

# Searching for consistent associations with a multi-environment knockoff filter

Shuangning Li<sup>\*1</sup>, Matteo Sesia<sup>\*2</sup>, Yaniv Romano<sup>3</sup>, Emmanuel Candès<sup>4</sup>, and Chiara Sabatti<sup>5</sup>

## Abstract

This paper develops a method based on model-X knockoffs to find conditional associations that are consistent across diverse environments, controlling the false discovery rate. The motivation for this problem is that large data sets may contain numerous associations that are statistically significant and yet misleading, as they are induced by confounders or sampling imperfections. However, associations consistently replicated under different conditions may be more interesting. In fact, consistency sometimes provably leads to valid causal inferences even if conditional associations do not. While the proposed method is flexible and can be deployed in a wide range of applications, this paper highlights its relevance to genome-wide association studies, in which consistency across populations with diverse ancestries mitigates confounding due to unmeasured variants. The effectiveness of this approach is demonstrated by simulations and applications to the UK Biobank data.

**Keywords**— Conditional independence, causality, false discovery rate, genome-wide association studies.

## 1 Introduction

A critical goal of statistics is to discover which variables, among the many measured in big-data applications, are meaningfully associated with an outcome of interest. The word “association” may have different connotations, ranging from *marginal* association, a tendency of two variables to vary together, to *causal* association, a relation ensuring interventions on one variable affect another. For example, a genome-wide association study may ascertain that some genetic variants occur more frequently among individuals with diabetes. An explanation for this marginal association may be that the discovered variants are biologically irrelevant but are shared, because of common ancestry, by a sub-population that happens to follow a less healthy diet [1]. Of course, it would be more actionable to identify variants involved in biological processes which, if modified by a drug, could influence the disease. Marginal associations are the simplest to recognize but also the least informative, while causal associations are more elusive, especially with high-dimensional observational data, although they better lend themselves to scientific interpretations.

Between marginal and causal associations one finds *conditional* association: the tendency of two variables to vary together when other quantities are fixed. Conditional testing has been traditionally tackled through parametric models; however, these require strong assumptions, which are not always justified. A new “model-X” strategy was proposed by [2], making no assumptions about the conditional distribution of the outcome and approximating instead the joint distribution of the predictors. This has led to two methods, *knockoffs* [2, 3] and the *conditional randomization test* [2], which can harness the power of any machine learning algorithm while providing type-I error guarantees in finite-samples. The model-X assumptions are particularly well-suited to genome-wide associations studies because reliable prior knowledge is available about the joint distribution of the explanatory variables [4–6], but the framework is quite robust to approximations [2, 7] and thus broadly applicable.

Despite the above advantages, conditional testing is not fully satisfactory for at least three reasons. First, it does not account for unmeasured *confounders*: missing variables which, if conditioned upon, would explain away the

<sup>\*</sup>Equal contribution.

<sup>1</sup>Department of Statistics, Stanford University, USA.

<sup>2</sup>Department of Data Sciences and Operations, University of Southern California, USA.

<sup>3</sup>Departments of Electrical Engineering and of Computer Science, Technion, Israel.

<sup>4</sup>Departments of Statistics and of Mathematics, Stanford University, USA.

<sup>5</sup>Departments of Statistics and of Biomedical Data Sciences, Stanford University, USA.

association [8]. For instance, one may discover that an irrelevant genetic marker is associated with a disease simply because it is physically close to, and thus inherited alongside, an unobserved causal variant [9]. Second, some data sets may be collected from a population that does not match exactly the target one, either because of accidental sampling bias [10], or for convenience [11]. Clearly, any conditional associations may be misleading in that case. Third, conditional testing methods typically assume individual samples are independent of one another, and this can lead to spurious associations in the presence of unexpected dependencies, such as network effects [12].

This paper extends the methodology of knockoffs to mitigate the three aforementioned limitations of conditional testing by analyzing data from many *environments*. The word “environment” is employed loosely here, referring to specific sub-populations, experimental settings, or data collection strategies depending on the context. Our work is motivated by the conjecture that the most informative associations are those which can be consistently reproduced under different environments, because these tend to enable more generally reliable predictions and may even reflect scientifically illuminating causal relations. This is an old idea, dating back at least to Hume [13].

“There is no phaenomenon in nature, but what is compounded and modified by so many different circumstances, that in order to arrive at the decisive point, we must carefully separate whatever is superfluous, and enquire by new experiments, if every particular circumstance of the first experiment was essential to it. These new experiments are liable to a discussion of the same kind; so that the utmost constancy is required to make us persevere in our enquiry, and the utmost sagacity to choose the right way among so many that present themselves.”  
(Hume [13])

We will translate the above logic into a practical method for the analysis of high-dimensional data, provably controlling the false discovery rate [14] for hypotheses of consistent conditional association. The proposed solution utilizes knockoffs because these are computationally efficient in very high dimensions and allow controlling the false discovery rate even if the predictors have strong dependencies [2]. However, an alternative approach based on the conditional randomization test [2] would be easy to implement; see Appendix A and Figure A1 therein.

The outline is the following. Section 2 states the problem. Section 3 proves that testing our hypotheses sometimes leads to causal inferences. Section 4 develops our methods to test the aforementioned hypotheses using knockoffs. Section 5 dives into genome-wide association studies [4], explaining how consistency across sub-populations mitigates confounding due to unobserved variants. Section 6 validates empirically our method with simulations. Section 7 applies it to analyze the genetic determinants of several phenotypes in the UK Biobank resource [15] using 600k genotypes from 500k individuals. Section 8 concludes by discussing some opportunities for future research.

## Related work

This paper was inspired by [16, 17], which advanced invariance across environments as a framework for causal inference. Departing from their work, however, we do not assume the existence of an invariant model with homogeneous effects; indeed, our approach is fully non-parametric and leads to meaningful inference without reference to a specific causal model. Further, we seek different guarantees: the method in [16] searches for a conservative confidence set of possible causal predictors, which may be desirable if the number of variables is small but does not scale well to high-dimensions, while ours controls the false discovery rate and can achieve high power even with hundreds of thousands of variables. Notions of invariance similar to that of [16] have also been utilized to improve predictive accuracy in the face of changes in the distribution of the explanatory variables [18–21], a phenomenon known also as *covariate shift* [22, 23]; although the problems are related, this paper concentrates on testing rather than prediction.

Our problem is related to *causal discovery* [8, 24–28], which is the challenge of learning the graph describing the relations between all variables in a system, without pre-specifying an outcome of interest as we do; this has also been extended to leverage invariance across environments [29]. Those methods can discover the direction of causal relations, as opposed to ours which solely tests conditional independence, but they require parametric assumptions and provide asymptotic rather than finite-sample guarantees. There has also been interest in invariance within the feature selection literature [30–32], though typically without seeking finite-sample inferences; see [33] for a review.

Knockoffs were introduced by [3] and later extended by [2] to the high-dimensional model-X framework. Subsequently, algorithms were developed to construct knockoffs for different distributions of predictors [4, 7, 34, 35], while others studied robustness to model misspecifications [36] and power [37–40]. Our work is orthogonal, as we extend the knockoff filter [3] to analyze data from many environments. We focus on applications to genome-wide association data [4], for which prior efforts addressed the problems of accounting for dependencies across the genotyped markers [5], population structure [6], and even other unmeasured confounders [41], but did not deal with missing variants. Knockoffs have also been deployed in other fields [42–45] for which our methodology may be helpful.

## 2 Conditional associations that hold across environments

Consider  $E$  environments, or experimental settings, from which one can sample observations  $(X, Y)$  consisting of  $p$  explanatory variables,  $X \in \mathcal{X}^p$ , and an outcome,  $Y \in \mathcal{Y}$ . Here,  $\mathcal{X}$  is the set of possible values for each variable and  $\mathcal{Y}$  denotes the possible outcomes; both sets may be either discrete or continuous. Assume the joint distribution of the explanatory variables within any environment  $e \in \{1, \dots, E\}$ ,  $P_X^e$ , is known. For simplicity, we imagine different samples as being independent of one another, although knockoffs can accommodate known patterns of dependency [6]. The model-X framework [2] provides practical methods to test the conditional independence hypothesis

$$\mathcal{H}_j^{\text{ci},e} : Y^e \perp\!\!\!\perp X_j^e \mid X_{-j}^e, \quad (1)$$

for any  $j \in \{1, \dots, p\}$ . Here,  $X_{-j}$  denotes all observable explanatory variables except  $X_j$ , and the superscript  $e$  clarifies we are focusing on the distribution of the data in the  $e$ -th environment.

Our goal is to powerfully test, for all  $j \in \{1, \dots, p\}$ , the following *consistent* conditional independence hypothesis,

$$\mathcal{H}_j^{\text{cst}} : \exists e \in \{1, \dots, E\} \text{ such that the null } \mathcal{H}_j^{\text{ci},e} \text{ in (1) is true,} \quad (2)$$

controlling the false discovery rate. Intuitively, we would interpret any findings by noting that, if  $\mathcal{H}_j^{\text{cst}}$  (2) is false, the conditional association of  $X_j$  with  $Y$  *must hold across all environments*.

It may be tempting to test  $\mathcal{H}_j^{\text{ci},e}$  (1) separately environment by environment [2] and then report the set of common discoveries. Unfortunately, this *intersection* heuristic would not control the false discovery rate for  $\mathcal{H}_j^{\text{cst}}$  (2) even if all environment-specific tests control it for  $\mathcal{H}_j^{\text{ci},e}$  (1) [46]. Alternatively, one may apply the standard knockoffs methodology [2] on the *pooled* data from all environments. This would control the false discovery rate for  $\mathcal{H}_j^{\text{ci}}$  (1) defined in the broader population obtained by taking the union of all environments, but is not a test of consistency. Indeed, the problem is non-trivial, as illustrated by the simulations in Figure 1. These preview that our proposed method controls the false discovery rate for  $\mathcal{H}_j^{\text{cst}}$  (2), unlike the two aforementioned heuristics, and achieves relatively high power. These simulations will be explained with more details in Section 6.1, after we develop our method.

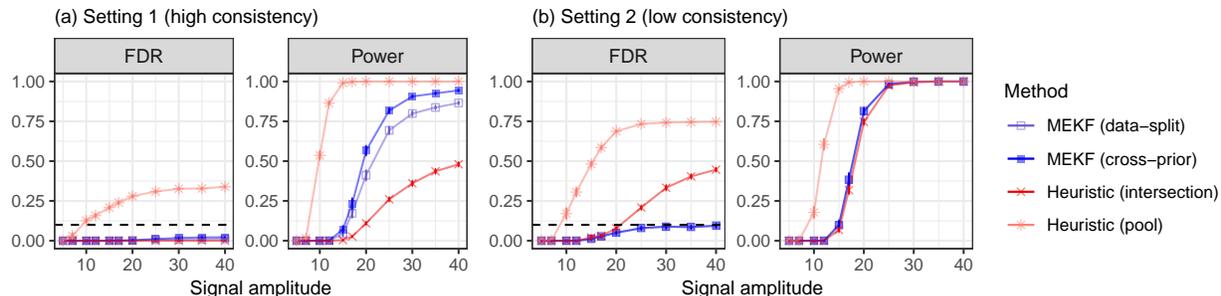


Figure 1: Performance of the multi-environment knockoff filter (MEKF) on simulated data from many environments, implemented with two alternative statistics. The performance is measured in terms of consistent conditional associations and is compared to that of two heuristics. The nominal false discovery rate is 0.1 (dashed line). (a) Data in which most conditional associations are consistent. (b) Data in which most conditional associations are not consistent.

A possible limitation of  $\mathcal{H}_j^{\text{cst}}$  (2) is that it becomes harder to reject if the number of environments grows but the sample size in each remains constant, because every one must provide evidence against the null. Such intransigence is not always necessary; sometimes, it may be satisfactory to find an association in most environments, especially if some have smaller sample sizes. This idea motivates *partial consistency* testing, also known as *partial conjunction* [47–50]. For any variable  $j \in \{1, \dots, p\}$  and parameter  $1 \leq r \leq E$ , we define the partial consistency null hypothesis as:

$$\mathcal{H}_j^{\text{pcst},r} : \left| \left\{ e \in \{1, \dots, E\} : \mathcal{H}_j^{\text{ci},e} \text{ is true} \right\} \right| > E - r. \quad (3)$$

In words,  $\mathcal{H}_j^{\text{pcst},r}$  states there are strictly less than  $r$  environments  $e$  in which  $\mathcal{H}_j^{\text{ci},e}$  (1) is false; thus, a rejection suggests  $X_j$  is associated with  $Y$  in at least  $r$  environments. This generalizes  $\mathcal{H}_j^{\text{cst}}$  (2), which is recovered if  $r = E$ . Note that, unlike [48], we will not account for multiplicity over different possible  $r$ , which we take instead as fixed.

Before outlining the methodology we propose to test the consistent conditional independence hypotheses  $\mathcal{H}_j^{\text{cst}}$  (2), we pause to motivate their interest. With this in mind, the next section will establish a link between  $\mathcal{H}_j^{\text{cst}}$  (2) and

causal inference, under some specific assumptions. We want to underscore, however, that identifying variables which are not independent from an outcome conditionally on all other observed sources of variation can be a powerful tool, even outside of the causal framework we will describe.

Consider, as an example, the case where  $Y$  is a measure of the quantitative competences acquired by students in public schools K-12, and let  $X$  collect a number of variables that could possibly be related to the outcome, as student/teacher ratio, class size, availability of tutoring, school meals, type of curriculum, teacher qualifications, family size, attendance, etc. It is entirely possible that these variables interact in different ways to influence the final student competences in different environments (e.g. urban vs. rural, low vs. high income neighborhoods, ethnically diverse vs. homogeneous population, and so on). Without hoping to estimate constant causal effects, it would be of interest to be able to identify those variables which show a conditional association with the outcomes that holds across all environments. It is these “robust” associations that one looks to in order to design interventions with the goal of improving outcomes, or detecting early signs of academic difficulties. While “one size fits all” is certainly wishful thinking, considerations of practicality, transparency and fairness make it preferable to focus on policies that have good chances of being effective across the board. Moreover, constant associations are also more likely to be robust to changes in environments or covariate shifts, such as those that may be expected with the passing of time.

More generally, an “environment” can be understood as a specific population, as that typically observed in a data set. The goal of science is to make conclusions that are not valid only in one data set, but that hold more generally. Yet, it is well known that when looking for patterns among a very large number of variables, collected without much prior selection, it is possible to fit models that relate  $X$  to  $Y$  with a precision that is hard to replicate in other data sets. We routinely carry out “cross-validation” to mitigate this problem, but this still only guarantees a form of *internal* consistency. It might very well be that among some Facebook users taken at a certain point in time, liking curly fries is a good predictor of IQ [51] but this association is probably not robust to different times and settings. Testing for  $\mathcal{H}_j^{\text{cst}}$  (2) can be considered as a way to do “*external* cross-validation” [52], thereby identifying models that are less ephemeral [53].

### 3 From consistent associations to causal inferences

We now explicitly investigate the connection between consistent associations and causal inference. We will assume a constant causal model across environments—an assumption that is by no means necessary for the interest or validity of the tests we propose, but that is helpful to illustrate a number of ways in which focusing on conditional associations that are consistent across environments facilitates the discovery of variables with causal effects.

#### 3.1 A constant causal model

Assume a structural equation model [54] to describe the relation between the outcome ( $Y$ ) and  $p_z$  explanatory variables ( $Z$ ), of which  $p$  are observed ( $X$ ) and  $p_c$  are unobserved ( $C$ ). Thus, we can write  $Z = (X, C) \in \mathcal{X}^{p_z}$ , with  $p_z = p + p_c$ . According to this model, which we assume to be constant across environments, the  $i$ -th individual outcome,  $Y^{(i)}$ , is determined as a function of  $Z$  through:

$$Y^{(i)} = \bar{f}(Z^{(i)}, V^{(i)}), \tag{4}$$

where  $\bar{f}$  is unknown and  $V$  is exogenous noise from a standard uniform distribution, for example. Assume the causal direction in this model is known:  $Y$  is determined by some combination of  $Z$  and  $V$ , not the other way around. Although this simplification does not always make sense, it is appropriate in genetic studies, for example, because the genotypes predate the phenotype. With this in place, consider the goal of discovering which variables are causal, in the sense that they have a direct effect on the outcome and do not satisfy the following *sharp* causal null hypothesis

$$\mathcal{H}_j^{\text{causal}} : \bar{f}((z_1, \dots, z_{j-1}, z_j, z_{j+1}, \dots, z_{p_z}), v) = \bar{f}((z_1, \dots, z_{j-1}, z'_j, z_{j+1}, \dots, z_{p_z}), v), \quad \forall z, z', v. \tag{5}$$

Intuitively, this tells us that intervening on  $Z_j$  while holding all other variables fixed would have no effect on  $Y$  at all; hence the adjective “sharp”. Alternatively, one may think of  $\mathcal{H}_j^{\text{causal}}$  as saying the potential outcomes are identical under all  $Z_j$  [55]. Sharp hypotheses have received some criticism in the causal inference literature [56], specifically for not describing heterogeneous treatment effects [57], but these concerns are more relevant from an estimation perspective. From a testing perspective, sharp hypotheses are helpful to discover which variables are more likely to be causal, especially for an exploratory analysis informing the design of follow-up studies [58]. Further, the model in (4) remains very flexible despite being constant across environments because it is fully non-parametric. In

particular, as it allows any causal relations to be complex and involve interactions, this model does not exclude that a variable may appear to have different *linear* effects on the outcome across environments with covariate shifts. Indeed, the sole purpose of the model in (4) is to allow a manageable definition of causality; in practice, our method simply tests whether a variable has *some* influence on the outcome in all environments.

In the following, we will assume we have data samples from  $E$  environments, each corresponding to a distribution  $P_Z^e$  of explanatory variables, for  $e \in \{1, \dots, E\}$ . Conditional on  $Z$ , the outcome  $Y$  is generated by the constant model in (4). For simplicity, we equivalently rewrite this model as a function only of the causal variables, which we list as  $\text{Pa}(Y) \subseteq \{1, \dots, p_z\}$  (standing for *parents of Y*):

$$\text{Pa}(Y) = \left\{ j \in \{1, \dots, p_z\} : \mathcal{H}_j^{\text{causal}} \text{ in (5) is false} \right\}. \quad (6)$$

In particular, we write

$$Y^{(i)} = f \left( Z_{\text{Pa}(Y)}^{(i)}, V^{(i)} \right), \quad (7)$$

where  $f$  is the restriction of  $\bar{f}$  on  $\text{Pa}(Y)$ . To lighten the subsequent notation, we partition  $\text{Pa}(Y)$  into:

$$\text{Pa}_x(Y) := \{j \in \{1, \dots, p\} : j \in \text{Pa}(Y)\}, \quad \text{Pa}_c(Y) := \{j \in \{1, \dots, p_c\} : (p+j) \in \text{Pa}(Y)\};$$

in words, these are the observed and unobserved causal variables, respectively. Of course, in practice we analyze only the former, seeking to make inferences about  $\text{Pa}_x(Y)$ , because we have no data involving the latter.

### 3.2 The gap between conditional testing and causal inference

We begin by focusing on a single environment  $e$ . The result below states that testing  $\mathcal{H}_j^{\text{ci},e}$  in (1) amounts to making causal inferences if there is no confounding, in the sense that either no unobserved variables are causal, or the unobserved causal variables are independent of the observed ones.

**Proposition 1.** *Fix an environment  $e$  and an observable variable  $j$ . Assume either (i)  $\text{Pa}_c(Y) = \emptyset$  or (ii)  $X_j^e \perp\!\!\!\perp C_{\text{Pa}_c(Y)}^e \mid X_{-j}^e$ . Then, under the structural equation model and the data sampling scheme described in Section 3.1, the causal null hypothesis  $\mathcal{H}_j^{\text{causal}}$  (5) implies the conditional independence null hypothesis  $\mathcal{H}_j^{\text{ci},e}$  (1).*

*Proof.* If  $\mathcal{H}_j^{\text{causal}}$  (5) is true,  $Y$  is a function of  $Z_{\text{Pa}(Y)}$  and  $V$  (7), with  $j \notin \text{Pa}(Y)$ . (We drop the superscript  $e$  for simplicity.) Suppose assumption (i) holds:  $Z_{\text{Pa}(Y)} = X_{\text{Pa}(Y)}$ . As  $Y$  is a function of  $X_{\text{Pa}(Y)}$  and  $V$ , it is independent of  $X_j \mid X_{-j}$  because  $j \notin \text{Pa}(Y)$  and  $V \perp\!\!\!\perp X$ ; thus  $\mathcal{H}_j^{\text{ci},e}$  (1) is true. Suppose instead (ii) holds. We may still write  $Y$  as a function of  $C_{\text{Pa}_c(Y)}$ ,  $V$ , and  $X_{\text{Pa}_x(Y)}$ . Thus,  $Y \perp\!\!\!\perp X_j \mid X_{-j}, C_{\text{Pa}_c(Y)}, V$  because  $j \notin \{1, \dots, p\} \cap \text{Pa}(Y)$ . Then, it follows from the contraction property of conditional independence that  $(Y, C_{\text{Pa}_c(Y)}, V) \perp\!\!\!\perp X_j \mid X_{-j}$ , implying  $\mathcal{H}_j^{\text{ci},e}$  (1).  $\square$

The assumptions of Proposition 1 are strong. The first one would require one to have measured every variable with a possible effect on the outcome; this is often unrealistic, especially when studying complex phenomena. The second one holds in randomized experiments, and may be justified in certain observational studies such as those involving genetic parents-child trio data [41]. However, without suitable domain knowledge as in [41], it is generally unclear why all potentially relevant missing variables should be conditionally independent of the observed ones. Therefore, even if the structural equation model in Section 3.1 is acceptable, it remains very challenging to draw causal inferences from conditional associations. Further, true random samples from the population of interest are not always easy to obtain; in fact, the process of gathering data is often far from ideal, as it may involve unknown sampling biases [10, 11, 59] or network effects [12, 60]. Searching for consistent associations will not fully resolve these difficulties, but it can mitigate them if the environments are sufficiently different from one another, as discussed next.

### 3.3 Consistency improves robustness to missing variables

The next simple result demonstrates that the assumptions under which conditional associations yield causal inferences can be relaxed if the associations are consistent across many environments. The intuition is that covariate shifts (changes in  $P_Z^e$ ) may induce different observed variables to pick up spurious associations in different environments, while causal associations tend to be consistent. Figure 2 visualizes this concept with a toy example involving two causal variables, one of which is unmeasured. Here, the two environments differ in  $P_Z^e$  to a sufficient extent that their spurious associations have no overlap and can thus be perfectly winnowed down through consistency. Section 5 will explain the relevance of this idea to genome-wide association studies involving individuals with different ancestries.

