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

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

January 15, 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 knockoff substantially outperforms Fixed-X and 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. However, knockoff filters using reflection knockoffs and aggregation revealed that a majority of these associations are hitchhikers. When applied to genetic fine mapping, knockoff filters using reflection knockoffs and aggregation outperform a state-of-the-art method. We discuss a promising 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:

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

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 oblivious to the phenotypes.

Currently there are two main methods to generate knockoffs, Fixed-X and Model-X. A critical component in both methods 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. 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 much more powerful knockoffs, and its computation does not require optimization. We call the knockoffs constructed via Householder reflection *reflection knockoffs* (ReKo).

Due to the inherent randomness in constructing knockoffs, the knockoff filter may produce inconsistent results when applied to different copies of knockoffs. A secondary 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 both Fixed-X and Model-X knockoffs have little power due to complex and highly correlated features. 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 demonstrated 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, the off-diagonal elements of $\text{cov}(\mathbf{X}, \mathbf{Y})$ match the off-diagonal elements of $\text{cov}(\mathbf{X})$ (Lemma 1 in Materials and Methods), a necessary condition for $\mathbf{Y}$ to be a contributing part of the knockoff. A linear combination of $\mathbf{X}$ and $\mathbf{Y}$ will have the correct cross covariance, but not the correct variance. The variance, however, can be fixed by adding a term that is uncorrelated with both $\mathbf{X}$ and $\mathbf{Y}$, which is the essence of a formulation provided by Barber and Candès (2015).

The Householder reflection regresses $\mathbf{X}_j$ over all other features $\mathbf{X}_{,-j}$ to obtain a fitted value $\hat{\mathbf{X}}_j$, then uses it to compute the reflection $\mathbf{Y}_j = 2\hat{\mathbf{X}}_j - \mathbf{X}_j$, and the residual $\mathbf{Z}_j = \frac{1}{2}(\mathbf{X}_j - \mathbf{Y}_j) = \mathbf{X}_j - \hat{\mathbf{X}}_j$. Therefore $\mathbf{Z}_j$ is orthogonal to $\mathbf{X}_j$ and uncorrelated with all other features. Thus, $\mathbf{T} = \text{cov}(\mathbf{X}, \mathbf{Z}) > \mathbf{0}$ and is a diagonal matrix (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 covariance matrix $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$. The knockoffs can be computed following Equation (8) in Materials and Methods. Notably, constructing reflection knockoffs requires no optimization.

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 $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$ is similar to an identity matrix, with diagonal entries equal to 1 and small off-diagonal entries. By Lemma 1 we have $\text{var}(\mathbf{Z}_j) = \mathbf{T}_{jj}$ and since $\mathbf{T} > \mathbf{0}$ is diagonal, we have $\text{diag}(\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})) = \text{diag}(\mathbf{I}_p)$.

The off-diagonal element $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})_{jk}$ is the correlation between residuals $\mathbf{Z}_j$ and $\mathbf{Z}_k$. Assuming $\text{cov}(\mathbf{X}_j, \mathbf{X}_k) = r$, and that both are uncorrelated with other features, a straightforward computation yields $\text{cov}(\mathbf{Z}_j, \mathbf{Z}_k) = r(r^2 - 1)$. Therefore both large (close to 1 or -1) and small (close to 0) correlations produce small $\text{cov}(\mathbf{Z}_j, \mathbf{Z}_k)$. This calculation also applies when r is the conditional correlation, which can be obtained from the precision matrix $\text{cov}(\mathbf{X})^{-1}$. When features are correlated, the conditional correlation between two features tends to be small. Then by the Gershgorin circle theorem, the eigenvalues satisfy $|\Lambda_j - 1| \leq \sum_{k \neq j} |\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})_{jk}|$. So as long as the off diagonal entries are small and linearly diminishing with the dimension of $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$, the leading eigenvalue Λ_1 will be small, and $\hat{\alpha}$ will be bounded away from 0.

By $\mathbf{T}_{jj} = \text{var}(\mathbf{Z}_j)$, 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, $\text{var}(\mathbf{Z}_j)$ 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 $\text{var}(\mathbf{Z}_j)$ 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 by just the right amount so that the image appears farther in the mirror without introducing obvious distortion. The priors also enable reflection knockoffs to be constructed when the number of features is larger than the number of samples, significantly broadening their scope of applications.

