False discovery rate control in genome-wide association studies with population structure

Matteo Sesia[a,1], Stephen Bates[b,c], Emmanuel Candès[d,e,1], Jonathan Marchini[f], and Chiara Sabatti[d,g]

[a]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA 90089; [b]Department of Statistics, University of California, Berkeley, CA 94720; [c]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720; [d]Department of Statistics, Stanford University, Stanford, CA 94305; [e]Department of Mathematics, Stanford University, Stanford, CA 94305; [f]Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY 10591; and [g]Department of Biomedical Data Sciences, Stanford University, Stanford, CA 94305

Contributed by Emmanuel Candès, July 20, 2021 (sent for review March 27, 2021; reviewed by Dan Nicolae and Saharon Rosset)

We present a comprehensive statistical framework to analyze data from genome-wide association studies of polygenic traits, producing interpretable findings while controlling the false discovery rate. In contrast with standard approaches, our method can leverage sophisticated multivariate algorithms but makes no parametric assumptions about the unknown relation between genotypes and phenotype. Instead, we recognize that genotypes can be considered as a random sample from an appropriate model, encapsulating our knowledge of genetic inheritance and human populations. This allows the generation of imperfect copies (knockoffs) of these variables that serve as ideal negative controls, correcting for linkage disequilibrium and accounting for unknown population structure, which may be due to diverse ancestries or familial relatedness. The validity and effectiveness of our method are demonstrated by extensive simulations and by applications to the UK Biobank data. These analyses confirm our method is powerful relative to state-of-the-art alternatives, while comparisons with other studies validate most of our discoveries. Finally, fast software is made available for researchers to analyze Biobank-scale datasets.

genome-wide association studies | false discovery rate | knockoffs | population structure | hidden Markov models

Genome-wide association studies shaped the research into the genetic basis of human traits for the past 15 y. While family studies had previously been the cornerstone of genetics, Risch and Merikangas (1) in 1996 described the power that large population samples held for the study of polygenic phenotypes, those influenced by many loci, each with relatively small effect. Ten years of biotech development resulted in the capacity to genotype hundreds of thousands of single-nucleotide polymorphisms (SNPs) in thousands of individuals, and in 2007 the first large-scale association studies were published (2). As of 2021, the National Human Genome Research Institute - European Bioinformatics Institute genome-wide association studies (GWAS) catalog (3) contains over 4,800 publications and 240,000 associations, implicating almost 150,000 SNPs for a diverse set of traits. The predictions in ref. 1 have been confirmed: There are thousands of associations with traits such as high cholesterol or autism, whose inheritance appeared hard to explain, that are now candidates for follow-up studies. Additional challenges have thus emerged: How does one sort through all these genetic variants? How does one identify those that are more likely to be causal? How does one select variants that can be used to construct robust prediction scores, maintaining validity across human populations? GWAS were designed based on statistical considerations and a closer look at the inferential methods used today to analyze these data shows progress, many success stories, and some open challenges.

Statistical Analysis of GWAS Data

The Requisites. An effective GWAS analysis should account for the role played by all relevant genetic variants and adjust for multiplicity. These studies were intended to uncover the genetic basis of polygenic traits; therefore, the analysis should use multivariate models. Multivariate models have two additional benefits. First, they can explain a higher fraction of phenotypic variance, which benefits prediction (witness the rise of machine-learning algorithms), and can facilitate the discovery of new loci. Second, they bring us closer to the identification of variants with causal effects (4, 5).

Indeed, the main purpose of a GWAS is not to construct a black-box model that predicts phenotypic outcomes given genotype information, but to identify precisely which genetic variants have an impact on the phenotype, uncovering the underlying biological pathway. A meaningful error measure should be based on the number of falsely discovered genetic loci and, as we are exploring the potential effects of hundreds of thousands of variables, a multiplicity adjustment is needed to guarantee the reproducibility of any findings. The false discovery rate (FDR) is a particularly appropriate control target: As we expect to make hundreds or thousands of true discoveries (corresponding to the polygenic nature of the trait), we can certainly tolerate a few false ones (6, 7).

