Controlling the rate of GWAS false discoveries

Damian Brzyski[a,b] Christine B. Peterson[c] Piotr Sobczyk[d] Emmanuel J. Candés [c]
Malgorzata Bogdan[e] Chiara Sabatti[c*]

[a] Jagiellonian University, [b] Indiana University, [c] Stanford University, [d] Wroclaw University of Technology, [e] University of Wroclaw

June 2016

Abstract

With the rise of both the number and the complexity of traits of interest, control of the false discovery rate (FDR) in genetic association studies has become an increasingly appealing and accepted target for multiple comparison adjustment. While a number of robust FDR controlling strategies exist, the nature of this error rate is intimately tied to the precise way in which discoveries are counted, and the performance of FDR controlling procedures is satisfactory only if there is a one-to-one correspondence between what scientists describe as unique discoveries and the number of rejected hypotheses. The presence of linkage disequilibrium between markers in genome-wide association studies (GWAS) often leads researchers to consider the signal associated to multiple neighboring SNPs as indicating the existence of a single genomic locus with possible influence on the phenotype. This *a posteriori* aggregation of rejected hypotheses results in inflation of the relevant FDR. We propose a novel approach to FDR control that is based on pre-screening to identify the level of resolution of distinct hypotheses. We show how FDR controlling strategies can be adapted to account for this initial selection both with theoretical results and simulations that mimic the dependence structure to be expected in GWAS. We demonstrate that our approach is versatile and useful when the data are analyzed using both tests based on single marker and multivariate regression. We provide an R package that allows practitioners to apply our procedure on standard GWAS format data, and illustrate its performance on lipid traits in the NFBC66 cohort study.

*corresponding author

Introduction

In the last decade, genome-wide association studies (GWAS) have been the preferential tool to investigate the genetic basis of complex diseases and traits, leading to the identification of an appreciable number of loci [1]. Soon after the first wave of studies, a pattern emerged: there exists a sizable discrepancy between, on the one hand, the number of loci that are declared significantly associated and the proportion of phenotypic variance they explain [2] and, on the other hand, the amount of information that the entire collection of genotyped single nucleotide polymorphisms (SNPs) appears to contain about the trait [3, 4]. In order to increase the number of loci discovered (and their explanatory power), substantial efforts have been made to obtain larger sample size by genotyping large cohorts [5, 6] and by relying on meta-analysis. However, the gap remains, although not as large as in the original reports. This parallels, in part, the discrepancy between the multivariate model that is used to define complex traits and the univariate approach to the discovery of associated SNPs which is standard practice, as underscored, for example, in [7–9].

Two approaches to bridge the gap emerge quite naturally: (a) an attempt to evaluate the role of genetic variants in the context of multivariate models, more closely matching the underlying biology, and (b) relaxing the very stringent significance criteria adopted by GWAS to control the false discovery rate (FDR) [10] rather than the family-wise error rate (FWER)—a strategy that has been shown attractive when prediction is considered as an end goal together with model selection [11]. Both strategies have been pursued, but have encountered a mix of success and challenges.

Multivariate models for the analysis of GWAS data have been proposed as early as 2008 [12, 13]: examining the distribution of their residuals, it is clear that they provide a more appropriate model for complex traits. However, their use to discover relevant genetic loci has encountered difficulties in terms of computational costs and interpretability of results. On the computational side, progress has been made using approaches based on convex optimization such as the lasso [14], developing accurate methods to screen variables [15–17], and relying on variational Bayes [18, 19]. There are, however, remaining challenges. Firstly, the genetics community is, correctly, very sensitive to the need of replicability, and finite samples guarantees for the selected variants are sought. Unfortunately, this has been difficult to achieve with techniques such as the lasso: [20] attempts to use stability selection, [21] does a simulation study of a variety of penalized methods, showing that tuning parameters play a crucial role and that standard selection methods for these do not work well, and [22] proposes some analytical approximation of FDR as an alternative to the lasso. Our recent

work [23] also explores alternative penalty functions that under some circumstances guarantee FDR control. Secondly, multivariate models encounter difficulties in dealing with correlated predictors, in that the selection among these is often arbitrary: this is challenging in the context of GWAS, when typically there is a substantial dependence between SNPs in the same genetic region.

The suggestion of controlling FDR rather then FWER in genetic mapping studies that expect to uncover a large number of loci was put forward over a decade ago [24–26] and is accepted in the expression quantitative trait loci (eQTL) community, where FDR is the standard error measure. The existence of strong local dependence between SNPs has also posed challenges for FDR controlling procedures. While the Benjamini-Hochberg [10] procedure (BH) might be robust to the correlation between tests that one observes in GWAS, the fact that the same biological association may be reflected in multiple closely located SNPs complicates both the definition and the counting of discoveries, so that it is not immediately evident how FDR should be defined. Prior works [27–29] underscore this problem and suggest solution for specific settings.

This paper proposes a phenotype-aware selective strategy to analyze GWAS data which enables precise FDR control and facilitates the application of multivariate regression methodology, by reducing the dependency between the SNPs included in final testing. The Methods section starts by briefly recapitulating the characteristics of GWAS, with reference to an appropriate count of discoveries and the identification of a meaningful FDR to control. We introduce our selective strategy and provide some general conditions under which it controls the target FDR. We then describe a specific selection procedure for GWAS analysis and describe how it can be coupled with standard BH for univariate tests, or with SLOPE [23] to fit multivariate regression. In the Results section we explore the performance of the proposed methodology with simulations and analyze a dataset collected in the study of the genetic basis of blood lipids. In both cases, the FDR-controlling procedures we propose allow us to explain a larger portion of the phenotype variability, without a substantial cost in terms of increased false discoveries.

1 Methods

1.1 The GWAS design, dependence and definition of discoveries

The goal of a GWAS study is to identify locations in the genome that harbor variability which influences the phenotype of interest. This is achieved using a sample of n individuals, for whom one acquires trait values y_i and genotypes at a collection of M SNPs that span the genome.

Following standard practice, we summarize genotypes by the count of copies of minor allele that each individual has at each site, resulting in a $n \times M$ matrix X, with entries $X_{ij} \in \{0, 1, 2\}$. The variant index j is taken to correspond to the order of the position of each SNP in the genome. The true relation between genetic variants and phenotypes can be quite complex. For simplicity, and in agreement with the literature, we assume a linear additive model, which postulates that the phenotype value y_i of subject i depends linearly on her/his allele counts at an unknown set $\mathcal{C}$ of causal variants. Since there is no guarantee a priori that the variants in $\mathcal{C}$ are part of the genotype set, we indicate their allele counts with Z_{ij}, letting

$$y_i = \sum_{j \in \mathcal{C}} b_j Z_{ij} + \epsilon_i.$$

Investigating the relation between y and X is helpful to learning information about the set of causal variants $\mathcal{C}$ and their effects b_j in two ways: (1) it is possible that some of the causal variants are actually genotyped, so that $Z_{ij} = X_{ik}$ for some k; (2) most importantly, the set of M genotyped SNP contains reasonable proxies for the variants in $\mathcal{C}$. To satisfy (2), GWAS are designed to capitalize on the local dependence between variable sites in the genome known as linkage disequilibrium (LD), which originates from the modality of transmission of chromosomes from parents to children, with modest recombination. The set of M genotyped SNPs is chosen with some redundancy, so that the correlation between X_j and X_{j+k} is expected to be non-zero for k in a certain range: this is to ensure that any non-typed casual variant Z_l will be appreciably correlated with one (or more) of the typed X_js which are located in the same genomic region. Any discovered association between a SNP X_j and the phenotype y is interpreted as an association between y and *some variant* in the genomic *neighborhood* of X_j. This design has a number of implications for statistical analysis:

1. Often, the existence of an association between y and each typed variant X_j is queried via a test statistic t_j which is a function of y and X_j only: these test statistics are "locally" dependent, with consequences for the choice of multiple comparison adjustment, that, for example, might not need to be as stringent as in the case of independence.

2. When multivariate regression models are used to investigate the relation between y and X, one encounters difficulties due to the correlation between regressors—the choice among which is somewhat arbitrary.

3. The fact that the true causal variants are not necessarily included among the genotyped SNPs makes the definition of a true/false association non-trivial.

