Reflection Knockoffs via Householder Reflection: Applications in Proteomics and Genetic Fine Mapping

Yongtao Guan and Daniel Levy
National Heart, Lung, and Blood Institute

May 29, 2025

Abstract

We present a novel knockoff construction method, and demonstrate its superior performance in two applications: identifying proteomic signatures of age and genetic fine mapping. Both applications involve datasets of highly correlated features, but they differ in the abundance of driver associations. Our primary contribution is the invention of the reflection knockoff, which is constructed from mirror images – obtained via Householder reflection – of the original features. The reflection knockoffs substantially outperform Model-X knockoffs in feature selection, particularly when features are highly correlated. Our secondary contribution is a simple method to aggregate multiple sets of identically distributed knockoff statistics to improve the consistency of knockoff filters. In the study of proteomic signatures of age, single feature tests showed overly abundant proteomic association with age. Knockoff filters using reflection knockoffs and aggregation, however, revealed that a majority of these associations are hitchhikers instead of drivers. When applied to genetic fine mapping, knockoff filters using reflection knockoffs and aggregation outperform a state-of-the-art method. We discuss a potentially exciting application of reflection knockoffs: sharing genetic data without raising concerns about privacy and regulatory violations.

Keywords: Knockoff filter; reflection knockoff (ReKo); proteomics; fine mapping

1 Introduction

Scientific research often deals with datasets containing strongly correlated features. One example is proteomics data, in which high correlation can be attributed to several factors including the inherent correlations among multiple proteins, protein-protein interactions, and the imperfect specificity of the assay. Another example is genetic fine mapping, where single nucleotide polymorphisms (SNPs) in a locus may be highly correlated due to linkage disequilibrium. For these types of highly correlated data, identifying features that are truly associated with a phenotype of interest is a critical but challenging task. One key difficulty is controlling for false discovery, i.e., "hitchhikers" that show association through correlation with genuinely associated "driver" features. The knockoff filter (Barber and Candès, 2015; Candès et al., 2018) is a statistical framework designed to address this challenge by providing a robust way to identify true associations while controlling the false discovery rate (FDR).

Knockoff filters consist of four components: 1) constructing knockoff features; 2) conducting association analysis by jointly analyzing the original features and their knockoffs. The joint analysis can be conducted with either glmnet (Friedman et al., 2010), which implements penalized regression methods, including LASSO (Tibshirani, 1996) and Elastic-Net (Zou and Hastie, 2005), or fastBVSR (Zhou and Guan, 2019), which implements an iterative complex factorization algorithm to fit Bayesian variable selection regression (Guan and Stephens, 2011), or with other machine learning algorithms, including support vector machines (Cortes and Vapnik, 1995) and random forests (Breiman, 2001); 3) computing *knockoff statistics* W: a vector of differences in the feature importance measure between the original features and their knockoffs, where the feature importance measure includes the magnitude of regression coefficients (Beta) and posterior inclusion probability (PIP); and 4) determining the knockoff threshold from W at a nominal FDR level (e.g.,10%) to select features.

The most important component of the knockoff filter is constructing knockoffs $\tilde{\mathbf{X}}$ conditioning on the original features $\mathbf{X}$. This is also the focus of our work. For the knockoffs to have the desired property of controlling FDR, $\mathbf{X}$ and $\tilde{\mathbf{X}}$ must follow:

$$\operatorname{cov}\left([\mathbf{X}, \tilde{\mathbf{X}}]\right) = \begin{bmatrix} \Sigma & \Sigma - \mathbf{S} \\ \Sigma - \mathbf{S} & \Sigma \end{bmatrix} \tag{1}$$

where $\mathbf{S} \geq \mathbf{0}$ is a diagonal matrix that satisfies

$$2\mathbf{S} - \mathbf{S}\Sigma^{-1}\mathbf{S} \geq \mathbf{0}. \tag{2}$$

$\mathbf{S}$ controls the degree of uncorrelated-ness between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$, and larger $\mathbf{S}$ values bring more power to the knockoff filter. Another key condition for $\tilde{\mathbf{X}}$ is that it must satisfy that, conditional on $\mathbf{X}$, $\tilde{\mathbf{X}}$ is independent of phenotypes of interest, which is automatically satisfied if the construction of $\tilde{\mathbf{X}}$ is agnostic to the phenotypes.

A critical component in generating both Fixed-X and Model-X knockoffs is estimating $\mathbf{S}$.

Three methods have been developed to estimate $\mathbf{S}$: 1) The equi-correlated method where $\mathbf{S}$ is a scaling of an identity matrix; 2) The SDP method where $\mathbf{S}$ is obtained by optimizing a natural objective function subject to the inherent constraint using a semidefinite program (SDP) (Barber and Candès, 2015); and 3) A two-stage method where the first stage produces an SDP-estimate of $\mathbf{S}$, but using an approximated feature covariance matrix, and the second stage linearly scales the SDP-estimated $\mathbf{S}$ (Candès et al., 2018). The SDP estimated $\mathbf{S}$ is more powerful than $\mathbf{S}$ estimated using the other two methods.

In this paper, we introduce a novel method to estimate $\mathbf{S}$ by first constructing mirror images $\mathbf{Y}$ of the original feature matrix $\mathbf{X}$ via Householder reflection. The difference, $\mathbf{Z} = \frac{1}{2}(\mathbf{X} - \mathbf{Y})$, is orthogonal to $\mathbf{X}$ and thus $\mathbf{T} = \text{cov}(\mathbf{X}, \mathbf{Z})$ is a diagonal matrix, where $\text{cov}(\mathbf{X}, \mathbf{Z})$ denotes a sample covariance matrix with $\mathbf{T}_{jk} = \text{cov}(\mathbf{X}_j, \mathbf{Z}_k)$ denoting sample covariance between two column vectors $\mathbf{X}_j$ and $\mathbf{Z}_k$. We define $\mathbf{S} = \alpha\mathbf{T}$, and estimate the scalar $\hat{\alpha}$ by examining the largest eigenvalues of the scaled $\mathbf{Z}$. Such $\mathbf{S} = \hat{\alpha}\mathbf{T}$ leads to more powerful knockoffs. We call the knockoffs constructed via Householder reflection *reflection knockoffs* (ReKo). Since its optimality comes from the leading eigenvalue and its computation does not require optimization, ReKo is numerically stable.

Due to the inherent randomness in constructing knockoffs, the knockoff filter may produce inconsistent results when applied to different copies of knockoffs. A second contribution of the paper is a simple method we propose to aggregate multiple sets of identically distributed knockoff statistics: pooling multiple knockoff statistics to determine the knockoff threshold and selecting features by comparing the average knockoff statistics with the knockoff threshold. This aggregation method produces more consistent results in real data analysis.

We apply reflection knockoffs and aggregate multiple sets of knockoff statistics to study two outstanding problems in genetics and proteomics, where the complex patterns of high correlation among features pose a challenge for existing methods. In the study of the proteomics signature of age, proteomic features appear to be overly abundant in association with age. Our study using reflection knockoffs revealed that the majority of single-feature associations observed between proteomic features and age are hitchhikers. In genetic fine mapping linkage, disequilibrium between nearby genetic variants makes it difficult to distinguish driver signals from hitchhikers. We demonstrate that reflection knockoffs significantly improve the power of knockoff filters for identifying driver variants.

2 Results

2.1 Reflection Knockoffs

A key observation leading to our invention of reflection knockoffs is that the reflection of a feature $\mathbf{X}_j$ over the hyperplane spanned by other features $\mathbf{X}_{,-j}$ produces a $\mathbf{Y}_j$ that is an almost perfect knockoff for that feature. Such $\mathbf{Y}_j$ can be computed using Householder reflection. Indeed, $\text{cov}(\mathbf{X}_j, \mathbf{Y}_k) = \text{cov}(\mathbf{X}_j, \mathbf{X}_k)$ for $j \neq k$ (Lemma 1 in Materials and Methods), where

cov denotes sample covariance. As a candidate for knockoff of $\mathbf{X}$, a linear combination of $\mathbf{X}$ and $\mathbf{Y}$ has the correct sample covariances between different column vectors, but not correct sample variances. The variance, however, can be fixed by adding a term that is uncorrelated with columns of $\mathbf{X}$ and $\mathbf{Y}$, and this is the essence of a formulation provided by Barber and Candès (2015).

The Householder reflection regresses $\mathbf{X}_j$ on $\mathbf{X}_{,-j}$ to obtain a fitted value $\hat{\mathbf{X}}_j$, then computes the reflection $\mathbf{Y}_j = 2\hat{\mathbf{X}}_j - \mathbf{X}_j$. The residual of regression is $\mathbf{Z}_j = \frac{1}{2}(\mathbf{X}_j - \mathbf{Y}_j)$. Therefore $\mathbf{Z}_j$ is orthogonal to $\hat{\mathbf{X}}_j$ and uncorrelated with all other features. Thus, the sample covariance matrix $\mathbf{T} = \mathrm{cov}(\mathbf{X}, \mathbf{Z})$, where $\mathbf{T}_{jk} = \mathrm{cov}(\mathbf{X}_j, \mathbf{Z}_k)$, is a diagonal matrix with $\mathbf{T} \geq \mathbf{0}$ (Materials and Methods). We define $\mathbf{S} = \alpha\mathbf{T}$ and seek the largest α such that $\mathbf{S}$ satisfies the constraint in Equation (2). The optimal $\hat{\alpha} = 2/\Lambda_1$, where Λ_1 is the leading eigenvalue of the sample covariance matrix $\mathbf{C} = \mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$, where $\mathbf{C}_{jk}$ is the sample covariance between j-th and k-th column of matrix $\mathbf{Z}\mathbf{T}^{-1/2}$. The knockoffs can be computed following Equation (8) in Materials and Methods. Notably, constructing reflection knockoffs requires no optimization and is numerically stable.