**Proposition 2.** Fix any observable variable  $j$  and consider  $E$  environments  $\{1, \dots, E\}$ . Assume either (i)  $\text{Pac}(Y) = \emptyset$  or (ii)  $\exists e \in \{1, \dots, E\} : X_j^e \perp\!\!\!\perp C_{\text{Pac}(Y)}^e \mid X_{-j}^e$ . Then, under the same setting as in Proposition 1, the causal null hypothesis  $\mathcal{H}_j^{\text{causal}}$  in (5) implies the consistent association null hypothesis  $\mathcal{H}_j^{\text{cst}}$  in (2).

*Proof.* Suppose either assumption (i) or (ii) is satisfied. Then, Proposition 1 implies  $\exists e \in \{1, \dots, E\}$  such that the conditional null  $\mathcal{H}_j^{\text{ci},e}$  (1) is true. In turn, this implies  $\mathcal{H}_j^{\text{cst}}$  (2), by definition of the latter.  $\square$

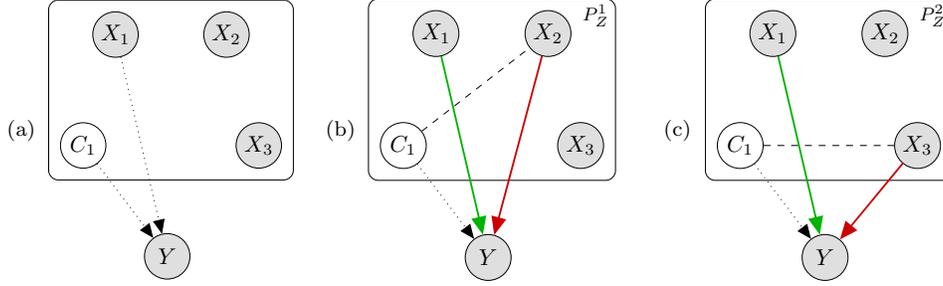


Figure 2: Graphical representation of consistency across environments improving robustness to missing variables. (a) Structural equation model for  $Y \mid X, C$ . (b) Conditional associations in environment one. (c) Conditional associations in environment two. The shaded nodes indicate the outcome or the observed variables, while the white node indicates the unobserved variable. The dotted arrows in (a) indicate causal links. The dashed segments represent graphically  $P_Z^e$ , which differs across environments and is such that  $C_1$  is associated with  $X_2$  in the first environment, and with  $X_3$  in the second one, while all other variables are conditionally independent. The solid arrows indicate conditional associations between the observable variables and the outcome in each environment (green if causal, red if spurious).

### 3.4 Consistency improves robustness to sampling biases

To understand why consistency can also improve robustness to sampling biases, consider Berkson’s paradox [61], which often arises in medicine [62] and social science [63], and is also known as the problem of “conditioning on a collider” [8]. We illustrate this idea with a classical example from [63]: imagine a school admitting students based on a composite score summarising their performance across different disciplines, and suppose only students whose composite score (standardized test + GPA + sports performance) exceeds a threshold are admitted. Even if GPA and sports performance were independent in the applicant population, one should expect a negative correlation between these variables among the admitted students: students with low GPA must perform well in sports, or else they would have not been observed. This bias makes it challenging to discover which variables may have a causal effect on GPA, especially if the admission criteria are unknown. Fortunately, consistency may help remove spurious associations if data are available from schools relying on different admission criteria.

To make our argument more general without sacrificing concreteness, consider a scenario in which the data are collected through a simple environment-specific rejection-sampling rule. Suppose a random individual from the population of interest is included in the sample for the  $e$ -th environment with probability proportional to some bias  $\phi^e$ , which is a function of a subset  $S^e \subseteq \{1, \dots, p_z\}$  of the variables and possibly also of the outcome:

$$\mathbb{P}[\text{include } (Z, Y) \text{ in the sample for environment } e] \propto \phi^e(Z_{S^e}, Y). \quad (8)$$

The effective joint distribution of  $(Z, Y)$  accessible from the  $e$ -th environment then becomes

$$P^e(Z, Y) \propto P_Z^e(Z) \cdot P^*(Y \mid Z) \cdot \phi^e(Z_{S^e}, Y),$$

where  $P^*(Y \mid Z)$  indicates the ideal conditional distribution of  $Y \mid Z$  determined by the causal model. Conditional testing on these data cannot lead to valid causal inferences even if no variables are missing, because the effective distribution of  $Y \mid Z$  is now  $P^e(Y \mid Z) \propto P^*(Y \mid Z) \cdot \phi^e(Z_{S^e}, Y)$ , which no longer corresponds to the model of interest. However, consistency can mitigate this issue, as long as the sampling biases (the subsets  $S^e$ ) are not constant.

**Proposition 3.** Fix any observable variable  $j$  and consider  $E$  environments. In the setting of Proposition 2, suppose the sampling mechanism in each environment  $e$  is biased by a subset  $S^e$  of explanatory variables as in (8). Assume that there are no unmeasured confounders ( $p_c = 0$ ), and  $\exists e \in \{1, \dots, E\} : j \notin S^e$ . Then,  $\mathcal{H}_j^{\text{causal}}$  (5) implies  $\mathcal{H}_j^{\text{cst}}$  (2), where each  $\mathcal{H}_j^{\text{ci},e}$  in (2) refers to the effective population induced by the biased sampling mechanism.

*Proof.* Focus on an environment  $e$  such that  $j \notin S^e$ , which is assumed to exist (we will drop the superscript  $e$  for simplicity). The effective conditional distribution of  $Y | Z$  is  $P(Y | Z) \propto P^*(Y | Z) \cdot \phi(Z_S, Y)$ , which does not depend on  $Z_j$  because  $j \notin S$  and  $P^*(Y | Z)$  only depends on  $\text{Pa}(Y) \not\ni j$ . Hence  $\mathcal{H}_j^{\text{ci},e}$  (1) is true, implying  $\mathcal{H}_j^{\text{cst}}$  (2).  $\square$

Figure 3 visualizes this concept with an example in which two environments have biased sampling depending on disjoint sets of variables, ensuring no spurious associations are consistent. The robustness of our proposed methods to sampling biases will be demonstrated with simulations in Section 6.3. Note that Proposition 3 presumes for simplicity that no variables are missing; a more general result relaxing this assumption could be obtained easily, although we do not explore that here because it would complicate the notation unnecessarily.

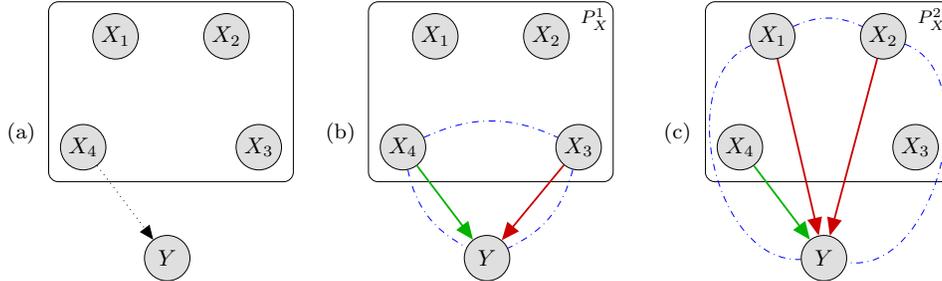


Figure 3: Graphical representation of consistency across environments improving robustness to sampling biases. The dash-dotted curves represent conditional dependence relations induced by biased sampling in each environment; the first one depends on  $X_3, X_4, Y$ , while the second one depends on  $X_1, X_2, Y$ . Other details are as in Figure 2.

### 3.5 Consistency improves robustness to homophily and contagion

Consistency can also improve robustness to homophily and contagion [60] or, more generally, to unaccounted dependencies among the observations within the same environment, which can lead to spurious associations [12]. Broadly speaking, we define homophily as the tendency of individuals sharing certain features to co-occur within one environment, and contagion as the mutual dependence of their outcomes. This is a well-known phenomenon, especially in the social sciences and in network studies [12, 60, 64–66], and it can be explained as follows. Consider a school in which students who enjoy reading are likely to join the book club (homophily). One student in the book club (patient zero) happens to be also interested in chess and convinces other members to learn how to play it (contagion). Later, a school-wide survey finds an association between reading and playing chess, even though there is no relation between the two at the population level. In particular, we do not expect to replicate this association in different schools (environments), as those may have different clubs and patient zeroes with other interests. Therefore, the risk of incorrectly reporting spurious findings can be mitigated by searching for consistent associations. This argument could be formalized as in Section 3.4, but we prefer to avoid introducing additional notation here, partially for lack of space, and partially because the main idea should at this point be already clear. The robustness of our proposed methods to homophily and contagion will be demonstrated with simulations in Section 6.4.

## 4 Methods

We start with a selective review of the knockoff methodology upon which our solution will be built. We will then develop a method for testing  $\mathcal{H}_j^{\text{cst}}$  (2) and finally extend it to  $\mathcal{H}_j^{\text{pcst},r}$  (3).

### 4.1 Review: the methodology of knockoffs

For an environment  $e$ , knockoffs enable testing  $\mathcal{H}_j^{\text{ci},e}$  (1) for all  $j \in \{1, \dots, p\}$ . The idea is to augment the data for each of the  $n$  observed individuals with  $p$  synthetic features, the *knockoffs*, which serve as negative control variables [2, 3]. The knockoffs are created by the statistician as a function of  $X$ , without looking at  $Y$ , and therefore they are conditionally independent of  $Y$  given  $X$ . Knockoffs are however pairwise exchangeable with the original variables in their joint distribution. That is, if  $[\mathbf{X}^e, \tilde{\mathbf{X}}^e] \in \mathbb{R}^{n \times 2p}$  is the matrix obtained by concatenating the variables  $\mathbf{X}^e \in \mathbb{R}^{n \times p}$

with the corresponding knockoffs  $\tilde{\mathbf{X}}^e \in \mathbb{R}^{n \times p}$  and, for any  $j \in \{1, \dots, p\}$ , the matrix  $[\mathbf{X}^e, \tilde{\mathbf{X}}^e]_{\text{swap}(j)}$  is obtained by swapping the  $j$ -th column of  $\mathbf{X}^e$  with the  $j$ -th column of  $\tilde{\mathbf{X}}^e$ , then

$$[\mathbf{X}^e, \tilde{\mathbf{X}}^e]_{\text{swap}(j)} \stackrel{d}{=} [\mathbf{X}^e, \tilde{\mathbf{X}}^e]. \quad (9)$$

The equation above says that swaps of a variable with its knockoff cannot be detected without looking at  $Y$ ; in fact, the only possible significant difference between  $X_j$  and  $\tilde{X}_j$  is the lack of conditional association of the latter with  $Y$ . The construction of knockoffs depends on the joint distribution  $P_X^e$  of all explanatory variables, and we refer to prior works for specific algorithms; see [2] for the multivariate Gaussian case, [4] for hidden Markov models, [35] for general graphical models, or [7] for fully nonparametric approximate methods. As our contribution does not concern this aspect of the analysis, we assume  $P_X^e$  is known and a knockoff generation algorithm is available.

The second step is to fit a model predicting  $Y$  given  $X$  and  $\tilde{X}$ , computing importance measures  $T_j^e$  and  $\tilde{T}_j^e$  for each  $X_j$  and  $\tilde{X}_j$ , respectively. Any model can be employed, as long as it does not unfairly discriminate between variables and knockoffs—swapping any  $X_j$  with  $\tilde{X}_j$  should result in  $T_j$  being swapped with  $\tilde{T}_j$  [2]. A typical choice is to fit a sparse generalized linear model (e.g., the lasso [67]), tuning its regularization parameter via cross-validation; then, the absolute values of the (scaled) regression coefficients are powerful importance measures [2]. For each  $j \in \{1, \dots, p\}$ ,  $T_j^e$  and  $\tilde{T}_j^e$  are combined into an anti-symmetric statistic, e.g.,  $W_j^e = T_j^e - \tilde{T}_j^e$ . This yields statistics  $W_j^e$  that are equally likely to be positive or negative if  $\mathcal{H}_j^{\text{ci},e}$  (1) is true [2]. By contrast, a large positive  $W_j^e$  is evidence against the null. Further, the signs of these statistics are mutually independent for all null indices  $j$  conditional on the absolute values,  $|W^e| = (|W_1^e|, \dots, |W_p^e|)$ . Formally, if  $\epsilon^e \in \{-1, +1\}^p$  is a random vector such that  $\epsilon_j^e = +1$  if  $\mathcal{H}_j^{\text{ci},e}$  is false and  $\mathbb{P}[\epsilon_j^e = +1] = 1/2$ , independently of everything else, otherwise, then  $W^e$  satisfies the following *coin-flip* property [2]:

$$W^e \stackrel{d}{=} W^e \odot \epsilon^e, \quad (10)$$

where  $\odot$  indicates element-wise multiplication. Therefore, the signs of  $W^e$  can be seen as one-bit conservative p-values [3] for  $\mathcal{H}_j^{\text{ci},e}$  (1), if we transform them as  $p_j^e = 1/2$  if  $W_j^e > 0$  and  $p_j^e = 1$  otherwise. The ordering provided by the absolute values of  $W^e$  allows one to powerfully test the above hypotheses with a sequential procedure [3].

Concretely, the knockoff filter [3] computes a significance threshold for the test statistics such that the rejection of  $\mathcal{H}_j^{\text{ci},e}$  (1) for all  $j$  with larger  $W_j^e$  controls the false discovery rate below the desired level  $\alpha$ . That is, one can prove

$$\text{FDR} := \mathbb{E} \left[ \frac{|\{j : \mathcal{H}_j^{\text{ci},e} \text{ is rejected}\} \cap \{j : \mathcal{H}_j^{\text{ci},e} \text{ is true}\}|}{|\{j : \mathcal{H}_j^{\text{ci},e} \text{ is rejected}\} \vee 1} \right] \leq \alpha,$$

where  $a \vee b := \max\{a, b\}$ . Equivalently, the knockoff filter can be seen as an instance of the selective SeqStep+ test [3] applied to the above ordered one-bit p-values. In the next section, we will extend this methodology to simultaneously analyze data from many environments, controlling the false discovery rate for the hypotheses  $\mathcal{H}_j^{\text{cst}}$  defined in (2).

## 4.2 Multi-environment knockoff statistics

Consider  $E$  environments  $e \in \{1, \dots, E\}$ , each corresponding to observations  $\mathbf{Y}^e \in \mathbb{R}^n$ ,  $\mathbf{X}^e \in \mathbb{R}^{n \times p}$ , and knockoffs  $\tilde{\mathbf{X}}^e \in \mathbb{R}^{n \times p}$ . Assume the sample size is  $n$  in all environments for ease of notation, although this is unnecessary. The first ingredient for testing  $\mathcal{H}_j^{\text{cst}}$  (2) are the multi-environment statistics defined below, which generalize those from [2].

**Definition 1** (Multi-environment knockoff statistics).  $\mathbf{W} \in \mathbb{R}^{E \times p}$  are valid multi-environment knockoff statistics if they satisfy  $\mathbf{W} \stackrel{d}{=} \mathbf{W} \odot \epsilon$ , and  $\epsilon \in \{\pm 1\}^{E \times p}$  is a random matrix with independent entries and rows  $\epsilon^e$  such that  $\epsilon_j^e = \pm 1$  with probability  $1/2$  if  $\mathcal{H}_j^{\text{ci},e}$  in (1) is true and  $\epsilon_j^e = +1$  otherwise, for  $j \in \{1, \dots, p\}$  and  $e \in \{1, \dots, E\}$ .

A simple solution to obtain such statistics is to analyze different environments separately with the existing method from the previous section, and then stack their output  $W^e$ . This approach, which we call *data-splitting*, is computationally efficient and private, as it allows researchers in different environments to collaborate without disclosing their data, but it is not the most powerful. For example, if all environments are identical, data splitting effectively divides the total sample size by  $E$  compared to a regular analysis of the pooled data, although the latter would test the same hypotheses in this scenario; this motivates a more general approach.

Let  $\mathbf{Y} \in \mathbb{R}^{En}$ ,  $\mathbf{X} \in \mathbb{R}^{En \times p}$ ,  $\tilde{\mathbf{X}} \in \mathbb{R}^{En \times p}$  indicate the data matrices obtained by stacking the observations or knockoffs from all environments. We define  $[\mathbf{T}, \tilde{\mathbf{T}}] \in \mathbb{R}^{E \times 2p}$  as a matrix of multi-environment importance measures for all variables and knockoffs, computed by applying a randomized function  $\tau$  to the full data set:

$$[\mathbf{T}, \tilde{\mathbf{T}}] = \tau(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]) := \left[ t(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]), \tilde{t}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]) \right]. \quad (11)$$

Above,  $\mathbf{t}$  (resp.  $\tilde{\mathbf{t}}$ ) are defined in terms of the matrix  $\boldsymbol{\tau}$ , as its first (resp. last)  $p$  columns. For any subset  $\mathcal{S} \subseteq \{1, \dots, E\} \times \{1, \dots, p\}$ , let  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})}$  be the matrix obtained from  $[\mathbf{X}, \tilde{\mathbf{X}}]$  after swapping the column  $\mathbf{X}_j^e$  for environment  $e$  with the corresponding  $\tilde{\mathbf{X}}_j^e$ , for all  $(e, j) \in \mathcal{S}$ . Note the slight change of notation compared to (9): there, swapping was defined only for one environment. The function  $\boldsymbol{\tau}$  may be almost anything, possibly involving sophisticated machine learning algorithms, as long as it satisfies

$$\boldsymbol{\tau} \left( \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})} \right) \stackrel{d}{=} \left[ \mathbf{t} \left( \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}] \right), \tilde{\mathbf{t}} \left( \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}] \right) \right]_{\text{swap}(\mathcal{S})} \mid \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]. \quad (12)$$

In words, this says that swapping variables and knockoffs in any one environment should have the only effect of swapping the corresponding importance measures in that environment, leaving all other elements of  $\boldsymbol{\tau}$  unchanged. Importantly, the equality only needs to hold in distribution conditional on the data and on the knockoffs, as  $\boldsymbol{\tau}$  may be randomized. For our method to be powerful, we also need  $\boldsymbol{\tau}$  to be such that a larger value in row  $e$  and column  $j$  indicates evidence from environment  $e$  that  $X_j$  is conditionally associated with  $Y$ , while a larger value in row  $e$  and column  $j + p$  points to a corresponding association of  $\tilde{X}_j$  with  $Y$ , which we know must be spurious because all knockoffs are null. Concrete examples of powerful multi-environment importance measures will be presented later.

Given any matrix  $[\mathbf{T}, \tilde{\mathbf{T}}]$  (11) of importance measures satisfying (12), we construct multi-environment statistics  $\mathbf{W} \in \mathbb{R}^{E \times p}$  by computing its rows  $W^e \in \mathbb{R}^p$  with the standard approach reviewed in Section 4.1, separately environment by environment. Precisely, we contrast each  $T_j^e$  with  $\tilde{T}_j^e$ , pairwise for all variables and knockoffs; e.g.,

$$W_j^e = T_j^e - \tilde{T}_j^e. \quad (13)$$

This ensures the signs of  $\mathbf{W}$  corresponding to null hypotheses are independent coin flips conditional on the absolute values of all entries, generalizing the result in (10) from [2].

**Proposition 4.** *If the statistics  $\mathbf{W} \in \mathbb{R}^{E \times p}$  are computed based on (12)–(13), then they satisfy Definition 1.*

*Proof.* Let  $\mathbf{w}$  denote the matrix-valued (randomized) function computing the knockoff statistics in all environments, consistently with (12)–(13), i.e.,  $\mathbf{W} = \mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}])$ . With  $\epsilon$  as in Definition 1, define the set  $\mathcal{S} = \{(e, j) : \epsilon_j^e = -1\}$ . It follows from (12)–(13) that  $\mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})}) \stackrel{d}{=} \epsilon \odot \mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}])$  conditional on  $\mathbf{Y}$  and  $[\mathbf{X}, \tilde{\mathbf{X}}]$ , which implies  $\mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})}) \stackrel{d}{=} \epsilon \odot \mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}])$  marginally. We also know that  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})} \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}] \mid \mathbf{Y}$  (Lemma 3.2 in [2]), which implies  $\mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})}) \stackrel{d}{=} \mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}])$ . Therefore,  $\mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]) = \epsilon \odot \mathbf{w}(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}])$ .  $\square$

An intuitive algorithm to compute powerful multi-environment statistics is the following, which we call the *empirical cross-prior* approach. The idea is that the requirements of Proposition 4 are satisfied even if each row  $W^e$  leverages the observations from other environments  $\{1, \dots, E\} \setminus \{e\}$ , as long as the latter are perturbed through suitable random column swaps. This perturbation only acts on a copy of  $[\mathbf{X}, \tilde{\mathbf{X}}]$  and all subsequent steps of the analysis will begin with the original data, so the order in which the environments are processed is irrelevant. Concretely, we perturb the data by randomly swapping each variable with its own knockoff based on a coin flip, independently observation by observation. Then, we estimate importance weights that are symmetric with respect to variables and knockoffs; these will serve as prior information for the subsequent analysis step (hence the name for this approach). Figure 4 provides a schematic of the full procedure, while the details are explained below.

Let  $\mathbf{V} \in \{0, 1\}^{E \times p}$  be a random matrix of i.i.d. coin flips, and  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathbf{V})}$  be the perturbed matrix obtained by swapping the  $i$ -th observation of  $X_j$  with the corresponding knockoffs if and only if  $V_{ij} = 1$ . We compute prior importance measures  $T^{\text{prior}}$  (resp.  $\tilde{T}^{\text{prior}}$ ) for all variables (resp. knockoffs) as the absolute values of the regression coefficients estimated by fitting a sparse generalized linear regression model (e.g., the lasso) to predict  $\mathbf{Y}$  given  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathbf{V})}$ , tuning the regularization parameter by cross-validation. The results are combined symmetrically into a prior weight  $\pi_j$  for each  $j$ , e.g.,  $\pi_j = \zeta(T_j^{\text{prior}} + \tilde{T}_j^{\text{prior}})$  for some positive and decreasing function  $\zeta$ , such as  $\zeta(t) = 1/(0.05 + t)$ . Finally, we compute the importance measures  $T^e$  and  $\tilde{T}^e$  based on the unperturbed data in the  $e$ -th environment by combining the above prior with the same lasso-based approach as in the data-splitting case. In particular, the regularization penalty is now feature-specific and depends on two parameters,  $\lambda^e > 0$  and  $\gamma^e \in [0, 1]$ , both tuned by cross-validation, as well as on the prior weights. Specifically, the penalty for the  $j$ -th variable is  $\lambda_j^e = \lambda^e(1 - \gamma^e) + \gamma^e \pi_j$ . This reduces to data splitting if we fix  $\gamma^e = 0$ . In general, larger values of  $\gamma^e$  may improve power by making it less likely for the environment-specific models to select spurious variables. However, the prior may not always be very informative and, even if it is, it is unclear how much weight it should be given; thus we also tune  $\gamma^e$  by cross-validation. To avoid a two-dimensional grid search, in practice we first tune  $\lambda^e$  and then  $\gamma^e$ .

