Multilayer Knockoff Filter:  
Controlled variable selection at multiple resolutions

Eugene Katsevich and Chiara Sabatti

June 27, 2017

Abstract

We tackle the problem of selecting from among a large number of variables those that are 'important' for an outcome. We consider situations where groups of variables are also of interest in their own right. For example, each variable might be a genetic polymorphism and we might want to study how a trait depends on variability in genes, segments of DNA that typically contain multiple such polymorphisms. Or, variables might quantify various aspects of the functioning of individual internet servers owned by a company, and we might be interested in assessing the importance of each server as a whole on the average download speed for the company’s customers. In this context, to discover that a variable is relevant for the outcome implies discovering that the larger entity it represents is also important. To guarantee meaningful and reproducible results, we suggest controlling the rate of false discoveries for findings at the level of individual variables and at the level of groups. Building on the knockoff construction of Barber and Candès (2015) and the multilayer testing framework of Barber and Ramdas (2016), we introduce the multilayer knockoff filter (MKF). We prove that MKF simultaneously controls the FDR at each resolution and use simulations to show that it incurs little power loss compared to methods that provide guarantees only for the discoveries of individual variables. We apply MKF to analyze a genetic dataset and find that it successfully reduces the number of false gene discoveries without a significant reduction in power.

Introduction

A motivating example

During the last twenty years, the biotechnology that allows us to identify the locations where the genome of an individual is different from a reference sequence has experienced a dramatic increase in speed and decrease in costs. Scientists have used the resulting wealth of information to investigate empirically how variations in our DNA translate into different measurable phenotypes. While we still know little about the causal mechanisms behind many traits, geneticists agree on the usefulness of a multivariate (generalized) linear model to capture at least as a first approximation the nature of the relation between genetic variation and complex phenotypes. If $y \in \mathbb{R}^{N\times1}$ is the vector collecting the values of a quantitative trait in $N$ subjects, and $X \in \mathbb{R}^{N\times n}$ the matrix storing, column by column,
their genotypes at \( n \) polymorphic sites in the genome, a starting model for their relation is

\[
y = X \beta + \epsilon,
\]

where the coefficients \( \beta \in \mathbb{R}^{n \times 1} \) represent the contributions of measured genetic variations to the trait of interest. We remark on a few characteristics of this motivating genetic application. (1) The adjective ‘complex’ referring to a trait is to be interpreted as non-Mendelian, that is influenced by many different genetic variants: we expect a number of the elements of \( \beta \) to be nonzero and our approaches to data analysis should reflect this to be most effective. (2) The main goal of these studies is the identification of which \( \beta_j \neq 0 \). In other words, the focus is not on developing a predictive model for \( y \), but on selecting important variables in a reproducible manner. (3) Recognizing that \( \beta_j \neq 0 \) corresponds to scientific discoveries at multiple levels: each column of \( X \) represents a single genetic variant, but these are organized spatially in meaningful ways and their coefficients also give us information about coarser units of variation. For example, a number of adjacent polymorphisms might all map to the same gene, the portion of DNA coding for a protein. If the coefficient for any of these polymorphisms is different from zero, then we can conclude that the gene is important for the trait under study. This type of discovery is relevant to advancing our understanding of the biology behind a trait. At the same time, knowing which specific variant influences the phenotype is also relevant: this is the type of information we need for precise clinical testing and genetic counseling.

In summary, an ideal solution would identify important genetic variants accounting for their interdependence, and provide error control guarantees for the discovery of both variants and genes. The work in this paper attempts to achieve this goal. Before we give an overview of the difficulties involved and of the tools we will leverage, we want to underscore that similar problems occur in contexts other than genetics. Modern methods of data acquisition often provide information on an exhaustive collection of possible explanatory variables, even if we know a priori that a large proportion of these are not relevant for our outcome of interest. In such cases, we rely on statistical analysis to identify the important variables in a reproducible and yet powerful manner. Moreover, it is often the case that we measure variables at a very fine resolution, and need to aggregate these measurements for meaningful interpretation. Consider, for example, studies that investigate the role of different brain structures. With magnetic resonance imaging (MRI) we measure on the order of a million voxels and we might use these in a model of neurocognitive ability. Usually, nearby voxel measurements are aggregated and mapped to recognizable larger-scale brain structures. It then becomes important to make sure that the statistical methods we adopt for the analysis of voxels guarantee reproducibility also of these interpretable findings.

### Controlled variable selection in high dimensional regression

In a typical genome wide association study (GWAS), the number of subjects \( N \) is in the order of tens of thousands and the number \( n \) of genetic variants (in this case single nucleotide polymorphisms, or SNPs) is on the order of a million. To provide finite sample guarantees of global error, geneticists typically analyze the relation between \( X \) and \( y \) using a series of univariate regressions of \( y \) on each of the columns of \( X \), obtain the p-values for the corresponding t-tests, and threshold them to achieve family wise error rate (FWER) control. This analysis is at odds with the polygenic nature of the traits and the choice of FWER as a measure of global error makes it difficult to recover a substantial portion of the genetic contribution to the phenotype \(^1\). Using a multiple regression model for analysis and targeting false discovery rate (FDR) \(^2\) are promising alternatives.
Unfortunately, in a context where \( n > N \) these are difficult to implement. Regularized regression, including the lasso \([3]\) and various generalizations, e.g. \([4, 5]\) have proven to be very versatile tools with nice prediction properties, but they do not come with model selection guarantees in finite samples (for examples of asymptotic properties see \([6, 7]\) ). Recent years have seen progress on this front. While in general it is difficult to obtain p-values for high-dimensional regression, \([8]\) propose a construction valid under certain sparsity assumptions. Alternatively, conditional inference after selection \([9, 10]\) can also be used in this context: the idea is to first reduce dimensionality by a screening method, and then apply the Benjamini Hochberg (BH) procedure to p-values that have been adjusted for selection \([11]\). Other approaches have been proposed that bypass the construction of p-values entirely. SLOPE is a modification of the lasso procedure which provably controls the FDR under orthogonal design matrices \([12]\) and has been applied to GWAS with some success \([13]\). A particularly promising approach is the knockoff filter \([14, 15]\), which is based on the construction of artificial “knockoff” variables. This method guarantees FDR control for variable selection in a wide range of settings: reviewed in Section \([1.2]\) it will be one ingredient of our proposal.

**Controlling the false discovery rate at multiple resolutions**

While the standard GWAS analysis results in the identification of SNPs associated with a trait, geneticists also routinely rely on gene level tests, based on the signal coming from multiple variants associated to the same coding region \([16]\) is a recent review), as well as other forms of aggregate tests, based on pathways (see \([17]\) for example). Unfortunately, each of these approaches represents a distinct analysis of the data and the results they provide are not necessarily consistent with each other: we might find association with a SNP, but not with the gene to which it belongs, or with a gene but with none of the variants typed in it. Moreover, multiple layers of analysis increase the burden of multiple testing, often without this being properly accounted for.

In contrast, we want to investigate all the interesting levels of resolution simultaneously (e.g. \([18]\)) and in a consistent fashion, providing meaningful error control guarantees for all the findings. The fact that we have chosen FDR as a measure of error rate makes this a non-trivial endeavor: unlike FWER, FDR is a relative measure of global error and its control depends crucially on how one defines discoveries. A procedure that guarantees FDR control for the discovery of single variants does not guarantee FDR control for the discoveries of genes, as we discuss in Section \([1.1]\). This has been noted before in contexts where there is a well defined spatial relationship between the hypotheses and the discoveries of scientific interest are at coarser resolution than the hypotheses tested (e.g. MRI studies \([19, 20]\), genome scan statistics \([21, 13]\), eQTL mapping \([22]\)). Proposed solutions to these difficulties vary depending on the scientific context, the definition of relevant discoveries, and the way the space of hypotheses is explored; \([23, 24, 25, 26]\) explore hypotheses hierarchically, while \([27]\) considers all levels simultaneously.

This last viewpoint—implemented in a multiple testing procedure called the p-filter and reviewed in Section \([1.2]\)—appears to be particularly well-suited to our context, where we want to rely on multiple regression to simultaneously identify important SNPs and genes.

**Our contribution**

To tackle problems like the one described in the motivating example, we develop the *multilayer knockoff filter* (MKF), a first link between multi-resolution testing and model selection approaches.
We bring together the innovative ideas of knockoffs [14, 15] and p-filter [27] in a new construction that allows us to select important variables and important groups of variables with FDR control. Our methodology—which requires a novel proof technique—does not rely on p-values and provides a great deal of flexibility in the choice of the analysis strategy at each resolution of interest, leading to promising results in genetic applications.

The rest of the paper is structured as follows. Section 1 precisely describes the multilayer variable selection problem and reviews the knockoffs and the p-filter. Section 2 introduces the multilayer knockoff filter, formulates its FDR control guarantees, and uses the new framework to provide a new multiple testing strategy. Section 3 reports the results of simulations illustrating the FDR control and power of MKF. Section 4 summarizes a case study on the genetic bases of HDL cholesterol, where MKF appears to successfully reduce false positives with no substantial power loss. The appendices contain the proofs of our results and details on our genetic findings.

1 Problem setup and background

1.1 Controlled multilayer variable selection

In what follows we assume that \( y \) and \( X \) are linked with a linear model

\[
y = X\beta + \epsilon,\]

where \( \epsilon \sim \mathcal{N}(0, \sigma^2 I) \), \( \beta \) is sparse, and \( N \geq n \). These modeling and distributional assumptions are entirely for ease of exposition: we discuss in the conclusion how our methodology is valid in a more general context. Our goal is to select a set of variables \( S \subset [n] = \{1, \ldots, n\} \) to approximate the underlying set \( \{j : \beta_j \neq 0\} \) of “important” variables. Following [27], we consider situations when \( S \) is interpreted at \( M \) different “resolutions” or layers of inference. The \( m \)th layer corresponds to a partition of \( [n] \), denoted by \( \{A^m_g\}_{g \in [G_m]} \), which divides the set of variables into groups representing units of interest. Our motivating example corresponds to \( M = 2 \): in the first partition, each SNP is a group of its own and in the second, SNPs are grouped by genes. Other meaningful ways to group SNPs in this context could be by functional units or by chromosomes. The selected variables \( S \) induce group selections \( S_m \subset [G_m] \) at each layer via

\[
S_m = \{g \in [G_m] : S \cap A^m_g \neq \emptyset\} :\]
i.e. a group is selected if at least one variable belonging to that group is selected.