We want to underscore the last point. To be concrete, let's assume the role of each variant X_j is examined with t_j, the t-statistic for H_0^j: $\beta_j = 0$, with β_j defined in the univariate regression $y_i = \alpha + \beta_j X_{ij} + \epsilon_i$. Even if none of the M genotyped variants are causal, a number of them will have a coefficient $\beta_j \neq 0$ in these reduced models: whenever X_j is correlated with one of the variants in $\mathcal{C}$, H_0^j should be rejected. Indeed, simulation studies that investigate the power and global error control of different statistical approaches routinely adopt definitions of "true positive" that account for correlation between the known causal variant and the genotyped SNPs (see [21] for a recent example). At the same time, a rejection of H_0^j should not be interpreted as evidence of a causal role for X_j: in fact, geneticists equate discovery with the identification of a genomic location rather then with the identification of a variant. The rejection of H_0^j for a number of correlated neighboring SNPs in a GWAS is described in terms of the discovery of one single locus associated with the trait of interest. The number of reported discoveries, then, corresponds to the number of distinct genomic regions (whose variants are uncorrelated) where an association has been established. This discrepancy between the number of rejected hypotheses and the number of discoveries has important implications for FDR controlling strategies, which have received only a modest attention in the literature. Siegmund and Zhang [29] suggest that in situations similar to those of GWAS, neighboring rejections should be grouped and counted as a single rejection and that the global error of interest should be the expected value of the "proportion of clusters that are falsely declared among all declared clusters". This FDR of clusters—a notion first introduced in [28]—is not the error rate controlled by the Benjamini-Hochberg [10] procedure on the p-values for the H_0^j hypotheses. Indeed, because FDR is the expected value of the ratio of the random number of discoveries, its control depends crucially on how one decides to count discoveries. In [30] we give another example of how controlling FDR for a collection of hypotheses does not extend to controlling FDR for a smaller group hypotheses logically derived from the initial set. Both in the setting described here and in [30], targeting FWER would have resulted in less surprising behavior: assuring that the probability of rejecting at least one null H_0^j is smaller than a level α would also guarantee that the probability of rejecting a null cluster of hypotheses is smaller than α. Siegmund and Zhang [29] study a setting that is close to our problem and propose a methodology to control their target FDR relying on a Poisson process distribution for the number of false discoveries. We investigate here a different approach: one that is more tightly linked to the GWAS design, is adapted to the variable extent of LD across the genome, and capitalizes on results in selective inference [31].

1.2 Controlling the FDR of interesting discoveries by selecting hypotheses

The approach we study emerged from our interest in using multivariate linear models to analyze the relation between y and X, so it is useful to motivate it in this context. Suppose both X_j and X_{j+1} are strongly correlated with the untyped causal variant Z_k. When univariate regression is used as the analysis strategy, both the test statistics t_j and t_{j+1} would have large values, resulting in the discovery of this locus. Instead, the marginal p-values for the coefficients of X_j and X_{j+1} derived from a multivariate model that includes both would be large; and model selection strategies would rather arbitrarily lead to the inclusion of one or the other regressor, leading to an underestimate of their importance when resampling methods are used to evaluate significance. If using multivariate linear models, one would achieve the best performance if, from the start, only one of X_j and X_{j+1} (the most strongly correlated with Z_k) is included among the possible regressors. A natural strategy is to prune the set of M typed SNPs to obtain a subset of m quasi-orthogonal ones and supply these to the model selection procedure of choice. However, this encounters the difficulty that the best proxy for some of the causal variants might have been pruned, resulting in a loss of power. It seems that ideally one would select from a group of correlated SNPs the one that has the strongest correlation with the trait to include among the potential regressors. Unfortunately, this initial screening for association would invalidate any guarantees of the model selection strategy, which operates now not on m variables, but on m *selected* ones. The emerging literature of selective inference, however, suggests that we might be able to appropriately account for this initial selection step, preserving guarantees on error rate control.

Abstracting from the specifics of multivariate regression, consider the setting where a collection $\mathcal{H}$ of M hypotheses $H_0^1, \ldots, H_0^M$ with some redundancy is tested to uncover an underlying structure of interest. The hypotheses in $\mathcal{H}$ can be organized linearly or spatially and are chosen because a priori they provide a convenient and general way of probing the structure; however, it is expected that a large portion of these will be true, and that when one H_0^j is false, a number of neighboring ones would be also false. In case of GWAS, these clusters of false hypotheses would correspond to markers correlated with causal mutations. Because of the mismatch between $\mathcal{H}$ and the underlying structure, the number of scientifically interesting discoveries does not correspond to the number of rejected H_0^js and strategies that control the FDR defined in terms of these might not lead to satisfactory inference. Specifically, as noted in [29], "a possibly large number of correct rejections at some location can inflate the denominator in the definition of false discovery rate, hence artificially

creating a small false discovery rate, and lowering the barrier to possible false detections at distant locations". This problem was recognized already in [27] and [28], who introduce the notion of cluster FDR and suggest defining *a priori* clusters of hypotheses, corresponding to signals of interest and apply FDR controlling strategies to hypotheses relative to these clusters. We take here a different approach, where "clusters" of hypotheses are defined *after looking at the data*, and used to select a subset of representative hypotheses. Only this subset is then tested, with a procedure that accounts for this initial selection.

Formally, let y indicate the data used to test the hypotheses in $\mathcal{H}$ and let $\mathcal{S}(y)$ be a selection procedure that, on the basis of the data, identifies a subset $\mathcal{H}^s$ of s representative hypotheses. Let $S = \{i : 1 \leq i \leq M \;\; \& \;\; H_0^i \in \mathcal{H}^s\}$ be the set of their indexes, so that it is relevant to control the following FDR_s:

$$\text{FDR}_s = E\left(\frac{\sum_{j\in S} 1(H_0^j \text{rejected}) 1(H_0^j \text{true})}{\sum_{j\in S} 1(H_0^j \text{rejected}) \vee 1}\right). \tag{1}$$

In other words, the decision of acceptance/rejection is made only for the hypotheses in the selected set. The work of [32] and [31] suggests a possible strategy to control FDR_s at level q: apply BH to the p-values $p_{[S]}$ corresponding to the subset of hypotheses $\mathcal{H}^s$, targeting the more stringent level $q|S|/M$ to penalize for the initial selection. According to this strategy, the smallest p-value $p_{[S](1)}$ for $\mathcal{H}^s$ would be compared to $|S|q/M \times 1/|S| = q/M$, and $p_{[S](i)}$ would be compared to qi/M: the p-value thresholds are identical to those implied by BH on $\mathcal{H}$, but the number of hypotheses tested is smaller and the hypotheses are more clearly separated. This prevents the excessive deflation of the BH threshold that results when each true discovery is represented by many rejected hypotheses, and therefore helps to control the number of false discoveries.

The results in [31] imply that if $\mathcal{S}(y)$ is a simple selection rule, the procedure described above controls the selective FDR

$$\text{selective FDR} = E\left(\frac{\sum_{j\in S} 1(H_0^j \text{rejected}) 1(H_0^j \text{true})}{|S| \vee 1}\right),$$

whenever BH applied to $\mathcal{H}$ would control the standard FDR, $E\left(\frac{\sum_{j=1}^{M} 1(H_0^j \text{rejected}) 1(H_0^j \text{true})}{\sum_{j=1}^{M} 1(H_0^j \text{rejected}) \vee 1}\right)$. If the selection is stringent enough, controlling the selective FDR might be meaningful. Building on the results obtained in [33], we can also prove that for a general class of selection rules the same procedure leads to control of FDR_s at level q, as long as the distribution of p-values follows the condition of positive regression dependence on a subset (PRDS), described in [34]. As noted in [25], PRDS condition can be loosely interpreted as the requirement of the positive correlation between

p-values at linked markers, and in this way it corresponds well to the practice of GWAS.

Theorem 1 *FDR control for selected hypotheses. Let $\mathcal{S}(y)$ be a selection procedure, and let R^S be the number of rejections derived by applying BH with target $q|S|/M$ on the selected hypotheses $\mathcal{H}^S$. If the p-values are PRDS and the selection procedure is such that $R^S(p_1, \ldots, p_M)$ is non-increasing in each of the p-values p_i , rejecting R^S guarantees control of* FDR_s.