**Proposition 5.** *The statistics  $\mathbf{W}$  obtained by applying (13) to the empirical cross-prior importance measures  $\mathbf{T}$  and  $\tilde{\mathbf{T}}$  described above satisfy Definition 1.*

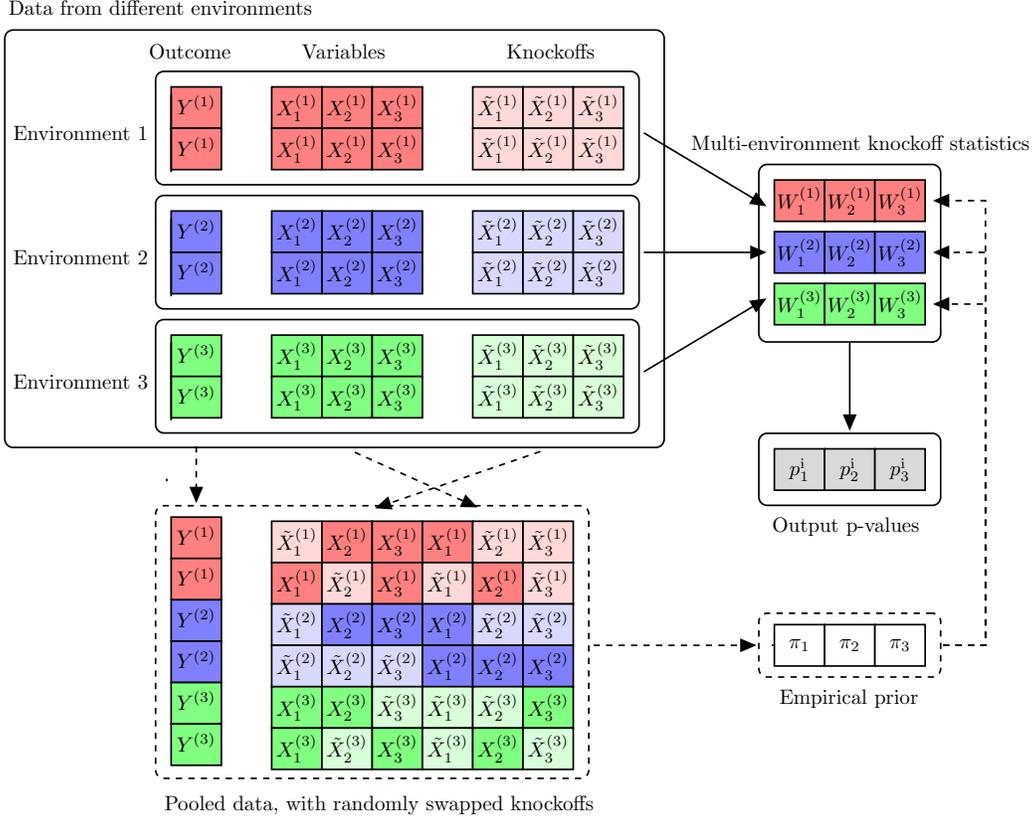


Figure 4: Schematics for a multi-environment knockoff analysis. In this example, there are 3 variables, 3 environments, and 2 observations per environment. The solid arrows represent the analysis based on data-splitting statistics, in which the environments are analyzed separately before combining the resulting knockoff statistics. The dashed arrows represent the additional steps corresponding to our empirical cross-prior statistics, which analyze jointly the data from all environments. The darker blocks indicate the real data, while the lighter ones indicate the knockoffs.

*Proof.* By Proposition 4, it suffices to show  $[\mathbf{T}, \tilde{\mathbf{T}}]$  satisfy (12) because (11) is trivial. For any  $\mathcal{S} \subseteq \{1, \dots, E\} \times \{1, \dots, p\}$  consider how  $[\mathbf{T}, \tilde{\mathbf{T}}]$  would change if  $[\mathbf{X}, \tilde{\mathbf{X}}]$  were replaced by  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{S})}$ . The joint distribution of the prior weights is invariant because  $\pi$  depends on  $[\mathbf{X}, \tilde{\mathbf{X}}]$  through  $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathbf{V})}$ , and the composition of  $\text{swap}(\mathcal{S})$  with  $\text{swap}(\mathbf{V})$  is statistically indistinguishable from  $\text{swap}(\mathbf{V})$ . Similarly, the cross-validation errors within each environment are invariant to the ordering of variables and knockoffs, and so are the optimal values of  $\lambda^e$  and  $\gamma^e$ . Therefore, the only change is that the final importance measure  $T_j^e$  is swapped with the corresponding  $\tilde{T}_j^e$  if and only if  $(e, j) \in \mathcal{S}$ .  $\square$

Perturbing the data by randomly swapping variables and knockoffs is essential (see the proof of Proposition 5) and, although it weakens the signals, it does not destroy them entirely. In fact, the empirical prior can learn which *pairs* of variables and knockoffs are likely to be important, although the effective sample for this task is cut in half. In any case, the prior cannot do much harm because its influence on the output is controlled by  $\gamma^e$ , which is tuned by cross-validation. Therefore, we would expect at worst to select  $\gamma^e \approx 0$  and thus approximately recover the data-splitting solution.

Statistics computed by naively looking at the full unperturbed data from all environments are generally invalid. For example, one could imagine using all data to compute the magnitudes of  $W_j^e$ , similarly to how we estimate the above empirical prior but without the initial perturbation, and then leveraging only the observations in environment  $e$  to determine the signs. This would yield statistics that are separately valid in each environment but are not mutually independent for different  $e$ , compromising our analysis downstream; see Appendix B and Figure A2 therein.

### 4.3 The multi-environment knockoff filter

Starting from a matrix  $\mathbf{W}$  of multi-environment knockoff statistics, we can combine its rows to obtain one-bit conservative p-values [3] for testing  $\mathcal{H}_j^{\text{cst}}$  in (2). Precisely, for each  $j \in \{1, \dots, p\}$ , we compute

$$p_j^{\text{cst}} = \begin{cases} 1/2, & \text{if } \min \{ \text{sign}(W_j^e) \}_{e=1}^E = +1, \\ 1, & \text{otherwise.} \end{cases} \quad (14)$$

The order in which these hypotheses will be tested depends on the absolute values of  $\mathbf{W}$ , which we combine column-wise with some symmetric function  $\bar{w}$  to obtain invariant statistics  $|W_j^{\text{cst}}|$  for all  $j \in \{1, \dots, p\}$ :

$$|W_j^{\text{cst}}| = \bar{w} \left( |W_j^1|, \dots, |W_j^E| \right). \quad (15)$$

We will adopt here  $\bar{w}(|W_j^1|, \dots, |W_j^E|) = \prod_{e=1}^E |W_j^e|$ , although other options are possible. The next result states that, conditional on the ordering determined by (15), the one-bit p-values in (14) are valid for  $\mathcal{H}_j^{\text{cst}}$  (2).

**Proposition 6** (Multi-environment p-values). *If the statistics  $\mathbf{W}$  satisfy Definition 1, the p-values  $p_j^{\text{cst}}$  in (14) corresponding to true  $\mathcal{H}_j^{\text{cst}}$  (2) stochastically dominate the uniform distribution conditional on  $|W^{\text{cst}}| = (|W_1^{\text{cst}}|, \dots, |W_p^{\text{cst}}|)$ . Further, these p-values are “almost independent”, i.e.,  $\mathbb{P}[p_j^{\text{cst}} \leq \alpha \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}}] \leq \alpha, \forall \alpha \in [0, 1]$  if  $\mathcal{H}_j^{\text{cst}}$  is true.*

*Proof.* It suffices to prove the second part of this statement because that implies the first one. Take any  $j \in \{1, \dots, p\}$  corresponding to a true  $\mathcal{H}_j^{\text{cst}}$  (2), and assume for simplicity  $|W_j^{\text{cst}}| > 0$  (if not,  $p_j^{\text{cst}} = 1$  by definition). By definition of  $\mathcal{H}_j^{\text{cst}}$ , we know  $\exists e \in \{1, \dots, E\}$  such that the environment-specific null  $\mathcal{H}_j^{\text{ci},e}$  in (1) is true. Therefore,

$$\begin{aligned} \mathbb{P}[p_j^{\text{cst}} = 1/2 \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}}] &= \mathbb{P}\left[ W_j^l \geq 0 \forall l \in \{1, \dots, E\} \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}} \right] \\ &\leq \mathbb{P}\left[ W_j^e > 0 \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}} \right] = 1/2. \end{aligned}$$

The above inequality follows from  $|W_j^{\text{cst}}| > 0$  and hence  $W_j^e \neq 0$  for all  $e$ , while the last equality follows from Definition 1 because  $|W^{\text{cst}}|$  is a function of  $|\mathbf{W}|$  and  $p_{-j}^{\text{cst}}$  of  $\mathbf{W}_{-j}$ . The proof is completed by  $p_j^{\text{cst}} \in \{1/2, 1\}$ .  $\square$

Proposition 6 suggests applying a sequential testing approach, such as selective SeqStep+ (the knockoff filter) [3], to the p-values in (14). However, this is not obviously valid because the null p-values  $p_j^{\text{cst}}$  are not independent [3]. In fact, each of them is also affected by the signs of entries of  $\mathbf{W}$  corresponding to non-null environments, which need not be independent;  $\mathcal{H}_j^{\text{cst}}$  only says there exists a null entry in the  $j$ -th column of  $\mathbf{W}$ , but  $p_j^{\text{cst}}$  also counts the signs of the others. Fortunately, the next result, proved in Appendix C.1, shows the “almost-independence” established by Proposition (6) is sufficient. The intuition is that this dependence at worse makes our p-values conservative.

**Theorem 1** (Multi-environment knockoff filter). *The selective SeqStep+ procedure of [3] applied to p-values  $p_j^{\text{cst}}$  ordered by  $|W_j^{\text{cst}}|$  and satisfying the “almost-independence” property of Proposition (6), i.e.,  $\mathbb{P}[p_j^{\text{cst}} \leq \alpha \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}}] \leq \alpha$  for any  $\alpha \in [0, 1]$ , controls the false discovery rate below the nominal level.*

### 4.4 Testing partially consistent conditional associations

Starting from multi-environment knockoff statistics  $\mathbf{W}$ , we summarise them column-wise as follows. For each  $j \in \{1, \dots, p\}$ , let  $n_j^-$  count the negative signs in the  $j$ -th column, and let  $n_j^0$  be the number of zeros. Then, compute

$$p_j^{\text{pcst},r} := \Psi \left( n_j^- - 1, (E - r + 1 - n_j^0) \vee 0, \frac{1}{2} \right) + U_j \cdot \psi \left( n_j^-, (E - r + 1 - n_j^0) \vee 0, \frac{1}{2} \right), \quad (16)$$

where  $\Psi(x, m, \pi)$  is the binomial cumulative distribution function evaluated at  $x$ ,  $\psi(x, m, \pi)$  is the corresponding probability mass, and  $U_j \sim \text{Uniform}[0, 1]$  independently of all else. Below, we show these are conservative p-values for  $\mathcal{H}_j^{\text{pcst},r}$  (3), for any fixed  $r$ . As in the previous section, we will filter them sequentially in decreasing order of

$$|W_j^{\text{pcst},r}| = \bar{w} \left( |W_j^1|, \dots, |W_j^E| \right), \quad (17)$$

for some symmetric function  $\bar{w}$ . Concretely, we will focus on the  $\bar{w}$  that multiplies the top  $r$  largest entries in that column by absolute value, thus extending the solution from the previous section:

$$w \left( |W_j^1|, \dots, |W_j^E| \right) = \prod_{e=1}^r |W_j^e|^{(E-e+1)}.$$

Above,  $|W_j|^{(e)}$  are the order statistics for the absolute values in the  $j$ -th column of  $\mathbf{W}$ . The next result proves these ordered p-values are conservative for  $\mathcal{H}_j^{\text{pcst},r}$  (3), and “almost independent” of each other in the sense of Proposition 6. Combined with Theorem 1, this will guarantee false discovery rate control using selective SeqStep+.

**Proposition 7** (Partial conjunction multi-environment p-values). *If the multi-environment statistics  $\mathbf{W}$  satisfy Definition 1, for any fixed  $r \in \{1, \dots, E\}$ , the p-values  $p_j^{\text{pcst}}$  (16) corresponding to true  $\mathcal{H}_j^{\text{pcst},r}$  in (3) stochastically dominate the uniform distribution conditional on  $|W^{\text{pcst},r}| = (|W_1^{\text{pcst},r}|, \dots, |W_p^{\text{pcst},r}|)$  (17). Further, these p-values are “almost independent”, in the sense that  $\mathbb{P}[p_j^{\text{pcst},r} \leq \alpha \mid |W^{\text{pcst},r}|, p_{-j}^{\text{pcst},r}] \leq \alpha, \forall \alpha \in [0, 1]$  if  $\mathcal{H}_j^{\text{pcst},r}$  is true.*

*Proof.* As in the proof of Proposition 6, it suffices to establish the second part of this statement. Take any  $j \in \{1, \dots, p\}$  such that  $\mathcal{H}_j^{\text{pcst},r}$  in (3) holds, so that there are at least  $(E - r + 1 - n_j^0) \vee 0$  environments in which  $\mathcal{H}_j^{\text{ci},e}$  is true and  $W_j^e \neq 0$ . Without loss of generality, assume these environments are those indexed by  $C_j := \{1, \dots, (E - r + 1 - n_j^0) \vee 0\}$ , with  $C_j = \emptyset$  if  $E - r + 1 - n_j^0 \leq 0$ . Then, define the random variable  $n_j^{*-}$  as the number of negative signs in  $\mathbf{W}_j^{C_j}$ . As  $\mathbf{W}$  satisfies the flip-sign property from Definition 1,  $\mathbf{W} \stackrel{d}{=} \mathbf{W} \odot \boldsymbol{\epsilon}$ , it follows the  $n_j^{*-}$ ’s are independent Binomial( $|C_j|, 1/2$ ), with  $|C_j| = (E - r + 1 - n_j^0) \vee 0$ , conditional on  $|\mathbf{W}|$  and hence also conditional on  $|W^{\text{pcst},r}|$ . Now, define  $p_j^{*\text{pcst},r}$  as the imaginary p-value obtained by replacing  $n_j^-$  with  $n_j^{*-}$  in (3). Because  $n_j^{*-} \leq n_j^-, \forall \alpha \in [0, 1]$ ,

$$\mathbb{P}[p_j^{\text{pcst},r} \leq \alpha \mid |W^{\text{pcst},r}|, p_{-j}^{\text{pcst},r}] \leq \mathbb{P}[p_j^{*\text{pcst},r} \leq \alpha \mid |W^{\text{pcst},r}|, p_{-j}^{\text{pcst},r}] = \mathbb{P}[p_j^{*\text{pcst},r} \leq \alpha] \leq \alpha.$$

Above, the last inequality follows from the fact that  $p_j^{*\text{pcst},r}$  is the standard randomized binomial p-value [68].  $\square$

If  $r = E$ , the p-value  $p_j^{\text{pcst},E}$  in (16) does not become identical to the  $p_j^{\text{cst}}$  defined earlier in (14). The difference between  $p_j^{\text{cst}}$  and  $p_j^{\text{pcst},E}$  is that the former is always equal to 1 when  $n_j^0 > 0$  and it is not randomized. This discrepancy is due to expository convenience, as we prefer to keep the notation simple in the previous section, and it is practically irrelevant because it would not make sense to reject  $\mathcal{H}_j^{\text{cst}}$  (2) if  $n_j^0 > 0$ . While randomization is useful here because it allows us to deal as powerfully as possible with the case in which  $n_j^0 > 0$  (it might make sense to reject  $\mathcal{H}_j^{\text{pcst},r}$  even if  $n_j^0 > 0$ , as long as  $r < E$ ), it would have not helped in the previous section. In fact, the p-values  $p_j^{\text{cst}}$  in (14) only have one bit of information, and selective SeqStep+ only cares about whether they are greater than  $1/2$ .

If  $r = 1$ , the hypothesis  $\mathcal{H}_j^{\text{pcst},1}$  (3) states  $X_j$  is conditionally associated with  $Y$  in at least one environment. In that case, a traditional knockoff analysis of the pooled data would test the same hypothesis, and it might be more powerful because it allows more flexibility in the test statistics. Therefore, we only recommend applying the method proposed here with  $r > 1$ .

A limitation of selective SeqStep+ is that it is unclear how to choose its parameter  $c$  (the baseline rejection threshold) [3] to maximize power because the p-values in (16) can take several possible values; by contrast,  $c = 1/2$  is the only option for one-bit p-values. Within our partial consistency problem, different  $c$  may result in higher power depending on the data. Therefore, we also consider filtering  $p_j^{\text{pcst},r}$  (16) with the accumulation test [69], which requires less tuning. If the p-values are independent, this test controls a modified version of the false discovery rate,

$$\text{mFDR}_q := \mathbb{E} \left[ \frac{|\{j : \mathcal{H}_j^{\text{ci},e} \text{ is rejected}\} \cap \{j : \mathcal{H}_j^{\text{ci},e} \text{ is true}\}|}{|\{j : \mathcal{H}_j^{\text{ci},e} \text{ is rejected}\}| + q} \right], \quad (18)$$

for some constant  $q$  whose value will be specified later. If  $q$  is not too large and the discoveries are sufficiently numerous, the above definition is not very different from the false discovery rate. Although our p-values  $p_j^{\text{pcst},r}$  in (16) are dependent, the next result proves the accumulation test is still valid if we add a little extra randomness.

Starting from a matrix  $\mathbf{W}$  of multi-environment knockoff statistics (Definition 1), randomly assign imaginary positive or negative signs to any zero entry by flipping independent fair coins. Then, define  $n_j^-$  as the number of negative entries in the  $j$ -th column of the resulting *tie-breaking*  $\mathbf{W}$ , and compute, with the same notation as in (16),

$$p_j^{\text{pcst},r} := \Psi \left( E - r + 1, \frac{1}{2}, n_j^- - 1 \right) + U_j \cdot \psi \left( E - r + 1, \frac{1}{2}, n_j^- \right). \quad (19)$$

**Theorem 2** (Multi-environment knockoff filter with accumulation test). *The accumulation test of [69] with an increasing accumulation function (e.g., HingeExp with parameter  $C = 2$ ) applied to the p-values defined in (19) controls the modified false discovery rate (18) (e.g., with  $q = C/\alpha$ ), as in [69]. That is, Theorem 1 of [69] still holds for the p-values  $p_j^{\text{pcst},r}$  in (19) even though they are not independent.*

The proof of Theorem 2 is in Appendix C.2. It is worth emphasizing Theorem 1 can accommodate replacing the p-values  $p_j^{\text{pcst},r}$  from (16) with those in (19), although the converse does not hold for Theorem 2. The issue is that  $n_j^0$  may vary arbitrarily across  $j$ , breaking the symmetry needed by our martingale proof. This may be a limitation of our proof, as the simulations in Section 6 suggest the accumulation test works well with the p-values in (16).

## 5 Consistent genome-wide associations across diverse ancestries

Before investigating the effectiveness of the proposed methods in practice, we present in some detail a genetic problem that has motivated our work; this will inform the design of our simulations and will be the subject of the subsequent data analysis. Genome-wide association studies aim to identify genetic variants with biological effects on some phenotype. Because the DNA of individuals is determined prior to any of their phenotypes, it is relatively easy to attribute a causal interpretation to observed links between genetic variation and traits. Furthermore, it is reasonable to assume the same biological pathways are involved in influencing the traits across different ethnicities or human populations: modulo the expected variations in “environment”, it is therefore meaningful to think about one common causal mechanism linking genetic variation to traits and to attempt to uncover it by testing  $\mathcal{H}_j^{\text{cst}}$  (2). We now describe the variables observed in genome-wide association studies and the changes expected across environments corresponding to different human populations.

### 5.1 Missing variants and knockoffs in genome-wide association studies

Genome-wide association studies are carried out in practice by measuring (*genotyping*) a subset of a few hundred thousands variants across the genome, as full sequencing is expensive. Such relatively few markers are sufficient to capture most genetic variation because nearby alleles on the same chromosome have strong dependencies and can thus be quite accurately inferred from one another. This property, known as *linkage disequilibrium* [9, 70], facilitates the localization of broad regions, or *loci*, containing associations with the phenotype, but at the same time it complicates the attribution of distinct signals to specific variants. Indeed, many genotypes can be *marginally associated* with the phenotype simply because they are in linkage disequilibrium with the same causal variant; this issue is alleviated by a conditional testing approach, but even conditional associations do not account for possible confounding due to unobserved variants.

The traditional analysis of genome-wide association data imputes the missing variants using models of linkage disequilibrium estimated on smaller, fully-sequenced, reference samples [71]. Then, the imputed variants are analyzed alongside the typed ones to localize significant associations, through either genome-wide marginal tests or multivariate linear models within narrow genomic regions [72]. However, this is not fully satisfactory because imputation is not as informative as a direct measurement; in fact, imputed variants carry no information in addition to that contained in the typed ones, as they are a function of the latter, conditionally independent of the phenotype. This may pass unnoticed if one fully trusts a multivariate linear model for the outcome—imputed variants may be significant within such models because their dependence with the measured variants is nonlinear—but it makes it impossible to find explicit evidence that a missing variant is causal within our nonparametric model-X perspective [5, 6]. Without repeating the arguments supporting a model-X analysis of genome-wide association data, which were explained in [4] and expanded in [5, 6], we will focus here on leveraging consistency to gather indirect evidence of causal associations within this framework. First though, we must briefly recall some relevant technical details of the current methodology.

The existing knockoff-based analysis partitions the genome into contiguous segments and then tests whether any of these include variants conditionally associated with the phenotype [5, 6]. Let  $G \subset \{1, \dots, p\}$  denote the indices of the typed variants in a particular segment. Then, knockoffs can be utilized to test a slightly more general version of the conditional hypothesis  $\mathcal{H}_j^{\text{ci},e}$  (1) for each environment (sub-population)  $e \in \{1, \dots, E\}$ ; namely,

$$\mathcal{H}_G^{\text{ci},e} : Y^e \perp\!\!\!\perp X_G^e \mid X_{-G}^e, \quad (20)$$

where  $X_G = \{X_j : j \in G\}$ ,  $X_{-G} = \{X_j : j \notin G\}$ . That is,  $\mathcal{H}_G^{\text{ci},e}$  (20) asserts that the variants in group  $G$  are conditionally independent of the phenotype given all other measured variants. This analysis can be performed at different levels of resolution, separately controlling the false discovery rate for increasingly refined genomic partitions to balance between power and the value of each discovery [5]. In short, hypotheses  $\mathcal{H}_G^{\text{ci},e}$  (20) involving smaller groups of genotypes (higher resolution) are harder to reject because the variables have strong local dependencies, making it difficult to distinguish the signal of one variant from those of its neighbors. At the same time, high-resolution hypotheses are more specific and thus their rejections more informative. Although  $\mathcal{H}_G^{\text{ci},e}$  (20) is not asking whether a physical genetic segment contains causal variants, it is a reasonable statistical approximation of that scientific

question. Intuitively, we expect these hypotheses to be more robust to missing variants at low resolution; this will be verified empirically later. By contrast, the robustness of  $\mathcal{H}_G^{ci,e}$  (20) is less clear at high resolution because there each tested segment contains few measured genotypes; this is where consistency will be most useful.