To ensure reproducibility of the findings with respect to each layer of interpretation, we seek methods for which \( S_m \) has a low false discovery rate for each \( m \). If \( H_0 = \{j \in [n] : \beta_j = 0\} \) is the set of null variables, then the set of null groups at layer \( m \) is

\[
H^m_0 = \{g \in [G_m] : A^m_g \subset H_0\}.
\]

Then, the number of false discoveries at layer \( m \) is

\[
V_m(S) = |S_m \cap H^m_0|,
\]

which we abbreviate with \( V_m \) whenever possible without incurring in confusion. The corresponding false discovery rate is defined setting

\[
\text{FDP}_m(S) = \frac{V_m(S)}{|S_m|}, \quad \text{FDR}_m = \mathbb{E}[\text{FDP}_m(S)],
\]
Figure 1: Demonstration that small FDP\textsubscript{ind} does not guarantee small FDP\textsubscript{grp}. Each square represents a variable and columns contain variables in the same group. The left-most panel illustrates the true status of the hypotheses: a black square corresponds to non-null and a white square to a null variable. In the second and third panels, black squares represent selected variables by KF and MKF respectively with $q_{\text{ind}} = q_{\text{grp}} = 0.2$. The KF has FDP\textsubscript{ind} = 0.21 and FDP\textsubscript{grp} = 0.58, while MKF has FDP\textsubscript{ind} = 0.05 and FDP\textsubscript{grp} = 0.17.

using the convention that 0/0 = 0 (the FDP of no discoveries is 0). A selection procedure obeys *multilayer FDR control* \cite{27} at levels $q_1, \ldots, q_M$ for each of the layers if

$$FDR_m \leq q_m$$

(2)  

It might be surprising that the guarantee of FDR control for the selection of individual variables $X_1, \ldots, X_n$ does not extend to the control FDR$\_m$. Figure 1 provides an illustration of this fact for the simple case of $M = 2$, with one layer corresponding to the individual variables (denoted below by the subscript “ind”) and one group layer (denoted by the subscript “grp”). We generate a matrix $X$ ($n = 500, N = 1200$) sampling each entry independently from a standard normal distribution. The 500 variables are organized into 50 groups of 10 elements each. The outcome $y$ is generated from $X$ using a linear model with 70 nonzero coefficients, evenly spread across 10 groups. The middle panel of Figure 1 shows the results of applying the knockoff filter \cite{14} with a target FDR level of 0.2. While the false discovery proportion for the individual layer is near the nominal level (FDP\textsubscript{ind} = 0.21), the FDP at the group layer is unacceptably high (FDP\textsubscript{grp} = 0.58). The middle panel of Figure 1 guides our intuition: false discoveries occur roughly uniformly across null variables and are then dispersed across groups (instead of being clustered in a small number of groups). When the number of null groups is comparable to or larger than the number of false discoveries, we have $V_{\text{grp}} \approx V_{\text{ind}}$ and we can write roughly

$$\text{FDP}_{\text{grp}} = \frac{V_{\text{grp}}}{|S_{\text{grp}}|} \approx \frac{V_{\text{ind}}}{|S_{\text{grp}}|} = \frac{|S_{\text{ind}}|}{|S_{\text{grp}}|} \frac{V_{\text{ind}}}{|S_{\text{ind}}|} = \frac{|S_{\text{ind}}|}{|S_{\text{grp}}|} \text{FDP}_{\text{ind}}.$$  

Hence, the group FDP is inflated roughly by a factor of $|S_{\text{ind}}|/|S_{\text{grp}}|$ compared to the individual FDP. This factor is high when we make several discoveries per group, as we expect when a non-null group has a high number of non-null elements (high *saturation*).

To summarize, as discussed in \cite{27}, if we want to to make reproducible discoveries at $M$ layers, we need to develop model selection approaches that explicitly target (2). The multilayer knockoff filter precisely achieves this goal, as is illustrated in the third panel of Figure 1. Since MKF builds on the knockoff filter and the p-filter, we pause to review both methods in more detail.
1.2 Knockoffs: Selecting individual variables with FDR control

Consider the linear model in [1]: the idea in [14, 15] is to artificially construct knockoff variables \( \widetilde{X} \in \mathbb{R}^{N \times n} \) that mimic the original variables \( X \), but are not related to \( y \). Then, \([X \ \widetilde{X}]\) are assessed jointly for association with \( y \) (e.g. via a penalized regression), and one selects those variables that appear more strongly related to \( y \) than their knockoffs. A more precise outline follows.

### Framework 1: Knockoff Filter

- **Data:** \( X, y \), FDR target level \( q \)
- **1** Construct knockoff features \( \tilde{X} \);
- **2** Construct knockoff statistics \((W_1, \ldots, W_n)\), which jointly satisfy the sign-flip property, such that the magnitude of \( W_j \) reflects the strength of association between \( X_j \) and \( y \) and its sign reflects whether \( X_j \) is more strongly associated to \( y \) than is \( \tilde{X}_j \);
- **3** For each \( t \geq 0 \), define \( S(t) = \{ j : W_j \geq t \} \);
- **4** Let \( \hat{V}(t) \) be an estimate of \( V(S(t)) \) and define \( \hat{FDP}(t) = \frac{\hat{V}(t)}{|S(t)|} \);
- **5** Find \( t^* = \min\{ t : \hat{FDP}(t) \leq q \} \);

**Result:** Selection set \( S = S(t^*) \).

**Step 1: constructing knockoffs.** The papers [14] and [15] provide two different constructions of knockoffs, based on different assumptions on \( X \). The original proposal [14] considers the case where \( N \geq n \) and \( X \) is a fixed design matrix, while [15] deals with the situation where rows of \( y \) and \( X \) are sampled i.i.d. from a population while making no assumptions on the dimensions \( n \) and \( N \). The hypotheses in [15] are more appropriate for the applications to genetic association studies, but for ease of exposition, we adopt the construction in [14]. It is important to note, however, that our contribution works equally well with either construction.

Valid knockoff features (as constructed in [14]) are those such that the first two sample moments of the augmented design matrix \([X \ \widetilde{X}]\) are invariant when any subset \( \mathcal{C} \subset [n] \) of variables are swapped with their knockoffs. We denote the result of this swapping operation by \([X \ \widetilde{X}]_{\text{swap}(\mathcal{C})}\). The invariance of first moments to swapping can be ensuring simply by centering each variable and its knockoff. The swap invariance of the second moment amounts to

\[
[X \ \widetilde{X}]_{\text{swap}(\mathcal{C})}^T[X \ \widetilde{X}]_{\text{swap}(\mathcal{C})} = [X \ \widetilde{X}]^T[X \ \widetilde{X}].
\]  

(3)

**Step 2: obtaining knockoff statistics.** The first step is to construct statistics \( Z \) and \( \tilde{Z} \) for each variable and knockoff variable, respectively. For example, this can be done by running a lasso regression [3] for each value of the regularization parameter \( \lambda \):

\[
\hat{b}^*(\lambda), \tilde{b}^*(\lambda) = \arg\min_{b, \tilde{b}} \frac{1}{2} \left\| y - [X \ \widetilde{X}] \begin{bmatrix} b \\ \tilde{b} \end{bmatrix} \right\|_2^2 + \lambda \left( \|b\|_1 + \|\tilde{b}\|_1 \right),
\]  

(4)

and then defining \( Z_j (\tilde{Z}_j) \) as the value of \( \lambda \) for which \( X_j (\tilde{X}_j) \) first enters the lasso path:

\[
Z_j = \sup \{ \lambda : b_j^*(\lambda) \neq 0 \}, \quad \tilde{Z}_j = \sup \{ \lambda : \tilde{b}_j^*(\lambda) \neq 0 \}.
\]  

(5)

Other ways of defining these statistics are possible, but the knockoff filter framework requires at least that the function \( \varepsilon([X \ \widetilde{X}], y) \) defining \((Z, \tilde{Z})\) should be swap-equivariant, i.e. \( \varepsilon([X \ \widetilde{X}]_{\text{swap}(\mathcal{C})}, y) = \)
(Z, ˘Z)swap(C) for any set C. Then, we define Wj as a “difference” fj(Zj, ˘Zj) between Zj and ˘Zj, where fj is any antisymmetric function, i.e. fj(zj, ˘zj) = −fj(˘zj, zj). For example, we may use fj(zj, ˘zj) = zj − ˘zj or fj(zj, ˘zj) = max(zj, ˘zj) · sign(zj − ˘zj), the signed max. Taken together, these steps lead to a function w mapping the augmented design matrix and response vector to a vector of knockoff statistics:

\[ W = (W_1, \ldots, W_n) = w([X \ X], y). \]

In particular, W defined via (4), (5), and the signed max function are called lasso signed max statistics.

**Definition: sign-flip property** The knockoff statistics W satisfy the sign-flip property if for any vector \( \epsilon \in \{-1, +1\}^n \) such that \( \epsilon_j = +1 \) for non-null Hj, we have

\[ \epsilon \odot W \overset{d}{=} W, \]

where \( \odot \) denotes element-wise multiplication.

Under the fixed-design linear model (1) with homoskedastic Gaussian noise, [14] show that any swap-equivariant choice of z that also satisfies the sufficiency property will lead to knockoff statistics with the sign-flip property. The sufficiency property is the requirement that z operates on the data only through the sufficient statistics \([X \ X]^T y\) and \([X \ X]^T[X \ X] \).

**Step 3: defining a selection set.** By our construction, large positive values of Wj provide evidence that variable j is non-null, so the selection set will take the form of all variables whose Wj is above a threshold t.

**Step 4: estimating FDP.** To estimate the FDP for each threshold t, we must estimate the number of false discoveries \( V(S(t)) \). Motivated by the sign-flip property, Procedures 1 and 2 define \( \hat{V}(t) \) via the reflection of S(t) about the origin.

| Procedure 1: KF                  |
|----------------------------------|
| 1 Framework 1 with \( \hat{V}(t) = |\{j : W_j \leq -t\}| \). |

| Procedure 2: KF+                |
|----------------------------------|
| 1 Framework 1 with \( \hat{V}(t) = 1 + |\{j : W_j \leq -t\}| \). |

**Step 5: choosing a threshold.** Once \( \hat{FDP}(t) \) is defined for each t, the threshold \( t^* \) is chosen as liberally as possible subject to this estimate being below the target level q. By leveraging the sign-flip property, [14] show that the KF+ procedure controls FDR at level q, and that the KF procedure controls the modified FDR at level q, where the modified FDR is defined as

\[ \text{mFDR} = \mathbb{E}\left[ \frac{|S \cap \mathcal{H}_0|}{|S| + q^{-1}} \right]. \]

**1.3 Group knockoff filter: Selecting groups of variables with FDR control**

A group knockoff filter [28] has been proposed to control the FDR of group discoveries defined by one partition \( \{A_g\}_{g \in [G]} \). This is essentially the same framework as the knockoff filter with the exception that groups, rather than individual variables, are the object of inference.
Crucially, group FDR control requires the invariance property (3) to hold only for all sets \( C \) that can be expressed as the union of a collection of groups. While the knockoffs in [14] satisfy this property, one can capitalize on the less stringent requirement to construct a different \( \tilde{X} \), leading to a more powerful procedure when variables within groups are highly correlated with each other. This reflects the fact that, in presence of correlation, it is easier to determine whether an entire group of variables is non-null than to determine which specific variables in this group are non-null. Among the many possible constructions, “equicorrelated knockoffs” are the most tractable: originally proposed by [14] and extended to the group case by [28], they are derived from an explicit formula in terms of \( X \) and computed in \( O(n^3) \) time (the bottleneck being an inversion of an \( n \times n \) matrix).

The authors of [28] suggest constructing group knockoff statistics \( W_1, \ldots, W_G \) in a way similar to (5), but relying on the path for the group lasso [4]:

\[
\arg \max_{b, \tilde{b}} \frac{1}{2} \| y - [X \; \tilde{X}] \| \left( b \; \tilde{b} \right) \|^2 + \lambda \sum_{g=1}^{G} \sqrt{|A_g|} \| b_{A_g} \|_2 + \lambda \sum_{g=1}^{G} \sqrt{|A_g|} \| \tilde{b}_{A_g} \|_2.
\]

(6)

Steps 3-5 of Framework 1 carry over naturally.

1.4 p-filter: Multilayer FDR control from p-values

The p-filter is a multiple testing procedure that takes as input a set of p-values for hypotheses \( H = (H_1, \ldots, H_n) \) and a collection of partitions of these hypotheses \( \{A_g^m\}_{g,m} \) with \( g = 1, \ldots, G_m \) and \( m = 1, \ldots, M \), and results in the identification of rejections that guarantee multilayer FDR control with respect to these partitions, as long as the p-values \( p_1, \ldots, p_n \) are PRDS [29]. The p-filter can actually be seen as a generalization of the BH procedure, with layers specific p-value cutoffs \( t = (t_1, \ldots, t_M) \). For each \( m \), the p-values for the intersection hypotheses corresponding to groups \( g \) are obtained by combining the p-values for the corresponding hypotheses in \( H \) using Simes’s rule [30]. The selection set \( S(t) \) is then defined as the set of hypotheses \( j \) whose corresponding group passes the threshold at each layer.