The Challenges. While striving to achieve the above desiderata, the analysis of GWAS data encountered several obstacles. A first challenge arises from the problem dimensions. The

Significance

Genome-wide association studies compare a phenotype to thousands of genetic variants, searching for associations of potential biological interest. Standard analyses rely on linear models of the phenotype given one variable at a time. However, their assumptions are difficult to verify and their univariate approaches make it hard to recognize interesting associations from spurious ones. Our work takes a different path: We analyze all variants simultaneously, modelling the randomness in the genotypes, which is better understood, instead of the phenotype. Our solution accounts for linkage disequilibrium and population structure, controls the false discovery rate, and leverages powerful machine-learning tools. Applications to the UK Biobank data indicate increased power compared to state-of-the-art alternatives and high replicability.

Author contributions: M.S., S.B., E.C., J.M., and C.S. designed research; M.S., S.B., E.C., J.M., and C.S. performed research; M.S. and S.B. contributed new reagents/analytic tools; M.S. analyzed data; and M.S., S.B., E.C., J.M., and C.S. wrote the paper.

Reviewers: D.N., University of Chicago; and S.R., Tel Aviv University.

The authors declare no competing interest.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: candes@stanford.edu or sesia@marshall.usc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2105841118/-/DCSupplemental.

Published September 27, 2021.

simplest polygenic model (8) describes a phenotype as the result of additive genetic effects, suggesting linear regression of the trait on the genotyped SNPs as an inferential method. However, given that the number of explanatory variables (hundreds of thousands) is larger than the sample size (historically in the thousands, nowadays routinely in the tens of thousands), classical linear regression is not a viable approach. Penalized regression and Bayesian models have been proposed as alternatives (9–11), but they also have shortcomings. The Lasso (12) is geared toward optimizing prediction and does not allow researchers to make statements on error control for the selected variables. While the recent literature on inference after selection (13) attempts to remedy this limitation, the most practical solutions in the context of GWAS are based on sample splitting (14, 15), with consequent loss of power, and do not guarantee FDR control in finite samples. Bayesian procedures encounter substantial computational complexity; the efficient sampling strategies that have been proposed over the years (16, 17) often rely on approximations of the posterior distribution that are too crude for precise variable selection.

A second difficulty arises from the strong local dependence between SNPs on the same chromosome, known as linkage disequilibrium (LD) (18). The genotyped SNPs are chosen as to "cover" the genome: Even though the true causal variants may not be directly observed, data on sufficiently close SNPs can act as proxy. Genotyping platforms thus rely on many closely spaced SNPs whose alleles are strongly dependent. This design poses some challenges for the statistical analysis. In multivariate regression models, dependencies between variables make it difficult to select any one of them, as others provide very similar signals. Marginal tests for independence between the trait and individual neighboring variants correspond to redundant hypotheses; therefore, such discoveries are more difficult to interpret, and straightforward applications of FDR-controlling methods (19) become particularly problematic (20).

A third challenge is the lack of independence over samples, or population structure, due to the presence of individuals sharing some degree of relatedness and common ancestry or otherwise belonging to identifiable subgroups. This induces long-range LD across the entire genome (21), which also gives rise to misleading associations. The problem is accentuated if the sample size is large, as in that case even relatively weak spurious associations can become statistically significant.