## 5.2 Linkage disequilibrium in populations with different ancestries

Linkage disequilibrium is explained by the inheritance of long and randomly cut genetic segments from parents to offspring, with occasional mutations. Generation after generation, the genotype distribution thus comes to resemble an imperfect mosaic of motifs inherited from the common ancestors, which can be encoded as a hidden Markov model [73]; this is at the heart of the imputation methods mentioned in the preceding section [71], as well as of the algorithms for generating knockoffs [4]. The block-like patterns of linkage disequilibrium vary across human sub-populations because these share different recent ancestors, and so their mosaics involve different patterns, and possibly also different transition (*recombination*) rates [70, 74, 75]. In other words, different sub-populations differ by covariate shift. This heterogeneity has already been factored into the generation of valid knockoffs to test *pooled* conditional associations [6], and it will be leveraged here to help highlight causal variants through consistency.

Figure 5 visualizes why covariate shift helps localize causal variants within an example with three sub-populations, four observed and five missing variables, one of which is causal. This toy genome is partitioned into segments containing one or two typed variants each; three segments are highlighted. Linkage disequilibrium is described by hidden Markov models yielding blocks of variables separated by high-recombination spots [76, 77]. Treating alleles across these spots as approximately independent, we can see that the only consistent association is that of the segment containing the causal variant. Of course, reality is more complicated. First, linkage disequilibrium is not perfectly organized into independent blocks, although this is a common simplification [78]. Second, we can only study a finite number of human sub-populations, so it may be unrealistic to expect all spurious associations to be removed. Nonetheless, consistency enables a step forward in an otherwise challenging problem, and we will verify empirically that our approach is indeed useful.

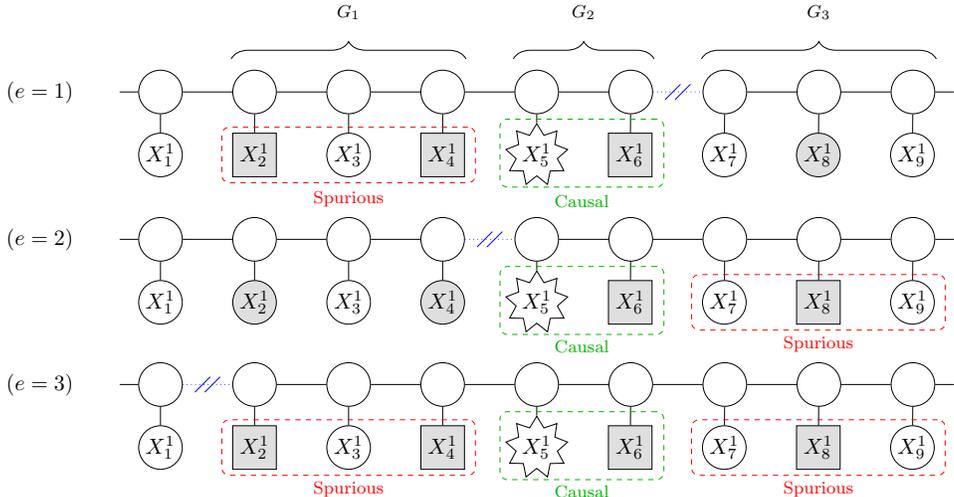


Figure 5: Schematic visualization of consistency in a genome-wide association study involving three sub-populations. The unobserved causal variant (star-shaped node) induces different spurious associations depending on the patterns of linkage disequilibrium, described by population-specific hidden Markov models. Shaded nodes: measured variables. Squares: variables that are not independent of the causal one conditional on the other measured variables. The broken segments in the Markov chain symbolize the boundaries of linkage disequilibrium blocks within a population.

## 6 Numerical experiments

We apply the multi-environment knockoff filter with the data-splitting and empirical cross-prior alternative statistics. The benchmarks are the two heuristics from Section 2: intersection and pooling. All test statistics are computed

with the R package `glmnet` [79], or `bigstatsr` [80] for genetic data. Software for our method is available from [https://github.com/lsn235711/MEKF\\_code](https://github.com/lsn235711/MEKF_code), along with code to reproduce the analyses.

## 6.1 Testing for full consistency with synthetic data

In each environment,  $p$  variables  $X$  are generated from an autoregressive model of order one with correlation  $\rho = 0.2$ . We will leverage the known  $P_X$  to construct exact knockoffs with the Gaussian semi-definite optimization algorithm from [2]. Knockoffs are typically quite robust to model misspecification, and the estimation of  $P_X$  is orthogonal to our problem. The distribution of  $Y^e \mid X^e$  in the  $e$ -th environment is given by a logistic model with log-odds equal to logit  $\mathbb{P}[Y^e = 1 \mid X^e] = X^e \beta^e$ , where  $\beta^e \in \mathbb{R}^p$  is an environment-specific effect parameter vector. We consider two settings corresponding to different numbers of environments  $E$ , explanatory variables  $p$ , and effect vectors  $\beta^e$ .

In the first setting,  $E = 4$ ,  $p = 500$ , and the number of observations per environment is  $n = 1000$ . First, 100 entries of  $\beta^e$  are randomly chosen to be non-zero in all environments and these are the consistent associations we seek; then, for each environment  $e$ , 10 of the remaining ones are set to be non-zero in all but the  $e$ -th environment, and these 40 associations are thus not consistent. See Figure A3 (a) in Appendix D.1 for a visualization of this setup. The absolute values of the 100 consistent non-null elements of  $\beta^e$  are equal to  $a/\sqrt{n}$ , where  $a$  is a signal amplitude parameter which we will vary, while the remaining non-zero values are equal to  $0.5a/\sqrt{n}$ . The signs of non-null elements of  $\beta^e$  are determined by independent coin flips. In the second setting,  $E = 3$ ,  $p = 200$ , and  $n = 2000$ . The coefficients  $\beta^e$  are determined as follows. First, 50 entries of  $\beta^e$  are randomly chosen to be non-zero in all environments; then, for each  $e$ , 50 of the remaining ones are set to be non-zero in all but the  $e$ -th environment; see Figure A3 (b) in Appendix D.1. Again, the 100 consistent non-null elements of  $\beta^e$  are equal to  $a/\sqrt{n}$  in absolute value, and the remaining non-zero entries are  $0.5a/\sqrt{n}$ . The signs of  $\beta^e$  are determined by independent coin flips.

Our goal is to discover the subset of consistent non-nulls, controlling the false discovery rate below 10%. Figure 1, previewed earlier, compares the performance of our method to those of the benchmarks, averaging over 100 experiments. The results confirm our method controls the false discovery rate, as anticipated by the theory. The cross-prior statistics are more powerful than the data-splitting ones in the first setting, in which most associations are consistent, and equivalent to the latter in the second setting, where most associations are not consistent. Pooling yields too many false discoveries because it reports all associations regardless of whether they are consistent, while the intersection heuristic may lead to either low power (first setting) or high false discovery rate (second setting).

## 6.2 Testing for partial consistency with synthetic data

We consider partial consistency testing as in Section 4.4. The goal is to discover which variables are non-null in at least  $r = 2$  out of  $E = 5$  environments. The *intersection* benchmark reports all discoveries found in at least  $r$  environments by separate analyses. The data are generated from the same model as in the previous section, although with  $p = 200$  variables and utilizing different model parameters  $\beta$ . The number of samples per environment is  $n = 600$ . Again, we consider two settings with alternative  $\beta$  patterns; see Figure A4 in Appendix D.1. In the first setting, 50 entries of  $\beta^e$  are randomly chosen to be non-zero in all environments and these are the consistent associations we seek; then, 5 unique additional variables are set to be non-zero in each environment; see Figure A4 (a). The absolute values of all non-zero elements of  $\beta^e$  are  $a/\sqrt{n}$ , where  $a$  is a signal amplitude parameter which we will vary. The signs of non-null elements of  $\beta^e$  are determined by independent coin flips. In the second setting, 100 entries of  $\beta^e$  are randomly chosen to be non-zero in the first four environments; then, the remaining 100 variables are set to be non-zero in the last environment; see Figure A4 (b). The values of the non-zero  $\beta^e$  coefficients are determined as in the first setting.

The multi-environment knockoff filter is applied with the empirical cross-prior statistics. The p-values are filtered either by the accumulation test with HingeExp accumulation function, in which case any zero-sign ties are broken as in (16), or with the selective SeqStep+ filter with cutoff parameter  $c = 0.5$ . The nominal false discovery rate is 10%. Figure 6 compares the performance of our method to that of the two heuristic benchmarks, as in the previous section. The results show our method always controls the false discovery rate when applied with the selective SeqStep+ filter, and typically also does so in combination with the accumulation test. Even though the accumulation test theoretically requires p-values with random tie breaking (19), Figure A5 shows the original ones in (19) often lead to higher power, without noticeable loss of type-I error control. Figure 6 highlights that pooling is overly liberal, while the intersection heuristic may be either underpowered (setting 1) or too liberal (setting 2).

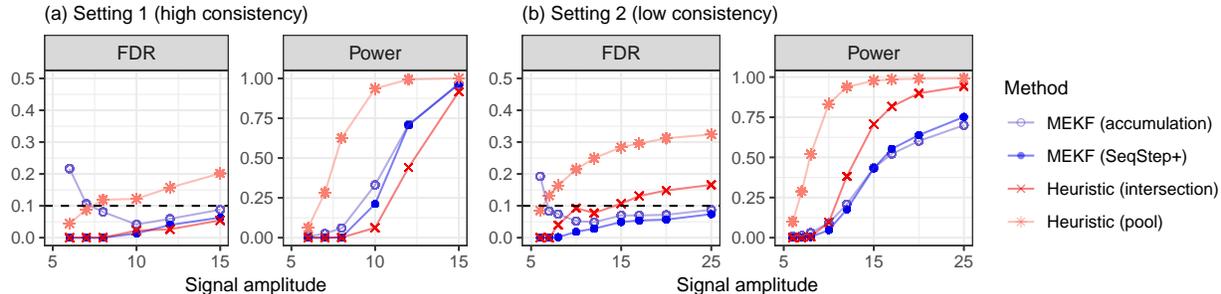


Figure 6: Performance of the multi-environment knockoff filter (MEKF) on simulated data, compared to two heuristics. The goal is to discover which variables are non-null in at least 2 out of 5 environments. Other details are as in Figure 1. Note that the accumulation version of our method is only guaranteed to control a modified version of the false discovery rate, which differs noticeably from our true objective only if the number of discoveries is very small.

### 6.3 Causal inference with biased data

We revisit the school admission example of Section 3.4 by considering a population in which the  $p = 200$  features  $X \in \mathbb{R}^p$  of each applicant are generated from an autoregressive model of order one, with  $\rho^e = 0.6 - 0.1e$  in the  $e$ -th school (environment), for  $e \in \{1, 2\}$ . A constant causal model describes the conditional GPA distribution:  $Y = X\beta + \epsilon$ , where  $\epsilon$  is i.i.d. Gaussian noise, and  $\beta \in \mathbb{R}^p$  has 25 non-zero entries and is constant. The non-zero entries of  $\beta$  are equal to  $2/15$ . The sample size  $n$  is varied as a control parameter. Our goal is to make inferences about the causal model by analyzing data collected through environment-specific biased sampling mechanisms. In particular, a random sample from the population is observed in the  $e$ -th environment if and only if  $Y_i/\sigma + \sum_{i \in S_e} X_i > 0$ , for two subsets  $S_1, S_2 \subset \{1, \dots, p\}$  with  $|S_1| = |S_2| = 25$ , and  $\sigma = 2\sqrt{2}$ . This is a much more ambitious objective compared to Sections 6.1–6.2. In fact, our method explicitly tests the consistency hypotheses  $\mathcal{H}_j^{\text{cst}}$  (2), while the validity of its inferences about  $\mathcal{H}_j^{\text{causal}}$  (5) depends on the coherence across environments of the biased sampling mechanisms. Therefore, we anticipate our causal inferences will be valid only if  $S_1$  and  $S_2$  are sufficiently different from one another.

We compare our method applied with data-splitting statistics to the *pool* and *intersection* benchmarks. The knockoffs are based in each environment on the feature distribution estimated from a much larger sample with the same bias as the data we analyze; this separates the problem of constructing knockoffs from that of testing for consistent associations. Figure 7(a) compares power and false discovery rate, defined in terms of the causal hypotheses  $\mathcal{H}_j^{\text{causal}}$  (5), for disjoint  $S_1, S_2$  as a function of the sample size  $n$ . Figure 7(b) plots analogous results with  $n = 1200$ , as a function of the overlap between  $S_1$  and  $S_2$ , which ranges from 0% (disjoint) to 100% (identical). The results show our causal inferences are valid if the overlap between  $S_1$  and  $S_2$  is small; if it is large, the spurious associations due to sampling bias also become consistent. The intersection heuristic is not as powerful as our method here, while pooling leads to many more non-causal discoveries because it does not account for the sampling biases at all.

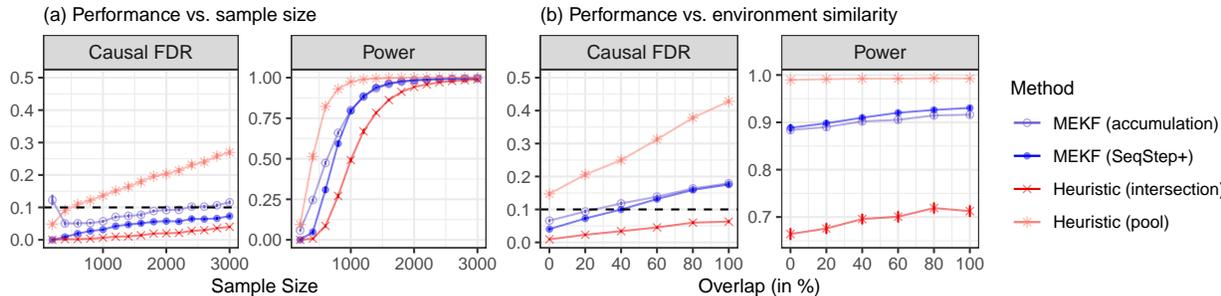


Figure 7: Performance of the multi-environment knockoff filter in a simulated study with environment-specific sample biases. The false discovery rate and power are defined in terms of causal hypotheses. Other details are as in Figure 6.

## 6.4 Causal inference with dependent data

Imagine a population in which the  $p = 200$  features  $X \in \mathbb{R}^p$  of each individual are generated from an autoregressive model of order one, with  $\rho^e = 0.6 - 0.1e$  in the  $e$ -th environment, for  $e \in \{1, 2\}$ . The causal model is constant:  $Y = X\beta + \epsilon$ , where  $\epsilon$  is standard i.i.d. Gaussian noise and  $\beta \in \mathbb{R}^p$  has 40 non-zero entries equal to 0.25. Our goal is to identify the causal variables; however, we cannot collect independent samples from this population. As in the example from Section 3.5, we say an individual  $i$  belongs to a club if  $Y^{(i)} > 0$  (homophily). For any  $i$  belonging to a club in the  $e$ -th environment, we observe modified features  $X_j^{(i)} \leftarrow X_j^{(i)} + I_j^e$  (contagion) for all  $j \in S^e \subseteq \{1, \dots, p\}$ , where each  $I_j^e \in \{0, 1\}$  is an independent coin flip shared by all individuals in environment  $e$ . We fix  $|S_1| = |S_2| = 25$ .

Figure 8(a) compares the performance of our method, with data-splitting statistics, to that of the usual two benchmarks, as a function of the sample size  $n$  in each environment. The nominal false discovery rate is 10%. Here,  $S_1, S_2$  are randomly chosen and disjoint. The knockoffs are based on larger sets of independent samples, as in the previous section. Again, we define the false discovery rate and power in terms of  $\mathcal{H}_j^{\text{causal}}$  (5); therefore, the causal validity of our inferences is only guaranteed if  $S_1 \cap S_2 = \emptyset$ . Indeed, here the multi-environment knockoff filter is more powerful than the intersection heuristic and controls the causal false discovery rate, unlike pooling. Figure 8(b) compares the performances of these methods as a function of  $|S_1 \cap S_2|$ , fixing  $n = 600$ . This shows the multi-environment knockoff filter yields valid causal inferences if two environments are sufficiently different.

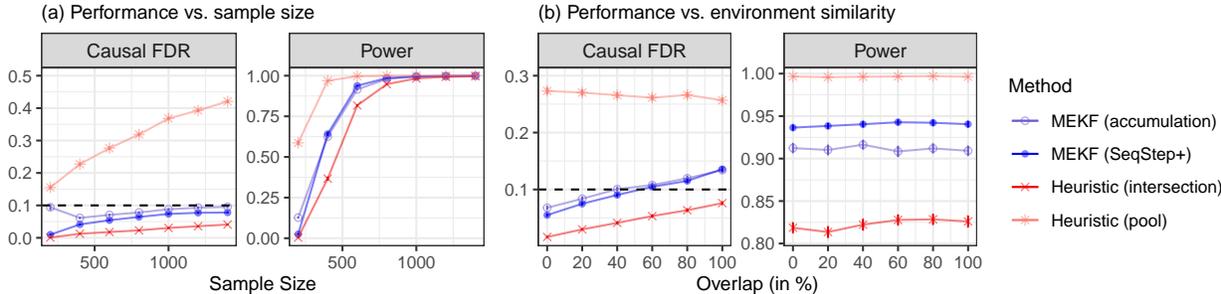


Figure 8: Performance of the proposed method in a simulated study with environment-specific homophily and contagion. Other details are as in Figure 6.

## 6.5 Causal inference in a simulated genome-wide association study

We analyze simulated yet realistic genetic association data involving different sub-populations, based on the haplotypes in the 1000 Genomes Project, Phase 3 [81]. This resource contains phased haplotypes for individuals belonging to one of five possible sub-populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). We utilize these real haplotypes to simulate genetic data from a hidden Markov model for 50,000 individuals belonging to one of the five sub-populations (10,000 individuals per sub-population); see Appendix D.2 for details. This approach ensures the genotype distribution is known exactly, allowing us to focus fully on the problem of accounting for missing variants. In particular, we can apply the algorithm from [5] to generate perfectly valid knockoffs separately within each sub-population, without having to estimate a hidden Markov model [4] or account for population structure [6]. We construct group-level knockoffs [5] for testing  $\mathcal{H}_G^{\text{ci},e}$  (20) at two levels of resolutions, with genetic segments of median lengths 233 kb and 15 kb, respectively. In the interest of time, we only analyze 359,811 biallelic single-nucleotide polymorphisms on chromosome 22 rather than the full genome.

Conditional on the genotypes, we simulate a continuous trait for all 50,000 individuals from a linear model with independent Gaussian errors and 50 causal variants. This model is constant across all populations (environments). The causal variables are randomly chosen among the 359,811 possible genotypes, ensuring each sub-population has at least 10 causal variants with minor allele frequency above 0.1. The signs of the causal effects are independent coin flips, and their sizes are inversely proportional to the standard deviation of the allele count, so that rarer variants have larger effects. The total heritability of the trait is varied as a control parameter. All causal variants are unmeasured, so that their exact identification from the data is impossible; instead, the goal is to localize as precisely as possible which genetic segments are likely to contain causal variants [5], controlling the false discovery rate. The proportion of typed variants is varied between 1% and 10%. In each case, we construct knockoffs only for the measured variants. This setup is likely to be even more challenging than a real genome-wide association study from the point of view of missing variants, because our genotyping is completely random. By contrast, genotyping arrays in real studies

are carefully designed. For example, the UK Biobank [15] data were collected using the UK Biobank Axiom™ Array, which specifically targets potentially causal coding variations, genomic regions of interest, and markers known to be associated with various phenotypes [15]. Thus, confounding may be a less severe problem in practice compared to what we shall experience here.

We perform (consistent) conditional testing at the 10% nominal false discovery rate level but measure performance in stricter terms, based on the causal false discovery rate and power: a discovery is counted as true if and only if it reports a genetic segment containing a causal variant. The power is defined as the average proportion of segments containing causal variants that are discovered. All results are averaged over 100 experiments with independent traits. In theory, the genotypes should also be resampled to ensure false discovery rate control because the knockoffs treat them as random; however, that would be computationally expensive with such large data.

Figure 9 (a) summarises the results of separate analyses in each sub-population as a function of the trait heritability, in the case in which only 1% of all variants are observed. These analyses do not lead to valid causal inferences, although they correctly test conditional association, demonstrating the need for consistency especially at high resolution. The multi-environment knockoff filter is applied to test whether any associations are significant in at least 3 environments, utilizing the data-splitting statistics due to the large size of this data set. Statistical significance is determined either with the accumulation test or with the selective SeqStep+ approach. Our method is compared to the *intersection* and *pool* benchmarks. The results indicate the multi-environment knockoff filter controls the causal false discovery rate and, when applied with the accumulation test, is only slightly less powerful than pooling at low resolution. The selective SeqStep+ approach tends to yield lower power, plausibly because it cannot extract as much information from the p-values. At higher resolution, our method is not as powerful as pooling, while the causal false discovery rate inflation of the latter becomes more severe. Unsurprisingly, all methods are less powerful at high resolution. The causal false discovery rate violation of the intersection heuristic is smaller but noticeable.

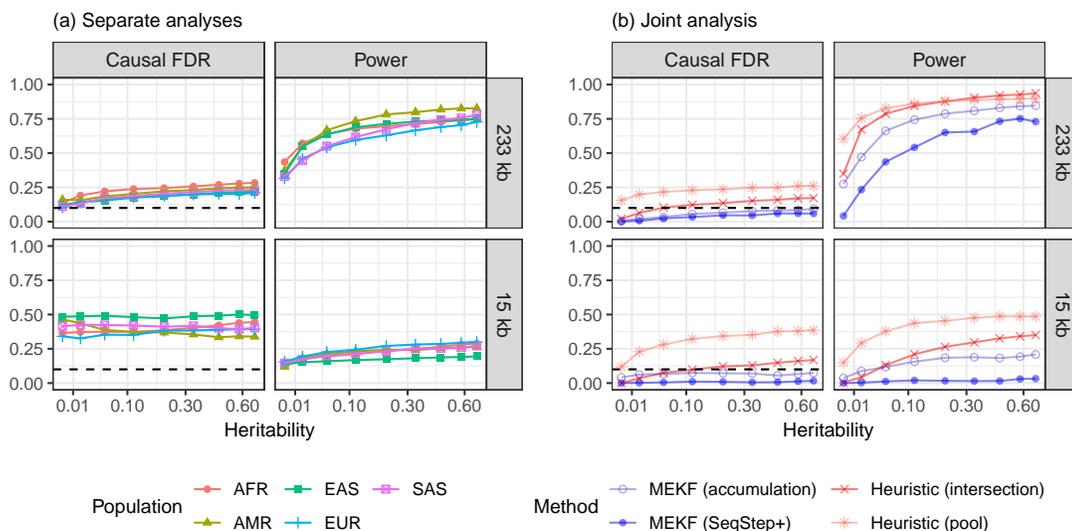


Figure 9: Analysis of a simulated multi-population genome-wide association study in which the causal variants are missing. The genotyping density is 1%. Top: low-resolution analysis (233 kb); bottom: high-resolution analysis (15 kb). The empirical false discovery rate and power are defined in a strict causal sense. The multi-environment knockoff filter seeks associations supported by the data from at least 3 populations. The nominal false discovery rate is 10%.