Procedure 3: p-filter

Data: p-values \( p_1, \ldots, p_n \) for hypotheses in \( H \); \( M \) partitions of \( H \): \( \{A_g^m\}_{g,m} \) with \( g = 1, \ldots, G_m \) and \( m = 1, \ldots, M \), and target levels \( q_1, \ldots, q_M \)

1. Derive group p-values via the Simes test: \( p_g^m = \text{Simes} \{ p_j : j \in A_g^m \} \);
2. For \( t \in [0, 1]^M \), define

\[
S(t) = \{ j : p_g^m(j, m) \leq t_m \text{ for all } m \},
\]

where \( g(j, m) \) is the group at layer \( m \) to which hypothesis \( j \) belongs;

3. For each \( m \), define \( \widetilde{FDP}_m(t) = \frac{G_m t_m}{S_m(t)} \) where \( S_m(t) = \{ g \in 1, \ldots, G_m : S(t) \cap A_g^m \neq \emptyset \} \);
4. Find \( t^* = \max \{ t : \widetilde{FDP}_m(t) \leq q_m \text{ for all } m \} \);

Result: Selection set \( S = S(t^*) \).

Step 4 above needs clarification, since the maximum of a set in \( M \) dimensions is not well defined in general. Consider the set of valid thresholds

\[
\mathcal{T}(q_1, \ldots, q_M) = \{ t : \widetilde{FDP}_m(t) \leq q_m \text{ for all } m \}.
\]

(7)
Then, [27] show that this set has the “upper right-hand corner property,” which means that the point 
\( t^* = (t_1^*, \ldots, t_M^*) \) defined by 
\[
  t_m^* = \max \{ t_m : (t_1, \ldots, t_M) \in \mathcal{T} \text{ for some } t_1, \ldots, t_{m-1}, t_{m+1}, \ldots, t_M \}
\]
belongs to \( \mathcal{T} \). Hence, the point \( t^* \) is the upper right-hand corner of \( \mathcal{T} \) and is the maximum in step 4 of Procedure [3]. It can be found efficiently using an iterative coordinate-wise search (see Algorithm 1 in [27]).

## 2 Multilayer Knockoff Filter

The knockoff filter provides a way to construct statistics, endowed with rigorous inferential properties, to quantify the significance of variables in a model selection context. The p-filter suggests a framework to combine p-values across multiple layers while ensuring FDR control at each layer. It is quite natural to integrate these two sets of ideas to tackle the problem of multilayer FDR control for model selection and we do so by modifying the p-filter procedure to handle knockoff statistics instead of p-values.

Importantly, we depart from the p-filter by allowing any valid constructions of knockoff statistics at each layer and assuming nothing about the between-layer dependencies of these statistics, gaining additional flexibility and power. While formulating this methodological extension is fairly easy, the main challenge lies in proving multilayer FDR control for the resulting procedure.

### 2.1 Procedure

We introduce first the multilayer knockoff filter FDR controlling framework and then fill in the details to make the procedure concrete.

| Framework 2: Multilayer Knockoff Filter |
|----------------------------------------|
| **Data:** \( X, y \), partitions \( \{A_g^m\}_{g,m} \) with \( g = 1, \ldots, G_m \) and \( m = 1, \ldots, M \), FDR target levels \( q_1, \ldots, q_M \) |
| **for** \( m = 1 \) **to** \( M \) **do** |
| 1. Construct group knockoff features \( \tilde{X}^m \); |
| 2. Construct group knockoff statistics \( W^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([X \tilde{X}^m], y) \) satisfying the sign-flip property; |
| **end** |
| **Define** \( S(t) = \{ j : W_{g(j,m)}^m \geq t_m \ \forall m \} \); |
| **For each** \( m \), let \( \hat{V}_m(t_m) \) be an estimate of \( V_m(S(t)) \) and define \( \hat{\text{FDP}}_m(t) = \frac{\hat{V}_m(t_m)}{|S_m(t)|} \); |
| **Find** \( t^* = \min \{ t : \hat{\text{FDP}}_m(t) \leq q_m \ \forall m \} \); |
| **Result:** Selection set \( S = S(t^*) \). |

**Step 2: Constructing group knockoffs.** To maximize our power, we construct group knockoffs \( \tilde{X}^m \) at each layer as in [28]; see Section 1.3.
Step 3: Constructing group knockoff statistics. Valid group knockoff statistics $W^m$ have to satisfy the the sign-flip property and, as in Framework 1, they can be constructed using an antisymmetric function of feature statistics $(Z^m, \tilde{Z}^m)$ that is group-swap equivariant and obeys the sufficiency property. Specifically, $z^m([X \tilde{X}^m], y) = (Z^m_1, \ldots, Z^m_{G_m}, \tilde{Z}^m_1, \ldots, \tilde{Z}^m_{G_m})$ is group-swap equivariant if for each $C_m \subset [G_m]$ and $C = \cup_{g \in C_m} A^m_g$, we have

$$z^m([X \tilde{X}^m]_{\text{swap}(C)}, y) = (Z^m_1, \ldots, Z^m_{G_m}, \tilde{Z}^m_1, \ldots, \tilde{Z}^m_{G_m}_{\text{swap}(C_m)}).$$

In words, this means that swapping entire groups in $[X \tilde{X}^m]$ translates to swapping the entries of $(Z^m, \tilde{Z}^m)$ corresponding to those groups.

Generalizing the proposal of [28]—based on the group lasso—we consider a class of feature statistics $(Z^m, \tilde{Z}^m)$ defined as in (5) starting from the penalized regression

$$\arg\max_{b, \bar{b}} \frac{1}{2} \|y - [X \tilde{X}^m](b)\|^2 + \lambda \left( \sum_{g=1}^{G_m} \ell^m_g(b_{A^m_g}) + \sum_{g=1}^{G_m} \ell^m_g(\bar{b}_{A^m_g}) \right),$$

where $\ell^m_g$ are arbitrary penalty functions (the group lasso corresponds to $\ell^m_g(u) = \sqrt{|A^m_g| \|u\|_2}$). For each $m$, the regularization in (8) is defined on subsets of $b$ corresponding to groups at the $m$th layers and it is separable with respect to the $m$th partition $\{A^m_g\}$. While this guarantees group-swap invariance, other constructions are certainly possible.

Different choices of $\ell^m_g$ allow us to adapt to the available information on signal structure and potentially gain power. The 2-norm that defines the group lasso has equal contributions from each variable in the group, resulting in equal $Z_j$s. Drawing an analogy to global testing, this definition is similar to the Fisher combination test or the chi-squared test, which are known to be powerful in regimes when the signal is weak and distributed. In our case, we should apply group lasso based statistics if we suspect that each group has many non-nulls. Taking this analogy further, we can also construct a Simes-like statistic that is more suited to the case when we believe each group has a few strong signals. This test statistic is defined by letting $\ell^m_g(u) = \|u\|_1$; i.e. running a regular lasso regression. This will allow for each variable to come in on its own, and the knockoff statistic $W^m_g$ will be driven by the strongest predictor in the corresponding group.

Step 6: Definition of $\hat{V}_m(t_m)$. As for the knockoff filter, the FDR guarantee for the method depends crucially on the estimate $\hat{V}_m(t_m)$. Intuitively, the larger $\hat{V}_m(t_m)$ is, the stronger the FDR guarantee. Similar to the knockoff filter, we consider methods MKF and MKF+ based on two definitions of $\hat{V}_m(t_m)$. However, we shall also consider the effect of adding a constant multiplier $c$ to this estimate as well; see Procedures 4 and 5.

**Procedure 4: MKF($c$)**

1. Framework 2 with $\hat{V}_m(t_m) = c \cdot |\{g : W^m_g \leq -t_m\}|$.

**Procedure 5: MKF($c$)**

1. Framework 2 with $\hat{V}_m(t_m) = c \cdot (1 + |\{g : W^m_g \leq -t_m\}|)$.
Definition and computation of $t^*$. As in the p-filter, the definition of $t^*$ is an optimization over an $M$-dimensional set. Luckily, this step is well-defined for the multilayer knockoff filter as well.

Lemma 1. For any definition of $\hat{V}_m(\cdot)$ in Framework 3 depending only on $t_m$ (as opposed to the entire vector $t$), the set $T(q_1, \ldots, q_M)$ defined in (7) will have the lower left-hand corner property.

Note that when basing the procedure on test statistics instead of p-values, we are looking for the lower left-hand corner instead of the upper right-hand corner. The proof of this lemma is the same as that of Theorem 3 in [27]. In addition to being well-defined, the threshold $t^*$ can be computed efficiently using the same iterative coordinate-wise search algorithm proposed in [27].

Figure 2 provides an illustration of the multilayer knockoff filter on simulated data. The 2000 hypotheses are broken into 200 groups, each of size 10. (More details about this example are in Section 3.1; it corresponds to the high saturation setting, with SNR = 0.5 and feature correlation 0.1.) The multilayer knockoff filter for the individual layer and the group layer results in the selection region in the upper right-hand corner and enjoys high power. For comparison, the (single layer) knockoff filter selects all points to the right of the broken light blue line; among these there are several nulls not selected by MKF as their group signal is not strong enough. In this simulation, MKF reduces false positives without losing any power.

Figure 2: Illustration of one run of the multilayer knockoff filter. Each point represents an individual-layer hypothesis $j$, with coordinates $W_{1j}$ and $W_{2g(j,2)}$: circles indicate nulls, asterisks non-nulls. The solid blue lines are the thresholds for multilayer knockoff filter, while reflected transparent blue lines are used in the definition of $\hat{V}_m$. The broken cyan line represents the threshold for knockoff filter. The darkly shaded upper right corner represents the selection set of the multilayer knockoff filter, and the lightly shaded left half-plane represents the area used to define $\hat{V}^{\text{ind}}$. 
2.2 Theoretical results

Our main theoretical result guarantees that the MKF\((c_{kn})+\) has multilayer FDR control at the target levels, and that MKF\((c_{kn})\) has multilayer mFDR control at the target levels, where \(c_{kn} = 1.93\).

**Theorem 1.** Suppose \(W^m\) obeys the sign-flip property for all \(m\). Then, the MKF\((c)\) method satisfies

\[
\text{FDR}_m \leq \frac{c_{kn}}{c} q_m \quad \text{for all } m,
\]

where \(c_{kn} = 1.93\). The MKF\((c)\) method satisfies

\[
\text{mFDR}_m \leq \frac{c_{kn}}{c} q_m \quad \text{for all } m.
\]

In particular, the MKF\((c_{kn})+\) and MKF\((c_{kn})\) methods have multilayer FDR control and multilayer mFDR control, respectively, at the target levels.

While deferring technical lemmas to the appendix, we outline here the essential steps of the proof as they differ fundamentally from those of KF and p-filter. The proof of FDR control for the knockoff filter relies on a clever martingale argument that depends heavily on the fact that the threshold \(t\) is one-dimensional: the cutoff \(t^*\) can be viewed as a stopping time with respect to a certain stochastic process. Instead, we are dealing with an \(M\)-dimensional threshold \(t^*\) whose entries depend on the values of \(W^m\) for all \(m\). As the knockoff statistics have complex dependencies with each other, we cannot represent \(t_m\) as a stopping time with respect to a process that depends only on \(W^m\). The p-filter being a multilayer method, the proof of FDR control deals with the complex nature of the threshold \(t^*\). However, by defining the \(p\)-values at each layer from the individual hypotheses \(p^m_{g}\) across layers and use this crucially in the proof. In contrast, we intentionally avoid specifying the relations between \(W^m\) for different \(m\).