For the knockoff to have power, we need $\mathbf{S} = \hat{\alpha}\mathbf{T}$ to be bounded away from 0. Let's examine $\hat{\alpha}$ and $\mathbf{T}$ in turn. Numerically $\mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$ is similar to an identity matrix, with diagonal entries equal to 1 and small off-diagonal entries. By Lemma 1 we have $\mathrm{var}(\mathbf{Z}_j) = \mathbf{T}_{jj}$ where var denote sample variance. Since $\mathbf{T} > \mathbf{0}$ is diagonal, we have $\mathrm{diag}(\mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})) = \mathrm{diag}(\mathbf{I}_p)$. The off-diagonal element $\mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})_{jk}$ is the sample correlation between residuals $\mathbf{Z}_j$ and $\mathbf{Z}_k$. Assuming $\mathrm{cov}(\mathbf{X}_j, \mathbf{X}_k) = r$, and that both are uncorrelated with other features, a straightforward computation yields $\mathrm{cov}(\mathbf{Z}_j, \mathbf{Z}_k) = r(r^2 - 1)$. Therefore both large (close to 1 or -1) and small (close to 0) correlations produce small $\mathrm{cov}(\mathbf{Z}_j, \mathbf{Z}_k)$. This calculation also applies when r is the conditional correlation, which can be obtained from the sample precision matrix $\mathrm{cov}(\mathbf{X})^{-1}$. When features are correlated, the conditional correlation between two features tends to be small. Then by the Gershgorin circle theorem, the eigenvalues satisfy $|\Lambda_j - 1| \leq \sum_{k \neq j} |\mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})_{jk}|$. So as long as the off diagonal entries are small and linearly diminishing with the dimension of $\mathrm{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$, the leading eigenvalue Λ_1 will be small, and $\hat{\alpha}$ will be bounded away from 0.

By $\mathbf{T}_{jj} = \mathrm{var}(\mathbf{Z}_j)$, for the reflection knockoff to have power the size of residual $\mathbf{Z}_j$ needs to be bounded away from 0. When features are highly correlated, however, $\mathrm{var}(\mathbf{Z}_j)$ tends to be small, as $\mathbf{X}_j$ is well explained by $\mathbf{X}_{,-j}$. Introducing shrinkage priors via ghost samples (Materials and Methods) proves to be an effective way to keep $\mathrm{var}(\mathbf{Z}_j)$ away from 0. This is a rare example where both sides of the bias and variance tradeoff are beneficial – we have more consistent estimates of $\mathbf{Z}$ while keeping the size of $\mathbf{Z}$ away from 0. Intuitively, when the mirror image is too close to the original feature, the priors bend the mirror so that the image appears farther away in the mirror without introducing obvious distortion. The ghost samples also enable reflection knockoffs to be constructed when the number of features is larger than the number of samples, significantly broadening the scope of applications.

2.2 Combined knockoff filter

We designated a knockoff filter by how the knockoffs were constructed, how the associations were analyzed (either using statistical methods or software packages that implement them), and what feature importance measure was used. For example, Model-X⊗glmnet⊗Beta means a knockoff filter that constructs Model-X knockoffs, uses glmnet (a R package) to conduct joint association analysis, and chooses the regression coefficient as the importance measure; ReKo⊗fastBVSR⊗PIP means a knockoff filter that constructs reflection knockoffs, performs association analysis using fastBVSR, and uses PIP (posterior inclusion probability) as the importance measure.

When comparing two knockoffs with everything else being equal, we designate knockoff filters in a manner such as ⊗glment⊗Beta. Thus, by saying we are comparing Model-X and ReKo based on ⊗glment⊗Beta, we mean comparing power and type I errors between the knockoff filters Model-X⊗glmnet⊗Beta and ReKo⊗glment⊗Beta. We may omit the importance measure from the notation of knockoff filters when it is clear from the context. For example, ⊗glment means ⊗glment⊗Beta, and ⊗fastBVSR means ⊗fastBVSR⊗PIP.

A *combined knockoff filter* ⊗combined combines results from knockoff filters ⊗glmnet and ⊗fastBVSR, such that a feature is selected as positive by the combined filter if and only if it is selected by both ⊗glmnet and ⊗fastBVSR. The combined knockoff filter provides a means to combine results from Frequentist and Bayesian inferences.

2.3 Aggregating multiple knockoff statistics

Due to random component in their construction, knockoffs are inherently random. Consequently, inferences based on a single knockoff statistics can be inconsistent. Generating multiple knockoffs and combining multiple knockoff statistics can effectively increase consistency and power of a knockoff filter (Gimenez and Zou, 2019; He et al., 2021a). The existing methods that combine multiple knockoffs assume that feature importance measures can be obtained for each copy of knockoffs independently of the original features (He et al., 2021a), while for our target applications, it is important to jointly fit a model that contains both origin features and a copy of their knockoffs. We developed a new method that combines multiple sets of knockoff statistics (Material and Method).

Our aggregating method concatenates multiple sets of knockoff statistics to identify a knockoff threshold at a nominal FDR level, and calls positives by comparing average knockoff statistics of each feature with the knockoff threshold. Intuitively, concatenation increases the sample size used to estimate thresholds and provides a more stable empirical estimate of the distribution of null and non-null features. Concatenation reduces variance in the estimation of threshold-relevant ratios, making the threshold estimation more stable while preserving the original FDR control. Under the null hypothesis, if m-th knockoff statistic of j-th feature $W_j^{(m)}$ is symmetric around zero for each m, their average $\bar{W}_j$ remains symmetric around zero. Consequently, $\bar{W}_j$ retains the crucial null property necessary for FDR control. Averaging also

reduces the variance of knockoff statistics for each feature, providing more stable and reproducible feature selection while maintaining theoretical FDR guarantees. Figure 1 provides an example to demonstrate that aggregating multiple copies of knockoffs increases power. For the numerical analysis conducted in this paper, all results were obtained by aggregating 10 knockoff statistics, unless otherwise noted.

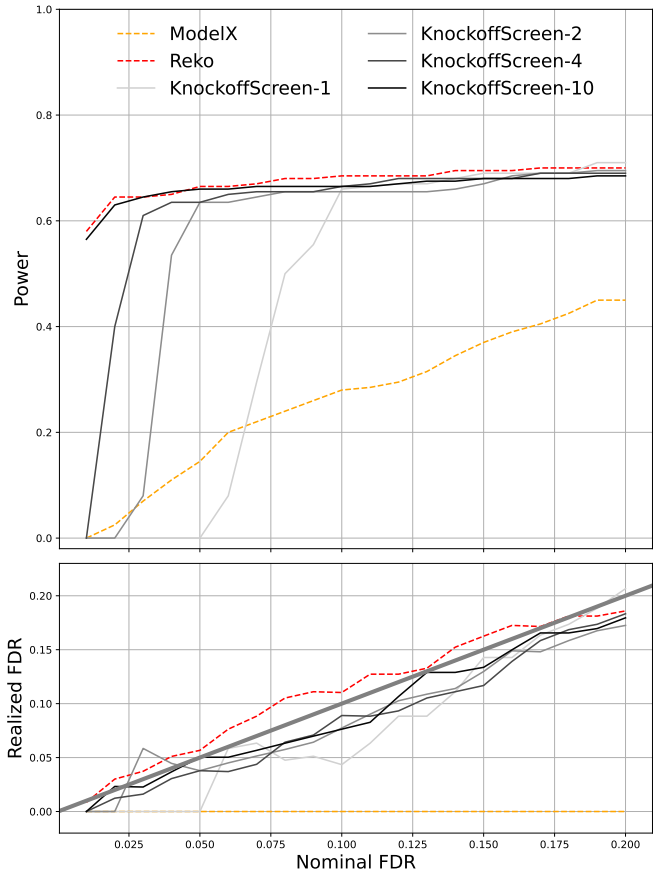

Figure 1: Aggregating knockoff statistics improve power. Top panel is nominal FDR vs power; Bottom panel is nominal FDR vs realized FDR. The knockoff filter used to call positives is ⊗glmnet for all. Results of Model-X and ReKo were obtained by aggregating 10 copies of knockoff statistics. For KnockoffScreen, we produced results by using a single copy of knockoff statistics and by aggregating $2, 4$, and 10 copies of knockoff statistics. The results are based on simulated features with $n = 2000$ and $p = 800$. More details of simulation can be found in Supplementary Section 2.

2.4 Comparing ReKo with other knockoffs

We first compared ReKo with Fixed-X and Model-X knockoffs using synthetic features (Supplementary Tab. S2). Building upon knowledge gained, we then focus on comparing ReKo with Model-X and KnockoffScreen (He et al., 2021a) using synthetic features of various simulation parameters. These include high and low content of association signals (Supplementary Fig. S2 and S3.), different levels of feature correlations (Supplementary Fig. S4), and non-

normal feature distributions (Supplementary Fig. S5). With synthetic features, we observed that ReKo is on par or slightly better than KnockoffScreen, and both are noticeably outperform Model-X. Since synthetic feature may not capture intricate correlation in the real datasets, we use real features for comparison.

