_{rG} \leq \min\left(h(\alpha, m), \min_{i=1,\ldots,m_k, i\neq r} P_{iG}\right), R = k \mid \mathcal{A}\right) \\
&\leq \sum_{r=1}^{R_k} n_{0r} \sum_{k=1}^{n} \frac{1}{k} Pr\left(\phi_{r1} = 1, R = k \mid \mathcal{A}\right) \\
&= \sum_{r=1}^{R_k} n_{0r} \sum_{k=1}^{n} \frac{1}{k} Pr\left(\phi_{r1} = 1, R = k \mid P_{rG} \leq h(\alpha, m)\right) \qquad \text{(C.1)} \\
&\leq \sum_{r=1}^{R_k} n_{0r} \frac{\alpha / R_k}{n} \leq \alpha \qquad \text{(C.2)}
\end{aligned}$$

where the equality in (C.1) follows since the p-values in row i are independent of the global null p-values in the other rows, and the first inequality in (C.2) follows from the inequality (A.2) at level α/R_k instead of α.

□

References

Foygel Barber, R. and Ramdas, A. The p-filter: multi-layer FDR control for grouped hypotheses *arXiv: 1512.03397*, 2015.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society, series B*, 76 (1): 297–318.

Benjamini, Y. and Heller, R. (2007). False Discovery Rates for Spatial Signals. *Journal of the American Statistical Association*, 102 (480): 1272–1281.

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, 64 (4): 1215–1222.

Benjamini, Y. and Hochberg, Y. (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, Y., Krieger, M., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93 (3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29 (4): 1165–1188.

Bhattacharjee, S. and Rajaraman, P. and Jacobs, K. and Wheeler, W. and William, A. and Melin, B. and Hartge, P. and Yeager, M. and Chung, C. and Chanock, S. and Chatterjee, N. and others (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90 (5): 821–835, 2012.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871.

Efron, B. Increasing Properties of Poly Frequency Function. *The Annals of Mathematical Statistics*, 36 (1): 272–279.

Fithian, W. and Sun, D. and Taylor, J. Optimal Inference After Model Selection. *arXiv: 1410.2597*, 2015.

Goeman, J.J., and Solari, A., Multiple Testing for Exploratory Research. *Statistical Science*, 26, 4, 2011.

Hua, X. and Goedert, J.J. and Pu, A. and Yu, G. and Shi, J. (2016). Allergy associations with the adult fecal microbiota: Analysis of the American Gut Project. *EBioMedicine*, 3 :172–179.

Lee, J.D. and Sun, D.L. and Sun, Y. and Taylor, J.E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44 (3): 907–927.

Li, G. and Shabalin, A.A. and Rusyn, I. and Wright, F.A. and Nobel, A.B., An Empirical Bayes Approach for Multiple Tissue eQTL Analysis. *arXiv: 1311.2948*, 2016.

Y. Li, and D. Ghosh, A two-step hierarchical hypothesis set testing framework, with application to gene expression data on ordered categories. *BMC Bioinformatics*, 15, Article 108, 2014.

Loughin, T. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47 :467–485.

Owen, A. Karl Pearson's meta-analysis revisited. *The Annals of Statistics*, 37 (6B): 3867–3892.

Peterson, C. and Bogomolov, M. and Benjamini, Y. and Sabatti, C. Many phenotypes wihtout many false discoveries: error controlling strategies for multi-traits association studies (2015). *Genetic Epidemiology*, 40 (1): 45–56.

Peterson, C. and Bogomolov, M. and Benjamini, Y. and Sabatti, C. (2016). TreeQTL: hierarchical error control for eQTL findings. *Bioinformatics*, bioRxiv. DOI: 10.1101/021170.

Reid, S. and Taylor, J. and Tibshirani, R. Post-selection point and interval estimation of signal sizes in Gaussian samples. *arXiv: 1405.3340*, 2015.

Rivas, M.A. and et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348 (6235) :666–669.

Schildknecht, K. and Tabelow, K. and Dickhaus, T. More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses. *WIAS Preprint No. 2127, ISSN 2198-5855* , 2015.

Singh, A. K. and Phillips, S. Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. *NeuroImage*, 50 (1) : 40–47, 2010.

Schweder, P. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28 (10):1353–1358.

Storey, J. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035.

Wu, M. C. and Kraft, P. and Epstein, M. P. and Taylor, D.M. and Chanock, S.J. and Hunter, D.J. and Lin, X. (2010). Powerful SNP Set Analysis for Case-Control GenomeWide Association Studies. *American Journal of Human Genetics*, 86 :929–942.

Wu, M. C. and Kraft, P. and Epstein, M. P. and Taylor, D.M. and Chanock, S.J. and Hunter, D.J. and Lin, X. (2010). Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). *American Journal of Human Genetics*, 89 :82–93.

Yekutieli, D. and Reiner-Benaim, A. and Benjamini, Y. and Elmer, G. I. and Kafkafi, N. and Letwin, N. E. and Lee, N. H. Approaches to multiplicity issues in complex research in microarray analysis. *Stat. Neerl.* 60 (4): 414–437, 2006.

Yekutieli, D. Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17 (3):458460, 2008.

Yekutieli, D. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103 :309316, 2008.

Yu, K. and Li, Q. and Bergen, A.W. and Pfeiffer, R.M. and Resenberg, P.S. and Caporaso, N. and Kraft, P. and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. *Genet Epidemiol.*, 33 (8):700–709.