Figure A6 in Appendix D.2 shows confounding due to unmeasured variants decreases as the genotyping density increases, until all methods control the causal false discovery rate. This is intuitive because there is no confounding in the limit of high genotyping density. Meanwhile, as the genotyping density increases, the multi-environment knockoff filter becomes unnecessarily conservative compared to pooling [5, 6]; this may seem unavoidable but the results in Figure 1 suggest the relative performance of our method would improve if we applied the cross-prior statistics from Section 4.2 instead of the data-splitting ones adopted here for convenience. Figure A7 shows qualitatively similar results corresponding to analyses at the 20% nominal false discovery rate level, which better highlight the causal type-I error inflation incurred by the heuristics. Finally, Figure A8 shows our method performs similarly regardless of whether the accumulation test is applied the p-values computed with random tie breaking (19) or without it (16).

## 7 Analysis of UK Biobank genome-wide association data

### 7.1 Data pre-processing

We study four continuous traits (body mass index, height, platelet count, and systolic blood pressure) and four diseases (cardiovascular disease, diabetes, hypothyroidism, and respiratory disease) using the UK Biobank [15] data; see Table A1 in Appendix E for more details. This analysis is based on the same quality control filtering and knockoffs for 486,975 genotyped and phased subjects in the UK Biobank (application 27837) as in [6]. The knockoffs preserve both the population structure and the kinship of the 136,818 individuals with close relatives; this accounts for most possible confounders except missing variants [6]. Our goal is to address this remaining limitation with the multi-environment knockoff filter. As in previous work [6], we only analyze 591,513 biallelic single nucleotide polymorphisms with minor allele frequency above 0.1% and in Hardy-Weinberg equilibrium ( $10^{-6}$ ) among the subset of 350,119 unrelated British individuals previously analyzed in [5]. The genome is then partitioned into contiguous groups at 7 levels of resolution, ranging from that of single polymorphisms to that of 425 kb-wide groups, as in [6]. The resolution of each genomic partition we consider is defined as the median width of its genetic segments.

The UK Biobank subjects who passed the above quality control are divided into five sub-populations based on their self-reported ancestry (African: 7,635; Asian: 3,284; British: 429,934; Non-British European: 28,994; and Indian: 7,628). We exclude subjects with missing ancestry information, as well as those falling outside these five broad categories; this leaves us with a total of 477,475 individuals; see Table A2, Appendix E, for additional details.

### 7.2 Searching for consistent associations

We apply the multi-environment knockoff filter to discover genetic segments containing distinct association with the phenotype in at least  $r$  environments, with  $r$  ranging from 2 to 5. In all cases, the significance threshold is computed by applying the accumulation test to the p-values in (16). In fact, the accumulation test without the random tie breaking (19) tends to be more powerful than selective SeqStep+ (Section 6), and tie breaking seems practically unnecessary; see Figure A5 in Appendix D. The analysis is performed at the 10% false discovery rate level, separately for each level of resolution [5]. We adopt the data-splitting statistics because the data set is very large. The intersection heuristic and the pool analysis on all UK Biobank samples from [6] will serve as benchmarks.

We repeat all tests with 100 independent realizations of the  $U_j$  variables in (16); this allows some understanding and a possible reduction of the variability of any findings, as our method is randomized. Alternatively, one may repeat the entire analysis starting from the generation of the knockoffs [82]; however, that would be impractical for a data set of this size. In comparison, the cost of resampling the  $U_j$  variables many times is negligible. Table 1 reports the numbers of discoveries for *height* and *platelet count* thus obtained in at least 51 out of 100 randomizations. The results for other phenotypes are in Table A3, Appendix E, for lack of space. Unfortunately, there are fewer consistent associations for the other phenotypes, consistently with previous observations that *height* and *platelet count* display the strongest signals [5, 6]. Our “stability selection” [83] reporting rule is not theoretically guaranteed to control the false discovery rate [2, 4]; however, we can empirically confirm it to be conservative; see Figure A9 in Appendix D. Figure A10 summarises the variability of the individual findings corresponding to different p-value randomizations.

Several consistent associations are discovered, although the power seems lower compared to pooling [6]. This is unsurprising, especially if  $r > 2$ , because the sample sizes are imbalanced: most individuals have either British or other European ancestry. Figure 10 visualizes some discoveries for platelet count through a Chicago plot [5], highlighting in different colors the numbers of environments across which the findings are consistent. It is not guaranteed that all discoveries corresponding to a fixed  $r \in \{2, \dots, E\}$  are also found with  $r' < r$ , although this occurs often; see Figure A11 in Appendix E. Table A4 in Appendix E summarises the findings obtained with selective SeqStep+ instead of the accumulation test, as well those obtained with the intersection heuristic. Clearly, it cannot be determined from Table A4 which approach is most effective at causal inference because the ground truth is unknown. Therefore, we will seek more evidence in support of our findings using prior domain knowledge.

### 7.3 Validation of genetic findings

Table 2 demonstrates almost all of our consistent discoveries for height and platelet count are confirmed by the NHGRI-EBI GWAS Catalog [84] (accessed on April 15, 2021). We say that a discovered genetic segment is confirmed if it spans a genomic region containing reported associations for the same phenotype. Relatively fewer discoveries obtained by pooling [6] are thus confirmed. Of course, this is not fully conclusive because the GWAS Catalog may

Table 1: Numbers of discoveries at different resolutions for two UK Biobank phenotypes. The second column indicates the numbers of populations (environments) across which the findings are consistent. The third column corresponds to the analysis of the pooled data from all populations [6]. The nominal false discovery rate is 10%.

Phenotype	Resolution (kb)	Number of environments				
		1	2	3	4	5
height	single-SNP	95	13	2	0	0
	3	570	9	6	0	0
	20	1503	33	0	0	0
	41	2384	42	7	7	2
	81	3006	48	24	0	0
	208	3339	103	23	7	3
	425	3073	68	26	3	0
platelet	single-SNP	53	9	3	0	0
	3	246	10	4	4	0
	20	1002	27	16	2	0
	41	1261	52	12	9	0
	81	1570	104	15	8	0
	208	1743	98	16	14	2
	425	1653	119	9	11	0

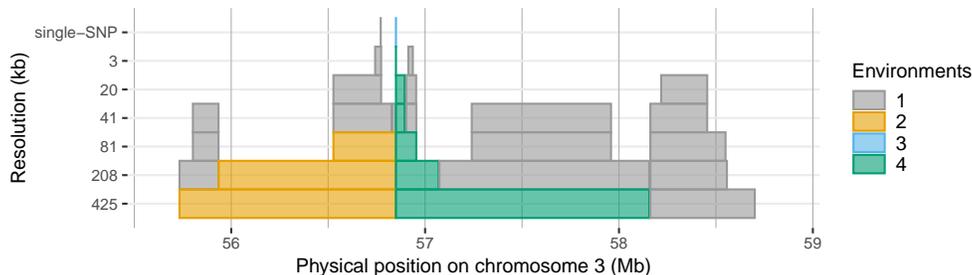


Figure 10: Chicago plot of some discoveries on chromosome three for platelet count based on UK Biobank data from individuals in five sub-populations (environments). Each block represents a genetic segment containing distinct associations; the colors indicate the numbers of environments across which they are consistent. The vertical position denotes the resolution of the discovery measured in millions of base pairs (Mb). Other details are as in Table 1.

include spurious associations and is likely to miss many causal ones, although it is a standard reference. Table 2 summarises the numbers of findings obtained with the intersection heuristic, as well as the proportions of those which are confirmed by the GWAS Catalog. This shows the intersection heuristic yields either fewer discoveries, or a (slightly) lower validation rate. This is consistent with our simulations suggesting this heuristic is often either underpowered or excessively liberal. Analogous information for the other phenotypes is in Table A5, Appendix E. Table A6 reports the names and associated genes of the genetic variants identified by our method at the single-nucleotide resolution. These results indicate all but two of our high-resolution consistent discoveries correspond to variants with known biological consequences, which are located on genes previously reported to be associated with the phenotypes of interest. The full list of discoveries is available online at <https://msesia.github.io/knockoffgwas/>.

## 8 Discussion

**Consistency, causal inference, and reliability.** This paper proposed a practical high-dimensional method to search for conditional associations that are consistent across environments, provably controlling the false discovery rate within the model-X framework [2]. While consistency can lead to valid causal inferences under certain conditions, the relevance of our method extends beyond the relatively narrow scope of the assumptions necessary for such formal connection. In fact, conditional associations and consistency are meaningful statistical concepts even in

Table 2: Proportions of discoveries confirmed by the GWAS Catalog, for two UK Biobank phenotypes. The multi-environment knockoff filter (MEKF) reports discoveries that are consistent in at least 2 populations, as in Table 1. Pooling refers to the analysis of [6]. The binomial p-value tests whether the proportion of our confirmed discoveries differs significantly from that corresponding to pooling. The intersection heuristic is the same as in Table A4.

Phenotype	Resolution (kb)	MEKF	Pooling [6]	Binomial p-value	Intersection
height	single-SNP	11 / 13 (85%)	63 / 95 (66%)	$3.11 \cdot 10^{-01}$	0 / 0
	3	9 / 9 (100%)	360 / 570 (63%)	$5.34 \cdot 10^{-02}$	0 / 0
	20	33 / 33 (100%)	1043 / 1503 (69%)	$3.12 \cdot 10^{-04}$	25 / 25 (100%)
	41	42 / 42 (100%)	1567 / 2384 (66%)	$6.99 \cdot 10^{-06}$	68 / 68 (100%)
	81	48 / 48 (100%)	1875 / 3006 (62%)	$1.94 \cdot 10^{-07}$	83 / 84 (99%)
	208	103 / 103 (100%)	1968 / 3339 (59%)	$1.21 \cdot 10^{-16}$	102 / 107 (95%)
	425	68 / 68 (100%)	1794 / 3073 (58%)	$1.16 \cdot 10^{-11}$	154 / 164 (94%)
platelet	single-SNP	8 / 9 (89%)	44 / 53 (83%)	1	0 / 0
	3	10 / 10 (100%)	200 / 246 (81%)	$2.76 \cdot 10^{-01}$	0 / 0
	20	27 / 27 (100%)	684 / 1002 (68%)	$9.31 \cdot 10^{-04}$	26 / 26 (100%)
	41	52 / 52 (100%)	804 / 1261 (64%)	$1.71 \cdot 10^{-07}$	50 / 50 (100%)
	81	101 / 104 (97%)	921 / 1570 (59%)	$1.54 \cdot 10^{-14}$	67 / 69 (97%)
	208	97 / 98 (99%)	937 / 1743 (54%)	$4.16 \cdot 10^{-18}$	57 / 58 (98%)
	425	119 / 119 (100%)	887 / 1653 (54%)	$1.67 \cdot 10^{-22}$	70 / 75 (93%)

situations where discussing causality would require more care, either because it is not obvious that the explanatory variables predate the outcome, as in the case of medical imaging data [85], or because there is no clear notion of possible interventions [86]. However, non-causal conditional associations can still be informative, especially if they are reproducible outside the data set in which they were discovered. For example, consistent conditional associations are useful to make reliable predictions, and fitting predictive models that can be accurate across different environments (*transfer learning*) is a well-known challenge in many fields, including genetics [87], machine learning [88], and econometrics [10], to name a few. Although this paper does not address prediction explicitly, the problem is related and our proposed method could be repurposed to select good predictors for transfer learning.

**Genome-wide association studies.** Genome-wide association studies are a natural application for our method because the model-X setup is supported by scientific knowledge of genetic inheritance [4]. Further, these studies are primarily exploratory, aiming to prioritize variants for follow-up investigations, which makes the false discovery rate a meaningful measure [89–91]. A constant causal model is also quite realistic here. First, there is little ambiguity about the causal direction because the genotypes are fixed at conception while the phenotype manifests itself later. Second, the biological mechanisms translating genotypes to phenotypes are likely the same for all humans: the differences across sub-populations lie in the genotypes. The paucity of non-British samples limits our power with the UK Biobank data, but the growing awareness that genetic studies should increase the representation of different ancestries [87, 92] suggests promising future opportunities, especially as some large diverse studies already exist [93].

**Opportunities for future work.** Our method may be useful in many fields, including the social sciences; there, it is easy to envision collecting data from multiple environments and consistency may help ensure the samples are truly random and free of network effects. Further, it is increasingly common to find high-dimensional data with many associations, for which controlling the false discovery rate is desirable. Regarding methodology, it may be possible to develop more powerful test statistics. Finally, we have proved selective SeqStep+ [3] and the accumulation test [69] are valid under a mild form of dependency which suggests broader applicability than previously known.

## Acknowledgements

We thank Stefan Wager for insightful comments about an earlier manuscript draft, as well as the Center for Advanced Research Computing at the University of Southern California and the Research Computing Center at Stanford University for computing resources. We are grateful to the participants and investigators of the UK Biobank.

## References

- [1] B. Devlin and K. Roeder. “Genomic control for association studies”. In: *Biometrics* 55.4 (1999), pp. 997–1004.
- [2] E. Candès, Y. Fan, L. Janson, and J. Lv. “Panning for gold: “model-X” knockoffs for high dimensional controlled variable selection”. In: *J. R. Stat. Soc. B* 80.3 (2018), pp. 551–577.
- [3] R. F. Barber and E. Candès. “Controlling the false discovery rate via knockoffs”. In: *Ann. Stat.* 43.5 (2015), pp. 2055–2085.
- [4] M. Sesia, C. Sabatti, and E. J. Candès. “Gene hunting with hidden Markov model knockoffs”. In: *Biometrika* 106.1 (Aug. 2018), pp. 1–18.
- [5] M. Sesia, E. Katsevich, S. Bates, E. Candès, and C. Sabatti. “Multi-resolution localization of causal variants across the genome”. In: *Nat. Commun.* 11.1 (2020), pp. 1–10.
- [6] M. Sesia, S. Bates, E. Candès, J. Marchini, and C. Sabatti. “FDR control in GWAS with population structure”. In: *preprint at bioRxiv* (2021).
- [7] Y. Romano, M. Sesia, and E. Candès. “Deep Knockoffs”. In: *J. Am. Stat. Assoc.* 0.ja (2019), pp. 1–27.
- [8] J. Pearl. *Causality*. Cambridge university press, 2009.
- [9] J. K. Pritchard and M. Przeworski. “Linkage disequilibrium in humans: models and data”. In: *Am. J. Hum. Genet* 69.1 (2001), pp. 1–14.
- [10] J. J. Heckman. “Sample selection bias as a specification error”. In: *Econometrica* (1979), pp. 153–161.
- [11] T. Harford. “Big data: A big mistake?” In: *Significance* 11.5 (2014), pp. 14–19.
- [12] Y. Lee and E. L. Ogburn. “Network dependence can lead to spurious associations and invalid inference”. In: *J. Am. Stat. Assoc* (2020), pp. 1–15.
- [13] D. Hume. *A Treatise of Human Nature: A Critical Edition*. London: John Noon, 1739.
- [14] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *J. R. Stat. Soc. B.* 57 (1995), pp. 289–300.
- [15] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562 (2018), pp. 203–209.
- [16] J. Peters, P. Bühlmann, and N. Meinshausen. “Causal inference by using invariant prediction: identification and confidence intervals”. In: *J. R. Stat. Soc. B* (2016), pp. 947–1012.
- [17] C. Heinze-Deml, J. Peters, and N. Meinshausen. “Invariant causal prediction for nonlinear models”. In: *Journal of Causal Inference* 6.2 (2018).
- [18] K. Zhang, M. Gong, and B. Schölkopf. “Multi-source domain adaptation: A causal view”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.
- [19] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. “Invariant models for causal transfer learning”. In: *J. Mach. Learn. Res.* 19.1 (2018), pp. 1309–1342.
- [20] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. “Invariant Risk Minimization”. In: *Stat* 1050 (2020), p. 27.
- [21] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. “Anchor regression: Heterogeneous data meet causality”. In: *J. R. Stat. Soc. B* (2021).
- [22] H. Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *J. Statist. Plann. Inference* 90.2 (2000), pp. 227–244.
- [23] M. Sugiyama, M. Krauledat, and K.-R. Müller. “Covariate shift adaptation by importance weighted cross validation.” In: *J. Mach. Learn. Res.* 8.5 (2007).
- [24] P. Spirtes, C. Meek, and T. Richardson. *An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery, 1999*. 1999.

- [25] D. M. Chickering. “Optimal structure identification with greedy search”. In: *J. Mach. Learn. Res.* 3.Nov (2002), pp. 507–554.
- [26] M. Koivisto and K. Sood. “Exact Bayesian structure discovery in Bayesian networks”. In: *J. Mach. Learn. Res.* 5 (2004), pp. 549–573.
- [27] J. Zhang. “On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias”. In: *Artificial Intelligence* 172.16-17 (2008), pp. 1873–1896.
- [28] C. Glymour, K. Zhang, and P. Spirtes. “Review of causal discovery methods based on graphical models”. In: *Frontiers in genetics* 10 (2019), p. 524.
- [29] J. M. Mooij, S. Magliacane, and T. Claassen. “Joint Causal Inference from Multiple Contexts”. In: *J. Mach. Learn. Res.* 21.99 (2020), pp. 1–108.
- [30] K. Yu, L. Liu, J. Li, W. Ding, and T. D. Le. “Multi-source causal feature selection”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.9 (2019), pp. 2240–2256.
- [31] Z. Ling, K. Yu, H. Wang, L. Liu, W. Ding, and X. Wu. “BAMB: A balanced Markov blanket discovery approach to feature selection”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.5 (2019), pp. 1–25.
- [32] H. Wang, Z. Ling, K. Yu, and X. Wu. “Towards efficient and effective discovery of Markov blankets for feature selection”. In: *Inf. Sci.* 509 (2020), pp. 227–242.
- [33] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu. “Causality-based feature selection: Methods and evaluations”. In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–36.
- [34] J. R. Gimenez, A. Ghorbani, and J. Zou. “Knockoffs for the mass: new feature importance statistics with false discovery guarantees”. In: *22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 2125–2133.
- [35] S. Bates, E. Candès, L. Janson, and W. Wang. “Metropolized knockoff sampling”. In: *J. Am. Stat. Assoc.* (2020), pp. 1–15.
- [36] R. F. Barber, E. Candès, and R. J. Samworth. “Robust inference with knockoffs”. In: *Ann. Stat.* 48.3 (2020), pp. 1409–1431.
- [37] J. Liu and P. Rigollet. “Power analysis of knockoff filters for correlated designs”. In: *NeurIPS* (2019).
- [38] E. Katsevich and A. Ramdas. “A theoretical treatment of conditional independence testing under model-X”. In: *preprint at arXiv:2005.05506* (2020).
- [39] W. Wang and L. Janson. “A Power Analysis of the Conditional Randomization Test and Knockoffs”. In: *preprint at arXiv:2010.02304* (2020).
- [40] A. Spector and L. Janson. “Powerful Knockoffs via Minimizing Reconstructability”. In: *preprint at arXiv:2011.14625* (2020).
- [41] S. Bates, M. Sesia, C. Sabatti, and E. Candès. “Causal inference in genetic trio studies”. In: *Proc. Natl. Acad. Sci. U.S.A* 117.39 (2020), pp. 24117–24126.
- [42] A. Shen, H. Fu, K. He, and H. Jiang. “False discovery rate control in cancer biomarker selection using knockoffs”. In: *Cancers* 11.6 (2019), p. 744.
- [43] C. Chia, M. Sesia, C.-S. Ho, S. Jeffrey, J. Dionne, E. Candès, and R. Howe. “Interpretable Classification of Bacterial Raman Spectra with Knockoff Wavelets”. In: *preprint at arXiv:2006.04937* (2021).
- [44] Y. Fan, J. Lv, M. Sharifvaghefi, and Y. Uematsu. “IPAD: stable interpretable forecasting with knockoffs inference”. In: *J. Am. Stat. Assoc.* 115.532 (2020), pp. 1822–1834.
- [45] A. Srinivasan, L. Xue, and X. Zhan. “Compositional knockoff filter for high-dimensional regression analysis of microbiome data”. In: *Biometrics* (2020).
- [46] E. Katsevich, C. Sabatti, and M. Bogomolov. “Filtering the rejection set while preserving false discovery rate control”. In: *J. Am. Stat. Assoc.* just-accepted (2021), pp. 1–27.
- [47] K. J. Friston, W. D. Penny, and D. E. Glaser. “Conjunction revisited”. In: *Neuroimage* 25.3 (2005), pp. 661–667.

- [48] Y. Benjamini and R. Heller. “Screening for partial conjunction hypotheses”. In: *Biometrics* 64.4 (2008), pp. 1215–1222.
- [49] R. Heller and D. Yekutieli. “Replicability analysis for genome-wide association studies”. In: *Ann. Appl. Stat.* 8.1 (2014), pp. 481–498.
- [50] J. Wang, W. Su, C. Sabatti, and A. B. Owen. “Detecting Replicating Signals using Adaptive Filtering Procedures with the Application in High-throughput Experiments”. In: *preprint at arXiv:1610.03330* (2016).
- [51] M. Kosinski, D. Stillwell, and T. Graepel. “Private traits and attributes are predictable from digital records of human behavior”. In: *Proc. Natl. Acad. Sci. U.S.A.* 110.15 (2013), pp. 5802–5805.
- [52] L. Waldron, B. Haibe-Kains, A. C. Culhane, M. Riester, J. Ding, X. V. Wang, M. Ahmadifar, S. Tyekucheva, C. Bernau, T. Risch, B. F. Ganzfried, C. Huttenhower, M. Birrer, and G. Parmigiani. “Comparative Meta-analysis of Prognostic Gene Signatures for Late-Stage Ovarian Cancer”. In: *JNCI: Journal of the National Cancer Institute* 106.5 (Apr. 2014). dju049.
- [53] B. Efron. “Prediction, Estimation, and Attribution”. In: *J. Am. Stat. Assoc* 115.530 (2020), pp. 636–655.
- [54] K. Boolen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.
- [55] D. B. Rubin. “Causal inference using potential outcomes: Design, modeling, decisions”. In: *J. Am. Stat. Assoc* 100.469 (2005), pp. 322–331.
- [56] J. Neyman and K. Iwazskiewicz. “Statistical problems in agricultural experimentation”. In: *Supplement to J. R. Stat. Soc.* 2.2 (1935), pp. 107–180.
- [57] L. Keele. “The statistics of causal inference: A view from political methodology”. In: *Political Analysis* (2015), pp. 313–335.
- [58] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- [59] E. Hargittai. “Is bigger always better? Potential biases of big data derived from social network sites”. In: *Ann. Am. Acad. Pol. Soc. Sci.* 659.1 (2015), pp. 63–76.
- [60] C. R. Shalizi and A. C. Thomas. “Homophily and contagion are generically confounded in observational social network studies”. In: *Sociol. Methods Res.* 40.2 (2011), pp. 211–239.
- [61] J. Berkson. “Limitations of the application of fourfold table analysis to hospital data”. In: *Biometrics Bulletin* 2.3 (1946), pp. 47–53.
- [62] A. Herbert, G. Griffith, G. Hemani, and L. Zuccolo. “The spectre of Berkson’s paradox: Collider bias in Covid-19 research”. In: *Significance* 17.4 (2020), pp. 6–7.
- [63] R. M. Dawes. “Graduate admission variables and future success”. In: *Science* 187.4178 (1975), pp. 721–723.
- [64] M. McPherson, L. Smith-Lovin, and J. M. Cook. “Birds of a feather: Homophily in social networks”. In: *Annu. Rev. Sociol.* 27.1 (2001), pp. 415–444.
- [65] J. H. Fowler and N. A. Christakis. “Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study”. In: *Bmj* 337 (2008).
- [66] S. Aral, L. Muchnik, and A. Sundararajan. “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”. In: *Proc. Natl. Acad. Sci. U.S.A.* 106.51 (2009), pp. 21544–21549.
- [67] R. Tibshirani. “Regression shrinkage and selection via the lasso: a retrospective”. In: *J. R. Stat. Soc. B* 73.3 (2011), pp. 273–282.
- [68] J. Neyman and E. S. Pearson. “On the problem of the most efficient tests of statistical hypotheses”. In: *Philos. Trans. R. Soc.* 231.694-706 (1933), pp. 289–337.
- [69] A. Li and R. F. Barber. “Accumulation tests for FDR control in ordered hypothesis testing”. In: *J. Am. Stat. Assoc* 112.518 (2017), pp. 837–849.