**Proof.** We prove FDR control for MKF\((c)\)+; the result for MKF\((c)\) follows from a very similar argument. We start introducing the following quantities:

\[
V^+_m(t_m) = |\{g : W^m_g \geq t_m\} \cap H^m_0|, \quad V^-_m(t_m) = |\{g : W^m_g \leq -t_m\} \cap H^m_0|.
\]

Note that both \(V^+_m(t_m)\) and \(V^-_m(t_m)\) are defined in terms of the \(m\)th layer only and that \(V^+_m(t_m) = V_m(S(0, \ldots, 0, t_m, 0, \ldots, 0))\), while \(V^-_m(t_m)\) is similar to \(\hat{V}_m(t_m)\). It is easy to verify that these two quantities satisfy

\[
V^+_m(t_m) \geq V_m(S(t)), \quad \hat{V}_m(t_m) \geq c(1 + V^-_m(t_m)).
\]

Then, for each \(m\), we have

\[
\begin{align*}
\text{FDR}_m &= \mathbb{E}[\text{FDP}_m(t^*)] \\
&= \mathbb{E} \left[ \frac{V_m(S(t^*))}{|S_m(t^*)|} | I(t^* < \infty) \right] \\
&= \mathbb{E} \left[ \frac{V_m(S(t^*))}{\hat{V}_m(t_m)} | I(t^* < \infty) \right] \\
&\leq q_m \cdot \frac{1}{c} \mathbb{E} \left[ \frac{V^+_m(t_m)}{1 + V^-_m(t_m)} \right].
\end{align*}
\]
Hence it suffices to show that
\[ E \left[ \sup_{t_m} \frac{V_m^+(t_m)}{1 + V_m^-(t_m)} \right] \leq c_{km}. \] (10)

The introduction of the supremum over \( t_m \) in the last equation is a key step in the proof: it makes the random variables in the expectation (10) depend only on the knockoff statistics at the \( m \)th layer, decoupling the problem across layers and allowing any type of dependence between statistics for different values of \( m \).

Given that we are working with quantities defined in a layer only, we can drop the subscript \( m \), and consider (10) as a statement about any set of knockoff statistics \((W_1, \ldots, W_G)\) satisfying the sign-flip property. Hence, \( W_g^m, V_m^+(t_m), V_m^-(t_m) \) become \( W_g, V^+(t), V^-(t) \), respectively, and so on. We are left with
\[ E \left[ \sup_{t} \frac{V^+(t)}{1 + V^-(t)} \right] = E \left[ \max_{k \leq G_0} \frac{|\{g : W_g \geq t \} \cap \mathcal{H}_0|}{1 + |\{g : W_g \leq -t \} \cap \mathcal{H}_0|} \left| W \right| \right]. \] (11)

Now, consider ordering \( \{W_g\} \in \mathcal{H}_0 \) by magnitude: \( |W_{(1)}| \geq \cdots \geq |W_{(G_0)}| \), where \( G_0 = |\mathcal{H}_0| \). Let \( \sigma_g = \text{sgn}(W_g) \). By the sign-flip property, \( \sigma_g \) are distributed as i.i.d. coin flips independently of \( |W| \). Hence, we can drop the conditioning in the last line of (11). Additionally rewriting the quantity inside the expectation in terms of \( \sigma_g \), we get
\[ E \left[ \sup_{t} \frac{V^+(t)}{1 + V^-(t)} \right] = E \left[ \max_{k \leq G_0} \frac{|\{g : \sigma_g = +1 \}|}{1 + |\{g : \sigma_g = -1 \}|} \right]. \]

We can think of \( \sigma_g \) as the increments of a simple symmetric random walk on \( \mathbb{Z} \). The numerator above represents the number of steps to the right this walk takes, and the denominator the number of steps to the left. The quantity we are bounding is essentially the maximum over all steps in the walk of the ratio of steps right to steps left, averaged over all realizations of the random walk. Let \( S_k = |\{g \leq k : \sigma_g = +1\}| \) be the number of steps right and \( k - S_k = |\{g \leq k : \sigma_g = -1\}| \) the number of steps left. It suffices to show that
\[ E \left[ \max_{k \geq 0} \frac{S_k}{1 + k - S_k} \right] \leq 1.93. \]

This is the content of Lemma 2 which is proved in the appendix.

\[ \Box \]

### 2.3 Discussion

#### Comparison with p-filter

In addition to using knockoff statistics instead of p-values, the multilayer knockoff filter differs from the p-filter in that it does not start from a set of individual-level statistics and construct group-level ones using specific functions of these: instead the statistics \( W^m \) are constructed starting directly from the original data. This decision involves a trade-off: we get a more general procedure (and theoretical result) at the cost of a looser bound.
By making no assumptions on the between-layer dependencies of $W^m_g$, the multilayer knockoff filter allows extra flexibility that can translate into greater power. For example, there might be different sources of prior information at the SNP and the gene levels: the analyst can use each source of information at its respective layer to define more powerful knockoff statistics (based on specific penalties) without worrying about coordinating these statistics in any way. Even if the same penalization is used in all layers, there is a potential power increase due to the fact that we can use group knockoff statistics rather than individual ones. This advantage is especially pronounced if none of the layers consists of singletons.

The price we pay for this generality is the multiplier $c_{kn} = 1.93$ in Theorem 1. To understand its effect, note that in Procedures 4 and 5, by analogy with KF, the natural choice is $c = 1$ and define $MKF = MKF(1)$ and $MKF_+ = MKF(1)+$. Then Theorem 1 states that $MKF_+ (MKF)$ has an FDR (mFDR) that is bound by $c_{kn} q_m$. Compare this to the theoretical result for the p-filter, which is shown to have exact multilayer FDR control: by explicitly leveraging the joint distribution of $p^m_g$, [27] get a handle on the complicated thresholds $t^*_m$ and get a tight result. Meanwhile, our constant multiplier comes from the introduction of the supremum in (9): this amounts to a worst-case analysis, which for most constructions of $W^m_g$ will not be tight.

Indeed, across all our simulations in Section 3, we find that $MKF_+$ has multilayer FDR control at the target levels (i.e. the constant is not necessary). Hence, we recommend that practitioners apply the MKF or $MKF_+$ methods, without worrying about the correction constants. We view our theoretical result as an assurance that even in the worst case, the FDRs of MKF at each layer will not be much higher than their nominal levels $q_m$s.

**Generalized p-filter**

On the heels of the above discussion, we define as generalized p-filter a procedure that is the same as the p-filter, except that the p-values $p^m_g$ are any valid p-values for the hypotheses in layer $m$.

**Theorem 2.** Suppose for each $m$, the null p-values among $\{p^m_g\}$ are independent and uniformly distributed. Then, the generalized p-filter satisfies

$$mFDR_m \leq c_{pf}(G_m) \cdot q_m \quad \text{for all } m,$$

where $c_{pf}(G) = 1 + \exp \left( G^{-1/2} + \frac{1}{2} G^{-1} \right) 0.42 + eG^{-1/4}$.

**Remark 2.1.** Unlike for the multilayer knockoff filter, note that we do not have one universal constant multiplier $c_{pf}$. Instead, we get a bound $c_{pf}(G_m)$ that depends on the number of groups at each layer. However, strong numerical evidence suggests that in fact we can replace $c_{pf}(G_m)$ in the theorem with its limiting value 1.42. See Remark B.1 in the appendix for additional comments. Moreover, the assumption of independent null p-values can potentially be relaxed to a PRDS assumption, but we have not explored this avenue.

**Proof.** By similar logic as in the proof of Theorem 1, it suffices to verify the sufficient condition

$$\mathbb{E} \left[ \sup_{t_m \in [0,1]} \frac{\left| \left\{ g \in \mathcal{H}_0^m : p^m_g \leq t_m \right\} \right|}{1 + G_m t_m} \right] \leq c_{pf}(G_m).$$

Again, note that the problem decouples across layers and we may drop the subscript $m$. Now, let $p_1, \ldots, p_G$ be a sequence of i.i.d. uniform random variables, and let $F_G(t)$ be their empirical CDF.
Then, it suffices to show that

\[
\mathbb{E} \left[ \sup_{t \in [0,1]} \frac{F_G(t)}{G^{-1/2} + t} \right] \leq 1 + \exp \left( G^{-1/2} + \frac{1}{2} G^{-1} \right) 0.42 + eG^{-1/4}.
\]

This is the content of Lemma 4, which is proved in the appendix.

**Power of multilayer methods**

By construction, the multilayer algorithms we propose are at most as powerful as their single-layer versions. For our purposes, groups of variables function as inferential units, and not as prior information used to boost power (e.g. as in [31]), although there is no reason groups cannot serve both functions within our framework. So while our methods are designed to provide more FDR guarantees, it is relevant to evaluate the cost in terms of power of these additional guarantees.

Consider controlling FDR for individual variables and for groups, compared to just controlling FDR for individual variables. When adding a group FDR guarantee, power loss depends on the group signal strength, the power of group statistics, and the desired group FDR control level \(q_{grp}\). Power will decrease to the extent that the signal strength at the group layer is weaker than the signal strength at the individual layer. Assuming for simplicity that non-null variables have comparable effect sizes, group signal is weak when saturation is low (recall from the introduction that saturation is the average number of non-null variables per non-null group). Also, if the sizes of the groups vary, then group signal will be weaker if the non-null hypotheses are buried inside very large groups. Even if group signal is not too weak, the power of multilayer procedures will depend on the way group statistics are chosen. In particular, power will be better if Simes (or Simes-like) statistics are used if groups have a small number of strong signals, and if Fisher (or Fisher-like) statistics are used in the case of weak distributed effects. Finally, it is clear that lowering \(q_{grp}\) will lower power.

As a final note, all of the multilayer methods discussed so far have a feature that might unnecessarily limit their power. This feature is the definition of \(\hat{V}_m = \hat{V}_m(t_m)\) in terms of only the threshold \(t_m\). Since the selection set \(S(t)\) is defined with respect to the \(M\)-dimensional vector of thresholds \(t\), a definition of \(\hat{V}_m\) depending on this entire vector would be a better estimate of \(V_m(S(t))\). In some situations, the procedures proposed might overestimate \(V_m(S(t))\) and thus pay a price in power. For a graphical illustration of this phenomenon, we revisit Figure 2. Note that we are using the number of points in the entire shaded left half-plane to estimate the number of false positives in just the shaded upper right quadrant. Unfortunately, this issue is not very easy to resolve. One challenge is that if we allow \(\hat{V}_m\) to depend on the entire vector \(t\), then the definition of \(t^*\) would be complicated by the fact that the upper right-hand corner property would no longer hold. Another challenge is that the dependencies between feature statistics across layers make it hard to come up with a better and yet tractable estimate of \(V_m(S(t))\). Despite this flaw, the multilayer knockoff filter (and the p-filter) enjoys very similar power to its single-layer counterpart, as we shall see in the next section.

## 3 Simulations

We rely on simulations to explore the FDR control and the power of the multilayer knockoff filter and the generalized p-filter across a range of scenarios, designed to capture the variability described in the previous section. All code is available at [http://web.stanford.edu/~ekatsevi/software.html](http://web.stanford.edu/~ekatsevi/software.html).
3.1 Performance of the multilayer knockoff filter

Matching the adopted expository strategy, we simulate data from the linear model with $N > n$. This also allows us to calculate p-values for the null hypotheses $\beta_j = 0$ and plug these into BH and p-filter: these two methods, in addition to KF, serve as points of comparison to MKF.