The first real dataset is a subset of 1459 proteomic features from 9914 UK Biobank samples. The largest correlation between different features is 0.98. We simulated phenotypes using 200 randomly selected features as drivers. Their effect sizes were independently drawn from the standard normal distribution, but rescaled so that the PVE, or Proportion of Variance Explained, was 0.90. The large PVE used here was motivated by the strong signals observed in real phenotypes. Power and realized FDR were computed by pooling results from 100 replicates. For each replicate, we aggregated knockoff statistics from 10 copies of knockoffs (independently generated for each replicate) to call true and false positives at nominal FDR levels ranging from 0.01 to 0.20.

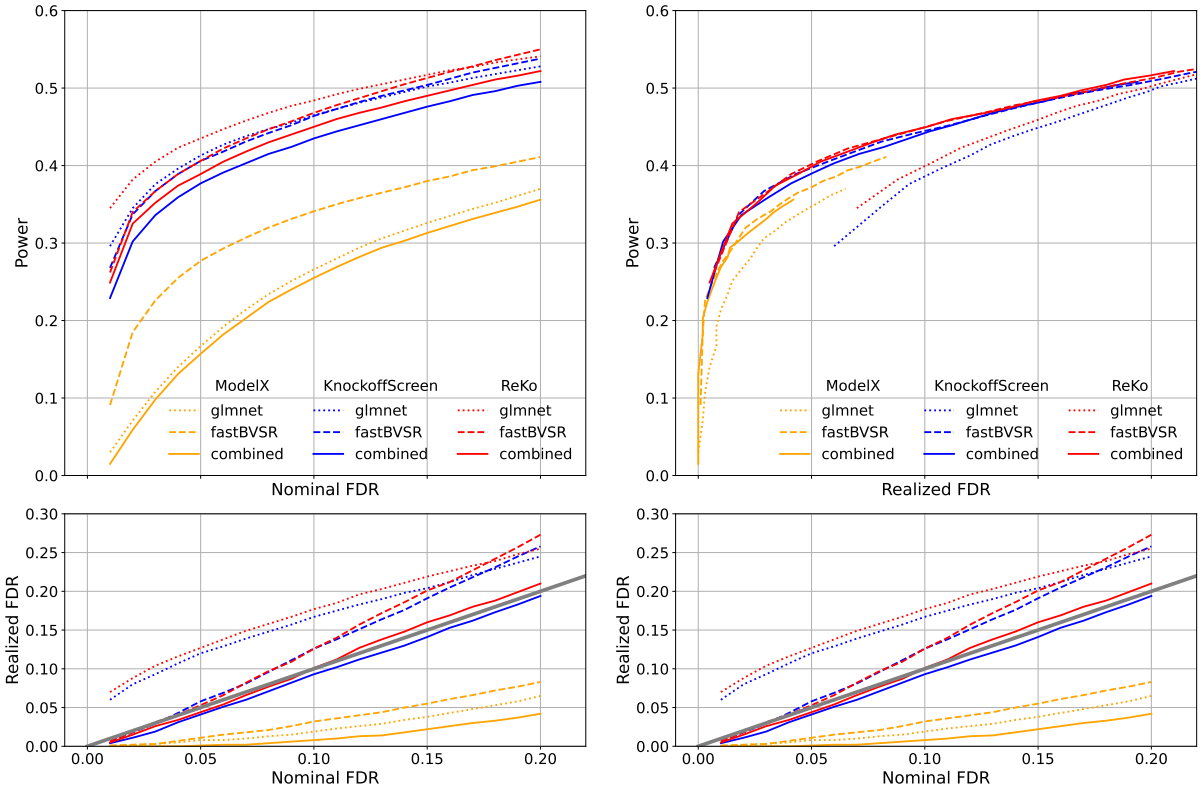

Figure 2: Power and realized FDR comparison between different knockoffs. The features are proteomics data from UK Biobank. . Left panel: the top plot is *nominal* FDR vs. power. Right panel: the top plot is *realized* FDR vs. power. The bottom plots of two panels are identical. Different line styles represent different knockoff filters and different colors represent different knockoff methods. The power and realized FDR were computed from 100 replicates, and each replicate aggregates 10 knockoff statistics.

Figure 2 presents the results, and we make the following observations: 1) Model-X appears to be overly conservative on the realized FDR, and its power suffers consequently. 2)

Both ReKo and KnockoffScreen showed inflated realized FDR for knockoff filters ⊗glmnet and ⊗fastBVSR. With ⊗combined, however, both ReKo and KnockoffScreen have their realized FDR under control. 3) ReKo performs on par with KnockoffScreen, the slightly better power observed is likely due to ReKo's slightly higher realized FDR . 4) Knockoff filter ⊗fastBVSR appear to be more powerful than ⊗glmnet. This is particularly true for Model-X, whose realized FDRs are under control for both knockoff filters. This echoes our prior studies (Guan and Stephens, 2011; Qi et al., 2018) that demonstrated advantage of Bayesian variable selection over panelized regression.

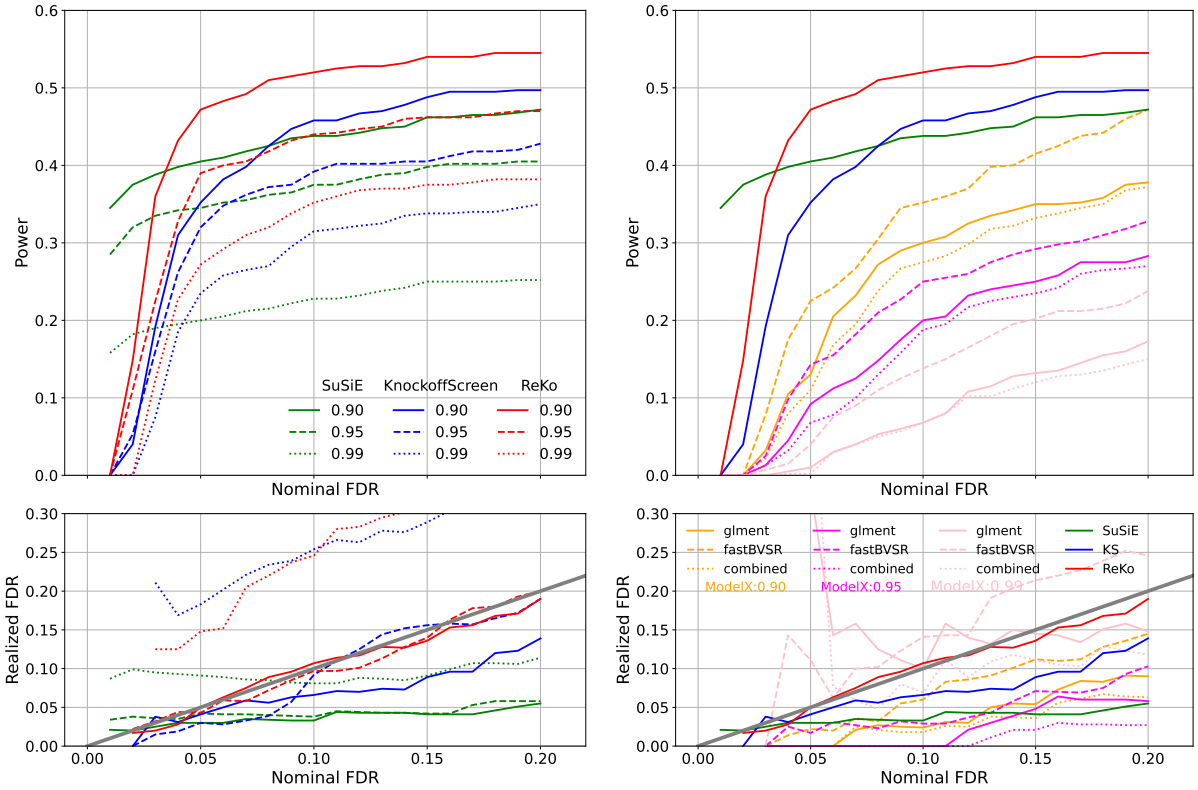

Figure 3: Left panel: Comparisons between SuSiE, KnockoffScreen⊗combined, and ReKo⊗combined for different correlation threshold with features were extracted from genomic region *HLA-DQB1*. Results of SuSiE in green; Results of KnockoffScreen in blue; Results of ReKo in red. Different line styles represent thresholds used to trim features. Right panel: Comparison of Model-X⊗glment (solid line), Model-X ⊗fastBVSR (dashed line), and Model-X⊗combined (dotted line) for different correlation threshold (orange for correlation threshold of 0.90, magenta for 0.95, and pink for 0.99) using the same features extracted from *HLA-DQB1*. Results for SuSiE, KnockoffScreen⊗combined, and ReKo⊗combined at the threshold of 0.90 were reproduced on the right panel for comparison. The power and realized FDR were computed from 50 replicates, and each replicate aggregates 10 knockoff statistics.

The second dataset of real features consists of genotypes. And our goal is to investigate how high correlation affects performance of knockoff filters. From a trio-phased dataset (Guan and Levy, 2024a) we extracted features of paternal and maternal alleles from genomic region

HLA-DQB1, which is known for its strong and complex pattern of linkage disequilibrium. We extracted three sets of features each having a different threshold for highest correlation among features: 0.90, 0.95, and 0.99. This is achieved by progressively thinning features in such a way that if a pair of features has a correlation that is higher than a threshold, we randomly remove one feature. To simulate phenotypes we randomly select 8 causal features, simulate each effect size with standard normal but rescale the effect sizes so that the PVE is 0.20. For each simulated phenotype, we independently generated knockoffs using Model-X, KnockoffScreen and ReKo.