Proof. Letting $\mathcal{H}_0$ be the collection of true null hypotheses in $\mathcal{H}$ and $\mathcal{H}_0^S$ the set of true null hypotheses in $\mathcal{H}^S$, we write FDR_s as

$$\mathrm{FDR}_s = E\left(\frac{\sum_{i\in\mathcal{H}_0^S} 1(H_0^i \text{rejected})}{R^S \vee 1}\right) = E\left(\frac{\sum_{i\in\mathcal{H}_0} 1(H_0^i \text{rejected}) 1(i \in S)}{R^S \vee 1}\right),$$

where R^S indicates the number of rejections resulting from applying the BH rule with target level $q|S|/M$ to the p-values $p_{[S]}$ of $\mathcal{H}^S$. Going forward, we write R^S instead of $(R^S \vee 1)$ for simplicity. Recalling that a hypothesis is rejected if its p-value is smaller than the BH threshold, exchanging the order of summation and expectation, and multiplying and dividing by q/M we have

$$\mathrm{FDR}_s = \frac{q}{M}\sum_{i\in\mathcal{H}_0} E\frac{1(p_i < R^S q/M)1(i \in S)}{R^S q/M} \leq \frac{q}{M}\sum_{i\in\mathcal{H}_0} E\frac{1(p_i < R^S q/M)}{R^S q/M}, \tag{2}$$

where the last inequality comes from relaxing a restriction. We now recall Lemma 1 from [33], which states that for a set of p-values that satisfy PRDS, when $f : (p_1, \ldots, p_M) \longrightarrow [0, 1]$ is non-increasing, $E\dfrac{1(p_i < f(p))}{f(p)} \leq 1$. Under the assumption that $f(p_1, \ldots, p_M) := R^S q/M$ is non-increasing we then have our result, as $\mathrm{FDR}_s \leq \frac{q}{M} \times M$. □

An example selection procedure that satisfies the assumptions of the theorem is as follows: the hypotheses $\mathcal{H}$ are separated in groups a priori and from each group, $\mathcal{S}(y)$ selects the hypothesis with the smallest associated p-value. In the next section, we describe a slightly more complicated selection procedure $\mathcal{S}(y)$, that appears appropriate for the case of GWAS, and where the separation of hypotheses into groups is data-driven. While this procedure may not satisfy the assumption that the number of rejections is a non-increasing function of the p-values, our extensive simulations studies suggests that its use in the context of Theorem 1 still leads to FDR_s control.

1.3 A GWAS selection procedure: phenotype-aware cluster representatives

In the context of genetic association studies, the selection function $\mathcal{S}(y)$ defined in Procedure 1 and illustrated in Figure 1 emerges quite naturally. One starts by evaluating the marginal association

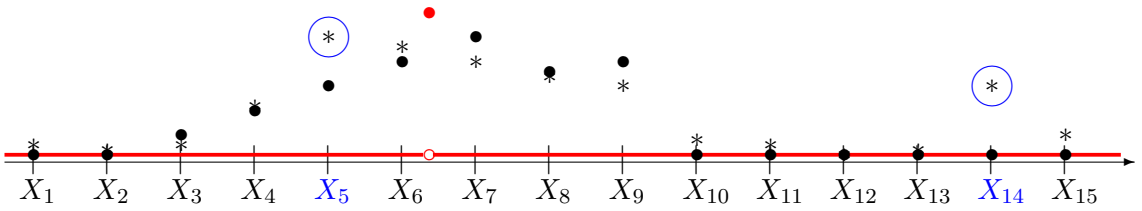


Figure 1: **Phenotype-aware cluster representatives**. *The x-axis represents the genome, with the locations of genotyped SNPs X_i indicated by tick-marks. The true causal effect of each position of the genome is indicated in red: there is only one causal variant in this region, between SNPs X_6 and X_7. Solid black circles indicate the value of β_i, coefficient of X_i in a linear approximation of the conditional expectation $E(y|X_i)$. Asterisks mark the estimated $\hat{\beta}_i$s in the sample. The SNPs X_5 and X_{14}, selected as cluster representatives in this schematic diagram, are indicated in blue.*

of each SNP to the phenotype using the p-value of the t-test for its coefficient in a univariate regression. Then, SNPs with a p-value larger than threshold π are removed from consideration. The collection of remaining SNPs is further pruned to obtain a selected set $\mathcal{S}$ with low correlation, so that each variant $X_i \in \mathcal{S}$ can be equated to a separate discovery. To achieve this, we define clusters of SNPs using their empirical correlation in our sample, starting from the variants with the strongest association to the phenotype, which are selected as cluster representatives.

Procedure 1 Selection function $\mathcal{S}(y)$ to identify cluster representatives

Input: $\rho \in (0, 1)$, $\pi \in (0, 1]$

Screen SNPs:

(1) Calculate the p-value for H_0^j: $\beta_j = 0$, with β_j defined in the univariate regression $y_i = \alpha + \beta_j X_{ij} + \epsilon_i$, as j varies across all SNPs.

(2) Retain in $\mathcal{B}$ only those SNPs whose p-values are smaller than π.

Cluster SNPs:

(3) Select the SNP j in $\mathcal{B}$ with the smallest p-value and find all SNPs whose Pearson correlation with this selected SNP is larger than or equal to ρ.

(4) Define this group as a cluster and SNP j as the representative of the cluster. Include SNP j in $\mathcal{S}$, and remove the entire cluster from $\mathcal{B}$.

(5) Repeat steps (3)-(4) until $\mathcal{B}$ is empty.

Procedure 1 has two tuning parameters: π and ρ, corresponding to the two steps of the selection. The screening in steps (1)-(2) is similar to that described in [13, 15] for model selection procedures, where the parameter π controls the stringency of the selection based on univariate association. Its influence is minimal if marginal tests are used to control FWER. However, large values of π result in a larger dimensions of selected clusters, leading to increased number of false discoveries when controlling FDR. On the other hand, in the context of multivariate regression, it is possible to uncover a role for variants that have weak marginal effects due to masking: to enable this, one must not be too stringent in the initial screening step. In all the simulations and data analyses presented here we have used $\pi = 0.05$, which seems to be a good compromise. The results in [13, 15] can provide additional guidance on the choice of π.

Steps (3)-(5) of Procedure 1 aim to "thin" the set of SNPs on account of the dependency among them. This is related to the selection of tag SNPs [35], for which there is an extensive literature, and is similar to correlation reduction approaches [36]. A defining characteristic of Procedure 1, however, is that both the SNP clusters and their representatives are selected with reference to the phenotype of interest. This ensures that the representatives maximize power, and that the location of the true signal is as close as possible to the center of the respective cluster. This also reduces the probability of the selection of more than one SNP per causal variant. The value of ρ needs to be set with reference to the sample size and the density of the available markers. Indeed, we suggest that researchers run a simple simulation (as in the one described in the first part of the Results section) to select appropriate values for ρ (see Discussion section for further remarks on this). Certainly, Procedure 1 is but one possibility for creating clusters. For example, one might want to include information on physical distance in the formation of clusters. In our experiments, however, this has not led to better performance.

We now consider two approaches to the analysis of GWAS data that can be adopted in conjunction with the selection of cluster representatives to control the FDR_s.

1.4 Univariate testing procedures after selection

By and large, the most common approach to the analysis of GWAS data relies on univariate tests of association between trait and variants. This has advantages in terms of computational costs, handling of missing data, and portability of results across studies. We therefore start by considering how to control relevant FDR in this context.

While most disease-related GWAS aim to control FWER, FDR has been the global error of

choice in eQTL studies, and that literature testifies to some of the challenges encountered, in particular, to difficulties related to dependence across tests and hypotheses (see [30] for a detailed description). Starting from [25], it was observed that BH seems to be able to deal with the type of dependence across test statistics induced by LD. However, the relatedness between hypotheses and the lack of one-to-one correspondence between hypotheses and meaningful scientific discoveries remains a problem. For example, when investigating the genetic basis of variation in gene expression, the authors in [37] change the unit of inference from SNPs to genes, so as to bypass the redundancy due to many SNPs in the same neighborhood. Here we address the problem by inviting the researchers to identify the resolution of discoveries prior to testing, but after having observed the data. We consider two different approaches to obtain the p-values for each of the H_0^j hypotheses: univariate linear regression (which we indicate with SMT for single marker test) and EMMAX [7], a mixed model which allows us to consider polygenic effects. To enable computational scaling, EMMAX only estimates the parameters of the variance component model once rather than for every marker. We use SMTs and EMMAXs to denote the procedures that consist in testing the set of hypotheses $\mathcal{H}^s$ corresponding to cluster representatives, using p-values obtained with SMT and EMMAX, respectively, and identifying rejections with the BH_s procedure described below.