2.2 Comparison with other knockoffs

We compared reflection knockoffs with Fixed-X and Model-X knockoffs using power and realized FDR generated from knockoff filters. Both Fixed-X and Model-X knockoffs

were generated by the R package *knockoff*, using $\mathbf{S}$ estimated by the SDP method. We generated two varieties of reflection knockoffs that differ in their random components: ReKo-Null that sampled from the null space of $\mathbf{X}$ (similar to Fixed-X knockoffs), and ReKo-Norm, which sampled from the standard normal distribution (similar to Model-X knockoffs).

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$\otimes$glmnet$\otimes$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-Norm$\otimes$fastBVSR$\otimes$PIP means a knockoff filter that constructs reflection knockoffs, performs association analysis using fastBVSR, and uses posterior inclusion probability as the importance measure.

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

We simulated correlated features, and randomly chose 20 driver features to simulate phenotypes with a PVE of 0.8. One way to simulate correlated features is to assume an equal correlation c between all pairs of features, and we examined $c = 0.2, 0.5$ and 0.8, representing low, medium, and high levels of correlation. Another approach is to simulate varying correlation, where the correlations between each pair of features

are drawn from a bell-shaped histogram, with the histogram peaking at $0.2, 0.5, 0.8$. Varying correlation approach better mimics real data, but the equal correlation method was used in the seminal knockoff paper (Barber and Candès, 2015), and we include it here for comparison. Complete simulation details can be found in the Supplementary Material.

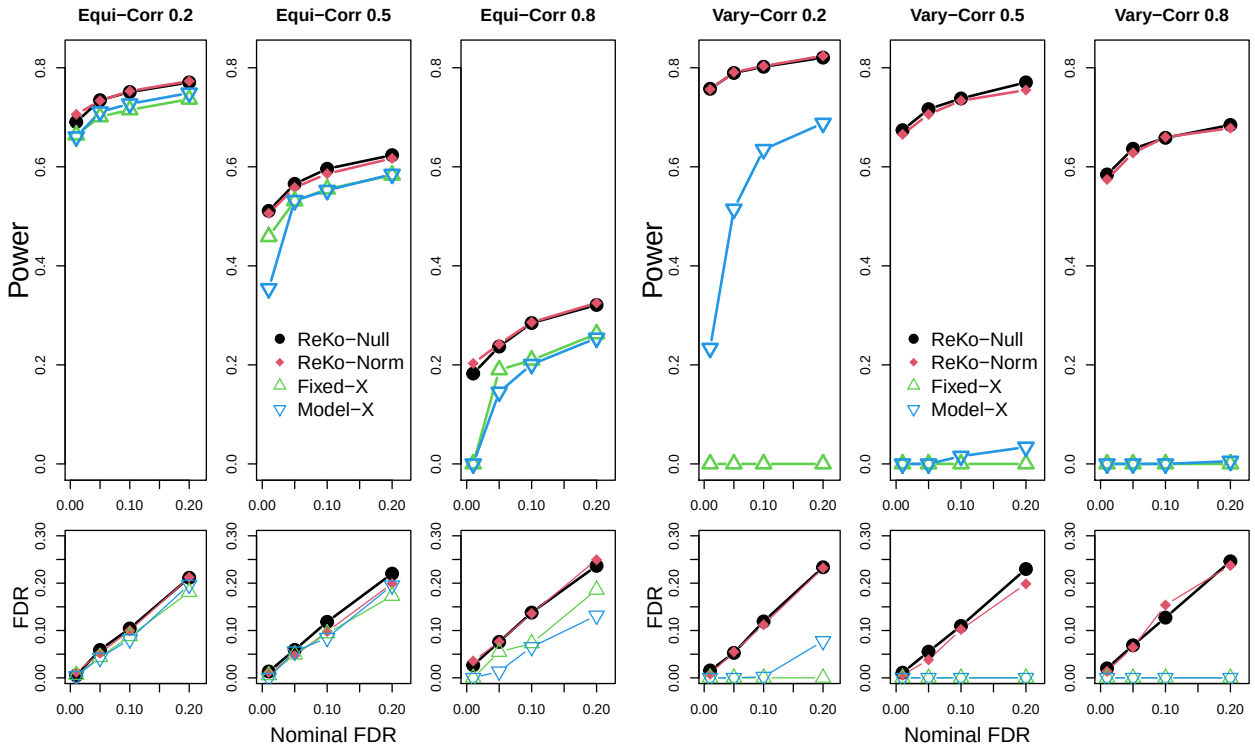

Figure 1: Power comparison between different knockoffs. Results shown was simulated with $n = 2000$ and $p = 800$. Phenotypes were simulated with 20 driver features whose effect sizes were drawn from standard normal but rescaled to achieve PVE of 0.8.