Since this simulation is designed to illustrate the application of knockoff filters to study fine mapping, we compared results with SuSiE (Wang et al., 2020), a Bayesian method for fine mapping. We used its PIP outputs to call positives for SuSiE: at a nominal FDR α, a feature is called positive if its PIP is greater than $1 - \alpha$.

Figure 3 left panel summarizes the results for KnockoffScreen⊗combined, ReKo⊗combined, and SuSiE. We make the following observations: 1) At correlation thresholds 0.90 and 0.95, both ReKo⊗combined and KnockoffScreen⊗combined appear to have well-controlled type I error. 2) ReKo appears to have more power than KnockoffScreen, and at less stringent nominal FDR, ReKo appears to have more power than SuSiE. 3) At correlation 0.99 both ReKo⊗combined and Knockoff⊗combined show inflated realized FDR. 4) As the correlation threshold increases, the power of all methods decreases. 5) SuSiE has almost flat realized FDR vs. nominal FDR, suggesting that the calibration of its PIP remains challenging in real data analysis, an observation also made by others (**?**). But its power increases with nominal FDR, and at a more stringent nominal FDR, SuSiE is more powerful than both ReKo and KnockoffScreen.

Figure 3 right panel compares Model-X under different knockoff filters using the same datasets. The first interesting observation is that at correlation 0.99, Model-X shows inflated realized FDR, a phenomenon not seen with proteomic features (Figure 2) and synthetic features (Supplementary Fig. S3 and Fig. S4). The second observation is that Model-X ⊗fastBVSR is more powerful than Model-X⊗glment and ⊗combined, an observation we made with proteomic features (Figure 2). We therefore use knockoff filters Model-X⊗fastBVSR, KnockoffScreen⊗combined, and ReKo⊗combined for real data analysis.

2.5 Proteomic association with age

Proteomic signatures of age are an important focus of proteomics and aging studies (Tanaka et al., 2018; Lehallier et al., 2019; Tanaka et al., 2020; Sun et al., 2023). In a recent UK Biobank study using the Olink Explore platform, $1,944$ proteins out of $2,923$ analyzed were significantly associated with age (Sun et al., 2023). Since protein features are highly correlated, a natural question is how many of these features are drivers and how many are hitchhikers. A driver association remains significant when conditioned on other associated features, while a hitchhiker does not. Thus this question can only be addressed by jointly analyzing all

features and cannot be answered by single feature tests. Knockoff filter is a framework to select (driver) features while keeping FDR under control. For the knockoff filter to work, however, it is important to construct knockoffs with high power, and the ReKo is designed for this requirement.

Using a subset of proteomic features from UK Biobank (Sun et al., 2023), we investigated how many proteomic features are drivers in their association with age. We first computed the realized kinship matrix for $14,752$ samples based on genotype data using Kindred (Guan and Levy, 2024c). Then we regressed out fixed effects (sex and BMI) and the random effect, whose covariance matrix is a linear scaling of the kinship, from age (the phenotype) and each of the $1,459$ proteomic features using IDUL (Guan and Levy, 2024b). The residuals of age and the proteomic features were used for further analysis. We first examined the marginal association. At a p-value threshold of 3×10^{-5}, set by Bonferroni correction, there were 902 out of $1,459$ (or 62%) proteomic features significantly associated with age. This proportion aligns with a previous report of $1,944$ out of $2,923$ (or 67%) (Sun et al., 2023).

We applied three knockoff filters to call positives at nominal FDR levels of 0.05 and 0.10. These knockoff filters include Model-X⊗glmnet, KnockoffScreen⊗combined, and ReKo ⊗combined. For each knockoff filter, we aggregated 10 sets of knockoff statistics to call positives (Materials and Methods). (The counts of positives called with a single set of knockoff statistics can be found in Supplementary Tab. S3.) Table 1 summarizes the results. At a nominal FDR of 0.05 Model-X⊗fastBVSR called 256 positives (or 16.9% of total number of features), KnockoffScreen⊗combined called 306 (or 21.0%), ReKo⊗combined called 309 (or 21.2%). At a nominal FDR of 0.10 Model-X⊗fastBVSR called 308 (or 19.9%), KnockoffScree ⊗combined called 361 (or 24.7%), ReKo⊗combined called 371 (or 25.4%).

Level	MX	KS	ReKo	Overlap
0.05	256	306	309	291 (90%)
0.10	308	361	371	345 (89%)

Table 1: Counts of significant association for proteomic data. Column MX contains results from Model-X⊗fastBVSR; Column KS contains results from KnockoffScreen ⊗combined; Column ReKo contains results from ReKo⊗combined. Denote S_K and S_R positive calls by KnockoffScreen⊗combined and ReKo⊗combined respectively, Column Overlap contains counts of intersect between S_K and S_R, and the ratios in the parenthesis are Jaccard similarity coefficient.

At nominal FDR level of 0.05 the number of overlapping positives between KnockoffScreen and ReKo is 291, the number of positives called by either KnockoffScreen or ReKo is 324, and the ratio between counts of intersection and union, known as Jaccard similarity coefficient, is 90%. At nominal FDR level of 0.10 the counts of intersection and union between positives called by KnockoffScreen and ReKo are 345 and 387, and the Jaccard similarity coefficient is 89%. The high Jaccard similarity coefficients signal consistency of positive calls between KnockoffScreen and ReKo. We thus conclude that the majority of proteomic features that

show significant association with age in a single feature test are hitchhikers rather than drivers.

2.6 Genetic fine mapping

Genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) analysis are effective at identifying genomic regions associated with phenotypes. Due to linkage disequilibrium, however, pinpointing the exact driver variants within these regions is challenging. Fine mapping methods (Hormozdiari et al., 2014; Benner et al., 2016; Wang et al., 2020) aim to distinguish the driver variants from hitchhiker variants whose association with the phenotype diminishes once conditioned on driver variants. Among current fine mapping methods SuSiE (Wang et al., 2020) is an elegant Bayesian regression approach that models multiple driver variants within genomic regions. Here we demonstrate that knockoff filters can be applied to tackle fine mapping.

In a recent study, we discovered abundant genes harboring eQTL (eGene) that have parent-of-origin effect (POE) (Guan and Levy, 2024a). When a eGene harbors both genotype eQTL and POE eQTL, one may wonder whether POE eQTL are driver signals. Fine mapping can help distinguish drivers from hitchhikers. A challenge in this analysis, however, is that collinearity is introduced into features when genotypes are separated into paternal and maternal alleles. We therefore removed genotypes and only use paternal and maternal alleles in fine mapping. Removing genotypes still allows us to distinguish between POE eQTL and genotype eQTL, because paternal and maternal alleles tend to have similar effects in the absence of POE.

We selected five example eGenes from the POE eQTL study to perform fine mapping. The selections include *PPIEL* and *GJB6* that harbor both genotype eQTL and POE eQTL, *CCR9* that harbors exclusively paternal eQTL, *COA8* that harbors exclusively maternal eQTL, and *NECAB3* that harbors opposing eQTL, whose paternal and maternal alleles having their effect sizes in opposite directions. For each eGene, we regressed out fixed effects including age, sex, BMI, and cell type composition, and random effects that account for relatedness using IDUL (Guan and Levy, 2024b). A consecutive genomic region in cis to the eGenes was visually identified by examining the eQTL peak for each selected eGene. Features in the genomic region were then trimmed such that the largest correlations among the remaining features did not exceed 0.90. If a pair of features displayed a correlation larger than 0.90, we kept the one with stronger marginal association. The trimming was conducted using genotypes, and in our analysis paternal and maternal alleles always present in pairs.

We performed fine mapping with the five selected eGene using four methods: three knockoff filters Model-X$\otimes$fastBVSR, KnockoffScreen$\otimes$combined, and ReKo$\otimes$combined, and SuSiE. Table 2 summarizes the results and from which we make the following observations: 1) *PPIEL* is known to have both genotype eQTL and paternal eQTL. All four methods identified two paternal alleles as drivers. 2) *GJB6* is also known to have both genotype eQTL and maternal

eGene	eQTL	MX	KS	ReKo	SuSiE	Type
PPIEL	rs79473113:C:A:1:39554348:1	1.00	1.00	1.00	1.00	1111
	rs7520271:A:C:1:39579041:1	0.80	0.82	0.87	0.97	1111
GJB6	rs12875121:T:C:13:20478719:1	0.95	0.97	0.98	0.99	1111
	rs12875121:T:C:13:20478719:0	0.76	0.78	0.79	0.89	1110
	rs9506450:G:C:13:20239478:0	-0.15	-0.27	0.61	0.44	0010
	rs9579911:C:G:13:20637437:0	0.23	0.26	0.38	0.20	0010
CCR9	rs9814578:T:G:3:45821775:1	0.85	0.93	0.98	0.98	1111
COA8	rs10143623:G:A:14:103720539:0	0.66	0.58	0.85	0.87	1110
NECAB3	rs2626556:A:G:20:33768509:0	0.79	0.80	0.88	0.95	1111