Procedure 2 Benjamini-Hochberg on selected hyotheses BH_s

Input:

M - total number of SNPs (before initial screening)

$\mathcal{H}^s$ - collection of selected hypotheses (cluster representatives)

$q \in (0, 1]$ - desired level for FDR_s

Let $|S|$ be the number of hypotheses in $\mathcal{H}^s$, and $p_{[S]}$ the vector of their p-values.

(1) Apply BH to $p_{[S]}$ with target level $|S|q/M$.

1.5 GeneSLOPE - FDR control in multivariate regression.

SLOPE [23] is a recently introduced extension of the lasso that achieves FDR control on the selection of relevant variables when the design is nearly orthogonal. Specifically, assume the following model

$$Y = X\beta + z,$$

where X is the design matrix of the dimension $n \times M$, $z \sim N(0, \sigma^2 I_{n\times n})$ is the n-dimensional vector of random errors, and β is the M-dimensional vector of regression coefficients, a significant portion of which is assumed to be zero. For a sequence of non-negative and non-increasing numbers $\lambda_1, \ldots, \lambda_M$, the SLOPE estimate of β is the solution to a convex optimization problem

$$\hat{\beta} = \arg\min_{b \in R^M} \quad \left\{ \frac{1}{2} \|y - Xb\|^2 + \sigma \sum_{i=1}^{M} \lambda_i |b|_{(i)} \right\}, \tag{3}$$

where $|b|_{(1)} \geq \ldots \geq |b|_{(M)}$ are sorted absolute values of the coordinates of b.

If we define a discovery as i such that the estimated $\hat{\beta}_i \neq 0$, and a false discovery as the case where $\hat{\beta}_i \neq 0$ but the true $\beta_i = 0$, [23] provides the sequence of λ_i (corresponding to the sequence of decreasing thresholds in BH), which provably controls FDR at a desired level if the design matrix X is orthogonal. Moreover, the modified sequence λ—described in Procedure 4 in the Appendix—has been shown in simulation studies to achieve FDR control in genetic studies when SNPs are nearly independent and the number of non zero β's is small or moderately large. Note that, as for other shrinkage methods [38, 39], the results of SLOPE depend on the scaling of explanatory variables: the values of the regularizing sequence in Procedure 4 assume that explanatory variables are "standardized" to have zero mean and a unit l_2 norm. Moreover, since in most cases the variance of the error term σ^2 is unknown and needs to be estimated, in [23] an iterative procedure for the joint estimation of σ and the vector of regression coefficients was proposed. This is described in the Appendix as Procedure 5 and follows closely the idea of *scaled lasso* [40]. All these data preprocessing and analysis steps are implemented in R package *SLOPE*, available on CRAN.

The fact that SLOPE comes with finite sample guarantees for the selected parameters makes it an attractive procedure for GWAS analysis. However, the presence of substantial dependence between SNPs (regressors X_j) presents challenges: on the one hand, the FDR-controlling properties have been confirmed so far only when the explanatory variables are quasi-independent; and on the other hand, the definition of FDR is problematic in a setting where the true causal variants are not measured and X contains a number of correlated proxies, similarly as for univariate procedures. The identification of a subset of variants with Procedure 1 takes care of both aspects : the regressors are not strongly correlated and, for sufficiently small ρ, they represent different locations in the genome, so that we can expect the projection of the true model in the space they span to be sparse and the number of $\hat{\beta}_i \neq 0$ to capture the number of scientifically relevant discoveries. We therefore propose as a potential analysis pipeline the application of Procedure 1 followed by Procedure 3, which outlines the application of SLOPE to the selected cluster representatives. Both procedures

have been implemented in the R package `GeneSLOPE`, which is available on CRAN and can handle typical GWAS data provided in PLINK format.

Procedure 3 GeneSLOPE

Input:

y - vector of trait values

M - total number of SNPs (before initial screening)

$X_{[S]}$ - selected SNPs (cluster representatives)

$q \in (0, 1]$ desired level for FDR_s

Initialize $\mathcal{A} = \emptyset$

(1) Center y by subtracting its mean, and standardize $X_{[S]}$ so that each column has a zero mean and unit l_2 norm.

(2) Calculate the sequence λ using Procedure 4, and retain the first $|S|$ elements of it.

(3) Compute the RSS obtained by regressing y onto variables in $\mathcal{A}$ and set $\widehat{\sigma}^2 = RSS/(n-|\mathcal{A}|-1)$, where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$.

(4) Compute the solution $\widehat{\beta}$ for SLOPE as in equation (3) explaining y as a linear function of $X_{[S]}$ with parameters $\widehat{\sigma}$ and λ. Set $\mathcal{A}^+ = supp(\widehat{\beta})$.

(5) If $\mathcal{A}^+ = \mathcal{A}$ stop; if not, set $\mathcal{A} = \mathcal{A}^+$ and iterate Steps (3)-(4).

2 Results

To test the performance of the proposed algorithms we relied on simulations and real data analysis. In both cases, genotype data came from the North Finland Birth Cohort (NFBC66) study [41], available in dbGaP under accession number phs000276.v2.p1 (`http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000276.v2.p1`). The raw genotype matrix contains 364,590 markers for 5,402 subjects. We filtered the data in PLINK to exclude copy number variants and SNPs with Hardy-Weinberg equilibrium p-value < 0.0001, minor allele frequency < 0.01, or call rate $< 95\%$. This resulted in an $n \times M$ predictor matrix with $n = 5,402$ and $M = 334,103$. When applying GeneSLOPE, missing genotype data were imputed as the SNP mean.

For simulations, the trait values are generated using the multiple regression model:

$$Y_i = \sum_{j \in C_k} \beta_j \tilde{X}_{ij} + \epsilon_i \ \ , \ i \in \{1, \ldots, n\} \ \ , \tag{4}$$

where $\tilde{X}$ is the standardized matrix of genotypes, C_k is the set of indices corresponding to "causal" mutations and $\epsilon_i \sim N(0,1)$. The number of causal mutations takes the value $k \in \{20, 50, 80, 100\}$, and in each replicate, the k "causal" features are selected at random from a subset of the M SNPs. For each k, the values of β_j are evenly spaced in the interval $[\texttt{SignalMin}, \texttt{SignalMax}]$, with $\texttt{SignalMin} := 0.6\sqrt{2\log p}$ and $\texttt{SignalMax} := 1.4\sqrt{2\log p}$. As a result, the smallest genetic effect is rather weak (heritability in a single QTL model $h^2 = 0.0017$), while the strongest effect is relatively large ($h^2 = 0.0091$). Each scenario is explored with 100 simulations.