Standard Pipeline. The standard statistical pipeline involves multiple steps, which evolved in response to the above challenges. Typically, one identifies promising signals by testing for the association of the phenotype with one variant at a time, through a simple linear model. To correct for population structure, the SNP-by-SNP marginal regression models may include additional covariates, such as the top principal components of the genotype matrix (21). Alternatively, random effects may be utilized to account for relatedness and, partially, for the effect of other genetic loci (22–25). The *P* values thus computed are thresholded to approximately control the familywise error rate (FWER). To eliminate the redundancy in the findings, variants associated with the phenotype and highly correlated with one another are "clumped" into distinct groups, utilizing procedures such as that implemented in PLINK (26). The results of these univariate tests are then taken as input by two different multivariate analyses: fine mapping (27, 28) and polygenic risk scores (29). The former aims to identify the causal variants among many similarly associated SNPs in LD. The latter seeks to construct a predictor of the trait for future samples based on a large number of genetic variants across the genome.

This pipeline is a patchwork of different approaches, based on strong assumptions (e.g., linear effects and Gaussian errors) and inconsistent models; in fact, the linear models often applied for locus discovery differ in various aspects from those employed for fine mapping, including in the choice of which variables correspond to fixed effects and which correspond to random ones. The issue is that, even though some fine-mapping methods can be very effective at teasing out distinct signals within one locus (28), they cannot be applied genome-wide without prior screening via marginal testing due to computational limitations. Unfortunately, although it is convenient, such a two-step approach does not guarantee the type-I errors are controlled in the final output (30). Despite the lack of rigorous theoretical grounding, the standard pipeline is well established, partly because it has led to the discovery of a number of loci that appear to be reproducibly associated with the traits of interest. Indeed, geneticists have identified more statistically significant loci than it is currently practical to investigate in follow-up studies. However, the limitations of the standard approach become more evident when one tries to identify causal variants and leverage them to predict disease risk. On the one hand, there have been few mechanistic validations of loci identified by GWAS (31): The outputs of the pipeline are hard to interpret directly and expensive to investigate in follow-up studies. On the other hand, the performance of polygenic risk scores is not robust across populations (32, 33), which highlights the difficulty of identifying causal variants and raises questions of equity and fairness (34, 35). A statistical method that pursues more directly the original GWAS requisites may improve performance in both the above tasks.

A Different Framework. We present a statistical approach, KnockoffGWAS, which accounts for the role of multiple variants, adjusting for multiplicity and population structure. This is the culmination of years of developments.

Our work begins with knockoffs, introduced by ref. 36 within the context of low-dimensional linear models and extended by ref. 37 to the high-dimensional "model-X" setting considered here, which requires no parametric assumptions about the distribution of the phenotype conditional on the genotypes. Knockoffs are randomly generated negative control variables, designed to be indistinguishable from the original null variables (those not directly associated with the trait), even with regard to their dependence with the causal ones. Such exchangeability allows us to tease apart variants that truly "influence" the trait, as those are the only ones whose association with the phenotype is significantly stronger than that of their knockoffs. This idea is implemented by the knockoff filter (36): the algorithm computing a knockoff-based significance threshold controlling the FDR. This filter can be applied to any association statistics (if they treat the original variables and the knockoffs fairly), which allows us to exploit the power of modern machine-learning algorithms while retaining valid inferences (37). The key ingredient for knockoffs is an accurate model for the distribution of the original variables, which fortunately is available for GWAS data.

This paper presents a series of technical advances, which cumulatively offer a complete analysis pipeline accounting for the most serious remaining source of confounding, population structure, thus bringing us closer to proper causal inferences (38). In particular, we improve on earlier works (30, 39) focused on knockoffs for individuals from a homogeneous population, by introducing methods to handle relatedness, diverse ancestries, and admixture, which finally allows us to analyze complex datasets in their entirety.

Methodology

Notation, Problem Statement, and Assumptions. Consider a dataset with genotype and phenotype information for n individuals, where $X^{(i)} \in \{0, 1, 2\}^p$ counts the minor alleles of the ith subject at each of p markers, and $Y^{(i)} \in \mathcal{Y}$ is the phenotype (taking either discrete or continuous values in $\mathcal{Y}$). The n individuals are divided into disjoint subsets (self-reported or inferred), $\{F\}_{F \in \mathcal{F}}$,

where $\mathcal{F}$ is a partition of $\{1,\dots,n\}$. We refer to these subsets as families, although the grouping is more flexible than in classical family studies. Concretely, we define families by grouping together individuals with observed kinship coefficient above 5%; see *Materials and Methods* for further details and a justification of this cutoff.