Table 2: Fine mapping of selected eGenes. MX: Model-X; KS: KnockoffScreen. Column eQTL contain variants information in the format of rs:ref:alt:chr:pos:pm where pm can be 1 for paternal alleles, and 0 for maternal alleles. Columns MX, KS, and ReKo contain average knockoff statistics from knockoff filter ⊗fastBVSR, aggregated from 10 replicates of knockoffs respectively. Column SuSiE contains PIP from SuSiE. Column Type contain four binary digits with 1 for positive call and 0 for negative call. From left to right the first digit is for Model-X⊗fastBVSR, the second digit is for KnockoffScreen ⊗combined, the third is for ReKo⊗combined, and the fourth digit is for SuSiE. The nominal FDR used to call positives is 0.10.

eQTL. Both paternal and maternal alleles of SNP *rs12875121* were found to be significant by all three knockoff methods. SuSiE missed the positive call of maternal allele by 0.01. This SNP is indeed a genotype eQTL. 3) ReKo called two additional maternal eQTL significant for *GJB6*, but Model-X, KnockoffScreen, and SuSiE did not. Regression analysis show that paternal and maternal alleles of SNP *rs12875121* alone had adjusted R-squared value of 0.081, and adding these two maternal eQTL the adjusted R-squared value increased to 0.103. This suggests that the two maternal maternal eQTL are likely true positives. 4) *CCR9* is known to harbor exclusively paternal eQTL, and all four methods call a paternal eQTL positive. 5) *COA8* provides another example where all three knockoff filters called positive, but the SuSiE missed by a small margin. 6) *NECAB* is known to harbor exclusively opposing eQTL; all four methods reported significant maternal eQTL without its companion paternal eQTL.

3 Discussion

In this paper, we presented a novel method to construct knockoffs via Householder reflection and demonstrated the power it brings to knockoff filters. We provided a simple method to aggregate multiple sets of identically distributed knockoff statistics to improve the consistency of knockoff filters. We also demonstrated that a combined knockoff filter, by combining Bayesian and Frequentist inference, can effectively bring the realized FDR under control when correlation is not extremely high. ReKo assumes feature matrix is fixed and known. In this sense, ReKo works within the framework of Fixed-X but provides a novel way to estimate $\mathbf{S}$ in Equation 1. Our current work, however, provides no theoretical guarantee of FDR control

under arbitrary covariance structure.

We applied reflection knockoffs to address two problems that involve highly correlated features. In the analysis of proteomic signatures of age using knockoff filters, we demonstrated that the majority of proteins showing significance in single-feature tests are hitchhikers, not drivers. In the analysis of genetic fine mapping, we demonstrated that ReKo can reliably make positive calls that are consistent with Model-X and KnockoffScreen. Two extra positive calls by ReKo alone were supported by high SuSiE PIPs. ReKo is computationally efficient, particularly when rSVD approximation is invoked. We documented the computational time required to generate knockoffs and to fit models for different knockoff filters in Supplementary Section 10.

In both simulated features and real datasets, ReKo improves the power of different knockoff filters. We believe ReKo can be beneficial to other knockoff-based methods, such as those developed for genetic association studies (Candès et al., 2018; He et al., 2021b). In Figure 1, we demonstrate that the aggregation helps KnockoffScreen (He et al., 2021a). We believe our method of aggregating multiple sets of identically distributed knockoff statistics can also benefit knockoffs constructed by other methods, including hidden Markov model knockoffs (Sesia et al., 2019), KnockoffZoom (Sesia et al., 2020), and GhostKnockoffs (He et al., 2022).

We used empirical Bayes estimated shrinkage priors to combat high correlation when constructing reflection knockoffs, an alternative approach is to use weighted regression in Equation (9) to down-weight covariates in $\mathbf{X}_{-j}$ that are highly correlated with $\mathbf{X}_j$. This is equivalent to putting a different prior on β_{-j} for different j. A natural prior for this purpose would be the partial correlation taken from $\mathrm{cov}(\mathbf{X})^{-1}$. We would still need a separate shrinkage prior to control the overall shrinkage (similar to the λ in LASSO), which could be determined empirically, for example, using cross-validation. This can be a fruitful future research topic to further improve reflection knockoffs.

Trimming highly correlated features was used in the current study to reduce high correlation in the data. As currently implemented, the trimming procedure aimed to keep features with more significant marginal association. There is certainly room for improvement, such as fitting models jointly to rank features, and using the ranks to inform feature trimming. Compared to trimming, it may be more fruitful to cluster highly correlated features or to perturb highly correlated features with carefully designed noises. This can be another fruitful future research topic to improve reflection knockoffs.

A clever approach to reduce high correlation in fine mapping is to pool genetic data from multiple ancestral backgrounds (Yuan et al., 2024; Cai et al., 2023). This practice takes advantage of different patterns of LD in different ancestral backgrounds to reduce correlation among genetic markers, which evidently increases power for fine mapping. Knockoff filters can also benefit from pooling genetic data from multiple ancestral backgrounds.

A Bayesian procedure such as fastBVSR may produce PIPs that are not calibrated due to, say, difficulties in estimating priors odds in applications involving highly correlated features. Our study suggests that knockoff filters can indeed improve calibration of PIP by recalibrating

it. Similar to computing p-values for Bayes factors (Zhou and Guan, 2018), using a knockoff filter – an intrinsic Frequentist procedure – with PIP as the importance measure can be perceived as another example of Bayesian / non-Bayesian compromise (Good, 1992). Another notable benefit of the knockoff filter is its ability to combine multiple knockoff filters, such as ⊗fastBVSR and ⊗glmnet, to improve calibration of the realized FDR. The combined knockoff filter can also be perceived as an example of Bayesian / non-Bayesian compromise.

Both knockoff filters ReKo⊗glmnet⊗Beta and ReKo⊗fastBVSR⊗PIP produced inflated realized FDR when features were highly correlated. The combined knockoff filter appeared to be effective in controlling the realized FDR when the correlation is not extremely high. The intuition behind this is as follows: Suppose there are four highly correlated features such that conditioning on any one of these features, the other three features become uncorrelated with the phenotype. One of the features is selected as a driver in simulation. Penalized regression will arbitrarily pick one feature among the four and call it a positive, while the other three are declared negative. In contrast, fastBVSR, by the virtue of model averaging, will assign each feature a PIP of about 0.25. If the knockoff threshold happens to be 0.20, then fastBVSR will produce three false positives, compare to at most one for penalized regression. On the other hand, suppose these four features are null features, only associated with the phenotype via hitchhiking. The model averaging of fastBVSR will make it less likely to produce false positives among these four features compared to glmnet. The combined filter will reduce false positive in both scenarios.

One potential application reflection knockoffs is data sharing for genetic research. By design, knockoffs ensure that the original genetic data cannot be reverse-engineered, thanks to their reliance on random processes and the inherently one-way nature of the construction method. Even with complete knowledge of the algorithm and access to the knockoff data, recovering the original data is mathematically infeasible. This opens a transformative pathway for sharing sensitive genetic or proteomic data while meeting stringent privacy and regulatory standards, such as HIPAA. Imagine a collaborative research ecosystem where scientists worldwide can freely access anonymized knockoff datasets to address critical challenges like fine mapping. This mirrors the ethos of the machine learning field, where shared data fuels innovation: the data are there, the goals are clear – innovate with dare! With reflection knockoffs, genetic research can embrace this model, empowering the community to collectively tackle the most pressing problems with creativity and bravery.

4 Materials and Methods

4.1 Reflection

We have the following lemma regarding Householder reflection (c.f. Golub and Van Loan, 2013) and some useful properties for constructing knockoffs.

Lemma 1. *Suppose the data matrix* $\mathbf{X}$ *is* $n \times p$ *with* $n > p$ *and* $\mathbf{X}$ *is column-centered, scaled, and has full rank* p*. Denote* $\mathbf{P}_j = \mathbf{X}_{,-j}(\mathbf{X}_{,-j}^t\mathbf{X}_{,-j})^{-1}\mathbf{X}_{,-j}^t$ *where* $\mathbf{X}_{,-j}$ *is matrix* $\mathbf{X}$ *with* j*-th column removed. Compute* $\mathbf{Y}$ *such that its* j*-th column is a Householder reflection of* $\mathbf{X}_j$ *with respect to* $\mathbf{P}_j$*,* $\mathbf{Y}_j = (2\mathbf{P}_j - \mathbf{I})\mathbf{X}_j$*. Define* $\mathbf{Z} = \frac{1}{2}(\mathbf{X} - \mathbf{Y})$*. Denote* $cov(\mathbf{X}_j, \mathbf{Y}_k)$ *as sample covariance between vector* $\mathbf{X}_j$ *and* $\mathbf{Y}_k$*. Then the following hold:*

a) $cov(\mathbf{Y}_j, \mathbf{X}_k) = cov(\mathbf{X}_j, \mathbf{X}_k)$ *for all* $j \neq k$ *and* $j, k \in (1, 2, \ldots, p)$*.*

b) $cov(\mathbf{Y}_j, \mathbf{X}_j) < var(\mathbf{X}_j)$ *for all* $j \in (1, 2, \ldots, p)$*.*

c) Let $\mathbf{T}$ *be a matrix such that* $\mathbf{T}_{jk} = cov(\mathbf{X}_j, \mathbf{Z}_k)$*, then* $\mathbf{T}_{jk} = 0$ *for all* $j \neq k$ *and* $\mathbf{T}_{jj} = cov(\mathbf{Z}_j, \mathbf{Z}_j) > 0$ *for all* j*.*

d) $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} = \mathbf{Y}$ *and* $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{Z} = \mathbf{Z}$*.*