Simulation setup. We generate the data from a linear model

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where $X \in \mathbb{R}^{N \times n}$, with $N = 4500$ observations on $n = 2000$ predictors. $X$ is constructed by sampling each row independently from $N(0, \Sigma)$, where $(\Sigma_{ij})_{ij} = \rho^{|i-j|}$ is the covariance matrix of an AR(1) process with correlation $\rho$. There are $M = 2$ layers: one comprising individual variables and one with $G = 200$ groups, each of 10 variables. The vector $\beta$ has 75 nonzero entries. The indices of the non-null elements are determined by firstly selecting $k$ groups uniformly at random, and then choosing, again uniformly at random, 75 elements of these $k$ groups. Here, $k$ controls the strength of the group signal; we considered three values: low saturation ($k = 40$), medium saturation ($k = 20$), and high saturation ($k = 10$). We generated these three sparsity patterns of $\beta$ once and fixed them across all simulations; see Figure 3. In all cases, the nonzero entries of $\beta$ are all equal, with a magnitude that satisfies

$$\text{SNR} = \frac{\|X\beta\|^2}{N}$$

for a given SNR value. For each saturation setting, we vary $\rho \in \{0.1, 0.3, \ldots, 0.9\}$ while keeping SNR fixed at 0.5, and vary SNR $\in \{0, 0.1, \ldots, 0.5\}$ while keeping $\rho$ fixed at 0.3. Across all experiments, we used nominal FDR levels $q_{\text{ind}} = q_{\text{grp}} = 0.2$.

This choice of simulation parameters captures some of the features of genetic data: the AR(1) process for the rows of $X$ is a first approximation for the local spatial correlations of genotypes and the signal is relatively sparse, as we would expect in GWAS. A notable difference between our simulations and common genetic data is the scale: a typical GWAS involves $n \approx 1,000,000$ variables.

Methods compared. We compare the following four methods on this simulated data:
(a) KF+ with lasso signed max statistics, targeting $q_{\text{ind}}$.

(b) MKF+ with Simes-like lasso signed max statistics, derived from the penalty $\ell_g^m(u) = \|u\|_1$, targeting $q_{\text{ind}}$ and $q_{\text{grp}}$. We find that this choice has better power across a range of saturation levels than the group lasso based construction of [28].

(c) Benjamini Hochberg procedure (BH) on the p-values based on t-statistics from linear regression, targeting $q_{\text{ind}}$.

(d) p-filter (PF) on the same set of p-values, targeting $q_{\text{ind}}$ and $q_{\text{grp}}$.

Note that the first two methods are knockoff-based and the last two are p-value based, and that methods (a) and (c) target only the FDR at the individual layer while methods (b) and (d) target the FDR at both layers.

Results. Figure 4 illustrates our findings. First, consider the FDR of the four methods. The multilayer knockoff filter achieves multilayer FDR control across all parameter settings: the constant $c_{\text{kn}} = 1.93$ from our proof does not appear to play a significant role in practice. The p-filter also has multilayer FDR control, even though the PRDS assumption is not satisfied by the two-sided p-values we are using. On the other hand, the knockoff filter and BH both violate FDR control at the group layer as the saturation level and power increase.

We also note that both the multilayer knockoff filter and regular knockoff filter have, on average, a realized FDP that is smaller than the target FDR. This is partly because we use the “knockoffs+” version of these methods, which is conservative when the power is low. In addition, we find that the multilayer knockoff filter is conservative at the individual layer even in high-power situations if the saturation is high. This is a consequence of our construction of $\hat{V}_m$: an estimate of the number of false discoveries that, as we have discussed, tends to be larger than needed. We see similar behavior for the p-filter, since it has an analogous construction of $\hat{V}_m$.

Next, we compare the power of the four methods. As expected, the power of all methods improves with SNR and degrades with $\rho$. We find that the knockoff-based approaches consistently outperform the p-value based approaches, with higher power despite having lower FDRs and the gap widening as saturation increases. This power difference is likely caused by the ability of the knockoff-based approaches to leverage the sparsity of the problem to construct more powerful test statistics for each variable. Finally, we compare the power of the multilayer knockoff filter to that of the regular knockoff filter: in most cases, the multilayer knockoff filter loses little or no power, despite providing an additional FDR guarantee. This holds even in the low saturation setting, where the groups are not very informative for the signal.

3.2 Performance of the generalized p-filter

We explore the possible advantages of the generalized p-filter in a setup when signals are expected to be weak and common within non-null groups, so one would want to define group p-values via the Fisher test instead of the Simes test. We consider two partitions of interest, both with groups of size 10 (thus no singleton layer). A situation similar to this might arise when scientists are interested in determining which functional genomic segments are associated with a trait. There exist several algorithms to split the genome into functional blocks (e.g. ChromHMM [32]), and segments in each of these can be partially overlapping.
Figure 4: Simulation results. From left to right, the saturation regime changes. The top panel varies signal-to-noise ratio while fixing $\rho = 0.3$. The bottom panel varies $\rho$ while fixing SNR = 0.5.

**Simulation setup.** We simulated $n = 2000$ hypotheses, with $M = 2$ layers. Each layer had 200 groups, each of size 10. The groups in the second layer were offset from the those in the first layer by 5. Hence, the groups for layer one are $\{1, \ldots, 10\}, \{11, \ldots, 20\}, \ldots$, while the groups for layer two are $\{6, \ldots, 15\}, \{16, \ldots, 25\}, \ldots$. The nonzero entries of $\beta$ are $\{1, \ldots, 200\}$. Hence, this is a “fully saturated” configuration. We generate $X_j \sim \mathcal{N}(\mu_j, 1)$, where $\mu_j = 0$ for null $j$ and $\mu_j = \mu$ for non-null $j$. We then derive two-sided p-values based on the z test. In this context, we define $\text{SNR} = \|\mu\|^2 / n$. The SNR varied in the range $\{0, 0.1, \ldots, 0.5\}$ and we targeted $q_{\text{ind}} = q_{\text{grp}} = 0.2$. 
Methods compared.
(a) The regular p-filter, which is based on the Simes test.
(b) The generalized p-filter with p-values based on the Fisher test.

Results. Figure 5 shows how both versions of the generalized p-filter have multilayer FDR control, with the Fisher version being more conservative. As with the multilayer knockoff filter, we see that the extra theoretical multiplicative factor is not necessary (at least in this simulation). In this case, Fisher has substantially higher power than Simes due to the weak distributed effects in each group.

Figure 5: Performance of the generalized p-filter: comparison of Fisher and Simes combination rules.

4 Case study: variants and genes influencing HDL

To understand which genes are involved in determining cholesterol levels and which genetic variants have an impact on its value, [33] carried out exome resequencing of 17 genetic loci identified in previous GWAS studies as linked to metabolic traits in about 5000 subjects. The original analysis is based on marginal tests for common variants and burden tests for the cumulative effect of the rare variants in a gene [34, 35]. Furthermore, to account for linkage disequilibrium and estimate the number of variants that influence cholesterol in each location, [33] uses model selection based on BIC. The original analysis, therefore, reports findings at the gene level and variant level, but these findings derive from multiple separate analyses and lack coordination. The data were re-analyzed in a Bayesian framework [36], focusing on the availability of multiple phenotypes. Here we deploy MKF to leverage multiple regression models, obtaining a coherent set of findings at both the variant and gene level, with approximate multilayer FDR control.
Data. The resequencing targeted the coding portion of 17 genetic loci, distributed over 10 chromosomes and containing 79 genes. We preprocessed the data as in [36]: we removed variants with minor allele counts below a threshold, and pruned the set of polymorphisms to assure that the empirical correlation between any pair is of at most 0.3 (this is a necessary step in multiple regression analysis to avoid collinearity [36, 13, 15]). After preprocessing, the data contained 5335 individuals and 768 variants. Since the study design was exome resequencing, every variant could be assigned to a gene. A special case is that of 18 SNPs that were typed in a previous study of these subjects and that were included in the final analysis as indicators of the original association signal: 12 of these are located in coding regions, but 6 are not.

While [33] studies the genetic basis of several metabolic traits, we focus our analysis here on HDL cholesterol: from the original measurement we regressed out the effects of sex, age, and the first five principal components of genomewide genotypes, representing population structure.

Methods compared. To focus on the effect of adding multilayer FDR guarantees (rather than on the consequences of different methods of analysis), we compare the results of the multilayer knockoff filter (MKF) and the knockoff filter (KF). The multilayer knockoff filter used a variant layer and a gene layer. We chose MKF and KF instead of MKF+ and KF+ for increased power, but otherwise used the same method settings as in Section 3. Each variant from the sequencing data and the 12 exonic GWAS SNPs were assigned to groups based on gene. The 6 intergenic GWAS SNPs are considered single members of 6 additional groups. Hence, our analysis has 85 (= 79 + 6) “genes” in total.

Results. Table 1 summarizes how many genes and SNPs each method discovers. KF has about twice as many discoveries at each layer, but how many of these are spurious? Unfortunately the identity of the variants truly associated with HDL is unknown, but we can get an approximation to the truth using the existing literature and online databases. At the variant level, this task is difficult because (1) linkage disequilibrium (i.e. correlations between nearby variants) makes the problem ill-posed and (2) rare variants, present in this sequencing dataset, are less well studied and cataloged. Instead, we focus on an annotation at the gene level. Comparing the two methods at the gene level is also meaningful because this is the layer at which the multilayer knockoff filter provides an extra FDR guarantee. See Appendix C for references supporting our annotations.

Table 2 shows the gene layer results: there are 5 true positive genes (ABCA1, CETP, GALNT2, LIPC, LPL) found by both methods, 1 false positive shared by both methods (PTPRJ), 1 true positive for KF that is missed by MKF (APOA5), and 4 false positives (NLRC5, SLC12A3, DYNC2LI1, SPI1) for KF that MKF correctly does not select. Hence, MKF reduced the number of false positives from 5 to 1 at the cost of 1 false negative.

Figure 6 shows a more detailed version of these association results, illustrating the signal at the variant level. Notably, the one extra false negative (APOA5) incurred by MKF just barely misses the cutoff for the gene layer. Aside from the extra false negative and the one false positive shared with KF, the additional horizontal cutoff induced by the need to control FDR at the gene level does a good job separating the genes associated with HDL from those that are not.
| Method | # SNPs found | # Genes found |
|--------|--------------|---------------|
| KF     | 23           | 11            |
| MKF    | 13           | 6             |

Table 1: Summary of association results on resequencing data.

| Gene        | Discovered by | Supported in literature |
|-------------|---------------|-------------------------|
| ABCA1       | KF, MKF       | yes                     |
| CETP        | KF, MKF       | yes                     |
| GALNT2      | KF, MKF       | yes                     |
| LIPC        | KF, MKF       | yes                     |
| LPL         | KF, MKF       | yes                     |
| PTPRJ       | KF, MKF       | no                      |
| APOA5       | KF            | yes                     |
| NLRC5       | KF            | no                      |
| SLC12A3     | KF            | no                      |
| DYNC2LI1    | KF            | no                      |
| SPI1        | KF            | no                      |

Table 2: Comparison of MKF and KF at the gene layer. False positives are highlighted in red.

5 Conclusions

With the multilayer knockoff filter, we have made a first step to equip model selection procedures with FDR guarantees for multiple types of reported discoveries, bridging results from the multi-resolution testing literature \cite{23, 24, 37, 27} with controlled selection methods \cite{14, 28, 15}. When tackling high dimensional variable selection, researchers have at their disposal a large bag of tricks based on regularized regression, with penalties that can reflect an array of sparse structures (see for example \cite{38, 39, 40}), corresponding to a multiplicity of possible resolutions for discoveries. While a number of these have been implemented in the context of genetic association studies \cite{18, 41}, their application has been hampered by the lack of inferential guarantees on the selection. It is our hope that the approach put forward with MKF will allow scientists to leverage these computationally attractive methods to obtain reproducible discoveries at multiple levels of granularity.