Figure 1 compares different knockoffs using knockoff filters $\otimes$glmnet. For the equi-correlation (left three panels), where all pairs of features have identical correlation, the power gain of reflection knockoffs (ReKo-Null and ReKo-Norm) over Fixed-X and Model-X knockoffs was less than in the vary-correlation (right three panels). Presumably, the semidefinite program used by both Fixed-X and Model-X knockoffs performs better under the equi-correlation scenario, where there are $p - 1$ identical small eigenvalues and one large eigenvalue for $\mathrm{cov}(\mathbf{X})$. For varying correlation, ReKo-Null

⊗glmnet and ReKo-Norm⊗glmnet produced higher power than both Fixed-X⊗glmnet and Model-X⊗glmnet. In particular, when correlations among features are large, both Fixed-X⊗glmnet and Model-X⊗glmnet have little to no power, while ReKo-Null ⊗glmnet and ReKo-Norm⊗glmnet maintain considerable power.

Next we used two sets of real features and simulated phenotypes to compare Model-X knockoffs and ReKo knockoffs (note: ReKo is short for ReKo-Norm). In the first dataset, we selected SNPs from the HLA-DQB1 region using a subset Framingham Heart Study samples. We randomly chose 8 SNPs as drivers to simulate phenotypes. The effect sizes were set to 1 but rescaled so that the PVE was 0.20. In the second dataset, we selected a subset of features from UK Biobank samples. We simulated phenotypes using 200 randomly selected features as drivers. The effect sizes were simulated from a standard normal distribution but rescaled so that the PVE was 0.90. Complete simulation details can be found in the Supplementary Material.

Data	Level	Model-X		ReKo	
		⊗glmnet	⊗fastBVSR	⊗glmnet	⊗fastBVSR
HLA-DQB1	0.05	0.48 (0.01)	0.66 (0.02)	0.87 (0.10)	0.92 (0.06)
	0.10	0.68 (0.06)	0.77 (0.04)	0.89 (0.15)	0.95 (0.14)
Proteomics	0.05	0.19 (0.01)	0.29 (0.01)	0.36 (0.05)	0.40 (0.06)
	0.10	0.30 (0.04)	0.35 (0.03)	0.42 (0.10)	0.45 (0.12)

Table 1: Comparison between Model-X and ReKo knockoffs. In each cell we present power and realized FDR (in parenthesis). Simulation results based on the genomic region of *HLA-DQB1* are at the top, where the number of driver SNPs is 8 and PVE is 0.20. Simulation results based on proteomic features are at the bottom, where the number of driver features is 200 and PVE is 0.90.

For each simulated set of phenotypes, we computed knockoff statistics using knockoff filters ⊗glment and ⊗fastBVSR with Model-X and ReKo knockoffs. We therefore compared power and type I error across four knockoff filters: Model-X⊗glment, Model-X⊗fastBVSR, ReKo⊗glmnet, and Reko⊗fastBVSR. Table 1 presents the results. We make the following observations: 1) knockoff filters using reflection knockoff produced

more power than knockoff filters using Model-X knockoffs in both datasets; 2) knockoff filters using fastBVSR produced more power than knockoff filters using glment in both datasets; 3) knockoff filters using reflection knockoffs produced inflated type I error in the HLA-DQB1 dataset. When applied to the UK Biobank dataset, knockoff filter ReKo⊗glmnet appears to have perfect realized FDR, while Reko⊗fastBVSR appears to have slightly inflated realized FDR (see Discussion).

2.3 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 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.