Proof. a) The projection matrix $\mathbf{P}_j = \mathbf{X}_{,-j}(\mathbf{X}_{,-j}^t\mathbf{X}_{,-j})^{-1}\mathbf{X}_{,-j}^t$ projects $\mathbf{X}_j$ onto the column space of $\mathbf{X}_{,-j}$. Since $(\mathbf{I}-\mathbf{P}_j)\mathbf{X}_j$ is orthogonal to $\mathbf{X}_k$ for $k \neq j$, we have $\text{cov}((\mathbf{I}-\mathbf{P}_j)\mathbf{X}_j, \mathbf{X}_k) = 0$. This implies $\text{cov}(\mathbf{P}_j\mathbf{X}_j, \mathbf{X}_k) = \text{cov}(\mathbf{X}_j, \mathbf{X}_k)$. Now $\mathbf{Y}_j = (2\mathbf{P}_j - \mathbf{I})\mathbf{X}_j$, and direct computation gives: $\text{cov}(\mathbf{Y}_j, \mathbf{X}_k) = \text{cov}((2\mathbf{P}_j - \mathbf{I})\mathbf{X}_j, \mathbf{X}_k) = 2\text{cov}(\mathbf{P}_j\mathbf{X}_j, \mathbf{X}_k) - \text{cov}(\mathbf{X}_j, \mathbf{X}_k)$. Substituting $\text{cov}(\mathbf{P}_j\mathbf{X}_j, \mathbf{X}_k) = \text{cov}(\mathbf{X}_j, \mathbf{X}_k)$, we obtain $\text{cov}(\mathbf{Y}_j, \mathbf{X}_k) = \text{cov}(\mathbf{X}_j, \mathbf{X}_k)$, proving (a).

b) The projection of $\mathbf{X}_j$ onto the space spanned by $\mathbf{X}_{,-j}$ as $\mathbf{X}_\mathbf{V} = \mathbf{P}_j\mathbf{X}_j$, and the orthogonal component as $\mathbf{X}_\perp = (\mathbf{I} - \mathbf{P}_j)\mathbf{X}_j$. Thus, $\mathbf{X}_j = \mathbf{X}_\mathbf{V} + \mathbf{X}_\perp$. The Householder reflection $\mathbf{Y}_j$ can then be written as $\mathbf{Y}_j = \mathbf{X}_\mathbf{V} - \mathbf{X}_\perp$. The sample covariance between $\mathbf{Y}_j$ and $\mathbf{X}_j$ is given by $\text{cov}(\mathbf{Y}_j, \mathbf{X}_j) = \text{cov}(\mathbf{X}_\mathbf{V} - \mathbf{X}_\perp, \mathbf{X}_\mathbf{V} + \mathbf{X}_\perp) = \text{cov}(\mathbf{X}_\mathbf{V}, \mathbf{X}_\mathbf{V}) - \text{cov}(\mathbf{X}_\perp, \mathbf{X}_\perp)$. That is, $\text{cov}(\mathbf{Y}_j, \mathbf{X}_j) = \text{var}(\mathbf{X}_\mathbf{V}) - \text{var}(\mathbf{X}_\perp)$. Since $\text{var}(\mathbf{X}_j) = \text{var}(\mathbf{X}_\mathbf{V}) + \text{var}(\mathbf{X}_\perp)$, it follows that: $\text{cov}(\mathbf{Y}_j, \mathbf{X}_j) = \text{var}(\mathbf{X}_j) - 2\text{var}(\mathbf{X}_\perp)$. Since $\text{var}(\mathbf{X}_\perp) \geq 0$ with equality holds only if $\mathbf{X}_\perp = 0$, which would violate the assumption that matrix $\mathbf{X}$ has full rank. Hence, $\text{cov}(\mathbf{Y}_j, \mathbf{X}_j) < \text{var}(\mathbf{X}_j)$, proving (b).

c) Statement c) follows directly from a) and b). Specifically, for any j, $4\text{var}(\mathbf{Z}_j) = \text{cov}(\mathbf{X}_j - \mathbf{Y}_j, \mathbf{X}_j - \mathbf{Y}_j) = \text{cov}(\mathbf{X}_j) + \text{cov}(\mathbf{Y}_j) - 2\text{cov}(\mathbf{X}_j, \mathbf{Y}_j) = 2 - 2\text{cov}(\mathbf{X}_j, \mathbf{X}_j - 2\mathbf{Z}_j) = 4\text{cov}(\mathbf{X}_j, \mathbf{Z}_j) = 4\mathbf{T}_{jj}$. From (b) $\text{cov}(\mathbf{X}_j, \mathbf{Y}_j) < \text{var}(\mathbf{X}_j)$, ensuring $\mathbf{T} = \text{cov}(\mathbf{X}, \mathbf{Z}) > 0$.

d) From (a), $\mathbf{Y}_j = (2\mathbf{P}_j - \mathbf{I})\mathbf{X}_j$, where $\mathbf{P}_j$ projects $\mathbf{X}_j$ onto the column space of $\mathbf{X}_{,-j}$. Thus, $\mathbf{Y}_j$ is a linear combination of columns of $\mathbf{X}$. Consequently, $\mathbf{Y}$ lies in the column space of $\mathbf{X}$. Similarly, $\mathbf{Z} = \frac{1}{2}(\mathbf{X} - \mathbf{Y})$ also lies in the column space of $\mathbf{X}$. Therefore: $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} = \mathbf{Y}$ and $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{Z} = \mathbf{Z}$. □

4.2 Construct knockoffs based on reflection

Following Equation (2.2) of (Barber and Candès, 2015) and Section 3.1.1 of (Candès et al., 2018), the mean $\mathbf{b}$ and covariance $\mathbf{C}^t\mathbf{C}$ of a knockoff $\tilde{\mathbf{X}}$ conditional on $\mathbf{X}$ are given by

$$\begin{aligned} \mathbf{b} &= \mathbf{X}(\mathbf{I} - \Sigma^{-1}\mathbf{S}) \\ \mathbf{C}^t\mathbf{C} &= 2\mathbf{S} - \mathbf{S}\Sigma^{-1}\mathbf{S}. \end{aligned} \tag{3}$$

Let $\mathbf{S} = \alpha\mathbf{T}$ for $\alpha \in (0, 1]$. Straightforward calculations using the identities in Lemma 1 parts (c) and (d) yields:

$$\begin{aligned} \mathbf{b} &= \mathbf{X} - \alpha\mathbf{Z} \\ \mathbf{C}^t\mathbf{C} &= 2\alpha\mathbf{T} - \alpha^2 \operatorname{cov}(\mathbf{Z}). \end{aligned} \tag{4}$$

To ensure $\mathbf{C}^t\mathbf{C} \geq \mathbf{0}$, we need to determine $\hat{\alpha}$ such that this condition holds. Noting that $\mathbf{T}$ is positive definite and diagonal, we can rewrite:

$$\mathbf{C}^t\mathbf{C} = \mathbf{T}^{1/2}\left[2\alpha\mathbf{I} - \alpha^2\operatorname{cov}(\mathbf{Z}\mathbf{T}^{-1/2})\right]\mathbf{T}^{1/2}. \tag{5}$$

Let $\mathbf{Q}\Lambda\mathbf{Q}^t$ be the eigendecomposition of $\operatorname{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$. Substituting this, we obtain

$$\mathbf{C}^t\mathbf{C} = \mathbf{T}^{1/2}\mathbf{Q}\left[2\alpha\mathbf{I} - \alpha^2\Lambda\right]\mathbf{Q}^t\mathbf{T}^{1/2}. \tag{6}$$

For $\mathbf{C}^t\mathbf{C}$ to be positive semi-definite, it is sufficient to have

$$\mathbf{B}^2(\alpha) \triangleq 2\alpha\mathbf{I} - \alpha^2\Lambda > \mathbf{0}. \tag{7}$$

Therefore it's sufficient to have $\alpha \leq 2/\Lambda_1$, where Λ_1 is the leading eigenvalue of $\operatorname{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$. Since Λ_1 is positive and finite, and $\alpha \in (0, 1]$, the optimal $\hat{\alpha}$ exists. This ensures $\mathbf{C}^t\mathbf{C}$ is positive semi-definite, with $\mathbf{C} = \mathbf{B}(\hat{\alpha})\mathbf{Q}^t\mathbf{T}^{1/2}$. Consequently, $\mathbf{S} = \hat{\alpha}\mathbf{T} \leq 2\operatorname{cov}(\mathbf{X})$, as guaranteed by the Schur complement condition.

We thus construct the knockoff as

$$\tilde{\mathbf{X}} = \mathbf{X} - \hat{\alpha}\mathbf{Z} + \mathbf{U}\mathbf{B}(\hat{\alpha})\mathbf{Q}^t\mathbf{T}^{1/2} \tag{8}$$

where: If $\mathbf{U}$ is chosen from the null space of $\mathbf{X}$, $\tilde{\mathbf{X}}$ is akin to Fixed-X knockoffs. If $\mathbf{U}_j \sim MVN(\mathbf{0}, \mathbf{I}_n)$, $\tilde{\mathbf{X}}$ is akin to Model-X second-order knockoff. Our numerical studies showed that their performances are highly similar (Supplementary Tab. S2), and in this paper we use the later for numerical convenience. As a special case, when the columns of $\mathbf{X}$ are uncorrelated, simple calculations show that $\tilde{\mathbf{X}} = \mathbf{U}$. This implies that for uncorrelated $\mathbf{X}$, $\tilde{\mathbf{X}}$ can be simulated directly from either a standard normal distribution or the basis of the null space of $\mathbf{X}$.