- [70] M. Slatkin. “Linkage disequilibrium in growing and stable populations.” In: *Genetics* 137.1 (1994), pp. 331–336.
- [71] J. Marchini and B. Howie. “Genotype imputation for genome-wide association studies”. In: *Nat. Rev. Genet.* 11 (2010), pp. 499–511.
- [72] D. J. Schaid, W. Chen, and N. B. Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nat. Rev. Genet.* 19.8 (2018), pp. 491–504.
- [73] N. Li and M. Stephens. “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data”. In: *Genetics* 165 (2003), pp. 2213–2233.
- [74] M. Laan and S. Pääbo. “Demographic history and linkage disequilibrium in human populations”. In: *Nat. Genet.* 17.4 (1997), pp. 435–438.
- [75] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. “High-resolution haplotype structure in the human genome”. In: *Nat. Genet.* 29.2 (2001), pp. 229–232.
- [76] J. D. Wall and J. K. Pritchard. “Haplotype blocks and linkage disequilibrium in the human genome”. In: *Nat. Rev. Genet.* 4.8 (2003), pp. 587–597.
- [77] M. Slatkin. “Linkage disequilibrium—understanding the evolutionary past and mapping the medical future”. In: *Nat. Rev. Genet.* 9.6 (2008), pp. 477–485.
- [78] T. Berisa and J. K. Pickrell. “Approximately independent linkage disequilibrium blocks in human populations”. In: *Bioinformatics* 32.2 (2016), p. 283.
- [79] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *J. Stat. Softw.* 33.1 (2010), p. 1.
- [80] F. Privé, H. Aschard, and M. G. Blum. “Efficient implementation of penalized regression for genetic risk prediction”. In: *Genetics* 212.1 (2019), pp. 65–74.
- [81] I. G. P. Consortium et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68.
- [82] Z. Ren, Y. Wei, and E. Candès. “Derandomizing Knockoffs”. In: *preprint at arXiv:2012.02717* (2020).
- [83] N. Meinshausen and P. Bühlmann. “Stability selection”. In: *J. R. Stat. Soc. B* 72.4 (2010), pp. 417–473.
- [84] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic Acids Res.* 47.D1 (2019), pp. D1005–D1012.
- [85] D. C. Castro, I. Walker, and B. Glocker. “Causality matters in medical imaging”. In: *Nat. Commun.* 11.1 (2020), pp. 1–10.
- [86] M. A. Hernán and S. L. Taubman. “Does obesity shorten life? The importance of well-defined interventions to answer causal questions”. In: *Int. J. Obes.* 32.3 (2008), S8–S14.
- [87] L. Duncan, H. Shen, B. Gelaye, J. Meijssen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. “Analysis of polygenic risk score usage and performance in diverse human populations”. In: *Nat. Commun.* 10.1 (July 2019), p. 3328.
- [88] S. J. Pan and Q. Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [89] J. D. Storey and R. Tibshirani. “Statistical significance for genomewide studies”. In: *Proc. Natl. Acad. Sci. U.S.A.* 100.16 (2003), pp. 9440–9445.
- [90] C. Sabatti, S. Service, and N. Freimer. “False discovery rate in linkage and association genome screens for complex disorders”. In: *Genetics* 164.2 (2003), pp. 829–833.
- [91] Y. Benjamini and D. Yekutieli. “Quantitative trait Loci analysis using the false discovery rate”. In: *Genetics* 171.2 (2005), pp. 783–790.

- [92] A. B. Popejoy, D. I. Ritter, K. Crooks, E. Currey, S. M. Fullerton, L. A. Hindorff, B. Koenig, E. M. Ramos, E. P. Sorokin, H. Wand, M. W. Wright, J. Zou, C. R. Gignoux, V. L. Bonham, S. E. Plon, and C. D. Bustamante. “The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics”. In: *Hum. Mutat.* (2018).
- [93] J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, et al. “Million Veteran Program: A mega-biobank to study genetic influences on health and disease”. In: *J. Clin. Epidemiol.* 70 (2016), pp. 214–223.
- [94] R. A. Fisher. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
- [95] M. Liu and L. Janson. “Fast and powerful conditional randomization testing via distillation”. In: *preprint at arXiv:2006.03980* (2020).
- [96] Y. Benjamini and D. Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *Ann. Stat.* (2001), pp. 1165–1188.
- [97] R. J. Simes. “An improved Bonferroni procedure for multiple tests of significance”. In: *Biometrika* 73.3 (1986), pp. 751–754.
- [98] R. F. Barber, E. Candès, and R. J. Samworth. “Robust inference with knockoffs”. In: *Ann. Stat.* 48.3 (2020), pp. 1409–1431.
- [99] J. O’Connell, K. Sharp, N. Shrine, L. Wain, I. Hall, M. Tobin, J.-F. Zagury, O. Delaneau, and J. Marchini. “Haplotype estimation for biobank-scale data sets”. In: *Nat. Genet.* 48.7 (2016), p. 817.

## A Testing for consistency with conditional randomizations

The conditional randomization test was proposed by [2] as an alternative to knockoffs, and it may be seen as an instance of Fisher’s randomization test [94] within the model-X framework. We discuss here how to utilize it to test the consistent hypothesis  $\mathcal{H}_j^{\text{cst}}$  (2), or the partially consistent  $\mathcal{H}_j^{\text{pcst},r}$  (3). We do not aim to control the false discovery rate over all variables; instead, we focus on a single  $j \in \{1, \dots, p\}$ . This problem is of separate interest because the conditional randomization test [2] gives (approximately) continuous p-values, while the single bit of information obtainable with knockoffs can only be significant at an aggregate level, within a multiple-testing procedure [3].

The conditional randomization test simulates independent realizations  $X'_j$  of the variable of interest,  $X_j$ , conditional on all other predictors,  $X_{-j}$ , independently of the outcome. Then, it compares importance statistics  $T_j$  based on the original  $X_j$  to the empirical distribution of the analogous quantities  $T'_j$  evaluated on the perturbed data set obtained by replacing  $X_j$  with the random  $X'_j$ . The output p-value is defined roughly as one minus the empirical percentile of  $T'_j$  in the aforementioned distribution; i.e., larger values of  $T'_j$  result in smaller p-values. Environment by environment, this procedure produces a conservative p-value  $p_j^e$  for  $\mathcal{H}_j^{\text{ci},e}$  (1) because the distribution of  $T'_j$  over multiple realizations of the random  $X'_j$  is equivalent to the true null distribution of  $T_j$  under  $\mathcal{H}_j^{\text{ci},e}$  (1) [2]. A strength of this test is that it is flexible and potentially powerful: it can accommodate any statistics, similarly to knockoffs [2]. Two limitations compared to the latter are: (i) conditional randomization tends to be computationally more expensive, depending on the statistics [95], because  $T'_j$  must be evaluated many times to obtain small p-values; (ii) the p-values for different  $j$  are not independent, complicating the control of the false discovery rate [96].

For any fixed variable  $j$ , let  $p_j^e$  be the conditional randomization p-value for testing  $\mathcal{H}_j^{\text{ci},e}$  (1) in environment  $e$ . Then, a conservative p-value for testing  $\mathcal{H}_j^{\text{cst}}$  (2) is simply given by

$$p_j^{\text{cst}} := \max \left\{ p_j^1, \dots, p_j^E \right\}. \quad (21)$$

Indeed,  $\mathcal{H}_j^{\text{cst}}$  (2) implies there exists at least one environment  $e$  such that  $\mathcal{H}_j^{\text{ci},e}$  (1) is true. Then,  $\forall \alpha \in (0, 1)$ ,

$$\mathbb{P} \left[ p_j^{\text{cst}} \leq \alpha \right] = \mathbb{P} \left[ \max \left\{ p_j^1, \dots, p_j^E \right\} \leq \alpha \right] \leq \mathbb{P} \left[ p_j^e \leq \alpha \right] \leq \alpha,$$

because  $p_j^e$  is a conservative p-value for  $\mathcal{H}_j^{\text{ci},e}$  (1).

To test the partial consistency hypotheses  $\mathcal{H}_j^{\text{pcst},r}$  (3), for any fixed  $r \leq E$ , one can combine the p-values as follows. First, sort the p-values for different environments in ascending order:  $p_j^{(1)} \leq \dots \leq p_j^{(E)}$ . Then, define

$$p_j^{\text{pcst},r} := \min_{r \leq e \leq E} \left\{ \frac{E - r + 1}{e - r + 1} p_j^{(e)} \right\}. \quad (22)$$

This is known as Simes' partial conjunction p-value, at it is valid for  $\mathcal{H}_j^{\text{pcst},r}$  (3) if the p-values  $p_j^e$  from different environments are mutually independent [48, 97]. Note that  $p_j^{\text{cst}}$  (21) is a special case of  $p_j^{\text{cst},r}$  (22) with  $r = E$ .

The following experiments demonstrate the performance of the above conditional randomization p-values for consistency testing. We simulate  $E = 3$  environments,  $p = 100$  variables, and  $n = 200$  observations per environment. The explanatory variables are generated from an autoregressive model of order one with correlation parameter  $\rho = 0.5$ . For each  $e$ , the distribution of  $Y^e | X^e$  is given by a linear model with Gaussian errors:  $Y^e = X^e \beta^e + \epsilon^e$ . The vector  $\beta^e \in \mathbb{R}^p$  is an environment-specific parameter, and  $\epsilon^e$  represents i.i.d. standard Gaussian noise. In each environment, the non-zero entries of  $\beta^e$  are equal to  $3/\sqrt{n}$ . The non-zero entries of  $\beta$  in each environment,  $S^1, S^2 \in \{1, \dots, p\}$ , are chosen at random such that  $S^3 = S^3 \cap S^2 = S^3 \cap S^2 \cap S^1$ ,  $S^2 = S^2 \cap S^1$ , while  $|S^3| = 20$ ,  $|S^2| = 40$  and  $|S^1| = 60$ .

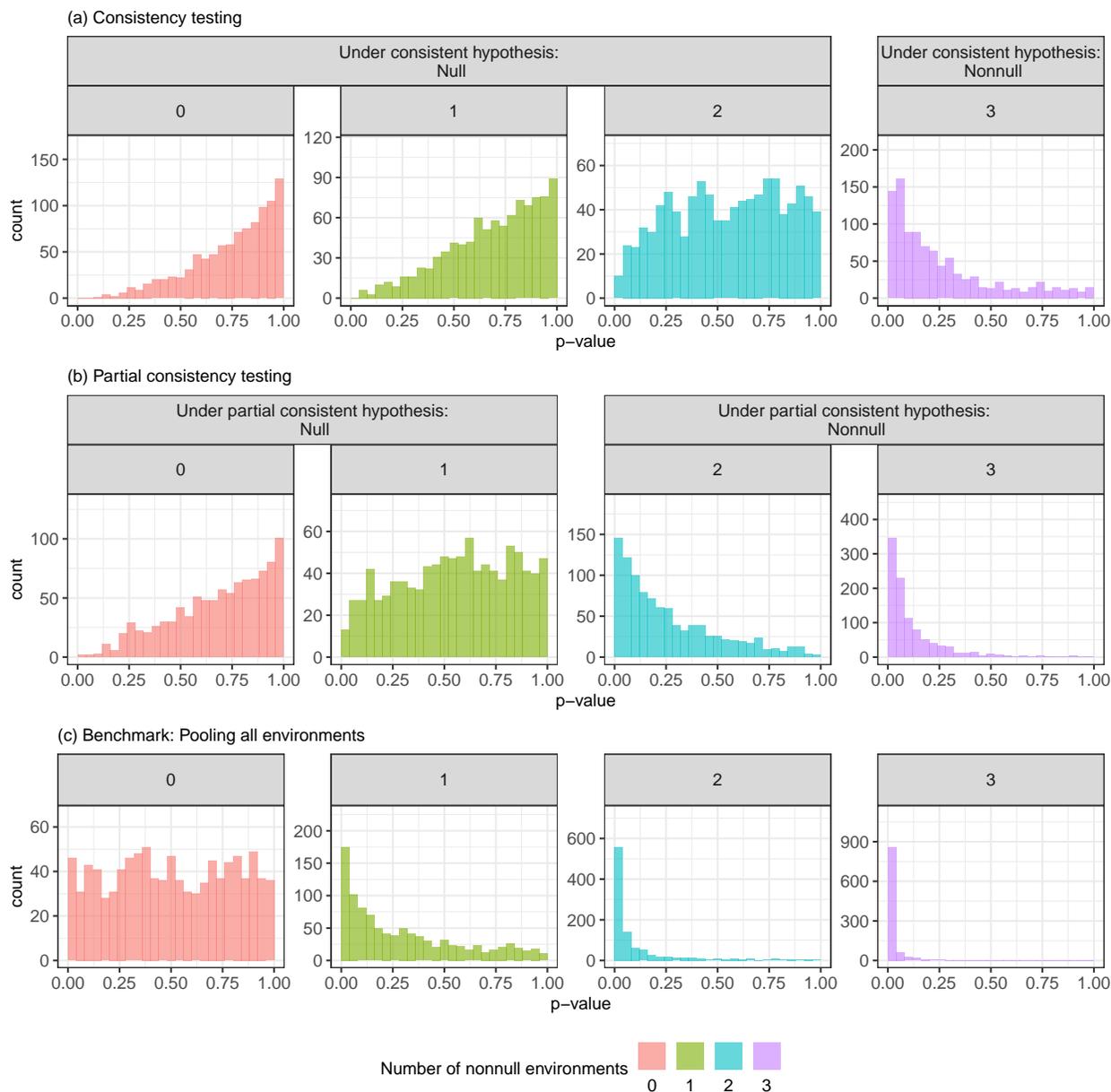


Figure A1: Distributions of p-values for testing consistent (a) and partially consistent (b) associations via conditional randomization, over 1000 experiments. Different columns correspond to different tested variables. The null  $\mathcal{H}_j^{\text{cst}}$  (2) is true for the first three variables in (a), while the null  $\mathcal{H}_j^{\text{pcst},r}$  (3) is true for the first two variables in (b). In (c), we show the distribution of p-values obtained by applying the conditional randomization test to the pooled data.

We focus on four variables indexed by  $j_0, j_1, j_2, j_3$ , which are non-null in the sense of  $\mathcal{H}_j^{\text{ci},e}$  (1) within exactly 0, 1, 2, 3 environments, respectively. The test is carried out separately environment by environment. We take the importance statistics to be the absolute values of the lasso coefficients tuned by 10-fold cross validation. The number of randomizations is 100. The resulting p-values are combined by applying (21), or (22) with  $r = 2$ . This leads to a p-value  $p_j^{\text{cst}}$ , or  $p_j^{\text{pcst},2}$ , for each  $j \in \{j_0, j_1, j_2, j_3\}$ . Figure A1 visualizes the distributions of  $p_j^{\text{cst}}$  and  $p_j^{\text{pcst},2}$  over experiments based on independent data. When the consistent conditional association is tested,  $p_{j_0}^{\text{cst}}, p_{j_1}^{\text{cst}}, p_{j_2}^{\text{cst}}$  are stochastically larger than Uniform[0, 1], as  $\mathcal{H}_j^{\text{cst}}$  (2) is true for the first three variables. At the same time,  $p_{j_3}^{\text{cst}}$  is stochastically smaller, suggesting the method can achieve non-trivial power. Similarly,  $p_{j_0}^{\text{pcst},2}$  and  $p_{j_1}^{\text{pcst},2}$  are stochastically larger than Uniform[0, 1], as  $\mathcal{H}_j^{\text{pcst},2}$  (3) is true for the first two variables, while  $p_{j_2}^{\text{pcst},2}$  and  $p_{j_3}^{\text{pcst},2}$  are clearly stochastically smaller than Uniform[0, 1]. As a heuristic benchmark, Figure A1 shows the distribution of p-values obtained by applying the conditional randomization test to the pooled data from all environments. Clearly, this is not a valid test of any consistency hypotheses. In fact, only  $p_{j_0}^{\text{pool}}$  is stochastically larger than Uniform[0, 1], while the other p-values are stochastically smaller. This demonstrates that  $p_{j_1}^{\text{pool}}$  and  $p_{j_2}^{\text{pool}}$  are not valid p-values for  $\mathcal{H}_j^{\text{cst}}$  (2), and  $p_{j_1}^{\text{pool}}$  is not a valid p-value for  $\mathcal{H}_j^{\text{pcst},2}$  (3).

## B An example of invalid multi-environment statistics

Following the discussion in Section 4.2, we include here an example of invalid multi-environment knockoff statistics that do not lead to false discovery rate control. These naive statistics are invalid because their magnitudes are computed by looking at the unperturbed data from all environments, although their signs only depend on the observations from the environment of interest. We simulate  $E = 2$  environments,  $p = 100$  variables, and  $n = 200$  observations per environment. The explanatory variables are generated from an autoregressive model of order one with correlation parameter  $\rho = 0.6$ . The distribution of  $Y^e | X^e$ , for each  $e \in \{1, 2\}$ , is given by a linear model with Gaussian errors:  $Y^e = X^e \beta^e + \epsilon^e$ . The vector  $\beta^e \in \mathbb{R}^p$  is an environment-specific parameter, and  $\epsilon^e$  represents i.i.d. standard Gaussian noise. In each environment, 70 entries of  $\beta^e$  are non-zero while the others are equal to  $a/\sqrt{n}$ , where  $a$  is a control parameter. The non-zero entries of  $\beta$  in each environment,  $S^1, S^2 \in \{1, \dots, p\}$ , are chosen at random such that  $|S^1 \cap S^2| = 40$ . The goal is to discover the set of consistent non-nulls, controlling the false discovery rate below 10%.

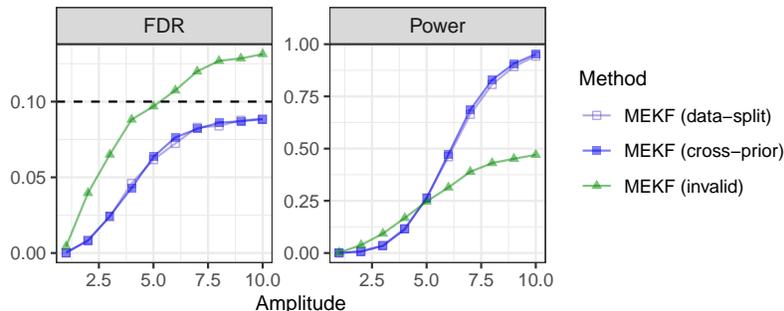


Figure A2: An example in which a naive implementation of the multi-environment knockoff filter with invalid statistics would not control the false discovery rate, while our method does. The false discovery rate and power evaluated over 500 experiments is shown as a function of the signal amplitude. The nominal false discovery rate is 10%.

Figure A2 compares the performances of three methods as a function of the signal amplitude  $a$ . The first two are our multi-environment knockoff filter with data-split and cross-prior statistics, respectively. The third one naively computes the magnitude of each  $W_j^e$  with the standard lasso coefficient difference statistics applied to the pooled data, and then it determines the sign of  $W_j^e$  by applying the same procedure on the data from environment  $e$ . The results show the naive approach does not control the false discovery rate, confirming the importance of our careful construction of multi-environment knockoff statistics (Definition 1).