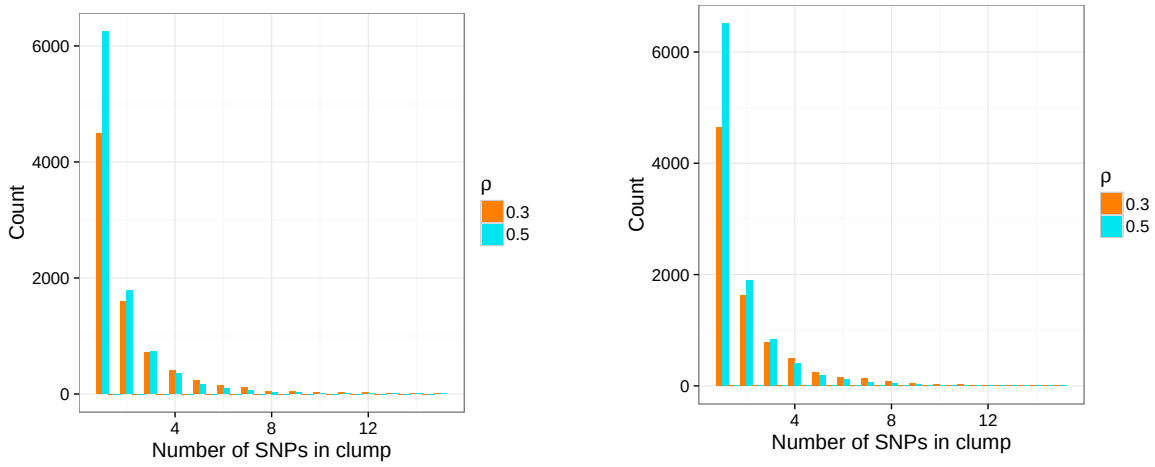
In evaluating FDRs and power, we adopt the following conventions, which we believe mimic closely the expectations of researchers in this field: the null hypothesis relative to a SNP/cluster representative is true if the SNP/cluster representative has a correlation less then 0.3 with any causal variant. Similarly, a causal variant is discovered if at least one of the variants in the rejection set has correlation of at least magnitude 0.3 with it.

In addition to evaluating performance in the context of simulated traits, we apply the proposed procedures to four lipid phenotypes available in NFBC66 [41]: high-density lipoproteins (HDL), low-density lipoproteins (LDL), triglycerides (TG), and total cholesterol (CHOL). We compare the discoveries obtained by the univariate and multivariate procedures on the NFBC data to those reported in [42], a much more powerful study based on 188,577 subjects.

2.1 Simulation study

Cluster sizes

We begin by exploring the distribution of the size of clusters created according to Procedure 1. Figure 2 illustrates the size of clusters when the trait was generated according to the model in equation (4) with $k = 80$ and genotypes from the NFBC dataset. We illustrate the results of both the original version of Procedure 1 using simple univariate regression (SMT) to obtain the p-values as well as a modification in which the initial p-value calculation is performed using EMMAX. It can be seen that most of the clusters are rather small and do not include more than 5 SNPs. There are no significant differences in the size of clusters created starting from EMMAX or SMT p-values. Of course, differences in the genotype density would result in a differences in the cluster sizes obtained.



(a) p-values from EMMAX

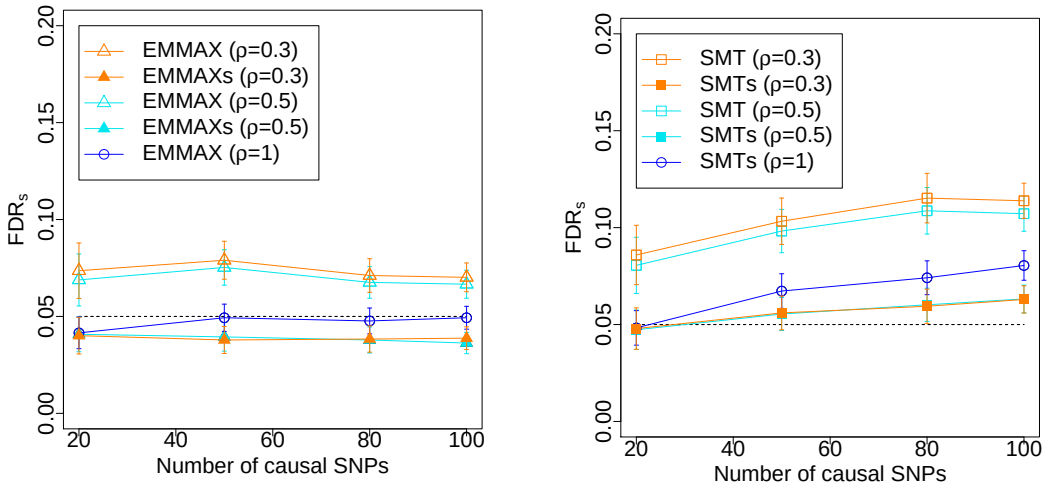
(b) p-values from SMT

Figure 2: Histograms of the number of SNPs included in each cluster when Procedure 1 is applied to p-values calculated from SMT and EMMAX with $\pi = 0.05$ and $\rho = 0.3$ or $\rho = 0.5$.

Error control with EMMAX and single marker tests

Figure 3 illustrates the results of simulations exploring the FDR_s control properties of BH applied the complete set of M p-values obtained from EMMAX or SMT (i.e. with no pre-screening or clustering of the hypotheses) and the corresponding two-step approaches we recommend (EMMAXs and SMTs), where cluster representatives are first chosen using Procedure 1 and then discoveries are identified with Procedure 2. The FDR_s for the traditional version of EMMAX and SMT is calculated mimicking what researchers typically do in practice to interpret GWAS results. Specifically, the SNPs for which the null hypotheses are rejected using BH are supplied to Procedure 1 to identify clusters. The realized FDR_s is defined as the average across 100 iterations of the fraction of falsely selected clusters over all clusters obtained.

Figure 3 illustrates that, in agreement with Theorem 1, EMMAXs controls FDR_s at all levels of ρ and for any number of causal SNPs. In contrast, BH applied to the full set of p-values obtained from EMMAX with post-hoc clustering of the discoveries results in a somewhat elevated FDR_s due to the deflation of the BH threshold. Moreover, EMMAXs offers better control of FDR_s than SMTs, particularly as the number of causal SNPs increases. This makes sense given that the model assumed by EMMAX is better able to account for polygenic effects than the single-marker test.



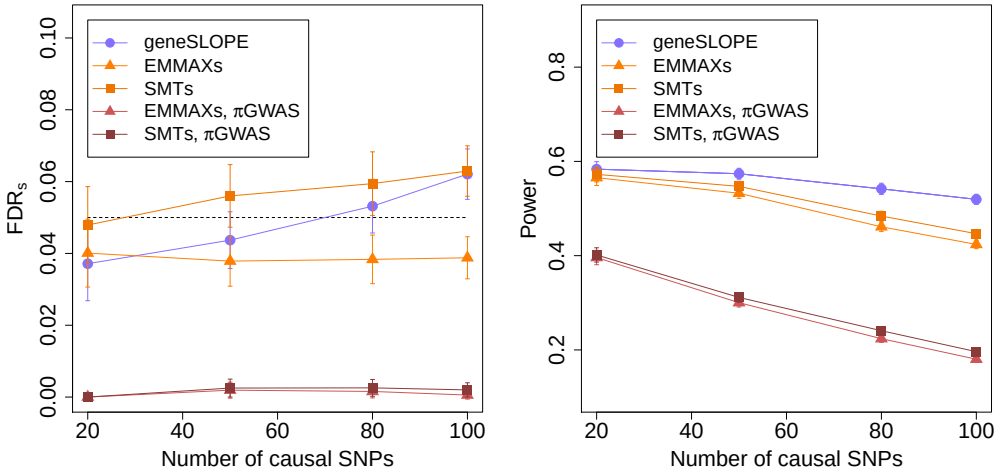
(a) FDR_s control for EMMAX

(b) FDR_s control for SMT

Figure 3: FDR_s for EMMAX and SMT and the corresponding procedures EMMAXs and SMTs which operate on cluster representatives. The dashed black line represents the target FDR_s level of 0.05. Note that EMMAX with $\rho = 1$ (i.e. with no clustering) coincides with EMMAXs, and that the FDR_s for this specific case corresponds to the regular FDR. Shapes indicate the procedures: hollow triangles for the application of BH to the collection of p-values from EMMAX for all hypotheses followed by clustering of the discoveries, filled triangles for the selective procedure EMMAXs, hollow squares for the application of BH to the collection of p-values from EMMAX for all hypotheses followed by clustering of the discoveries, filled squares for the selective procedure SMTs, and hollow circles for the application of BH to the full collection of p-values with no clustering. Colors indicate the parameters for clustering: orange for $\rho = 0.3$, turquoise for $\rho = 0.5$, and blue for $\rho = 1$.

GeneSLOPE error control and power

Figure 4 illustrates the performance of geneSLOPE in terms of FDR_s and power in the context of the performance of EMMAXs and SMTs for the same setting and range of k. For all procedures, power decreases as k increases, with a slower decay for geneSLOPE. Note that the average power of geneSLOPE is systematically larger than the power of SMTs, with the difference increasing with k, while the FDR_s of geneSLOPE is always smaller then that of SMTs. Figure 4 also demonstrates how using the standard genome-wide significance threshold setting $\pi = 5 \times 10^{-8}$ results in a very substantial loss of power as compared to procedures controlling FDR.