4.3 Shrinkage priors and data augmentation

In Lemma 1, computing $\mathbf{Y}_j$ requires the least square fit of the linear regression:

$$\mathbf{X}_j = \mathbf{X}_{,-j}\beta_{-j} + \mathbf{e} \tag{9}$$

to obtain $\hat{\beta}_{-j}$, which is then used to calculate $\mathbf{Y}_j = 2\mathbf{X}_{,-j}\hat{\beta}_{-j} - \mathbf{X}_j$. When p is large compare to n, and/or there exist highly correlated features, the model (9) is prone to overfitting, which can adversely affect the performance of the knockoff filter. While penalized regression is a

good choice to avoid overfitting, Bayesian regression offer a more elegant solution in this context, as it serves two purposes.

Suppse we put priors on β_{-j} such as

$$\beta_{-j} \sim MVN(0, \sigma^2 \mathbf{I}_{(p-1)}), \tag{10}$$

where MVN denotes multivariate normal distribution, $\mathbf{I}_m$ is the identity matrix of dimension m, and σ is the prior effect size to be specified. The Bayesian estimated effect sizes are given by

$$\hat{\beta}_{-j} = \left(\mathbf{X}_{,-j}^t \mathbf{X}_{,-j} + \sigma^{-2}\mathbf{I}_{(p-1)}\right)^{-1} \mathbf{X}_{,-j}^t \mathbf{X}_j. \tag{11}$$

The same solution can be achieved via augmenting $\mathbf{X}$ with ghost observations $\sigma^{-1}\mathbf{I}_p$ such that $\mathbf{X} \leftarrow [\mathbf{X}^t, \sigma^{-1}\mathbf{I}]^t$. A natural choice for σ is $\sigma^2 = 1/p$, as this prior ensures the posterior distribution remains well behaved for large p. That is, the posterior remain proper and without excessive variance inflation. But empirical Bayes estimated σ (to be discussed below) appears to perform better. This data augmentation enables constructing reflection knockoffs for $\mathbf{X}$ of any shape, as the augmented $\mathbf{X}$ will have dimension $(n+p) \times p$ and full rank p, which satisfies the conditions in Lemma 1. By augmenting $\mathbf{X}$ with ghost observations, we can extend this construction to handle cases where $p > n$.

4.4 Empirical Bayes estimates of shrinkage priors

Consider a Bayesian linear regression for model (9) and prior (10), we have

$$\begin{aligned} \mathbf{X}_j &= \mathbf{X}_{,-j}\beta + \mathbf{e} \\ \beta &\sim MNV(0, \tau^{-1}\sigma^2) \\ \mathbf{e} &\sim MNV(0, \tau^{-1}\mathbf{I}_n) \end{aligned} \tag{12}$$

Integrate out β to obtain the marginal likelihood, and we get

$$\mathbf{X}_j | \mathbf{X}_{,-j}, \sigma^2 \sim MNV(0, \tau^{-1}\sigma_j^2 \mathbf{X}_{,-j}\mathbf{X}_{,-j}^t + \tau^{-1}\mathbf{I}). \tag{13}$$

This is a special case of a linear mixed model and we can obtain maximum likelihood estimate of σ_j^2 using IDUL (Guan and Levy, 2024b). To proceed, we obtain eigeindecomposition $\mathbf{X}_{,-j}\mathbf{X}_{,-j}^t = \mathbf{Q}_j\mathbf{D}_j\mathbf{Q}_j^t$ where $\mathbf{Q}_j$ is orthonormal and $\mathbf{D}_j$ is diagonal. We rotate the system by multiplying (12) by $\mathbf{Q}^t$ from left we get

$$\mathbf{X}_j' | \mathbf{X}_{,-j}', \sigma^2 \sim MNV(0, \tau^{-1}\sigma_j^2 \mathbf{D}_j + \tau^{-1}). \tag{14}$$

For this particular application, we find that using a bisection method is more stable than IDUL, particularly for smaller p. In addition, a practical upper bound 1 can be imposed on σ_j^2 to facilitate the bisection. Therefore, we used the bisection method to find maximum

likelihood estimate of σ_j.

We estimated σ_j^2 for each feature, which requires eigendecomposition of $\mathbf{X}_{,-j}\mathbf{X}_{,-j}^t$ for each j, and naive computation can be expensive $O(np^3)$. Here we approximate eigenvectors of $\mathbf{X}_{-j}\mathbf{X}_{-j}^t$ with eigenvectors of $\mathbf{X}\mathbf{X}^t$, but update eigenvalues for each $\mathbf{X}_{-j}\mathbf{X}_{-j}^t$. The rank one update for eigenvalue requires one to solve secular equations, which can be solved in a constant time (for updating a single eigenvalue) using another bisection search method. The total complexity of to obtain eigenvalues for $\mathbf{X}_{-j}\mathbf{X}_{-j}^t$ using rank one update is $O(np + p^2)$ (more details in Supplementary Section 1). For eigen decomposition of $\mathbf{X}\mathbf{X}^t$, we implemented randomized singular value decomposition (rSVD) (Halko et al., 2011) to efficiently compute (say 200) approximate leading eigenpairs, and use them to compute empirical Bayes estimated priors.

After estimating σ_j for each j, we obtained $\Sigma = diag(\sigma_1, \ldots, \sigma_p)$. Instead of fitting Model (12) for each j, which would be prohibitively, we augmented $\mathbf{X}$ with Σ and compute reflection using Algorithm 1 detailed below. This appears to maintain the performance and computation efficiency. A comparison between data augmentation with empirical Bayes estimated priors and those with fixed priors can be found in Supplementary Fig. S1.

4.5 Computation

To speed up the computation and avoid repeatedly computing the matrix inverse needed for $\hat{\beta}_{-j}$ in Equation (11), we employ QR decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q}$ is orthonormal $n \times p$ and $\mathbf{R}$ is a $p \times p$ upper triangular matrix. Multiply both sides of the linear regression (9) by Q^t, we get $\mathbf{R}_j = \mathbf{R}_{,-j}\beta_{-j} + \mathbf{e}_{\mathbf{Q}}$, where $\mathbf{R}_j$ is the jth column of $\mathbf{R}$, and $\mathbf{R}_{,-j}$ is the matrix with its j-th column removed.

Efficient computation via Givens Rotation. The matrix $\mathbf{R}_{,-j}$ is close to being upper triangular and requires at most $(p - j)$ many Givens rotations to become triangular, which can be done efficiently. A Givens rotation corresponds to left-multiplying a sparse matrix, which we apply to both sides of the regression. Importantly, this rotation is unitary and does not affect mean and variance of the error term. Once the matrix is properly triangularized, we use backward substitution to solve the system for $\hat{\beta}_{-j}$. This value is then used to compute $\tilde{\mathbf{R}}_j$ and subsequently $\mathbf{Y}_j$.

Parallel Computing for Speed-Up. The algorithm below outlines the steps, where the for loop can be parallelized for efficiency. Our software implementation leverages this feature for high-performance computation.

Algorithm 1 QR decomposition for Householder reflection

Input: A matrix $\mathbf{X}$ of $n \times p$ with full rank p.
Output: $\mathbf{Y}$ of $n \times p$.
Augment $\mathbf{X}$ by appending Σ_p, update $n \leftarrow (n+p)$.
Perform QR decomposition to obtain $\mathbf{X} = \mathbf{QR}$.
for each column $\mathbf{R}_j$ of $\mathbf{R}$ **do**
    Denote $\mathbf{R}_{,-j}$ as matrix $\mathbf{R}$ with j-th column removed.
    Do Givens rotation on $\mathbf{R}_{,-j}$ to obtain an upper triangular matrix $\mathbf{U}$.
    Do the same sequence of Givens rotation on $\mathbf{R}_j$ to obtain feature $\mathbf{V}$.
    Solve $\mathbf{V} = \mathbf{U}\beta_{-j}$ with backward substitution to get $\hat{\beta}_{-j}$.
    Compute $\tilde{\mathbf{R}}_j = 2\mathbf{R}_{,-j}\hat{\beta}_{-j} - \mathbf{R}_j$, which is j-th column of $\tilde{\mathbf{R}}$.
end for
Compute $\mathbf{Y} = \mathbf{Q}\tilde{\mathbf{R}}$
return $\mathbf{Y}[1:n,]$

4.6 Aggregation

We present a method for aggregating multiple knockoff statistics. Let $W = (W_1, \ldots, W_p)$ represent the knockoff features importance statistics from a single reflection knockoff. Under the null, the distribution W_j is symmetric about zero, ensuring FDR control. The knockoff threshold T for W, controlling the FDR at nominal level q is defined as:

$$T = \min\left\{t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}. \tag{15}$$

Suppose we construct M copies of reflection knockoffs, the m-th copy of the knockoff lead to the knockoff statistics $W^{(m)}$ for $m = 1, \ldots, M$, which are identically distributed as W. This assumes the same knockoff construction mechanism, data, and model-fitting procedure, differing only by random seeds or knockoff randomness. The following lemma establishes the validity of aggregating these identically distributed knockoff statistics.