This question can only be addressed by jointly analyzing all features and cannot be answered by single feature tests, as it is best addressed by conditional association instead of marginal association. In other words, a driver association remains significant when conditioned on other associated features, while a hitchhiker does not. 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 reflection knockoff is motivated by such a demand. In the following, we first use simulations to investigate power and type I error of two knockoff filters ReKo⊗glmnet and ReKo⊗fastBVSR, then we analyze the real UK Biobank proteomics dataset.

The correlation structure observed in real proteomic data is difficult to simulate,

Drivers	Level	⊗glmnet	⊗fastBVSR	Combined-Filter
200	0.05	0.80 (0.08)	0.83 (0.13)	0.80 (0.05)
	0.10	0.83 (0.12)	0.86 (0.29)	0.82 (0.09)
400	0.05	0.62 (0.07)	0.67 (0.11)	0.61 (0.05)
	0.10	0.68 (0.11)	0.72 (0.19)	0.67 (0.10)

Table 2: Combined-Filter has calibrated realized FDR. With ReKo knockoffs, both knockoff filters ⊗glment and ⊗fastBVSR have inflated realized FDR, but Combined-Filter have calibrated realized FDR. Column Level denotes nominal FDR. Each cell contains power and realized FDR (in parenthesis).

and we elected to use real data from the UK Biobank (to be described in the next paragraph), and to simulate only phenotypes. We simulated ten replicates of phenotypes for 200 or 400 driver features, which were randomly selected from a total of $1,459$ proteomic features. Their effect sizes were simulated from a standard normal distribution, and rescaled such that the PVE was 0.99. Each simulated phenotype, along with the original features and ReKo knockoffs (independently constructed for each phenotype), were passed to knockoff filters ⊗glmnet and ⊗fastBVSR. The ten sets of knockoff statistics were pooled together, separately for each knockoff filters, to compute power and realized FDR. Table 2 shows that both knockoff filters exhibit inflated realized FDR (see Discussion). The Combined-Filter, which calls a feature positive if both knockoff filters call it positive, however, demonstrates calibrated realized FDR with reduced but comparable power.

Equipped with the knowledge that the Combined-Filter has decent power and well controlled realized FDR, we set out to investigate how many proteomic features are drivers in their association with age. We first computed the kinship matrix using Kindred (Guan and Levy, 2024c), for $14,752$ samples based on genotype data, then regressed out fixed effects (sex and BMI) and the random effect, whose covariance matrix is a linear scaling of the kinship, from both age and the $1,459$ proteomic features using a linear mixed model implemented in IDUL (Guan and Levy, 2024b). The residual

of the proteomic features and age were examined for 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. The proportion aligns with a previous report of $1,944$ out of $2,923$ (or 67%) (Sun et al., 2023).

We generated ten copies of reflection knockoffs, each knockoff copy, together with the original features and phenotype, was passed to two knockoff filters: $\otimes$glmnet and $\otimes$fastBVSR. For each knockoff filter, the ten sets of knockoff statistics were aggregated to call positives (Materials and Methods). The Combined-Filter then called a feature positive if it was called positive by both knockoff filters. According to the Combined-Filter, at a nominal FDR of 0.05 there are 301 out of $1,459$ (or 21%), at a nominal FDR of 0.10 there are 360 out of $1,459$ (or 25%) proteomic features that are significantly associated with age. Thus, we concluded that the majority of proteomic features – among those that are significantly associated with age in a single feature test – are hitchhikers, not drivers.

2.4 Genetic Fine-mapping

Genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) analysis are effective at identifying genomic regions associated with phenotypes. However, due to linkage disequilibrium, 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. Owing to its focus on single effect and approximate computation, SuSiE is also computation-

ally efficient. It has been shown, however, that the calibration of PIP of SuSiE remains challenging in real data analysis (Cui et al., 2024).

We begin by examining the power and calibration of SuSiE, and whether a knockoff filter using SuSiE as the association method ReKo⊗SuSiE⊗PIP outperforms SuSiE. Our simulation study used both simulated features and real SNPs extracted from the genomic region *HLA-DQB1*. Since SuSiE assumes a small number of driver variants in each locus, we simulated phenotypes with 8 drivers (more simulation details can be found in Supplementary).