## C Additional proofs

### C.1 Selective SeqStep+ test with dependent p-values

**Theorem 1** (Multi-environment knockoff filter). *The selective SeqStep+ procedure of [3] applied to p-values  $p_j^{\text{cst}}$  ordered by  $|W_j^{\text{cst}}|$  and satisfying the “almost-independence” property of Proposition (6), i.e.,  $\mathbb{P}[p_j^{\text{cst}} \leq \alpha \mid |W^{\text{cst}}|, p_{-j}^{\text{cst}}] \leq \alpha$  for any  $\alpha \in [0, 1]$ , controls the false discovery rate below the nominal level.*

*Proof of Theorem 1.* Our proof is based on a leave-one-out argument following closely that of Theorem 2 in [98]. The rejection threshold of selective SeqStep+ at level  $\alpha$  is:

$$\hat{\omega} = \min \left\{ \omega : \frac{1 + |\{j : |W_j^{\text{cst}}| \geq \omega, p_j^{\text{cst}} > c\}|}{|\{j : |W_j^{\text{cst}}| \geq \omega, p_j^{\text{cst}} \leq c\}| \vee 1} \leq \frac{1-c}{c} \alpha \right\}. \quad (23)$$

Define  $\mathcal{H}^{\text{cst}} \subseteq \{1, \dots, p\}$  as the subset of true null  $\mathcal{H}_j^{\text{cst}}$  (2), and  $\hat{\omega}_j$  as the rejection threshold resulting from  $p_j^{\text{cst}} \rightarrow 0$  in (23). Then, the false discovery rate is

$$\begin{aligned} \text{FDR} &= \mathbb{E} \left[ \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c ]}{|\{j : |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c\}| \vee 1} \right] \\ &= \mathbb{E} \left[ \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c ]}{1 + |\{j : |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} > c\}|} \cdot \frac{1 + |\{j : |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} > c\}|}{|\{j : |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c\}| \vee 1} \right] \\ &\leq \mathbb{E} \left[ \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c ]}{1 + |\{j : |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} > c\}|} \cdot \frac{1-c}{c} \alpha \right] \\ &\leq \mathbb{E} \left[ \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c ]}{1 + \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} > c ]} \right] \cdot \frac{1-c}{c} \alpha \\ &= \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{E} \left[ \frac{\mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}, p_j^{\text{cst}} \leq c ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}, p_l^{\text{cst}} > c ]} \right] \cdot \frac{1-c}{c} \alpha \\ &= \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{E} \left[ \frac{\mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j, p_j^{\text{cst}} \leq c ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_j, p_l^{\text{cst}} > c ]} \right] \cdot \frac{1-c}{c} \alpha, \end{aligned}$$

Above, the last equality follows with the same argument as in the proof of Theorem 2 of [98]: if  $p_j^{\text{cst}} \leq c$ , then  $\hat{\omega} = \hat{\omega}_j$ . Then, as  $\hat{\omega}_j$  is only a function of  $|W^{\text{cst}}|$  and  $p_{-j}^{\text{cst}}$ , we can write

$$\begin{aligned} \text{FDR} &\leq \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{E} \left[ \frac{\mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j ] \mathbb{P} [ p_j^{\text{cst}} \leq c \mid p_{-j}^{\text{cst}}, |W^{\text{cst}}| ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_j, p_l^{\text{cst}} > c ]} \right] \cdot \frac{1-c}{c} \alpha \\ &\leq \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{E} \left[ \frac{\mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j ] \mathbb{P} [ p_j^{\text{cst}} > c \mid p_{-j}^{\text{cst}}, |W^{\text{cst}}| ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_j, p_l^{\text{cst}} > c ]} \right] \cdot \alpha \\ &= \sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{E} \left[ \frac{\mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j, p_j^{\text{cst}} > c ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_j, p_l^{\text{cst}} > c ]} \right] \alpha, \end{aligned}$$

where the second inequality follows directly from the “almost-independence” property of Proposition (6) because

$$\frac{\mathbb{P} [ p_j^{\text{cst}} \leq c \mid p_{-j}^{\text{cst}}, |W^{\text{cst}}| ]}{\mathbb{P} [ p_j^{\text{cst}} > c \mid p_{-j}^{\text{cst}}, |W^{\text{cst}}| ]} \leq \frac{c}{1-c}.$$

Now, it follows from Lemma 6 of [98] that

$$\begin{aligned} \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j, p_j^{\text{cst}} > c ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_j, p_l^{\text{cst}} > c ]} &= \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j, p_j^{\text{cst}} > c ]}{1 + \sum_{l \in \mathcal{H}^{\text{cst}}, l \neq j} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_l, p_l^{\text{cst}} > c ]} \\ &= \frac{\sum_{j \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_j^{\text{cst}}| \geq \hat{\omega}_j, p_j^{\text{cst}} > c ]}{\sum_{l \in \mathcal{H}^{\text{cst}}} \mathbb{1} [ |W_l^{\text{cst}}| \geq \hat{\omega}_l, p_l^{\text{cst}} > c ]} = 1. \end{aligned}$$

Hence we proved  $\text{FDR} \leq \alpha$ .  $\square$

## C.2 Accumulation test with dependent p-values

Our proof of Theorem 2 follows the strategy of [69] (Theorem 2 therein), with some modifications to relax as needed their independence assumption. The new idea is to couple our p-values to *imaginary and mutually independent* p-values obtained by replacing  $n_j^-$  in (19) with the number of negative signs in a suitable subset of true null environments for column  $j$ . Our result leverages two lemmas: the first one is borrowed from [69], and the second one is a generalization of a similar result from [69] which we prove at the end of this section.

**Lemma A1** (Lemma B.3 from [69]). *Let  $B_1, \dots, B_m \in \{0, 1\}$  be independent, with  $B_j \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho)$  for all  $j \in \mathcal{H}_0$ , for some subset  $\mathcal{H}_0 \subseteq \{1, \dots, m\}$  and  $\rho \in (0, 1)$ . Let  $\{\mathcal{F}_k\}_{k=1, \dots, m}$  be any filtration in reverse time (i.e.,  $\mathcal{F}_{k+1} \subseteq \mathcal{F}_k$ ) such that:*

$$B_j \in \mathcal{F}_k \text{ for all } j \notin \mathcal{H}_0, \text{ and for all } j > k \text{ with } j \in \mathcal{H}_0, \quad (24)$$

$$\sum_{j \leq k, j \in \mathcal{H}_0} B_j \in \mathcal{F}_k, \text{ and} \quad (25)$$

$$\{B_j : j \leq k, j \in \mathcal{H}_0\} \text{ are exchangeable with respect to } \mathcal{F}_k, \quad (26)$$

for all  $k \in \{1, \dots, m\}$ . Then, the following is a supermartingale with respect to  $\{\mathcal{F}_k\}$  and  $\mathbb{E}[M_n] \leq 1/\rho$ :

$$M_k := \frac{1 + \#\{j \leq k : j \in \mathcal{H}_0\}}{1 + \sum_{j \leq k, j \in \mathcal{H}_0} B_j}. \quad (27)$$

**Lemma A2.** *Let  $p_1, \dots, p_m$  be the p-values defined in (19), sorted in decreasing order of  $|W_j^{\text{pcst}, r}|$ , for  $j \in \{1, \dots, m\}$ . Let  $\hat{k}$  denote the number of rejections obtained by applying the accumulation test of [69] at level  $\alpha$  to these p-values, with a monotone increasing accumulation function  $h : [0, 1] \mapsto [0, \infty)$  such that  $\int_0^1 h(t) dt = 1$ . That is,*

$$\hat{k} = \max \left\{ k \in \{1, \dots, m\} : \frac{1}{k} \sum_{j=1}^k h(p_j) \leq \alpha \right\}. \quad (28)$$

Then, for some fixed  $C > 0$ ,

$$\mathbb{E} \left[ \frac{\#\{j \leq \hat{k} : j \in \mathcal{H}^{\text{pcst}, r}\}}{C + \sum_{j=1}^{\hat{k}} h(p_j)} \right] \leq \frac{1}{\int_0^1 [h(t) \wedge C] dt}. \quad (29)$$

**Theorem 2** (Multi-environment knockoff filter with accumulation test). *The accumulation test of [69] with an increasing accumulation function (e.g., HingeExp with parameter  $C = 2$ ) applied to the p-values defined in (19) controls the modified false discovery rate (18) (e.g., with  $q = C/\alpha$ ), as in [69]. That is, Theorem 1 of [69] still holds for the p-values  $p_j^{\text{pcst}, r}$  in (19) even though they are not independent.*

*Proof of Theorem 2.* Assume the p-values are sorted in decreasing order of  $|W_j^{\text{pcst}, r}|$  and let  $\hat{k}$  be the number of rejections made by the accumulation test at a fixed nominal level  $\alpha$ , as defined in (28). To simplify the notation, we will write  $p_j$  instead of  $p_j^{\text{pcst}, r}$ . Proceeding as in [69], we see that

$$\begin{aligned} \mathbb{E} \left[ \text{mFDP}_{C/\alpha}(\hat{k}) \right] &= \mathbb{E} \left[ \frac{\#\{j \leq \hat{k} : j \in \mathcal{H}^{\text{pcst}, r}\}}{C/\alpha + \hat{k}} \right] \\ &= \mathbb{E} \left[ \frac{\#\{j \leq \hat{k} : j \in \mathcal{H}^{\text{pcst}, r}\}}{C + \sum_{j=1}^{\hat{k}} h(p_j)} \cdot \frac{C + \sum_{j=1}^{\hat{k}} h(p_j)}{C/\alpha + \hat{k}} \right] \\ &\leq \alpha \cdot \mathbb{E} \left[ \frac{\#\{j \leq \hat{k} : j \in \mathcal{H}^{\text{pcst}, r}\}}{C + \sum_{j=1}^{\hat{k}} h(p_j)} \right], \end{aligned}$$

where the inequality follows from the definition of  $\hat{k}$  (28). Lemma A2 completes the proof. The same argument can be repurposed to prove the stricter threshold  $\hat{k}^+$  defined in [69] controls the (unmodified) false discovery rate.  $\square$

*Proof of Lemma A2.* This proof follows a similar strategy as that of Lemma B.2 in [69], although we additionally need to leverage the special structure of our partial consistency p-values (19) to get around their lack of independence.

Take any  $j$  such that  $\mathcal{H}_j^{\text{pcst},r}$  is true. Because there must be at least  $K - r + 1$  null environments for this variable, we can define  $\tilde{n}_j^-$  as the number of negative signs among the first (in any arbitrary order)  $K - r + 1$  null entries of the  $j$ -th column of  $\mathbf{W}$ . We assume hereafter that any zero entries in  $\mathbf{W}$  have been randomly assigned a positive or negative sign by flipping independent fair coins. Concretely, let us define the list of indices for these environments as  $\mathcal{E}_j^0$ , as we will need to refer to them later. Define also  $\tilde{n}_j^- = n_j^- - \tilde{n}_j^- \geq 0$ , the number of negative signs from environments other than those in  $\mathcal{E}_j^0$ . Define then  $\tilde{p}_j$  as the p-value obtained by replacing  $n_j^-$  with  $\tilde{n}_j^-$  in (19), that is

$$\tilde{p}_j := \Psi \left( E - r + 1, \frac{1}{2}, \tilde{n}_j^- - 1 \right) + U_j \cdot \psi \left( E - r + 1, \frac{1}{2}, \tilde{n}_j^- \right).$$

The imaginary p-values  $\tilde{p}_j$  for  $j \in \mathcal{H}^{\text{pcst},r}$  are independent of each other because they are only affected by the signs of the true null entries in  $\mathbf{W}$ , which satisfies Definition 1. They are also exactly uniformly distributed,

$$\tilde{p}_j \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}[0, 1], \quad (30)$$

for  $j \in \mathcal{H}^{\text{pcst},r}$ , because they are the randomized binomial p-values corresponding to

$$\tilde{n}_j^- \stackrel{\text{i.i.d.}}{\sim} \text{Binomial}(E - r + 1, 1/2). \quad (31)$$

Further, note that  $\tilde{p}_j \leq p_j$  almost-surely because  $\tilde{n}_j^- \leq n_j^-$ ; the independent uniform random variables  $U_j$  are taken to be the same for both real and imaginary p-values. Therefore, as we assumed the accumulation function to be monotone increasing, we also have that  $h(\tilde{p}_j) \leq h(p_j)$  and, for all  $k \in \{1, \dots, p\}$ ,

$$M_k := \frac{\#\{j \leq k : j \in \mathcal{H}^{\text{pcst},r}\}}{C + \sum_{j=1}^k h(p_j)} \leq \frac{\#\{j \leq k : j \in \mathcal{H}^{\text{pcst},r}\}}{C + \sum_{j=1}^k h(\tilde{p}_j)} =: \tilde{M}_k.$$

Now we can deal with  $\tilde{M}_k$  with the same approach of [69]. Define an i.i.d. sequence  $V_j \sim \text{Uniform}[0, 1]$ , for  $j \in \{1, \dots, m\}$ , independent of everything else, and random variables  $B_j = \mathbb{1}[V_i \leq h(\tilde{p}_i)/C]$ . Conditional on  $\tilde{p}_1, \dots, \tilde{p}_m$ , the variables  $B_j$  are independent and  $B_j \sim \text{Bernoulli}([h(\tilde{p}_j)/C] \wedge 1)$ . Further, marginally,

$$\mathbb{E}[B_j] = \mathbb{E} \left[ \frac{h(\tilde{p}_j) \wedge C}{C} \right] = \frac{1}{C} \int_0^1 [h(t) \wedge C] dt =: \rho.$$

As in [69], we would like to apply Lemma A1 to bound

$$\mathbb{E} \left[ \frac{1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\}}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j} \right],$$

which requires a reverse-time filtration satisfying (24)–(26), and then show  $\hat{k}$  is a stopping time with respect to it.

First, let  $\mathcal{G}$  be the  $\sigma$ -algebra generated by the following variables:

- $|\mathbf{W}|$  (the absolute values of all entries in  $\mathbf{W}$ );
- $\text{sign}(W_j^e)$  for  $j \in \{1, \dots, m\}$  and  $e \notin \mathcal{H}_j$  (the signs of  $W_j^e$  for non-null environments);
- $\tilde{p}_j$  for  $j \notin \mathcal{H}^{\text{pcst},r}$  (the imaginary p-values corresponding to non-null consistent hypotheses).

Second, for any  $k \in \{1, \dots, m\}$ , let  $\mathcal{F}_k$  be the union of  $\mathcal{G}$  with the  $\sigma$ -algebra generated by:

- $(\tilde{p}_j, \tilde{n}_j^-, U_j, \tilde{n}_j^-)$  for all  $j > k$ ;
- $\{(\tilde{p}_j, \tilde{n}_j^-, U_j, \tilde{n}_j^-)\}_{j=1}^k$  (as an unordered set).

This is a reverse-time filtration and it satisfies (24)–(26). The key observations here are that the property of  $\mathbf{W}$  in Definition 1 implies (30)–(31) hold also conditional on  $\mathcal{G}$ , and that  $\tilde{n}_j^-$  and  $\tilde{n}_j^-$  are independent. Regarding  $\hat{k}$  (28), note that it is defined in terms of  $p_j$ , not  $\tilde{p}_j$  (as in practice one needs to evaluate  $\hat{k}$  without knowing a priori which environments correspond to true nulls). However, the real p-values  $p_j$  can be reconstructed exactly from knowledge of  $\tilde{n}_j^-, U_j, \tilde{n}_j^-$  and the information in  $\mathcal{G}$ . Therefore,  $\hat{k}$  is a stopping time with respect to  $\{\mathcal{F}_k\}$ . Thus, we can apply Lemma A1 in combination with the optional stopping theorem to conclude that

$$\mathbb{E} \left[ \frac{1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\}}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j} \right] \leq \frac{1}{\rho} = \frac{1}{\frac{1}{C} \int_0^1 [h(t) \wedge C] dt}. \quad (32)$$

The rest of the proof follows closely that of Lemma B.2 in [69] but it is nonetheless recalled here for completeness.

$$\begin{aligned}
\mathbb{E} \left[ \frac{1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\}}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j} \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \frac{1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\}}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j} \mid \tilde{p}_1, \dots, \tilde{p}_m \right] \right] \\
&= \mathbb{E} \left[ \left( 1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\} \right) \mathbb{E} \left[ \frac{1}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j} \mid \tilde{p}_1, \dots, \tilde{p}_m \right] \right] \\
&\geq \mathbb{E} \left[ \left( 1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\} \right) \frac{1}{\mathbb{E} \left[ 1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} B_j \mid \tilde{p}_1, \dots, \tilde{p}_m \right]} \right] \\
&= \mathbb{E} \left[ \left( 1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\} \right) \frac{1}{1 + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} \frac{h(\tilde{p}_j)^\wedge C}{C}} \right] \\
&\geq C \cdot \mathbb{E} \left[ \frac{1 + \#\{j \leq \hat{k} : j \in \mathcal{H}_0\}}{C + \sum_{j \leq \hat{k}, j \in \mathcal{H}_0} h(\tilde{p}_j)} \right].
\end{aligned}$$

Above, the first inequality is Jensen’s inequality. The proof of this lemma is then completed by (32).  $\square$

## D Additional details about numerical experiments

### D.1 Synthetic data

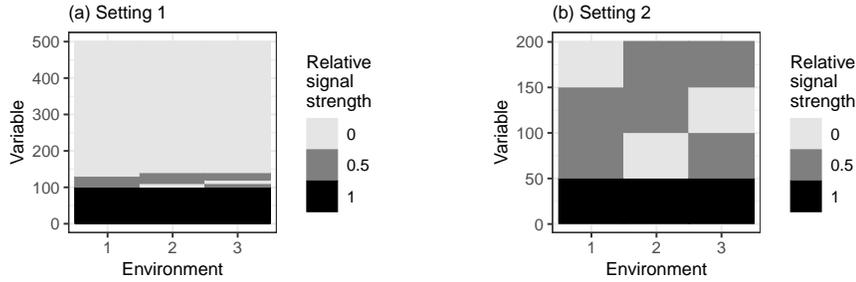


Figure A3: True hypothesis structure for the numerical experiments of Figure 1. The shaded rectangles indicate which variables are non-null in each environment. Darker shades indicate stronger associations. The variables are sorted here for ease of visualization; in the experiments, their order is randomized.

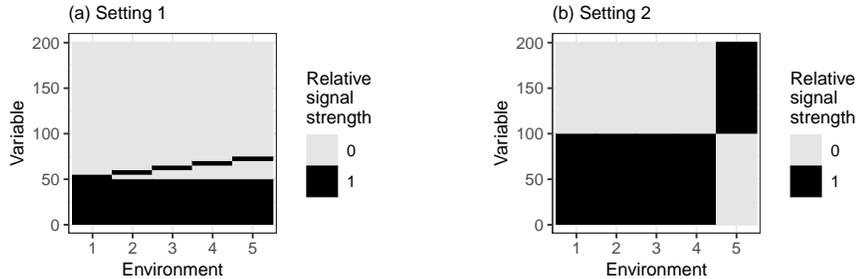


Figure A4: True hypothesis structure for the numerical experiments of Figure 6. Other details are as in Figure A3.

### D.2 Simulated genetic study

Synthetic genotypes from different sub-populations are generated based on the phased haplotypes in the 1000 Genomes Project as follows. First, 20 haplotype sequences are randomly picked from each of the five populations represented

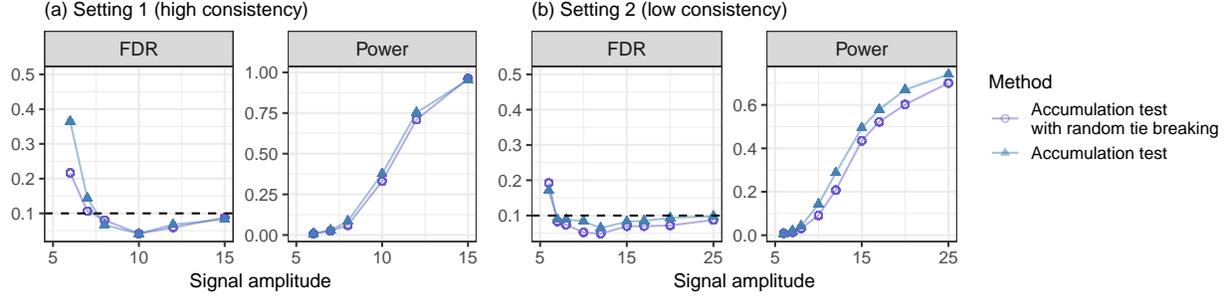


Figure A5: Performance of the accumulation test applied with two alternative implementations of our knockoff-based p-values: those in (19) (random tie breaking for zero statistics) and (16) (no random tie breaking).

therein (AFR, AMR, EAS, EUR, SAS); these haplotypes will serve as ancestral motifs in a hidden Markov model similar to that of SHAPEIT [99]. Our hidden Markov model is constant across populations, and in each of them it describes the distribution of new haplotypes sequences as an imperfect mosaic of the corresponding 1000 Genomes motifs. More precisely, letting  $H = (H_1, \dots, H_p) \in \{0, 1\}^p$  denote a sequence of haplotypes at  $p$  sites,  $H$  satisfies

$$\begin{cases} Z \sim \text{MC}(Q), & \text{(latent Markov chain),} \\ H_j | Z \stackrel{\text{ind.}}{\sim} f_j(H_j | Z_j), & \text{(emission distribution),} \end{cases} \quad (33)$$

where  $Z = (Z_1, \dots, Z_p)$  are latent random variables, each taking values in  $\{1, \dots, L\}$ , for  $L = 20$ . Above,  $\text{MC}(Q)$  is a Markov chain with initial probabilities  $Q_1$  and transition matrices  $(Q_2, \dots, Q_p)$ :

$$Q_1(l) = \frac{1}{L}, \quad Q_j(l' | l) = \begin{cases} (1 - e^{-\rho d_j}) \frac{1}{L} + e^{-\rho d_j}, & \text{if } l' = l, \\ (1 - e^{-\rho d_j}) \frac{1}{L}, & \text{if } l' \neq l, \end{cases} \quad (34)$$

for all  $j \in \{1, \dots, p\}$  and  $l \in \{1, \dots, L\}$ . Above,  $d_j$  indicates the genetic distance between loci  $j$  and  $j - 1$ , measured in cM and provided by the 1000 Genomes Project. In our simulations, we simply set  $\rho = 1$  and let the emission distribution of  $H_j | Z_j$  be such that  $H_j$  is equal to the reference (motif) haplotype indexed by  $Z_j$  with probability 0.999 (0.1% per-site mutation rate). Then, to obtain unphased genotypes for  $n$  individuals, we sample  $2n$  independent haplotype sequences from the above model and combine them pairwise by taking element-by-element sums. Finally, we can leverage our exact knowledge of this hidden Markov model to generate knockoff copies of the typed variants with the same approach as in [5]; in fact, it is immediate to show any subset of haplotypes from the above model still jointly follows a hidden Markov model from the same SHAPEIT family, with suitably modified genetic distances.

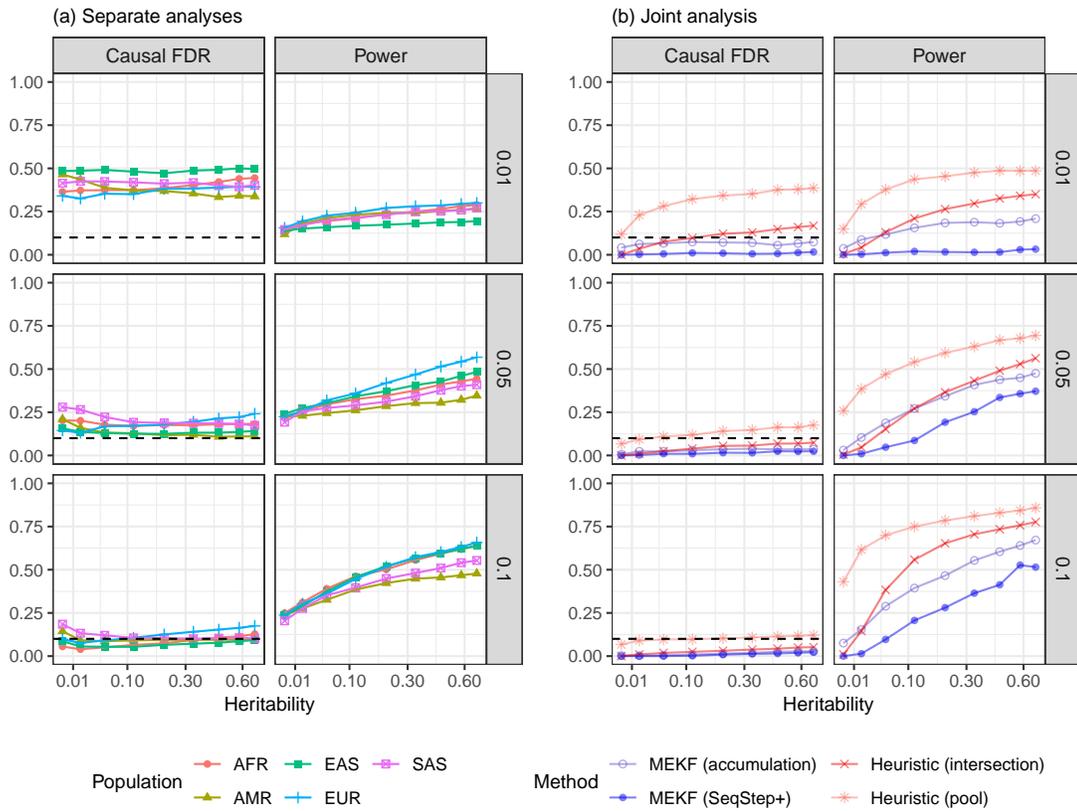


Figure A6: Analysis of a simulated multi-population genome-wide association study in which the true causal variants are missing, for different genotyping densities. Resolution: 15 kb. Other details are as in Figure 9.