Lemma 2. *Let $W_{cat} = (W^{(1)}, \ldots, W^{(M)})$, i.e., the pooled knockoff statistics from the M runs. If we apply the standard knockoff threshold selection rule* (15) *to W_{cat}, the resulting threshold T_{cat} also controls the FDR at level q. Define the average knockoff statistic $\bar{W}_j = \frac{1}{M}\sum_{m=1}^{M} W_j^{(m)}$. Selecting feature j such that $\bar{W}_j > T_{cat}$ controls the FDR at level q.*

Proof. For the m-th knockoff statistics $W^{(m)}$, the null features are symmetrically distributed around zero and the knockoff threshold selection 15 controls the FDR at level q for $W^{(m)}$. Since $W^{(m)}$ for $m = (1, \ldots, M)$ are identically distributed, the combined set W_{cat} of size Mp is drawn from the same underlying distribution as W. If the j-th feature is a null feature, then $W_j^{(m)}$ is symmetric about zero for all $m \in (1, \ldots, M)$. Conversely, if the j-th features

is non-null, then $W_j^{(m)} > 0$ for all $m \in (1, \ldots, M)$. The two counts in (15) have the same expectation for each j and for all t. Thus, applying the knockoff threshold rule to W_{cat} yields a threshold T_{cat} that controls FDR at level q.

For a null feature j, the distribution of $W_j^{(m)}$ is symmetric about zero. Formally, the pair $(W_j^{(m)}, -W_j^{(m)})$ is identically distributed. This symmetry holds for each copy of the knockoff. The average statistics $\bar{W}_j$ preserves this symmetry, ensuring that $\bar{W}_j$ is symmetry about zero. The super-martingale argument from the original FDR control proof applies directly (supplementary material of Barber and Candès, 2015). Under the null, features have no preference for positive or negative deviations. Thus, selecting feature j such that $\bar{W}_j > T_{cat}$ controls the FDR at level q. □

5 Acknowledgements

This research is supported by the Division of Intramural Research of the National Heart, Lung, and Blood Institute, Bethesda, MD (D.L. Principal Investigator). We thank editors and two anonymous reviewers for their constructive review comments that greatly improved clarity of the presentation. In particular, we thank one reviewer who brought our attention to KnockoffScreen, and the other for suggesting empirical Bayes estimates of the prior effects for reflection. Our revised manuscript benefitted greatly from both suggestions.

6 Declaration of interests

The authors declare no competing interests.

7 Data and code availability

This study analyzed existing datasets from public domain and simulated datasets. Data from the UK Biobank can be obtained from `https://www.ukbiobank.ac.uk/`, and datasets from Framingham Heart Study can be obtained via dbGaP. The software for ReKo and its source code are freely available at `https://github.com/haplotype/ReKo`.

8 Author contributions

Y.G. conceived the study, developed the methodology, implemented the computational software and performed experiments, analyzed results, and wrote the manuscript. D.L. supervised the work, provided access to data and computation, edited and approved the final manuscript.

References

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055 – 2085.

Benner, C., C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen (2016, May). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics 32*(10), 1493–1501.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Cai, M., Z. Wang, J. Xiao, X. Hu, G. Chen, and C. Yang (2023). Xmap: Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. *Nature Communications 14*(1), 6870.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018, 01). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 551–577.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Friedman, J., R. Tibshirani, and T. Hastie (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gimenez, J. R. and J. Zou (2019, 16–18 Apr). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Volume 89 of *Proceedings of Machine Learning Research*, pp. 2184–2192. PMLR.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations* (4th ed.). Baltimore, MD: Johns Hopkins University Press.

Good, I. J. (1992). The bayes/non-bayes compromise: A brief review. *Journal of the American Statistical Association 87*(419), 597–606.

Guan, Y. and D. Levy (2024a). Abundant parent-of-origin effect eqtl in humans: The framingham heart study. *bioRxiv*.

Guan, Y. and D. Levy (2024b, 08). Asymptotically Exact Fit for Linear Mixed Model in Genetic Association Studies. *Genetics*, iyae143.

Guan, Y. and D. Levy (2024c, 02). Estimation of inbreeding and kinship coefficients via latent identity-by-descent states. *Bioinformatics 40*(2), btae082.

Guan, Y. and M. Stephens (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics 5*(3), 1780 – 1815.

Halko, N., P. G. Martinsson, and J. A. Tropp (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review 53*(2), 217–288.

He, Z., L. Liu, M. E. Belloy, Y. Le Guen, A. Sossin, X. Liu, X. Qi, S. Ma, P. K. Gyawali, T. Wyss-Coray, H. Tang, C. Sabatti, E. Candès, M. D. Greicius, and I. Ionita-Laza (2022, November). GhostKnockoff inference empowers identification of putative causal variants in genome-wide association studies. *Nat. Commun. 13*(1), 7209.

He, Z., L. Liu, C. Wang, Y. Le Guen, J. Lee, S. Gogarten, F. Lu, S. Montgomery, H. Tang, E. K. Silverman, M. H. Cho, M. Greicius, and I. Ionita-Laza (2021a). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature Communications 12*(1), 3152.

He, Z., L. Liu, C. Wang, Y. Le Guen, J. Lee, S. Gogarten, F. Lu, S. Montgomery, H. Tang, E. K. Silverman, M. H. Cho, M. Greicius, and I. Ionita-Laza (2021b). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature Communications 12*(1), 3152.

Hormozdiari, F., E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin (2014, October). Identifying causal variants at loci with multiple signals of association. *Genetics 198*(2), 497–508.

Lehallier, B., D. Gate, N. Schaum, T. Nanasi, S. E. Lee, H. Yousef, P. Moran Losada, D. Berdnik, A. Keller, J. Verghese, S. Sathyan, C. Franceschi, S. Milman, N. Barzilai, and T. Wyss-Coray (2019, December). Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med. 25*(12), 1843–1850.

Qi, Z., S. Xie, R. Chen, H. A. Aisa, G. C. Hon, and Y. Guan (2018). Tissue-specific gene expression prediction associates vitiligo with suox through an active enhancer. *bioRxiv*.

Sesia, M., E. Katsevich, S. Bates, E. Candès, and C. Sabatti (2020). Multi-resolution localization of causal variants across the genome. *Nature communications 11*(1), 1093.

Sesia, M., C. Sabatti, and E. J. Candès (2019, March). Gene hunting with hidden markov model knockoffs. *Biometrika 106*(1), 1–18.

Sun, B. B., J. Chiou, M. Traylor, C. Benner, Y.-H. Hsu, T. G. Richardson, P. Surendran, A. Mahajan, C. Robins, S. G. Vasquez-Grinnell, L. Hou, E. M. Kvikstad, O. S. Burren, J. Davitte, K. L. Ferber, C. E. Gillies, A. K. Hedman, S. Hu, T. Lin, R. Mikkilineni, R. K. Pendergrass, C. Pickering, B. Prins, D. Baird, C.-Y. Chen, L. D. Ward, A. M.

Deaton, S. Welsh, C. M. Willis, N. Lehner, M. Arnold, M. A. Wörheide, K. Suhre, G. Kastenmüller, A. Sethi, M. Cule, A. Raj, Alnylam Human Genetics, AstraZeneca Genomics Initiative, Biogen Biobank Team, Bristol Myers Squibb, Genentech Human Genetics, GlaxoSmithKline Genomic Sciences, Pfizer Integrative Biology, Population Analytics of Janssen Data Sciences, Regeneron Genetics Center, L. Burkitt-Gray, E. Melamud, M. H. Black, E. B. Fauman, J. M. M. Howson, H. M. Kang, M. I. McCarthy, P. Nioi, S. Petrovski, R. A. Scott, E. N. Smith, S. Szalma, D. M. Waterworth, L. J. Mitnaul, J. D. Szustakowski, B. W. Gibson, M. R. Miller, and C. D. Whelan (2023, October). Plasma proteomic associations with genetics and health in the UK biobank. *Nature 622*(7982), 329–338.

Tanaka, T., N. Basisty, G. Fantoni, J. Candia, A. Z. Moore, A. Biancotto, B. Schilling, S. Bandinelli, and L. Ferrucci (2020, nov). Plasma proteomic biomarker signature of age predicts health and life span. *eLife 9*, e61073.

Tanaka, T., A. Biancotto, R. Moaddel, A. Z. Moore, M. Gonzalez-Freire, M. A. Aon, J. Candia, P. Zhang, F. Cheung, G. Fantoni, C. consortium, R. D. Semba, and L. Ferrucci (2018). Plasma proteomic signature of age in healthy humans. *Aging Cell 17*(5), e12799.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens (2020). A simple new approach to variable selection in regression, with application to genetic fine-mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(5), 1273–1300.

Yuan, K., R. J. Longchamps, A. F. Pardiñas, M. Yu, T.-T. Chen, S.-C. Lin, Y. Chen, M. Lam, R. Liu, Y. Xia, Z. Guo, W. Shi, C. Shen, M. J. Daly, B. M. Neale, Y.-C. A. Feng, Y.-F. Lin, C.-Y. Chen, M. C. O'Donovan, T. Ge, H. Huang, and C. The Schizophrenia Workgroup of Psychiatric Genomics (2024). Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nature Genetics 56*(9), 1841–1850.

Zhou, Q. and Y. Guan (2018). On the null distribution of bayes factors in linear regression. *Journal of the American Statistical Association 113*(523), 1362–1371.

Zhou, Q. and Y. Guan (2019). Fast Model-Fitting of Bayesian Variable Selection Regression Using the Iterative Complex Factorization Algorithm. *Bayesian Analysis 14*(2), 573 – 594.

Zou, H. and T. Hastie (2005, 03). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.