(a) FDR_s (b) Power

Figure 4: FDR_s and power for geneSLOPE when clustering is done with $\pi = 0.05$, $\rho = 0.3$, and target FDR_s level 0.05 (marked in slate blue). For comparison, we reproduce from Figure 3 the curves indicating the performance of EMMAXs and SMTs for the same setting (marked in shades of orange). We also include the values of FDR_s and power when EMMAXs and SMTs are carried out using cluster representatives selected with $\pi = 5 \times 10^{-8}$, the standard GWAS genome-wide significance threshold (marked in shades of red). Shapes indicate the procedures: filled circles for geneSLOPE, filled triangles for EMMAXs, and filled squares for SMTs.

2.2 Real Data Analysis

To analyze the lipid phenotypes, we adopted the same protocol described in [41]: subjects that had not fasted or were being treated for diabetes ($n = 487$) were excluded, leaving a set of $4,915$ subjects for further analysis. All phenotypes were adjusted for sex, pregnancy, oral contraceptive use, and population structure as captured by the first 5 genotype principal components (computed using EIGENSOFT [43]); the residuals were used as the trait values Y_i in the subsequent association analysis.

We compare the results of geneSLOPE, EMMAXs and classically applied EMMAX. GeneSLOPE (Procedure 1 followed by Procedure 3) was applied using $\pi = 0.05$, $\rho = 0.3$ or 0.5, and $q = 0.05$ or 0.1 (for a total of 4 versions) to a centered and normalized version of the genotype matrix where each column has mean 0 and ℓ_2 norm 1. EMMAXs (Procedure 1 followed by Procedure 2) was applied with $\pi = 0.05$, $\rho = 0.3$ or 0.5, and $q = 0.05$ or 0.1. To mimic the standard GWAS analysis, we ran EMMAX identifying as significant those SNPs with p-value $\leq 5 \times 10^{-8}$; to obtain

comparable numbers of discovered SNPs we applied Procedure 1 to cluster the results.

We compare the discoveries of these three methods on the NFBC data to those reported in [42], a much more powerful study based on 188,577 subjects. We compute the realized selected false discovery proportion FDP_s for each method assuming that SNPs within 1Mb of a discovery (defined as $p < 5 \times 10^{-8}$) in the comparison study are true positives (even if, of course, the biological truth for the given study population is not known, and the association statistics in [42] are based on univariate tests and may therefore not fully capture the genetic underpinnings of these complex traits). We also seek to understand what proportion of the trait heritability is captured by the selected SNPs: to this end, we estimate the proportion of phenotypic variance explained by the set of genome-wide autosomal SNPs using GCTA [44], and compare this to the adjusted r^2 obtained from a multiple regression model including the selected cluster representatives as predictors.

The estimated proportion of phenotypic variance explained by genome-wide SNPs is 0.34, 0.32, 0.10, and 0.29 for HDL, LDL, TG and CHOL, respectively. A comparison of the number of discoveries (i.e. the number of selected cluster representatives), number of true discoveries, FDP_s, and r^2 across methods is given in Figure 5. As an illustrative example, geneSLOPE selections with $\pi = 0.05$, $q = 0.1$ and $\rho = 0.5$ are shown in Figure 6 along with p-values obtained using EMMAX and those obtained in the more highly-powered comparison study [42].

The application on real data illustrates how FDR_s controlling procedures are more powerful than the standard practice of identifying significant SNPs using a p-value threshold of 5×10^{-8}. Both EMMAXs and geneSLOPE attain realized selected false discovery proportions that are consistent with the nominal targeted FDR_s. There does not appear to be an advantage of multivariate analysis (geneSLOPE) over univariate tests (EMMAXs) in this example: this is consistent with the results in our simulations, which indicate that multivariate analysis is really more powerful when there are many (detectable) signals contributing to the phenotype. While it is by now established that hundreds of different loci contribute to lipid levels, the signal strength in our dataset (which has a modest sample size) is such that only a handful can be identified: in this regime we find no evidence of an advantage for the multivariate linear model.

3 Discussion

Following up on an initial suggestion of [29] and reflecting elements of the standard practice, we argue that discoveries in a GWAS study should not be counted in terms of the number of SNPs for

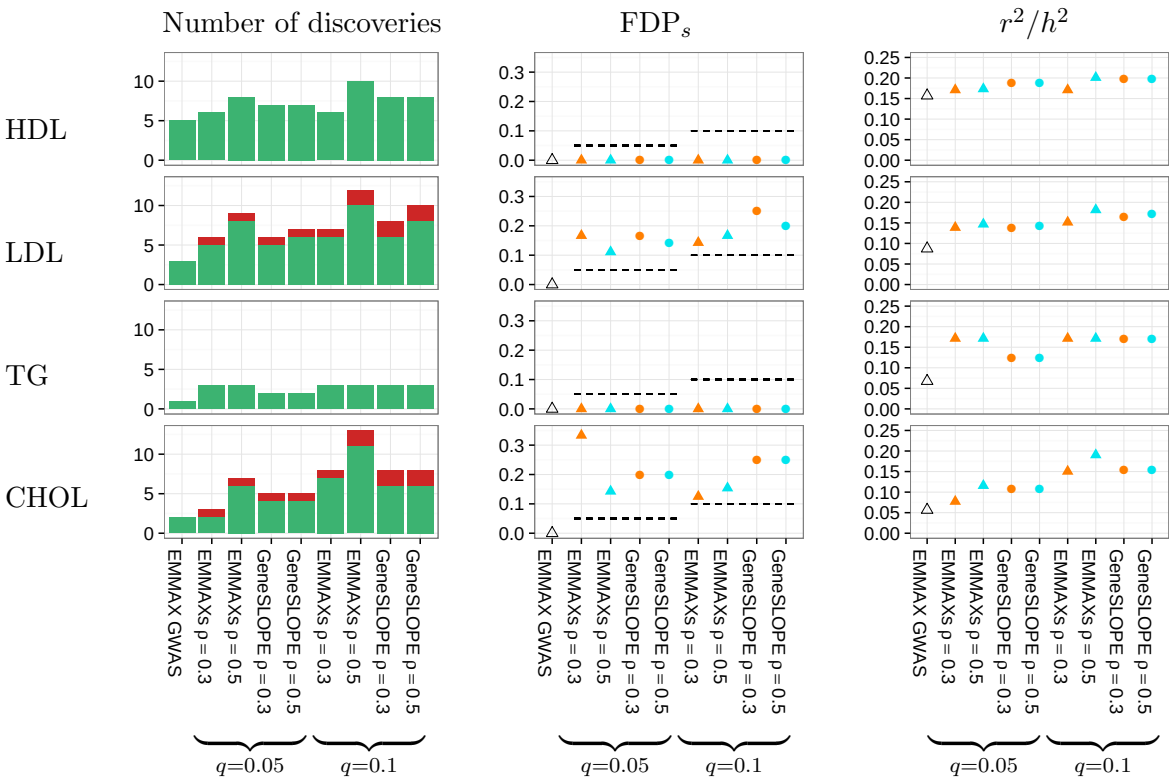


Figure 5: Comparison of methods on the phenotypes high-density lipoproteins (HDL), low-density lipoproteins (LDL), triglycerides (TG), and total cholesterol (CHOL). "Total discoveries" corresponds to the number of selected cluster representatives under each method; in the plot, true discoveries (those within 1Mb of a discovery in [42]) are marked in green, while false discoveries (those not within 1Mb of a discovery in [42]) are marked in red. FDP$_s$ is the realized selected false discovery proportion, and r^2/h^2 is the adjusted r^2 obtained when using the set of selected cluster representatives as predictors in a multiple regression model divided by the proportion of phenotype variance explained by genome-wide SNPs obtained using GCTA.

which the hypothesis of no association is rejected, but in terms of the number of "clusters" of such SNPs. We propose a strategy to control the FDR of these discoveries that consists in identifying groups of hypotheses on the basis of the observed data, selecting a representative for each group, and applying a modified FDR-controlling procedure to the p-values for the selected hypotheses. We present two articulations of this strategy: in one case we rely on marginal tests of association and modify the target rate of BH on the selected hypotheses; in the other case we build on our previous work on SLOPE to fit a multivariate regression model. We show with simulations and real data analysis that the suggested approaches appear to control FDR$_s$ and allow an increase in

power with respect to the standard analysis methods for GWAS.