Table 3 summarizes the results. We make the following observations: 1) Features extracted from *HLA-DQB1* have a smaller median absolute correlation of 0.10 compared to simulated features, whose median absolute correlations is 0.5. This largely accounts for the difference in power. 2) For simulated features, the knockoff filter ReKo⊗SuSiE ⊗PIP shows improved power over SuSiE. Without knockoffs, SuSiE outputs PIPs either near 1 or near 0, as if it has infinite amount of data. Consequently all realized FDRs for SuSiE are near 0 irrespective of the nominal FDR level. In constrast, the knockoff filter produced more relaxed realized FDR values, and this likely explains the power gain. 3) When it comes to the real dataset with features extracted from the *HLA-DQB1* region, however, the knockoff filter ReKo⊗SuSiE⊗PIP showed inflated realized FDR (see Discussion), and this inflation is more severe than that observed for ReKo⊗glmnet ⊗Beta and ReKo⊗fastBVSR⊗PIP (Table 1).

We know that the Combined-Filter is effective controlling realized FDR for proteomic features (Table 2). We thus investigated whether the Combined-Filter is effective for fine mapping. We used features extracted from two real genomic region *PPIEL* and *HLA-DQB1* to simulate phenotypes and examined power and realized FDR. Simulation details can be found in Supplementary. Figure 2 contains the results for *PPIEL*,

Region	Level	SuSiE	ReKo⊗SuSiE⊗PIP
Simulated	0.05	0.46 (0.00)	0.54 (0.03)
	0.10	0.47 (0.00)	0.57 (0.07)
HLA-DQB1	0.05	0.75 (0.03)	0.86 (0.16)
	0.10	0.78 (0.04)	0.87 (0.21)

Table 3: Comparison between SuSiE and knockoff filter ReKo⊗SuSiE⊗PIP. For the simulated features, we let $n = 2000$, $p = 800$, and the number of drivers is 8. The effect sizes are equal between selected features, but rescaled so that the PVE equals 0.2. For the real features extracted from *HLA-DQB1* region, we used the same simulated data with ReKo from Table 1.

and Supplementary Fig. S1 contains results for *HLA-DQB1*. We make the following observations: 1) both knockoff filters, ReKo⊗glmnet and Reko⊗fastBVSR, exhibit substantially inflated type I error; 2) the Combined-Filter is able to bring type I error under control, provided that the correlation threshold does not exceed 0.90 or 0.95; 3) the Combined-Filter indeed works for fine mapping, showing an appreciable power gain over SuSiE; 4) As the correlation threshold increases, the power of both Combined-Filter and SuSiE decreases. In addition, the Combined-Filter showed increasing type I error with the increasing correlation thresholds.

In a recent study, Guan and Levy (2024a) discovered abundant eGenes that harbor parent-of-origin effect (POE) eQTL. When an eGene harbors both genotype eQTL and POE eQTL, however, one may wonder whether POE eQTL are driver signals. Fine mapping can help identify drivers versus hitchhikers. An additional challenge in this analysis, however, is that collinearity is introduced into the features when genotypes are separated into paternal and maternal alleles. Our simulation studies suggest that all methods suffer from deflated power and inflated type I error in the presence of collinearity. We therefore elected to remove genotypes and only use paternal and maternal alleles in fine mapping to mitigate the collinearity.