Our goal is to detect variants (or groups thereof) containing distinct associations with Y; that is, we ask whether the conditional distribution of $Y \mid X$ depends on a variant X_j or a fixed group of variants $X_G = \{X_j\}_{j\in G}$. In practice, our analyses will be repeated for different choices of these groups at different levels of resolution; see *The Knockoff Filter* and *Preprocessing of the UK Biobank Data* for more details. For simplicity, here we focus on one genotype partition, which we assume to be fixed and such that all SNPs in each group X_G are physically contiguous; see ref. 30 for a justification of this simplification. Let then $\mathcal{G} = \{G\}_{G\in\mathcal{G}}$ be any partition of $\{1,\dots,p\}$ into contiguous groups. If X_G denotes the genotypes for all SNPs in group G, and X_{-G} denotes the genotypes for those outside it, we want to test conditional null hypotheses (37) of the form

$$\mathcal{H}_G : Y \perp\!\!\!\perp X_G \mid X_{-G}. \qquad [1]$$

In words, $\mathcal{H}_G$ is true if and only if knowing X_G provides no information about Y beyond what can be gathered from the rest of the genome.* Testing $\mathcal{H}_G$ in Eq. **1** accounts for population structure because the latter is determined by the genotypes and can be reconstructed almost exactly by looking at X_{-G}. (Since G is a relatively small set, X_{-G} collects almost all measured variants across the genome.) That is, by conditioning on X_{-G}, we are automatically conditioning on ethnicity, subpopulation information, and so on. This is akin to using principal components to capture population structure (21), but X_{-G} contains even more information: The top principal components can be essentially reconstructed from X_{-G}. Thus, testing Eq. **1** correctly addresses the requisite that the analysis of GWAS data should promote the discovery of interesting biological effects rather than just any association. This argument is formalized with a causal inference model in *SI Appendix*, section S1.a and Fig. S1.

Our approach draws strength from modeling the randomness we understand, not from making assumptions about the unknown relation between the genetic variants and the trait. In this sense, our method is Fisherian. For instance, we do not posit a (generalized) linear model, although it would be convenient, because we have a priori no way to tell whether it is realistic. Instead, we assume only the phenotypes in different families are conditionally independent of one another given the genotypes, while those in the same family may also be affected by shared environmental factors (*SI Appendix*, section S1.a). Our modeling concentrates on the genotypes, whose inheritance mechanisms are already well understood. Precisely, we jointly describe the distribution of all genotypes within the same family with hidden Markov models (HMMs) analogous to those used for phasing and imputation (41), taking into account the observed patterns of relatedness and the ancestry of each individual. These HMMs, which we assume to be conditionally independent across different families given the reconstructed ancestries of all individuals, are then leveraged to generate the negative controls, or knockoffs, defined below.

Exchangeable Negative Controls. A random matrix $\tilde{\mathbf{X}} \in \{0,1,2\}^{n\times p}$ is a knockoff of $\mathbf{X}$, with respect to a partition $\mathcal{G}$ of $\{1,\dots,p\}$, if it satisfies two properties. First, $\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}$, which says that $\tilde{\mathbf{X}}$ provides no additional information about $\mathbf{Y}$ (this is always true because $\tilde{\mathbf{X}}$ is generated looking at $\mathbf{X}$ but not at $\mathbf{Y}$). Second, the joint distribution of $[\mathbf{X}, \tilde{\mathbf{X}}] \in \{0,1,2\}^{n\times 2p}$ must be invariant upon swapping X_G with the corresponding knockoffs, simultaneously for all individuals in any family F:

$$\left[X^{(F)}, \tilde{X}^{(F)}\right]_{\mathrm{swap}(G)} \overset{d}{=} \left[X^{(F)}, \tilde{X}^{(F)}\right], \qquad [2]$$