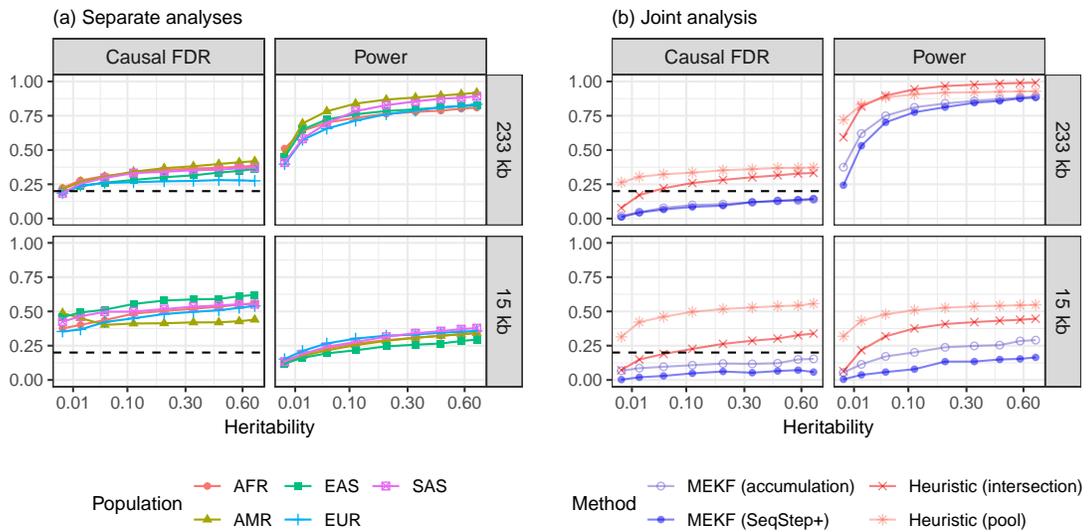


Figure A7: Analysis of a simulated multi-population genome-wide association study in which the true causal variants are missing. The nominal false discovery rate is 20%. Other details are as in Figure 9.

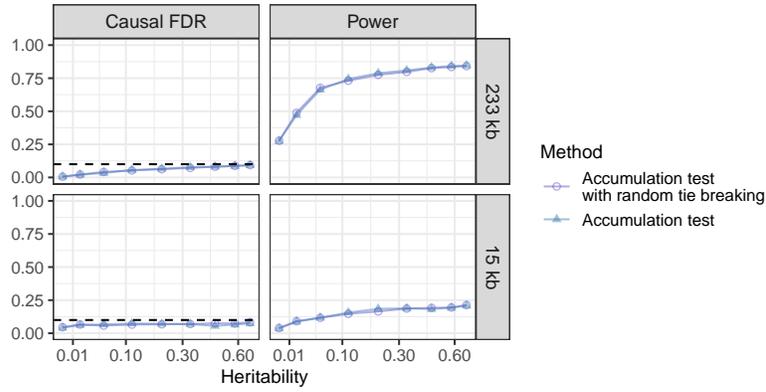


Figure A8: Performance in a simulated genome-wide association study of the accumulation test applied with two alternative implementations of our p-values: those in (19) (random tie breaking for zero statistics) and (16) (no random tie breaking). The two methods are essentially equivalent here. Other details are as in Figure 9 (a).

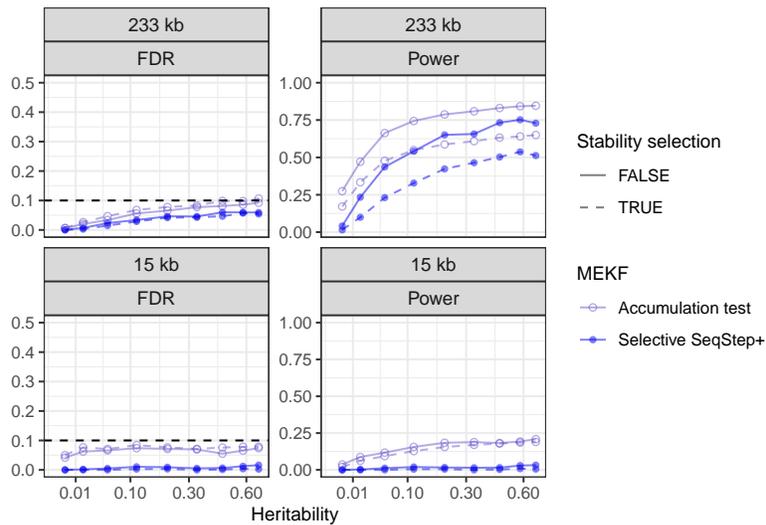


Figure A9: Performance in a simulated genome-wide association study of our methods with and without stability selection, as implemented in our analysis of the UK Biobank data of Section 7. Other details are as in Figure 9 (a).

## E Additional details about the analysis of the UK Biobank data

Table A1: Definitions of the UK Biobank phenotypes used in our analysis, which match those of [6]. For binary disease-status phenotypes, the number of cases refers to the subset of individuals that passed our quality control.

Name	Description	Number of cases	UK Biobank Fields	UK Biobank Codes
bmi	body mass index	continuous	21001-0.0	
cvd	cardiovascular disease	148715	20002-0.0–20002-0.32	1065, 1066, 1067, 1068, 1081, 1082, 1083, 1425, 1473, 1493
diabetes	diabetes	19897	20002-0.0–20002-0.32	1220
height	standing height	continuous	50-0.0	
hypothyroidism	hypothyroidism	22493	20002-0.0–20002-0.32	1226
platelet	platelet count	continuous	30080-0.0	
respiratory	respiratory disease	64945	20002-0.0–20002-0.32	1111, 1112, 1113, 1114, 1115, 1117, 1413, 1414, 1415, 1594
sbp	systolic blood pressure	continuous	4080-0.0, 4080-0.1	

Table A2: Definitions of environments for UK Biobank individuals in terms of self-reported ancestries.

Environment	Sample size	Self-reported ancestries
African	7,623	“African”, “Caribbean”, “Any other black background”, “Black or Black British”
Asian	3,284	“Asian or Asian British”, “Chinese”, “Any other Asian background”
British	429,934	“British”
European	28,994	“Any other white background”, “Irish”, “White”
Indian	7,628	“Indian”, “Pakistani”, “Bangladeshi”

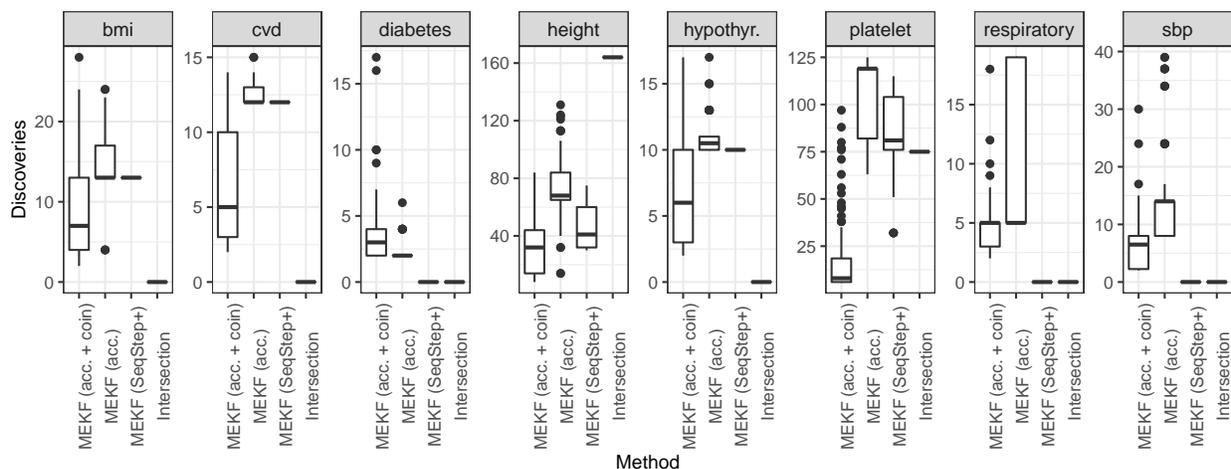


Figure A10: Numbers of low-resolution discoveries consistent across at least two populations, as obtained with different methods. For methods involving randomized p-values (multi-environment knockoff filter with the accumulation test or the selective SeqStep+), this figure shows the empirical distribution of the numbers of discoveries over 100 independent realizations of the p-values conditional on the knockoff test statistics. Other details are as in Table 1.

Table A3: Numbers of discoveries made with the multi-environment knockoff filter at different resolutions for several UK Biobank phenotypes. Other details are as in Table 1.

Phenotype	Resolution (kb)	Number of environments				
		1	2	3	4	5
bmi	single-SNP	0	0	0	0	0
	3	10	0	0	0	0
	20	343	8	3	2	0
	41	918	6	3	3	0
	81	1480	3	4	3	0
	208	2395	5	7	0	0
	425	2460	13	2	0	0
cvd	single-SNP	0	0	0	0	0
	3	22	0	0	0	0
	20	239	8	0	0	0
	41	339	0	0	0	0
	81	566	0	0	0	0
	208	940	2	0	0	0
	425	1089	12	0	0	0
diabetes	single-SNP	0	2	2	0	0
	3	21	5	0	0	0
	20	61	6	0	2	0
	41	109	4	3	0	0
	81	109	2	3	0	0
	208	113	5	0	0	0
	425	194	2	0	0	0
hypothyroidism	single-SNP	19	0	0	0	0
	3	40	2	0	0	0
	20	105	5	0	0	0
	41	222	5	0	0	0
	81	277	7	0	0	0
	208	295	11	0	0	0
	425	335	10	0	0	0
respiratory	single-SNP	0	0	0	0	0
	3	0	0	0	0	0
	20	83	4	0	0	0
	41	123	2	0	0	0
	81	193	15	0	0	0
	208	262	0	0	0	0
	425	383	5	0	0	0
sbp	single-SNP	0	0	0	0	0
	3	83	0	0	0	0
	20	191	2	0	0	0
	41	511	3	0	0	0
	81	830	6	0	0	0
	208	1183	4	0	0	0
	425	1543	14	0	0	0

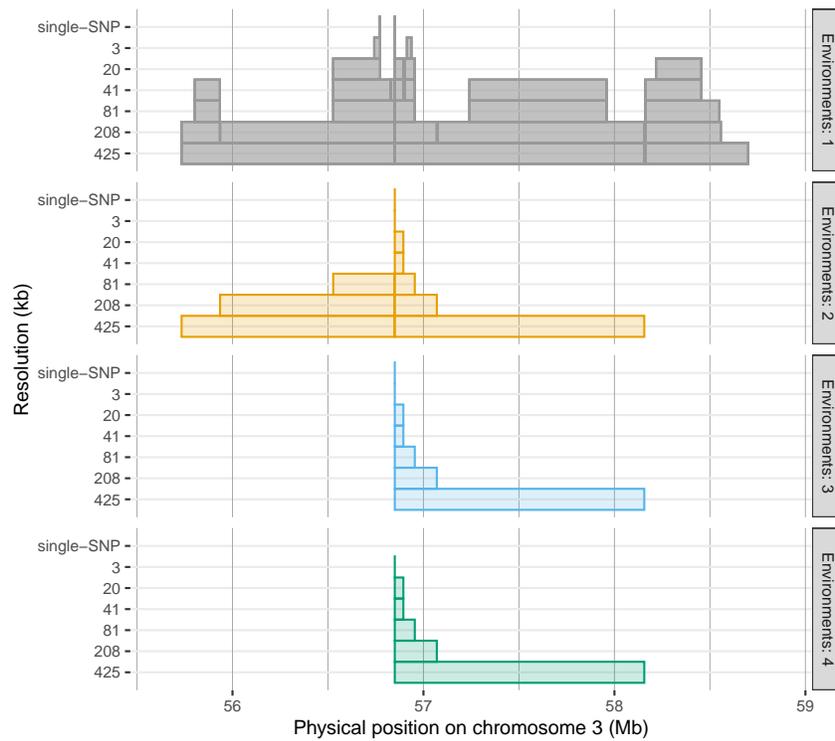


Figure A11: Chicago plot of some discoveries on chromosome three for platelet count based on UK Biobank data from individuals from five environments, as in Figure 10. Here, the findings are shown separately based on the numbers of environments across which they are consistent.

Table A4: Numbers of consistent discoveries for several UK Biobank phenotypes obtained using three alternative methods: the multi-environment knockoff filter implemented with the accumulation test (Acc.) or the selective SeqStep+ (SSStep), and the intersection heuristic (Int.). Other details are as in Table 1.

Phenotype	Resolution (kb)	Number of environments / Consistent testing method											
		2			3			4			5		
		Acc.	SSStep	Int.	Acc.	SSStep	Int.	Acc.	SSStep	Int.	Acc.	SSStep	Int.
bmi	single-SNP	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	20	8	0	0	3	0	0	2	0	0	0	0	0
	41	6	0	0	3	0	0	3	0	0	0	0	0
	81	3	0	0	4	0	0	3	0	0	0	0	0
	208	5	0	0	7	0	0	0	0	0	0	0	0
	425	13	0	0	2	0	0	0	0	0	0	0	0
cvd	single-SNP	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	20	8	0	0	0	0	0	0	0	0	0	0	0
	41	0	0	0	0	0	0	0	0	0	0	0	0
	81	0	0	0	0	0	0	0	0	0	0	0	0
	208	2	0	0	0	0	0	0	0	0	0	0	0
	425	12	12	0	0	0	0	0	0	0	0	0	0
diabetes	single-SNP	2	0	0	2	0	0	0	0	0	0	0	0
	3	5	0	0	0	0	0	0	0	0	0	0	0
	20	6	0	0	0	0	0	2	0	0	0	0	0
	41	4	0	0	3	0	0	0	0	0	0	0	0
	81	2	0	0	3	0	0	0	0	0	0	0	0
	208	5	0	0	0	0	0	0	0	0	0	0	0
	425	2	0	0	0	0	0	0	0	0	0	0	0
height	single-SNP	13	0	0	2	0	0	0	0	0	0	0	0
	3	9	0	0	6	0	0	0	0	0	0	0	0
	20	33	29	25	0	0	0	0	0	0	0	0	0
	41	42	20	68	7	0	0	7	0	0	2	0	0
	81	48	49	84	24	23	0	0	0	0	0	0	0
	208	103	33	107	23	16	0	7	0	0	3	0	0
	425	68	31	164	26	15	0	3	0	0	0	0	0
hypothyroidism	single-SNP	0	0	0	0	0	0	0	0	0	0	0	0
	3	2	0	0	0	0	0	0	0	0	0	0	0
	20	5	0	0	0	0	0	0	0	0	0	0	0
	41	5	0	0	0	0	0	0	0	0	0	0	0
	81	7	0	0	0	0	0	0	0	0	0	0	0
	208	11	11	0	0	0	0	0	0	0	0	0	0
	425	10	10	0	0	0	0	0	0	0	0	0	0
platelet	single-SNP	9	0	0	3	0	0	0	0	0	0	0	0
	3	10	10	0	4	0	0	4	0	0	0	0	0
	20	27	20	26	16	0	0	2	0	0	0	0	0
	41	52	0	50	12	12	0	9	0	0	0	0	0
	81	104	69	69	15	15	3	8	0	0	0	0	0
	208	98	58	58	16	14	8	14	0	0	2	0	0
	425	119	70	75	9	0	6	11	0	0	0	0	0
respiratory	single-SNP	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	20	4	0	0	0	0	0	0	0	0	0	0	0
	41	2	0	0	0	0	0	0	0	0	0	0	0
	81	15	0	0	0	0	0	0	0	0	0	0	0
	208	0	0	0	0	0	0	0	0	0	0	0	0
	425	5	0	0	0	0	0	0	0	0	0	0	0
sbp	single-SNP	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	20	2	0	0	0	0	0	0	0	0	0	0	0
	41	3	0	0	0	0	0	0	0	0	0	0	0
	81	6	0	0	0	0	0	0	0	0	0	0	0
	208	4	0	0	0	0	0	0	0	0	0	0	0
	425	14	0	0	0	0	0	0	0	0	0	0	0

Table A5: Proportions of discoveries confirmed by previously known associations in the GWAS Catalog, for several UK Biobank phenotypes. Other details are as in Table 2.

Phenotype	Resolution (kb)	MEKF	Pooling	Binomial p-value	Intersection
bmi	single-SNP	0 / 0	0 / 0		0 / 0
	3	0 / 0	10 / 10 (100%)		0 / 0
	20	8 / 8 (100%)	308 / 343 (90%)	$7.22 \cdot 10^{-01}$	0 / 0
	41	6 / 6 (100%)	658 / 918 (72%)	$2.79 \cdot 10^{-01}$	0 / 0
	81	3 / 3 (100%)	875 / 1480 (59%)	$3.95 \cdot 10^{-01}$	0 / 0
	208	5 / 5 (100%)	1113 / 2395 (46%)	$5.14 \cdot 10^{-02}$	0 / 0
	425	13 / 13 (100%)	1146 / 2460 (47%)	$3.56 \cdot 10^{-04}$	0 / 0
sbp	single-SNP	0 / 0	0 / 0		0 / 0
	3	0 / 0	69 / 83 (83%)		0 / 0
	20	2 / 2 (100%)	166 / 191 (87%)	1	0 / 0
	41	3 / 3 (100%)	358 / 511 (70%)	$6.19 \cdot 10^{-01}$	0 / 0
	81	6 / 6 (100%)	518 / 830 (62%)	$1.41 \cdot 10^{-01}$	0 / 0
	208	4 / 4 (100%)	645 / 1183 (55%)	$1.87 \cdot 10^{-01}$	0 / 0
	425	14 / 14 (100%)	714 / 1543 (46%)	$1.83 \cdot 10^{-04}$	0 / 0
cvd	single-SNP	0 / 0	0 / 0		0 / 0
	3	0 / 0	21 / 22 (95%)		0 / 0
	20	8 / 8 (100%)	178 / 239 (74%)	$2.19 \cdot 10^{-01}$	0 / 0
	41	0 / 0	248 / 339 (73%)		0 / 0
	81	0 / 0	365 / 566 (64%)		0 / 0
	208	2 / 2 (100%)	559 / 940 (59%)	$6.56 \cdot 10^{-01}$	0 / 0
	425	11 / 12 (92%)	717 / 1089 (66%)	$1.16 \cdot 10^{-01}$	0 / 0
respiratory	single-SNP	0 / 0	0 / 0		0 / 0
	3	0 / 0	0 / 0		0 / 0
	20	4 / 4 (100%)	74 / 83 (89%)	1	0 / 0
	41	2 / 2 (100%)	110 / 123 (89%)	1	0 / 0
	81	14 / 15 (93%)	156 / 193 (81%)	$3.90 \cdot 10^{-01}$	0 / 0
	208	0 / 0	196 / 262 (75%)		0 / 0
	425	5 / 5 (100%)	267 / 383 (70%)	$3.28 \cdot 10^{-01}$	0 / 0
hypothyroidism	single-SNP	0 / 0	7 / 19 (37%)		0 / 0
	3	2 / 2 (100%)	23 / 40 (57%)	$6.48 \cdot 10^{-01}$	0 / 0
	20	5 / 5 (100%)	71 / 105 (68%)	$3.00 \cdot 10^{-01}$	0 / 0
	41	5 / 5 (100%)	101 / 222 (45%)	$4.97 \cdot 10^{-02}$	0 / 0
	81	7 / 7 (100%)	126 / 277 (45%)	$1.35 \cdot 10^{-02}$	0 / 0
	208	8 / 11 (73%)	141 / 295 (48%)	$1.88 \cdot 10^{-01}$	0 / 0
	425	8 / 10 (80%)	140 / 335 (42%)	$3.74 \cdot 10^{-02}$	0 / 0
diabetes	single-SNP	2 / 2 (100%)	0 / 0		0 / 0
	3	3 / 5 (60%)	20 / 21 (95%)	$1.51 \cdot 10^{-01}$	0 / 0
	20	5 / 6 (83%)	54 / 61 (89%)	1	0 / 0
	41	4 / 4 (100%)	94 / 109 (86%)	$9.63 \cdot 10^{-01}$	0 / 0
	81	2 / 2 (100%)	95 / 109 (87%)	1	0 / 0
	208	5 / 5 (100%)	101 / 113 (89%)	$9.90 \cdot 10^{-01}$	0 / 0
	425	2 / 2 (100%)	157 / 194 (81%)	1	0 / 0

Table A6: Consistent discoveries at the single-nucleotide resolution for UK Biobank phenotypes, compared to associations previously reported in the GWAS Catalog. Asterisks indicate associations previously reported in the GWAS Catalog for the gene corresponding to our variant, rather than for the variant itself.

Phenotype	Chr.	Position (Mb)	SNP	Association	Env.	Consequence	Gene	
diabetes	10	114.758	rs7903146	diabetes	3	intron	TCF7L2	
	11	92.709	rs10830963	diabetes	2	intron	MTNR1B	
height	2	56.097	rs3791679	height	2	intron	EFEMP1	
	3	141.126	rs1344672	height*	2	intron	ZBTB38	
	4	18.025	rs2011603	height*	3	2KB upstream	LCORL	
	6	7.720	rs12198986	height	2	regulatory region		
	6	19.839	rs41271299	height	2	intron	ID4	
	8	130.726	rs6470764	height	2	intron	GSDMC	
	11	75.276	rs606452	height	2	intron	SERPINH1	
	12	66.360	rs8756	height	2	3 prime UTR	HMGA2	
	12	93.979	rs11107116	height	2	intron	SOCS2	
	15	99.195	rs2871865	height	3	intron	IGF1R	
	15	100.693	rs72755233	height	2	missense	ADAMTS17	
	19	55.880	rs4252548	height	2	missense	IL11	
	20	34.026	rs143384	height	2	3 prime UTR	GDF5	
	platelet	3	56.850	rs1354034	platelet	3	intron	ARHGEF3
		5	75.997	Affx-26978473		2		
6		135.419	rs7775698	platelet	2	intron	HBS1L	
9		4.763	rs385893	platelet	3	regulatory region	AL353151.2, ECM1P1	
10		65.028	rs10761731	platelet	2	intron	JMJD1C	
12		111.885	rs3184504	platelet	2	missense	SH2B3	
12		111.885	rs72650673	hematocrit	2	missense	SH2B3	
18		20.721	rs11082304	platelet	2	intron	CABLES1	
19		16.186	rs8109288	platelet	3	n.c. transcript exon	AC008894.3, TPM4	