From the same data used in (Guan and Levy, 2024a), we selected five eGene exam-

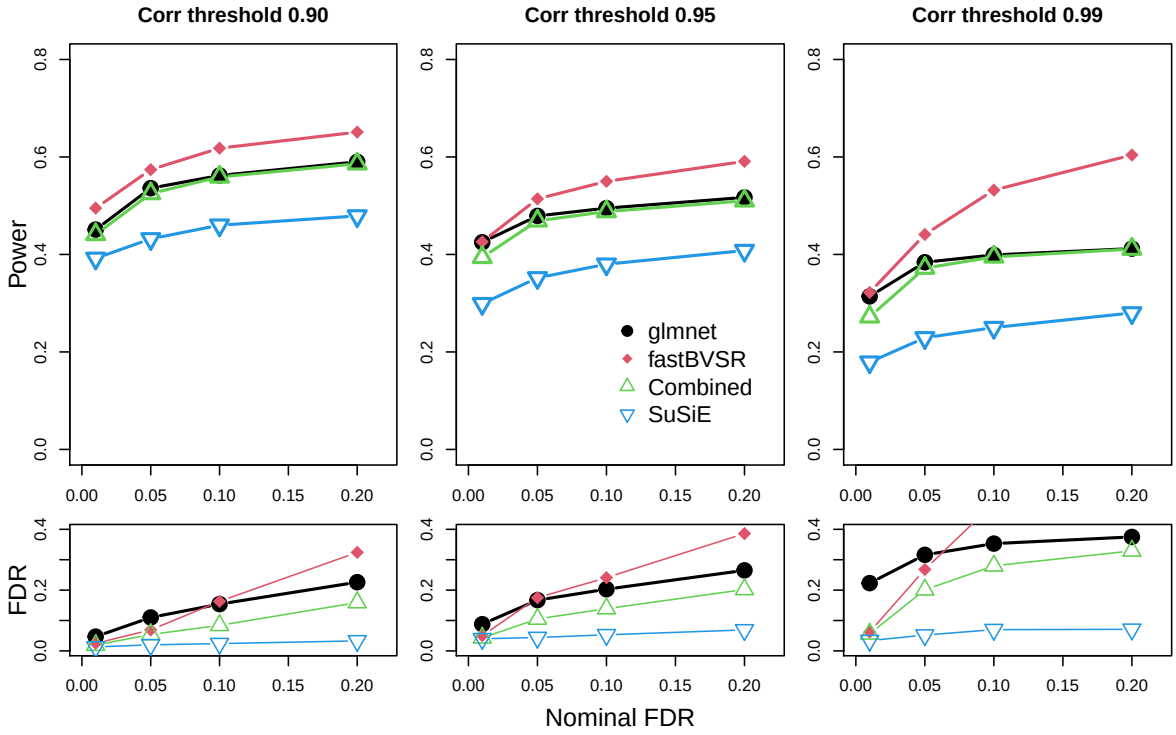

Figure 2: Combined-Filter vs SuSiE. Features were extracted from genomic region of *PPIEL*. Column Corr denote genotype correlation threshold used to trim features.

ples, *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*, which harbors opposing eQTL, with paternal and maternal alleles having opposite effect sizes. For each eGene, we regressed out fixed effects including age, sex, BMI, and cell 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 were always present as pairs.

eGene	eQTL	fastBVSR	glmnet	SuSiE	Type
PPIEL	rs79473113:C:A:1:39554348:1	1.00	0.19	1.00	1
	rs7520271:A:C:1:39579041:1	0.89	0.03	0.97	1
CCR9	rs4683146:G:A:3:45856729:1	0.84	0.05	0.98	1
GJB6	rs12875121:T:C:13:20478719:1	0.97	0.07	0.99	1
	rs12875121:T:C:13:20478719:0	0.80	0.07	0.88	1
	rs9506450:G:C:13:20239478:0	0.56	0.04	0.50	0
	rs9579911:C:G:13:20637437:0	0.33	0.03	0.15	0
	rs7991441:G:A:13:20475149:0	0.05	0.03	0.01	0
COA8	rs28513222:A:T:14:103561732:0	0.01	-0.00	0.05	2
	rs861528:C:T:14:103716661:0	0.01	-0.00	0.04	2
	rs10143623:G:A:14:103720539:0	0.56	0.03	0.87	2
NECAB3	rs4911352:T:G:20:33640756:0	0.00	0.00	0.01	2
	rs2626556:A:G:20:33768509:0	0.68	0.04	0.95	2

Table 4: Fine mapping of selected eGenes. 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 fastBVSR and glmnet contain average knockoff statistics from 10 replicates of ReKo knockoffs, for knockoff filters ⊗fastBVSR and ⊗glment respectively. Column SuSiE contains PIP from SuSiE. Column Type contain three possible values: 1 if an eQTL is significant in combined knockoff filter and in the credible sets reported by SuSiE; 0 if an eQTL is significant in combined knockoff filter, but not reported in SuSiE credible sets; 2 if an eQTL is reported in SuSiE credible sets, but not significant in combined knockoff filter.