The idea of identifying groups of hypotheses and somehow transferring the burden of FDR control from the single hypothesis level to the group one is not new [27, 28]. In particular, two recent contributions to the literature can be considered parallel to our suggestions. In the context of tests for marginal association, Foyel-Barber and Ramdas [33] propose a methodology to control FDR both at the level of single hypotheses and groups. In the context of multivariate regression, [45] extends SLOPE to control the FDR for the discoveries of groups of predictors. Both these contributions, however, are substantially different from ours in that they require a definition of groups prior to observation of the data. Instead, our "clusters" are adaptive to the signal, and identified starting from the data. This assures that the group of hypotheses are centered around the locations with strongest signal.

Defining cluster representatives that are input to a multivariate regression framework allows us to think more carefully about what FDR means in the context of a regression model that does not include among the regressor the true causal variants, where one is substantially looking for relevant proxies. In their recent work [46], Foyel-Barber and Candes take a different approach, deciding to focus on the directional FDR. The knock-off filter provides an attractive methodology to analyze GWAS data. However, it still requires an initial selection step: top performance can be achieved only when the selected features are optimally capturing the signal present in a given dataset. We believe that the cluster representatives approach has a substantial edge at this level over, for example, running LASSO with only a modest penalization parameter.

We consider here a fairly simple strategy to construct clusters of SNPs, exploring two possible levels of resolution, corresponding to $\rho = 0.3$ and $\rho = 0.5$. In reality, depending on sample size and genotype density, each dataset might have a different achievable level of resolution. The study of how this can be adaptively learned is deferred to future work.

It should be noted while we conduct formal testing only on the selected set of cluster representatives, when the null hypothesis of no association is rejected for a selected SNP, the entire cluster is implicated. In other words, in follow-up studies, the entire region spanned by the cluster should be considered associated with the trait in question.

Finally, we would like to underscore how, even if we have here focused on the case of GWAS, adopting a selective approach might have wide range applications whenever there is not an exact correspondence between the hypotheses conveniently tested and the granularity of the scientific discoveries. Further studies of the emerging literature on selective inference should lead to better

understanding of the theoretical properties of the method we propose as well as to the identification of other possible strategies.

Acknowledgements

D.B. would like to thank Professor Jerzy Ombach for significant help with the process of obtaining access to the data. This research is supported by the European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552, cofinanced by the Polish Ministry of Science and Higher Education under Grant Agreement 2932/7.PR/2013/2 and by NIH grants R01 HG006695, R01MH101782 and R01MH108467.

References

[1] "*GWAS Catalog*." `http://www.ebi.ac.uk/gwas/` [Accessed: 2016].

[2] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, Oct 2009. PMID: 19812666.

[3] S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, pp. 748–752, Aug 2009. PMID: 19571811.

[4] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common SNPs explain a large proportion of the heritability for human height," *Nat. Genet.*, vol. 42, pp. 565–569, Jul 2010. PMID: 20562875.

[5] M. N. Kvale, S. Hesselson, T. J. Hoffmann, Y. Cao, D. Chan, S. Connell, L. A. Croen, B. P. Dispensa, J. Eshragh, A. Finn, J. Gollub, C. Iribarren, E. Jorgenson, L. H. Kushi, R. Lao, Y. Lu, D. Ludwig, G. K. Mathauda, W. B. McGuire, G. Mei, S. Miles, M. Mittman, M. Patil, C. P. Quesenberry, D. Ranatunga, S. Rowell, M. Sadler, L. C. Sakoda, M. Shapero, L. Shen, T. Shenoy, D. Smethurst, C. P. Somkin, S. K. Van Den Eeden, L. Walter, E. Wan, T. Webster,

R. A. Whitmer, S. Wong, C. Zau, Y. Zhan, C. Schaefer, P. Y. Kwok, and N. Risch, "Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort," *Genetics*, vol. 200, pp. 1051–1060, Aug 2015. PMID: 26092718.

[6] "*UK biobank.*" `http://www.ukbiobank.ac.uk` [Accessed: 2016].

[7] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," *Nature Genetics*, vol. 42, no. 4, pp. 348–354, 2010. PMID: 20208533.

[8] S. Stringer, N. R. Wray, R. S. Kahn, and E. M. Derks, "Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes," *PLOS One*, vol. 6, no. 11, p. e27964, 2011. PMID: 22140493.

[9] C. Sabatti, "Multivariate linear models for GWAS," in *Advances in Statistical Bioinformatics*, pp. 188–208, Cambridge University Press, 2013.

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[11] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Ann. Statist.*, vol. 34, no. 2, pp. 584–653, 2006.

[12] C. J. Hoggart, J. C. Whittaker, M. D. Iorio, and D. J. Balding, "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies," *PLoS Genetics*, vol. 4, no. 7, p. e1000130, 2008. PMID: 18654633.

[13] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, pp. 714–721, Mar 2009. PMID: 19176549.

[14] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange, "Association screening of common and rare genetic variants by penalized regression," *Bioinformatics*, vol. 26, no. 19, pp. 2375–2382, 2010. PMID: 20693321.

[15] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. R. Stat. Soc. Ser. B*, vol. 70, pp. 849–911, 2008.

[16] J. Wu, B. Devlin, S. Ringquist, M. Trucco, and K. Roeder, "Screen and clean: a tool for identifying interactions in genome-wide association studies," *Genet Epidemiol*, vol. 34, pp. 275–285, 2010. PMID: 20088021.

[17] Q. He and D. Y. Lin, "A variable selection method for genome-wide association studies," *Bioinformatics*, vol. 27, pp. 1–8, Jan 2011. PMID: 21036813.

[18] B. A. Logsdon, G. E. Hoffman, and J. G. Mezey, "A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis," *BMC Bioinformatics*, vol. 11, p. 58, 2010. PMID: 20105321.

[19] P. Carbonetto and M. Stephens, "Scalable variational inference for Bayesian variable selection, and its accuracy in genetic association studies," *Bayesian Analysis*, vol. 6, pp. 1–42, 2011.

[20] D. H. Alexander and K. Lange, "Stability selection for genome-wide association," *Genet. Epidemiol.*, vol. 35, pp. 722–728, Nov 2011. PMID: 22009793.

[21] H. Yi, P. Breheny, N. Imam, Y. Liu, and I. Hoeschele, "Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits," *Genetics*, vol. 199, no. 1, pp. 205–222, 2015. PMID: 25354699.

[22] E. Dolejsi, B. Bodenstorfer, and F. Frommlet, "Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion," *PLOS One*, vol. 9, no. 7, p. e103322, 2014. PMID: 25061809.

[23] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candes, "SLOPE – adaptive variable selection via convex optimization," *Annals of Applied Statistics*, vol. 9, pp. 1103–1140, 2015. PMID: 26709357.

[24] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 9440–9445, Aug 2003. PMID: 12883005.

[25] C. Sabatti, S. Service, and N. Freimer, "False discovery rate in linkage and association genome screens for complex disorders," *Genetics*, vol. 164, no. 2, pp. 829–833, 2003. PMID: 12807801.

[26] Y. Benjamini and D. Yekutieli, "Quantitative trait Loci analysis using the false discovery rate," *Genetics*, vol. 171, pp. 783–790, 2005.

[27] M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, "False discovery control for random fields," *J. Amer. Statist. Assoc.*, vol. 99, no. 468, pp. 1002–1014, 2004.

[28] Y. Benjamini and R. Heller, "False discovery rates for spatial signals," *J. Amer. Statist. Assoc.*, vol. 102, no. 480, pp. 1272–1281, 2007.

[29] D. O. Siegmund, B. Yakir, and N. Zhang, "The false discovery rate for scan statistics," *Biometrika*, vol. 98, pp. 979–985, 2011.

[30] C. B. Peterson, M. Bogomolov, Y. Benjamini, and C. Sabatti, "Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies," *Genet. Epidemiol.*, vol. 40, pp. 45–56, Jan 2016. PMID: 26626037.