In the process of developing a framework for multilayer FDR control for variable selection, we have also generalized the p-filter multiple testing procedure. Our approach places no restrictions on the relations between the p-values used to test the hypotheses at different layers. By contrast, theoretical results for the p-filter rely heavily on the specific way in which p-values for individual hypotheses are aggregated to obtain p-values for groups. The constant $c_{pf}$ can be viewed as the price we pay for allowing these arbitrary dependencies. Nevertheless, simulations show that both $c_{pf}$ and the corresponding constant $c_{kn}$ for MKF appear to be inconsequential in practice.

When scientific hypotheses have a complex structure, even formulating inferential guarantees is nontrivial; the multilayer hypothesis testing approach proposed in \cite{27} and used in our work is one of several options. Approaches to testing hypotheses at multiple levels vary in two key features: the way the space of hypotheses is traversed and the way families to be tested are defined. We
Figure 6: Scatterplot of variant level and gene level knockoff statistics. Solid blue lines are the thresholds for MKF while the cyan broken line is the threshold for KF. Each dot corresponds to a variant, and variants selected by at least one of the methods are in color: dark green indicates selected variants that belong to genes that are true positives for both methods, light green a true positive found by KF but missed by MKF, red is for false positives of KF but true negatives of MKF, and magenta for the false positive shared by both methods. To facilitate comparison with Table 2, we indicate the names of the genes representing false positives or false negatives for at least one of the methods.

illustrate the different approaches using the two-layer setup considered in the introduction. If the hypotheses have a nested (i.e. tree) structure, then it is common to traverse them hierarchically: one starts by testing groups and then proceeds to test individual hypotheses within rejected groups. The procedures described in [23, 24, 26] follow this hierarchical approach. An alternative to hierarchical hypothesis traversal is to consider the multiple testing problem from the point of view of individual-level hypotheses: rejecting a set of individual-level hypotheses induces the rejections of the groups that contain them at each layer of interest. This is the approach taken by the p-filter [27]. By testing hypotheses only if their corresponding groups were rejected, hierarchical approaches have the advantage of a smaller multiplicity burden. On the other hand, defining selections at each layer via the individual-level hypotheses has the advantage that it applies equally well to non-hierarchical ways of grouping hypotheses. The second dichotomy in based on how one defines families to be tested: either each group is a family of its own, or each resolution is a family of its own. For instance, the former corresponds to SNPs being tested against other SNPs in the same gene, and the latter corresponds to testing all SNPs against each other as one family. The methods [24, 26] take the former approach,
and \[23, 27\] take the latter. Both choices can be meaningful, depending on the application. In this work, we define discoveries using individual-level hypotheses as this marries well with the multiple regression framework and does not limit us to nested groups. We treat each resolution (instead of each group) as a family because discoveries are often reported by type (e.g. as a list of SNPs or a list of genes), so FDR guarantees for each type are appropriate. These two choices align our testing framework with that of the p-filter.

Our approach relies on the construction of knockoff statistics for each group. For definiteness and to avoid confusion, from Section 1 onwards we have referred only to the construction of knockoff variables for the case of linear model with fixed design \[14\]. However, it is important to remark that the knockoff construction has recently been generalized to the high dimensional random design setting and to handle any conditional distribution \(y|X\), shifting distributional assumptions instead to the design matrix \(X\) \[15\]. These assumptions may be more appropriate in the context of association studies, where we have fairly accurate models of correlation patterns of genetic variants (e.g. \[42\]). These and any future advances in the construction of knockoff statistics can seamlessly be incorporated into the multilayer knockoff filter.

Finally, we remark that multi-task regression, the study of the impact of a set of predictor variables on multiple outcome variables, is a very promising area of application for the MKF. The multi-task regression problem is often reshaped into a larger single-task regression problem, in which the predictors have group structure based on which task they correspond to. For example, \[23\] takes this approach to multi-task regression alongside its development of the group knockoff filter. MKF can then provide a framework for FDR control in this settings, where group discoveries correspond to finding variables important for at least one of the outcomes, and individual discoveries correspond to the identification of variables important for a specific outcome. In the context of the linear model with \(N \leq n\) and independent errors, the MKF as described here provides the desired FDR guarantee. However, the general case is more challenging, and we are working on this direction and hope to report results in future work.

Acknowledgements

The authors are indebted to Emmanuel Candès and David Siegmund for help with theoretical aspects of this work. E.K. also thanks Subhabrata Sen for a helpful discussion. E.K. was supported by the Fannie and John Hertz Foundation and the National Defense Science and Engineering Graduate Fellowship. C.S. was supported by NIH grants HG006695 and HL113315.

References

[1] 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. Gudtmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, 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.
[2] 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 (Methodological)*, pp. 289–300, 1995.

[3] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[4] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[5] J. Friedman, T. Hastie, and R. Tibshirani, “A note on the group lasso and a sparse group lasso,” *arXiv preprint arXiv:1001.0736*, 2010.

[6] K. Knight and W. Fu, “Asymptotics for lasso-type estimators,” *Annals of Statistics*, pp. 1356–1378, 2000.

[7] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar, “A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers,” in *Advances in Neural Information Processing Systems*, pp. 1348–1356, 2009.

[8] A. Javanmard and A. Montanari, “Confidence intervals and hypothesis testing for high-dimensional regression,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[9] W. Fithian, D. Sun, and J. Taylor, “Optimal inference after model selection,” *ArXiv e-prints*, Oct. 2014.

[10] J. Taylor and R. J. Tibshirani, “Statistical learning and selective inference,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, pp. 7629–7634, Jun 2015.

[11] J. Markovic, L. Xia, and J. Taylor, “Adaptive p-values after cross-validation,” *arXiv preprint arXiv:1703.06559*, 2017.

[12] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, “SLOPE—adaptive variable selection via convex optimization,” *The Annals of Applied Statistics*, vol. 9, no. 3, p. 1103, 2015.

[13] D. Brzyski, C. Peterson, M. Bogdan, C. Sabatti, and P. Sobczyk, “geneSLOPE: genome-wide association study with SLOPE,” 2015.

[14] R. F. Barber and E. J. Candès, “Controlling the false discovery rate via knockoffs,” *The Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015.

[15] E. Candes, Y. Fan, L. Janson, and J. Lv, “Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection,” *arXiv preprint arXiv:1610.02351*, 2016.

[16] S. A. Santorico and A. E. Hendricks, “Progress in methods for rare variant association,” *BMC Genet.*, vol. 17 Suppl 2, p. 6, 2016.

[17] K. Wang, M. Li, and M. Bucan, “Pathway-based approaches for analysis of genomewide association studies,” *Am. J. Hum. Genet.*, vol. 81, pp. 1278–1283, Dec 2007.

[18] 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.

[19] Y. Benjamini and R. Heller, “False discovery rates for spatial signals,” *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1272–1281, 2007.
[20] R. A. Poldrack, “Region of interest analysis for fMRI,” *Social cognitive and affective neuroscience*, vol. 2, no. 1/4, p. 67, 2007.

[21] D. Siegmund, N. Zhang, and B. Yakir, “False discovery rate for scanning statistics,” *Biometrika*, vol. 98, no. 4, pp. 979–985, 2011.

[22] K. G. Ardlie, [...], 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.

[23] D. Yekutieli, “Hierarchical false discovery rate–controlling methodology,” *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 309–316, 2008.

[24] Y. Benjamini and M. Bogomolov, “Selective inference on multiple families of hypotheses,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 297–318, 2014.

[25] R. Heller, N. Chatterjee, A. Krieger, and J. Shi, “Post-selection inference following aggregate level hypothesis testing in large scale genomic data,” *bioRxiv*, p. 058404, 2016.

[26] M. Bogomolov, C. B. Peterson, Y. Benjamini, and C. Sabatti, “Testing hypotheses on a tree: new error rates and controlling strategies,” *arXiv preprint arXiv:1705.07529*, 2017.

[27] R. F. Barber and A. Ramdas, “The p-filter: Multilayer false discovery rate control for grouped hypotheses,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 2016.

[28] R. Dai and R. F. Barber, “The knockoff filter for FDR control in group-sparse and multitask regression,” *arXiv preprint arXiv:1602.03589*, 2016.

[29] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Annals of Statistics*, pp. 1165–1188, 2001.

[30] R. J. Simes, “An improved Bonferroni procedure for multiple tests of significance,” *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.

[31] A. Li and R. F. Barber, “Multiple testing with the structure adaptive Benjamini-Hochberg algorithm,” *arXiv preprint arXiv:1606.07926*, 2016.

[32] J. Ernst and M. Kellis, “ChromHMM: automating chromatin-state discovery and characterization,” *Nature methods*, vol. 9, no. 3, pp. 215–216, 2012.

[33] S. K. Service, T. M. Teslovich, C. Fuchsberger, V. Ramensky, P. Yajnik, D. C. Koboldt, D. E. Larson, Q. Zhang, L. Lin, R. Welch, L. Ding, M. D. McLellan, M. O’Laughlin, C. Fronick, L. L. Fulton, V. Magrini, A. Swift, P. Elliott, M.-R. Jarvelin, M. Kaakinen, M. I. McCarthy, L. Peltonen, A. Pouta, L. L. Bonnycastle, F. S. Collins, N. Narisu, H. M. Stringham, J. Tuomilehto, S. Ripatti, R. S. Fulton, C. Sabatti, R. K. Wilson, M. Boehnke, and N. B. Freimer, “Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci,” *PLoS Genet*, vol. 10, no. 1, p. e1004147, 2014.

[34] C. Li and H. Li, “Network-constrained regularization and variable selection for analysis of genomic data,” *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[35] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, “Rare-variant association testing for sequencing data with the sequence kernel association test,” *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.
[36] L. Stell and C. Sabatti, “Genetic variant selection: Learning across traits and sites,” *Genetics*, vol. 202, pp. 439–455, Feb 2016.

[37] 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.

[38] S. Kim and E. P. Xing, “Statistical estimation of correlated genome associations to a quantitative trait network,” *PLoS Genet*, vol. 5, no. 8, p. e1000587, 2009.

[39] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar, “A dirty model for multi-task learning,” in *Advances in Neural Information Processing Systems*, pp. 964–972, 2010.

[40] N. Rao, C. Cox, R. Nowak, and T. T. Rogers, “Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis,” in *Advances in neural information processing systems*, pp. 2202–2210, 2013.

[41] E. P. Xing, R. E. Curtis, G. Schoenherr, S. Lee, J. Yin, K. Puniyani, W. Wu, and P. Kinnaird, “Gwas in a box: statistical and visual analytics of structured associations via GenAMap,” *PLoS one*, vol. 9, no. 6, p. e97524, 2014.

[42] P. Scheet and M. Stephens, “A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase,” *The American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644, 2006.

[43] S. Karlin and H. E. Taylor, *A second course in stochastic processes*. Elsevier, 1981.

[44] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.

[45] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, *et al.*, “The NHGRI GWAS catalog, a curated resource of SNP-trait associations,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[46] X. Lu, J. Huang, Z. Mo, J. He, L. Wang, X. Yang, A. Tan, S. Chen, J. Chen, C. C. Gu, *et al.*, “Genetic susceptibility to lipid levels and lipid change over time and risk of incident hyperlipidemia in Chinese populations,” *Circulation: Cardiovascular Genetics*, vol. 9, no. 1, pp. 37–44, 2016.

[47] P. M. Ridker, G. Paré, A. N. Parker, R. Y. Zee, J. P. Miletich, and D. I. Chasman, “Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction,” *Circulation: Cardiovascular Genetics*, vol. 2, no. 1, pp. 26–33, 2009.
Appendices

A  FDR control for multilayer knockoff filter

**Lemma 2.** Let $S_k = \sum_{i=1}^k X_i$, where $X_i \sim \text{Ber}(1/2)$. Then,

$$a_{kn} = \mathbb{E} \left[ \max_{k \geq 0} \frac{S_k}{1 + k - S_k} \right] \leq 1.93.$$

First we prove the following lemma about the hitting time of a linear boundary by the random walk $S_k$.