We analyzed each eGene in the same manner as we analyzed the UK Biobank proteomics dataset. For each eGene, we constructed ten copies of reflection knockoffs to produce ten sets of knockoff statistics separately for two knockoff filters ReKo⊗glmnet and ReKo⊗fastBVSR. For each knockoff filter, the ten sets of knockoff statistics were aggregated to call positives at a nominal FDR level of 0.1 (Materials and Methods). The Combined-Filter called a feature positive when it was called by both knockoff filters as positive. We also ran SuSiE without knockoff as a comparison.

Table 4 contains the results of fine mapping. We make the following observations: 1) For genes *PPIEL* and *CCR9*, the Combined-Filter and SuSiE agreed, identifying paternal eQTL as drivers in both genes. 2) For the gene *GJB6*, both paternal and maternal alleles of SNP *rs12875121* were found to be significant in fine mapping, and

their effect sizes were close to each other. In fact, this SNP is known to be a genotype eQTL. 3) The Combined-Filter called three additional maternal eQTL significant for *GJB6*, but SuSiE missed these maternal signals. 4) For Genes *COA8* and *NECAB3*, the Combined-Filter returned no significant findings, but both genes had significant maternal eQTL identified by SuSiE. *NECAB* is known to harbor exclusively opposing eQTL, but SuSiE reported significant maternal eQTL without its companion paternal eQTL.

3 Discussion

In this paper, we presented a novel method to construct reflection knockoffs via Householder reflection and demonstrated the exceptional power that reflection knockoffs bring to knockoff filters. We also provided a simple method to aggregate multiple sets of identically distributed knockoff statistics to improve the consistency of knockoff filters. We applied reflection knockoffs to address two problems, both of which involve highly correlated features. In the analysis of proteomic signatures of age, using the Combined-Filter of ReKo⊗glmnet⊗Beta and ReKo⊗fastBVSR⊗PIP, we demonstrated that the majority of proteins showing significance in single-feature tests are hitchhikers, not drivers. In genetic fine mapping, we demonstrated that the Combined-Filter outperformed SuSiE and in three out of five selected eGenes, POE eQTL were found to be drivers. We documented the computational time required to generate knockoffs and fit models with different methods in the Supplementary Material.

Reflection knockoffs significantly enhance the power of knockoff filters. They will benefit other knockoff-based methods proposed for genetic association studies including (Candès et al., 2018; He et al., 2021). Our method for aggregating multiple sets of

identically distributed knockoff statistics can also benefit methods that employ different approaches to construct knockoffs, including (Sesia et al., 2019, 2020, 2021), and perhaps even (He et al., 2022).

For highly correlated features, Fixed-X and Model-X knockoffs have little power, but the current method to construct reflection knockoffs also faces difficulties. The ghost observations that we used to combat high correlation when computing the Householder reflection are in fact associated with a prior (Materials and Methods). If the prior is too strong, it may skew the knockoff distribution to a degree that violates the assumption of the knockoff filter, consequently producing inflated realized FDR, while the knockoff filter is designed to have calibrated FDR. On the other hand, if the prior associated with the ghost observation is too weak, the reflection knockoffs will lose power. As a future research direction, it might be worthwhile to optimize the prior associated with the ghost observation through cross-validation.

To combat high correlation when constructing reflection knockoffs, an alternative approach to using ghost observations 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 approach could provide another fruitful research topic to 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.

The knockoff filter is an intrinsic Frequentist procedure. In both applications, we incorporated fastBVSR into a knockoff filter in an attempt to improve its calibration. In both cases, we used PIP as a measure of importance. Similar to computing p-values for Bayes factors (Zhou and Guan, 2018), using a knockoff filter with PIP as the importance measure can be perceived as another example of a Bayesian non-Bayesian compromise (Good, 1992). A Bayesian procedure such as fastBVSR or SuSiE may produce PIPs that are not calibrated due to, say, difficulties in estimating priors odds in applications involving highly correlated features. It was our hope that the knockoff filter could improve the calibration of PIP. We demonstrated its effectiveness with simulated features, but encountered difficulties with real datasets. One notable benefit of the knockoff filter, however, is its ability to combine multiple knockoff filters, as demonstrated by the Combined-Filter.

Both knockoff filters ReKo$\otimes$glmnet$\otimes$Beta and ReKo$\otimes$fastBVSR$\otimes$PIP produced inflated realized FDR when features were highly correlated. However, the Combined-Filter appeared to be effective in bringing the realized FDR under control. 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 will assign each feature a PIP of about 0.25, by the virtue of model averaging. 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 thus has advantages in both scenarios for controlling false positives.