[31] Y. Benjamini and M. Bogomolov, "Selective inference on multiple families of hypotheses," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 76, no. 1, pp. 297–318, 2014.

[32] Y. Benjamini and D. Yekutieli, "False discovery rate-adjusted multiple confidence intervals for selected parameters," *J. Amer. Statist. Assoc.*, vol. 100, no. 469, pp. 71–93, 2005. PMID: With comments and a rejoinder by the authors.

[33] R. Foygel Barber and A. Ramdas, "The p-filter: multi-layer FDR control for grouped hypotheses," *ArXiv e-prints*, Dec. 2015.

[34] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[35] E. Halperin, G. Kimmel, and R. Shamir, "Tag SNP selection in genotype data for maximizing SNP prediction accuracy," *Bioinformatics*, vol. 21 Suppl 1, pp. 195–203, Jun 2005. PMID: 15961458.

[36] L. Stell and C. Sabatti, "Genetic Variant Selection: Learning Across Traits and Sites," *Genetics*, vol. 202, pp. 439–455, Feb 2016. PMID: 26680660.

[37] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng,

C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y. H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis, "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, pp. 648–660, May 2015. PMID: 25954001.

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, Feb. 1994.

[39] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[40] T. Sun and C. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[41] C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. R. Järvelin, N. B. Freimer, and L. Peltonen, "Genome-wide association analysis of metabolic

traits in a birth cohort from a founder population," *Nat Genet*, vol. 41, no. 1, pp. 35–46, 2009. PMID: 19060910.

[42] Global Lipids Genetics Consortium, "Discovery and refinement of loci associated with lipid levels," *Nature Genetics*, vol. 45, no. 11, pp. 1274–1283, 2013. PMID: 24097068.

[43] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006. PMID: 16862161.

[44] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011. PMID: 21167468.

[45] D. Brzyski, W. Su, and M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors," *ArXiv e-prints*, Nov. 2015.

[46] R. Foygel Barber and E. J. Candes, "A knockoff filter for high-dimensional selective inference," *ArXiv e-prints*, Feb. 2016.

Procedure 4 Sequence of penalties λ for SLOPE

Input: $q \in (0,1)$; $n, M \in \mathbb{N}$

1. set $\lambda_{BH} = [\lambda_{BH}(1), \ldots, \lambda_{BH}(M)]^T$, for $\lambda_{BH}(i) := \Phi^{-1}\left(1 - \frac{qi}{2M}\right)$;
2. define
$$\lambda_G(i) := \begin{cases} \lambda_{BH}(1) & , i = 1 \\ \lambda_{BH}(i)\sqrt{1 + \sum_{j<i} \frac{\lambda_G^2(j)}{n-i}} & , i > 1 \end{cases};$$
3. find the largest index, $k^\star$, such that $\lambda_G(1) \geq \ldots \geq \lambda_G(k^\star)$;
4. put
$$\lambda_i := \begin{cases} \lambda_G(i), & i \leq k^\star \\ \lambda_G(k^\star), & i > k^\star \end{cases}.$$

Procedure 5 Selecting λ when σ is unknown

Input: y, X and basic sequence λ

1. **initialize:** $S_+ = \emptyset$

repeat

2. $S = S_+$
3. compute RSS obtained by regressing y onto variables in S
4. set $\widehat{\sigma}^2 = RSS/(n - |S| - 1)$, where —S— is the number of elements in S
5. compute the solution $\widetilde{\beta}$ to SLOPE with parameter sequence $\widetilde{\sigma} \cdot \lambda_S$
6. set $S_+ = supp(\widetilde{\beta})$ (i.e. S_+ is the set of regressors selected by SLOPE in step 5).

until $S_+ = S$

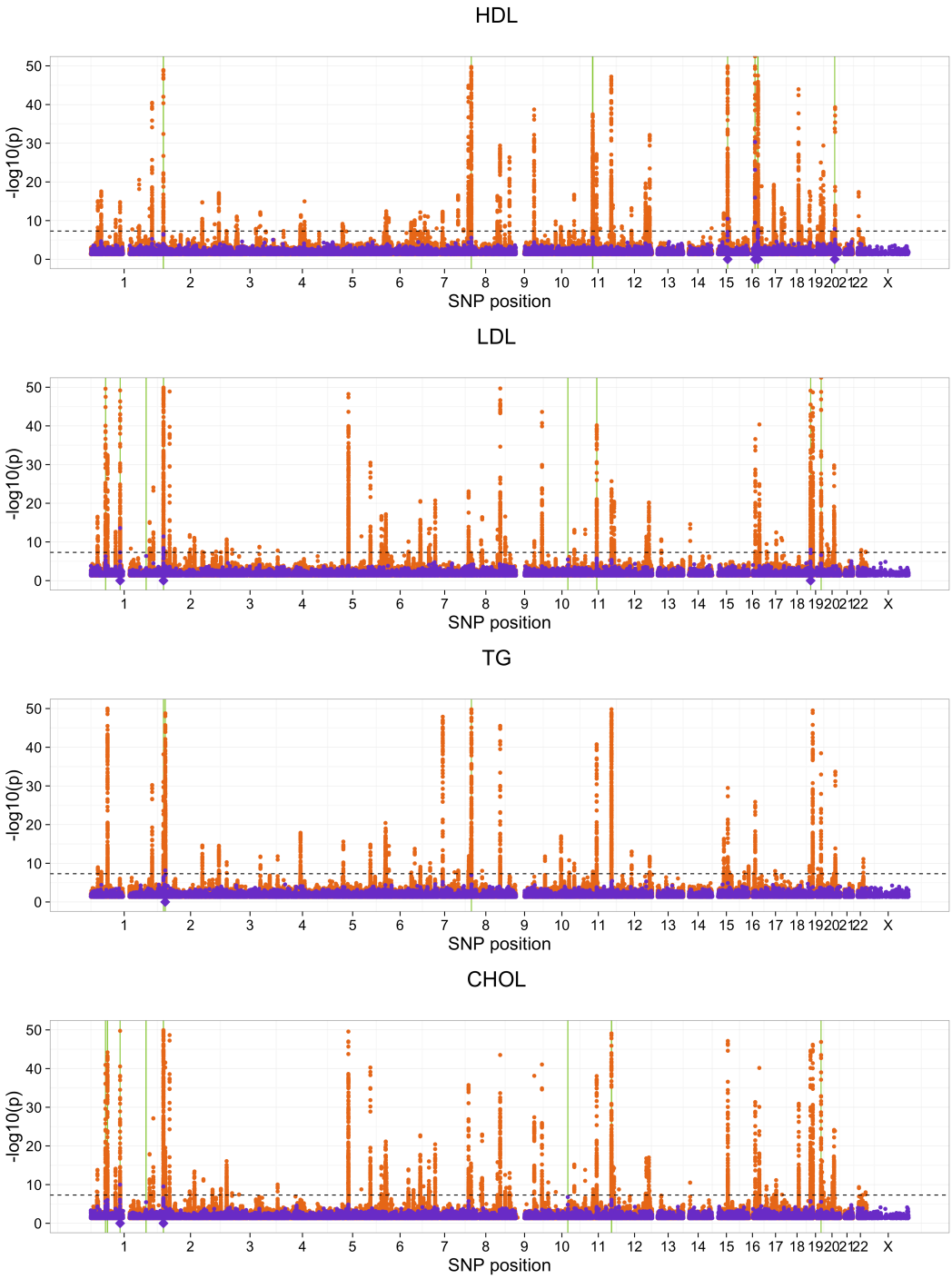


Figure 6: GeneSLOPE selections using $\pi = 0.05$, $\rho = 0.5$, and target FDR_s 0.1 are marked using solid green bars for cluster representatives and semi-transparent bars for the remaining members of the cluster. P-values from EMMAX (purple) and the Global Lipids Genetics Consortium comparison study (orange) are plotted on the -log10 scale. The horizontal dashed line marks a significance cut-off of 5×10^{-8}, and the purple diamonds below the x-axis represent selected cluster representatives under EMMAX using $\pi = 0.05$, $\rho = 0.3$, and a p-value threshold of 5×10^{-8}.