**Lemma 3.** Fix constants $c_1 > 1/2$ and $c_2$. We have

$$\mathbb{P}[S_k \geq c_1 k + c_2 \text{ for some } k \geq 0] \leq \exp(-\Theta(c_1) \cdot c_2),$$

where $\Theta(c_1)$ is the positive root of

$$\exp(\theta(1 - c_1)) + \exp(-\theta c_1) = 2.$$

**Proof.** For fixed $\theta > 0$, define the process

$$Z_k = \exp(\theta(S_k - c_1 k - c_2)).$$

Then,

$$\mathbb{P}[S_k \geq c_1 k + c_2 \text{ for some } k \geq 0] = \mathbb{P} \left[ \sup_{k \geq 0} Z_k \geq 1 \right].$$

The goal is to choose $\theta$ so that $Z_k$ becomes a martingale. For this, it is sufficient that

$$1 = \mathbb{E}[Z_k/Z_{k-1}] = \mathbb{E}[\exp(\theta(X_k - c_1))] = \frac{1}{2} \exp(\theta(1 - c_1)) + \frac{1}{2} \exp(-\theta c_1),$$

which is satisfied for $\theta = \Theta(c_1)$. Applying the maximal inequality for martingales, we obtain

$$\mathbb{P} \left[ \sup_{k \geq 0} Z_k \geq 1 \right] \leq \mathbb{E}[Z_0] = \exp(-\Theta(c_1) \cdot c_2). \quad (12)$$

**Proof of Lemma 2.** We write

$$a_{kn} = \mathbb{E} \left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \right] = \int_0^\infty \mathbb{P} \left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \right] dt = 1 + \int_1^\infty \mathbb{P} \left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \right] dt. \quad (13)$$

The last equality holds because $S_k/(1 + k - S_k) \overset{a.s.}{\to} 1$ by the law of large numbers. Note that $S_k/(1 + k - S_k) \geq t$ if and only if $S_k \geq \frac{t}{1+t} k + \frac{t}{1+t}$. Hence, the probability in the integrand can be bounded using Lemma 3. However, the resulting bound would not be very tight (we would get a bound of 2.1). Indeed, the inequality in (12) is somewhat loose, which is related to the fact that the
random walk $S_k$ sometimes overshoots the linear boundary. This overshoot can be fairly substantial at the beginning of the random walk.

To get a better estimate, we carry out the following first-step analysis. Fix $k_0 \geq 0$.

$$
\mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \right] = \mathbb{E} \left[ \mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \mid X_1, \ldots, X_{k_0} \right] \right]
$$

$$
= 2^{-k_0} \sum_{x_1, \ldots, x_{k_0}} \mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \mid X_1 = x_1, \ldots, X_{k_0} = x_{k_0} \right].
$$

Hence, we have a sum over all $2^{k_0}$ possible first $k_0$ steps of the path. Define

$$
R_{k_0}(x) = \max_{k \leq k_0} \frac{\sum_{i \leq k} x_i}{1 + k - \sum_{i \leq k} x_i}; \quad P_{k_0}(x) = \sum_{i \leq k_0} x_i.
$$

These are the maximum ratio and partial sum after $k_0$ steps, respectively. Then, we have

$$
\mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k}{1 + k - S_k} \geq t \right]
= 2^{-k_0} \sum_{x_1, \ldots, x_{k_0}} \left\{ I(R_{k_0}(x) \geq t) + \mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k + P_{k_0}(x)}{1 + k_0 + k - (S_k + P_{k_0}(x))} \geq t \mid R_{k_0}(x) < t \right] \right\}.
$$

Integrating the above expression and using [13], we get

$$
a_{kn} = 2^{-k_0} \sum_{x_1, \ldots, x_{k_0}} \left\{ \max(R_{k_0}(x), 1) + \int_{\max(R_{k_0}(x), 1)}^{\infty} \mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k + P_{k_0}(x)}{1 + k_0 + k - (S_k + P_{k_0}(x))} \geq t \right] dt \right\}.
$$

Note that

$$
\frac{S_k + P_{k_0}(x)}{1 + k_0 + k - (S_k + P_{k_0}(x))} \geq t \iff S_k \geq t + k(1 + k_0) - P_{k_0}(x).
$$

Hence, by Lemma 3, we have

$$
\mathbb{P}\left[ \sup_{k \geq 0} \frac{S_k + P_{k_0}(x)}{1 + k_0 + k - (S_k + P_{k_0}(x))} \geq t \right] \leq \exp\left( -\Theta \left( \frac{t}{1 + t} \right) \cdot \left( \frac{t}{1 + t} (1 + k_0) - P_{k_0}(x) \right) \right).
$$

Hence, we have

$$
a_{kn} \leq 2^{-k_0} \sum_{x_1, \ldots, x_{k_0}} \left\{ \max(R_{k_0}(x), 1) + \int_{\max(R_{k_0}(x), 1)}^{\infty} \exp\left( -\Theta \left( \frac{t}{1 + t} \right) \cdot \left( \frac{t}{1 + t} (1 + k_0) - P_{k_0}(x) \right) \right) dt \right\}.
$$

Recall that $\Theta(t/(1 + t))$ is defined as the root of the nonlinear equation

$$
\exp(\theta/(1 + t)) + \exp(-\theta t/(1 + t)) = 2.
$$
Define 
\[ \theta_t = \frac{1}{1 + t} \Theta \left( \frac{t}{1 + t} \right). \]

Then, \( \theta_t \) satisfies
\[ e^{\theta_t} + e^{-\theta_t} = 2. \] (14)

In terms of \( \theta_t \), we get
\[ a_{kn} \leq 2^{-k_0} \sum_{x_1, \ldots, x_{k_0}} \left\{ \max(R_{k_0}(x), 1) + \int_{\max(R_{k_0}(x), 1)}^{\infty} \exp \left( (1 + k_0 - P_{k_0}(x)) \log(2 - e^{\theta_t}) + \theta_t P_{k_0}(x) \right) dt \right\}. \] (15)

Solving (14) for \( t \) in terms of \( \theta_t \), we get
\[ t = \frac{-\log(2 - e^{\theta_t})}{\theta_t}. \]

At this point, we change variables in (15). Taking a derivative, the transformation is
\[ dt = \frac{e^{\theta_t} \theta_t + \log(2 - e^{\theta_t})}{\theta_t^2} d\theta_t. \]

This yields
\[ \int_{\max(R_{k_0}(x), 1)}^{\infty} \exp \left( (1 + k_0 - P_{k_0}(x)) \log(2 - e^{\theta_t}) + \theta_t P_{k_0}(x) \right) dt \]
\[ = \int_{\max(R_{k_0}(x), 1)}^{\log 2} \exp \left( (1 + k_0 - P_{k_0}(x)) \log(2 - e^{\theta_t}) + \theta_t P_{k_0}(x) \right) \frac{e^{\theta_t} \theta_t + \log(2 - e^{\theta_t})}{\theta_t^2} d\theta_t \] (16)
\[ = \int_{\max(R_{k_0}(x), 1)}^{\log 2} \exp \left( (k_0 - P_{k_0}(x)) \log(2 - e^{\theta_t}) + \theta_t P_{k_0}(x) \right) \frac{e^{\theta_t} \theta_t + (2 - e^{\theta_t}) \log(2 - e^{\theta_t})}{\theta_t^2} d\theta_t \]

Putting together (15) and (16), we can compute the desired bound using only numerical integration over bounded intervals.

Figure 7 shows the bounds we obtain from this approach for \( k_0 = 1, \ldots, 20 \). For instance, the bound obtained using \( k_0 = 20 \) is 1.922, which proves the lemma.

\[ \square \]

**B FDR control for generalized p-filter**

**Lemma 4.** Let \( F_n(t) \) be the empirical CDF of \( n \) independent uniform random variables. Then
\[ a_{pd}(n) = \mathbb{E} \left[ \sup_{t \in [0,1]} \frac{F_n(t)}{n^{-1} + t} \right] \leq 1 + \exp \left( n^{-1/2} + \frac{1}{2} n^{-1} \right) 0.42 + en^{-1/4}. \] (17)
Figure 7: Upper bounds on $a_{kn}$.

Note that $N_t = nF_n(t) \sim \mu$, where $\mu$ is the distribution of a Poisson process with rate $n$, conditioned on $N_1 = n$. Hence, we may equivalently bound the quantity

$$\mathbb{E}_\mu \left[ \sup_{t \in [0,1]} \frac{N_t}{1 + nt} \right].$$

We first state two auxiliary lemmas, and show how the main result follows.

**Lemma 5.** Let $t_0 = n^{-1/2}$. Then,

$$\mu \left[ \sup_{t \in [0,t_0]} \frac{N_t}{1 + nt} \geq x \right] \leq \exp \left( n^{-1/2} + \frac{1}{2} n^{-1} \right) \exp(-\gamma_x x),$$

where $\gamma_x$ is the positive root of

$$e^\gamma = 1 + \gamma x.$$

**Lemma 6.** Let $t_0 = n^{-1/2}$. Then,

$$\mu \left[ \sup_{t \in [t_0,1]} \frac{N_t}{1 + nt} \geq x \right] \leq \exp(1 - n^{1/4} (x - 1)).$$

**Proof of Lemma 4.** Taking $t_0 = n^{-1/2}$ and using the results of Lemmas 5 and 6, we find

$$a_{pl}(n) = \mathbb{E}_\mu \left[ \sup_{t \in [0,1]} \frac{N_t}{1 + nt} \right] = \int_0^\infty \mu \left[ \sup_{t \in [0,1]} \frac{N_t}{1 + nt} \geq x \right] dx \leq 1 + \int_1^\infty \mu \left[ \sup_{t \in [t_0,1]} \frac{N_t}{1 + nt} \geq x \right] dx + \int_1^\infty \mu \left[ \sup_{t \in [t_0,1]} \frac{N_t}{1 + nt} \geq x \right] dx$$

$$\leq 1 + \int_1^\infty \exp \left( n^{-1/2} + \frac{1}{2} n^{-1} \right) \exp(-\gamma_x x) dx + \int_1^\infty \exp(1 - n^{1/4} (x - 1)) dx$$

$$= 1 + \exp \left( n^{-1/2} + \frac{1}{2} n^{-1} \right) 0.42 + e n^{-1/4}.$$
Here, we use the fact (obtained by numerical integration), that \(\int_{1}^{\infty} \exp(-\gamma x)dx = 0.42\). 

Now, we prove the two auxiliary lemmas.