The true potential of reflection knockoffs lies in their ability to revolutionize 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 – go push the boundaries. With reflection knockoffs, genetic research can embrace this model, empowering the community to collectively tackle the most pressing problems with creativity and precision.

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* $\mathbf{Z} = \frac{1}{2}(\mathbf{X} - \mathbf{Y})$*. 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 $\mathbf{T} = cov(\mathbf{X}, \mathbf{Z})$*, then* $\mathbf{T} = diag(cov(\mathbf{Z})) > \mathbf{0}$*.*

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

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 covariance of $\mathbf{Y}_j$ and $\mathbf{X}_j$ is given by $\text{cov}(\mathbf{Y}_j, \mathbf{X}_j) = \text{cov}(\mathbf{X}_\mathbf{V} - \mathbf{X}_\perp, \mathbf{X}_\mathbf{V} + \mathbf{X}_\perp) = \text{cov}(\mathbf{X}_\mathbf{V}, \mathbf{X}_\mathbf{V}) - \text{cov}(\mathbf{X}_\perp, \mathbf{X}_\perp)$. Using orthogonality, $\text{cov}(\mathbf{X}_\mathbf{V}, \mathbf{X}_\perp) = \text{cov}(\mathbf{X}_\perp, \mathbf{X}_\mathbf{V}) = 0$, so: 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: $\mathbf{X}(\mathbf{X}^t\mathbf{X})^t\mathbf{X}\mathbf{Y} = \mathbf{Y}$ and $\mathbf{X}(\mathbf{X}^t\mathbf{X})^t\mathbf{X}\mathbf{Z} = \mathbf{Z}$. □

4.2 Construction

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:

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

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:

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

Let $\mathbf{Q}\Lambda\mathbf{Q}^t$ be the eigendecomposition of $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$. 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, the diagonal matrix

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

must also be positive semi-definite. This requires $\alpha \leq 2/\Lambda_1$, where Λ_1 is the leading eigenvalue of $\text{cov}(\mathbf{Z}\mathbf{T}^{-1/2})$. Since Λ_1 is positive and finite, and $\alpha \in (0, 1]$, and 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, $\mathbf{S} = \hat{\alpha}\mathbf{T} \leq 2\ \text{cov}(\mathbf{X})$, 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.

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 $\mathbf{X}$.

By introducing ghost observations that correspond to a shrinkage prior, we can extend this construction to handle cases where $p > n$. More details on this are provided below.

4.3 Computation

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, 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 $\sigma^2\mathbf{I}_p$ such that $\mathbf{X} \leftarrow [\mathbf{X}^t, \sigma^2\mathbf{I}]^t$. When columns of $\mathbf{X}$ are centered and scaled, we find $\sigma = \sqrt{p}$ to be a balanced choice. 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.

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 $\mathbf{R}_j = \mathbf{R}_{,-j}\beta_{-j} + \mathbf{e_Q}$, 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 beging 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 does not alter the error structure. 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 $\sigma^2\mathbf{I}_p$, 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 $\mathbf{Y}[1 : n,]$

4.4 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:

$$T = \min \left\{ t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}. \tag{12}$$

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* (12) *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 12 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 (12) 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. □

Intuition. Concatenation increases the effective sample size and provides a more stable empirical estimate of the distribution of null and non-null features. In expectation, 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 hypthesis, if $W_j^{(m)}$ is symmetric around zero for each m, their average $\bar{W}_j$ remains symmetric around zero. Consequently, $\bar{W}$ 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.

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).

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.

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.

Cui, R., R. A. Elzur, M. Kanai, J. C. Ulirsch, O. Weissbrod, M. J. Daly, B. M. Neale, Z. Fan, and H. K. Finucane (2024, January). Improving fine-mapping by modeling infinitesimal effects. *Nat. Genet. 56*(1), 162–169.

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.

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.

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 (2021). 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.

Sesia, M., S. Bates, E. Candès, J. Marchini, and C. Sabatti (2021, October). False discovery rate control in genome-wide association studies with population structure. *Proc. Natl. Acad. Sci. U. S. A. 118*(40), e2105841118.

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, Å. 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.

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.