**Proof of Lemma 3.** Let \(\nu\) be the unconditional distribution of a Poisson process with rate \(n\). The idea is that \(\nu\) is easier to work with than \(\mu\) due to the martingale properties of the unconditional Poisson process, and \(\nu\) and \(\mu\) are close on \([0, t_0]\). Hence, we change measure from \(\mu\) to \(\nu\):

\[
\mu\left[ \sup_{t \in [0, t_0]} \frac{N_t}{1 + nt} \geq x \right] = \mathbb{E}_\nu\left[ \frac{d\mu}{d\nu} I\left( \sup_{t \in [0, t_0]} \frac{N_t}{1 + nt} \geq x \right) \right]. \tag{18}
\]

Our first claim is that 

\[
\text{ess sup}_\nu \frac{d\mu}{d\nu} \leq C_n = \exp\left(n^{-1/2} + \frac{1}{2} n^{-1}\right), \tag{19}
\]

where \(\mu\) and \(\nu\) are viewed as measures on the space of stochastic processes on \([0, t_0]\) (that are right continuous with left limits). Let \(\mathcal{F}_t = \sigma(N_s: s \leq t)\). It suffices to show that

\[
\mu[A] \leq C_n \nu[A] \quad \text{for all } A \in \mathcal{F}_{t_0}. \tag{20}
\]

To show this, let

\[
\mathcal{C} = \{ A \in \mathcal{F}_{t_0} : \mu[A] \leq C_n \nu[A] \},
\]

and let

\[
\mathcal{S} = \{ \{ N_{t_1} \in B_1, \ldots, N_{t_k} \in B_k \} : 0 \leq t_1 < \ldots < t_k \leq t_0, k \geq 1, B_1, \ldots, B_k \subset \mathbb{N} \}.
\]

be the collection of finite-dimensional rectangles, which generates \(\mathcal{F}_{t_0}\). Let \(\mathcal{A}\) be the algebra generated by \(\mathcal{S}\), i.e. the collection of finite disjoint unions of sets in \(\mathcal{S}\). We claim that it suffices to show that \(\mathcal{S} \subset \mathcal{C}\). Indeed, if \(\mathcal{S} \subset \mathcal{C}\), then by additivity it is clear that \(\mathcal{A} \subset \mathcal{C}\). But since \(\mathcal{A}\) is an algebra and \(\mathcal{C}\) is a monotone class, the monotone class theorem implies that \(\mathcal{F}_{t_0} = \sigma(\mathcal{A}) \subset \mathcal{C}\), from which it follows that \(\mathcal{C} = \mathcal{F}_{t_0}\), which is the statement (20). Finally, to show that \(\mathcal{S} \subset \mathcal{C}\), by countable additivity it suffices to show that \(\{ N_{t_1} = y_1, \ldots, N_{t_k} = y_k \} \in \mathcal{C}\) for all \(y_1, \ldots, y_k \in \mathbb{N}\). Note that

\[
\frac{\mu[N_{t_1} = y_1, \ldots, N_{t_k} = y_k]}{\nu[N_{t_1} = y_1, \ldots, N_{t_k} = y_k]} = \frac{\mu[N_{t_k} = y_k] \mu[N_{t_1} = y_1, \ldots, N_{t_{k-1}} = y_{k-1} | N_{t_k} = y_k]}{\nu[N_{t_k} = y_k] \nu[N_{t_1} = y_1, \ldots, N_{t_{k-1}} = y_{k-1} | N_{t_k} = y_k]} = \frac{\mu[N_{t_k} = y_k]}{\nu[N_{t_k} = y_k]},
\]

where the conditional distributions in the numerator and denominator cancel due to the Markov property of Poisson processes. Hence, it suffices to show that for all \(t \leq t_0\) and all \(y \in \mathbb{N}\),

\[
\frac{\mu[N_t = y]}{\nu[N_t = y]} \leq C_n.
\]

31
Note that $\mu[N_t = y] = 0$ for $y > n$. For $y \leq n$, we write

$$
\log \left( \frac{\mu[N_t = y]}{\nu[N_t = y]} \right) = \log \left( \frac{(n/y)^y(1 - t)^{n-y}}{e^{-nt(n/y)^y}} \right)
$$

$$
= \log \left( \frac{n!}{n^y(n-y)!} (1 - t)^{n-y} e^{nt} \right)
$$

$$
= \log \left( \prod_{i=1}^{y-1} \left( 1 - \frac{i}{n} \right) \right) + (n-y) \log(1-t) + nt
$$

$$
= \sum_{i=1}^{y-1} \log \left( 1 - \frac{i}{n} \right) + (n-y) \log(1-t) + nt
$$

$$
\leq - \sum_{i=1}^{y-1} \frac{i}{n} - (n-y) \left( t + \frac{1}{2} t^2 \right) + nt
$$

$$
= -\frac{1}{2n} y(y-1) - (n-y) \left( t + \frac{1}{2} t^2 \right) + nt
$$

$$
= -\frac{1}{2n} y^2 + \left( \frac{1}{2n} + t + \frac{1}{2} t^2 \right) y - \frac{1}{2} nt^2.
$$

The value $y_* = \frac{1}{2} + nt + \frac{1}{2} nt^2$ maximizes the above expression, so we get

$$
\log \left( \frac{\mu[N_t = y]}{\nu[N_t = y]} \right) \leq -\frac{1}{2n} y_*^2 + \left( \frac{1}{2n} + t + \frac{1}{2} t^2 \right) y_* - \frac{1}{2} nt^2
$$

$$
= \frac{n}{2} \left( \frac{1}{2n} + t + \frac{1}{2} t^2 \right)^2 - \frac{1}{2} nt^2
$$

$$
= \frac{1}{8n} + \frac{nt^4}{8} + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3 n}{2}
$$

$$
\leq \frac{1}{8n} + \frac{nt^4}{8} + \frac{t_0}{2} + \frac{t_0^2}{4} + \frac{t_0^3 n}{2}
$$

$$
= \frac{1}{8n} + \frac{1}{8n} + \frac{1}{2} n^{-1/2} + \frac{1}{4n} + \frac{1}{2} n^{-1/2} = n^{-1/2} + \frac{1}{2} n^{-1}.
$$

This proves statement (19). Together with (18), this implies that

$$
\mu \left[ \sup_{t \in [0,t_0]} \frac{N_t}{1 + nt} \geq x \right] \leq \exp \left( n^{-1/2} + \frac{1}{2} n^{-1} \right) \nu \left[ \sup_{t \in [0,t_0]} \frac{N_t}{1 + nt} \geq x \right].
$$

(21)

Hence, we can now compute the probability under the measure $\nu$. Let

$$
Z_t = \exp(\gamma_x N_t - nt(e^{\gamma_x} - 1)),$$

where $\gamma_x$ is defined as in the statement of Lemma 5. Under $\nu$, this is an exponential martingale
associated with the Poisson process $N_t$. We have

$$\nu \left[ \sup_{t \in [0, t_0]} \frac{N_t}{1 + nt} \geq x \right] \leq \nu \left[ \sup_{t \geq 0} \frac{N_t}{1 + nt} \geq x \right]$$

$$= \nu \left[ \sup_{t \geq 0} (N_t - (1 + nt)x) \geq 0 \right]$$

$$= \nu \left[ \sup_{t \geq 0} (\exp(\gamma_x N_t - \gamma_x(1 + nt)x)) \geq 1 \right]$$

$$= \nu \left[ \sup_{t \geq 0} (\exp(\gamma_x N_t - nt(e^{\gamma_x} - 1) - \gamma_x x)) \geq 1 \right]$$

$$= \nu \left[ \sup_{t \geq 0} Z_t \geq \exp(\gamma_x x) \right]$$

$$\leq \exp(-\gamma_x x) E[Z_0] = \exp(-\gamma_x x),$$

where the last inequality follows from the maximal inequality for martingales. Putting this together with (21) completes the proof of the lemma.

Proof of Lemma 6. For any $\alpha > 0$, we have

$$\mu \left[ \sup_{t \in [t_0, 1]} \frac{N_t}{1 + nt} \geq x \right] \leq \mu \left[ \sup_{t \in [t_0, 1]} \frac{N_t}{nt} \geq x \right] \leq \mu \left[ \sup_{t \in [t_0, 1]} \exp \left( \frac{\alpha N_t}{nt} \right) \geq \exp(\alpha x) \right].$$

Now, since $N_t/nt$ is a backwards martingale under $\mu$, it follows that $\exp(\alpha N_t/nt)$ is a backwards submartingale, so from the maximal inequality we have

$$\mu \left[ \sup_{t \in [t_0, 1]} \exp \left( \frac{\alpha N_t}{nt} \right) \geq \exp(\alpha x) \right] \leq \exp(-\alpha x) E_{\mu} \left[ \exp \left( \frac{\alpha N_{t_0}}{nt_0} \right) \right]$$

$$= \exp(-\alpha x) ((1 - t_0) + t_0 \exp(\alpha/nt_0))^n$$

$$= \exp(-\alpha x) (1 + t_0(\exp(\alpha/nt_0) - 1))^n.$$

At this stage, let us take $\alpha = n^{1/4}$ and $t_0 = n^{-1/2}$. Then, the bound becomes

$$\exp(-n^{1/4} x) \left( 1 + n^{-1/2}(\exp(n^{-1/4}) - 1) \right)^n$$

Note that $\exp(n^{-1/4}) - 1 \leq n^{-1/4} + \frac{1}{2} n^{-1/2} \exp(n^{-1/4}) \leq n^{-1/4} + n^{-1/2}$ for $n \geq 5$. Hence,

$$\mu \left[ \sup_{t \in [t_0, 1]} \frac{N_t}{1 + nt} \geq x \right] \leq \exp(-n^{1/4} x) \left( 1 + n^{-1/2}(n^{-1/4} + n^{-1/2}) \right)^n$$

$$\leq \exp(-n^{1/4} x)(n^{1/4} + 1)$$

$$= \exp(1 - n^{1/4}(x - 1)).$$

This completes the proof of the lemma.
Remark B.1. Using a formula from Karlin and Taylor [43] for the probability of an empirical process hitting a linear boundary, we can actually get the following exact expression for any $x > 1$:

$$\mu \left[ \sup_{t \in [0,1]} \frac{N_t}{1 + nt} \geq x \right] = \sum_{i=0}^{n-x} \binom{n}{i} \left( \frac{n-i-x}{nx} \right)^{n-i} \left( 1 - \frac{n-i-x}{nx} \right)^{i-1} \left( 1 + \frac{1}{n} - \frac{1}{x} \right).$$

Hence, we find that

$$a_{kn} = \mathbb{E}_\mu \left[ \sup_{t \in [0,1]} \frac{N_t}{1 + nt} \geq x \right] = 1 + \int_1^\infty \left\{ \sum_{i=0}^{n-x} \binom{n}{i} \left( \frac{n-i-x}{nx} \right)^{n-i} \left( 1 - \frac{n-i-x}{nx} \right)^{i-1} \left( 1 + \frac{1}{n} - \frac{1}{x} \right) \right\} dx. \quad (22)$$

While this quantity is hard to work with theoretically, we may compute it. Figure 8 shows this quantity, along with the limiting value (equal to 1.42) from Lemma 4. This plot strongly suggests that $a_{pf}(n)$ is an increasing function of $n$, which leads us to conjecture that $a_{pf}(n) \leq \limsup_{n \to \infty} a_{pf}(n) \leq 1.42$ for all $n$.

Figure 8: Points represent the exact expression (22), and the horizontal line represents the limiting bound from Lemma 4.

C Evidence for gene annotations

In this appendix, we discuss the evidence in literature and databases for our annotations of association or no association with HDL cholesterol. We use the Online Mendelian Inheritance in Man (OMIM) database [44] and NHGRI/EBI GWAS catalog [45].

Likely Associated with HDL.

- The genes ABCA1, CETP, GALNT2, LIPC, LPL are all well-known to be associated with HDL. These associations are documented in OMIM and in the GWAS catalog.
• We also conclude that APOA5 is likely associated with HDL, although the primary trait for which this gene is known is triglycerides. This association with HDL is documented in the GWAS catalog and in [46].

Likely not associated with HDL.

• The genes PTPRJ, DYNC2LI1, and SPI1 show no evidence in literature or databases of association with HDL.

• The gene NLRC5 has no reported association with HDL in OMIM or the GWAS catalog. There is one paper (Charlesworth et al 2009) that predicts an association using a gene-based analysis. However, this gene is very near CETP. Moreover, this paper states that “Interestingly, the list also prioritizes a number of genes of little-known function, such as NLRC5 (NLR family CARD domain containing 5)…, which would not be selected by any form of candidate gene approach.”

• SLC12A3 is a gene in the CETP locus that does contain a GWAS hit. However, a paper reporting this association [17] states that SLC12A3 is not an obvious candidate for association with HDL and hypothesizes that the effect is “mediated by long range linkage disequilibrium to a causal variant nearer the CETP gene.”