$\forall G \in \mathcal{G}, F \in \mathcal{F}$. Above, $\mathrm{swap}(G)$ swaps all columns of $X^{(F)}$ indexed by G with the corresponding columns of $\tilde{X}^{(F)}$ (*SI Appendix*, Fig. S2). This means that, upon seeing a list of unordered pairs $\{X_G^{(F)}, \tilde{X}_G^{(F)}\}_{G\in\mathcal{G}}$, we have no way to tell which genetic variants are original and which are knockoffs. Of course, looking at Y may allow us to tell some variables apart because, conditional on Y, the symmetry between X_G and $\tilde{X}_G$ is lost for nonnulls, and this is the whole point of knockoff testing.

It is difficult to construct valid knockoffs because the exchangeability property in Eq. **2** does not simply say the knockoffs should have the same distribution as the genotypes. For example, permuting the rows of $\mathbf{X}$ would lead to dummy variables $\mathbf{X}'$ with the same distribution as $\mathbf{X}$, but not in LD with them and hence not satisfying Eq. **2**; indeed, any swap of X_G with X'_G would be noticeable if we compare them to the real variants from neighboring groups. By contrast, our knockoffs $\tilde{X}_G$ will be in LD with $\tilde{X}_{-G}$ and with X_{-G}. In summary, to satisfy Eq. **2**, knockoffs need to preserve short- as well as long-range LD, both among themselves and with the original genotypes, consistent with the ancestries and family structures reconstructed from all remaining variants. This is highly nontrivial, especially if the population structure is unknown a priori, and it requires a different approach; see *Modeling Genotypes and Constructing Knockoffs* for an overview of the key ideas developed here, while the details are in *Materials and Methods* and *SI Appendix*, sections S1.b and S1.c.

Knockoffs for a genotype partition $\mathcal{G}$ are specifically designed to test $\mathcal{H}_G$ in Eq. **1** as powerfully as possible (30). Smaller groups of SNPs allow us to test more informative but fundamentally more challenging hypotheses. As a result, higher-resolution knockoffs must satisfy stronger exchangeability in Eq. **2** and tend to be individually more similar to the real genotypes (*SI Appendix*, Fig. S3), which reduces power. Larger groups relax the constraint in Eq. **2**, increasing power but making any discoveries less informative.

We demonstrate empirically that we can construct knockoffs that are nearly indistinguishable, in the above sense, from the real genotypes in the UK Biobank data. We explain in *The Knockoff Filter* how this exchangeability theoretically guarantees our method can control the FDR for any phenotype, regardless of what its genetic architecture may be (37). Therefore, in principle there is no need to carry out simulation studies based on synthetic phenotypes to empirically validate the FDR control of our method, although we will nonetheless utilize this approach later to compare its power to that of some benchmarks. Fig. 1 visualizes different measures of exchangeability comparing our knockoffs to real genotypes in terms of principal components, familial relatedness, and LD. For simplicity, we focus on the partition containing exactly one SNP per group, which must follow the strictest constraints; analogous diagnostics for low-resolution knockoffs are in *SI Appendix*, Fig. S4. Fig. 1*A* shows that a principal component analysis (PCA) on 10,000 individuals with extremely diverse ancestries (*SI Appendix*, Table S1) gives very similar results when performed on either $\mathbf{X}$ or $\tilde{\mathbf{X}}$; this is consistent with the requirement that knockoffs preserve population structure and cannot be recognized even by someone knowing the top principal components. By contrast, knockoffs constructed by earlier methods did not preserve population structure (30, 37, 39) (*SI Appendix*, Fig. S5). Fig. 1*B* demonstrates the

*If we were to posit a linear model, $Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon$, which we do not, then the null hypothesis in Eq. **1** would be equivalent to saying that $\beta_j = 0$ for all $j \in G$ (37, 40).

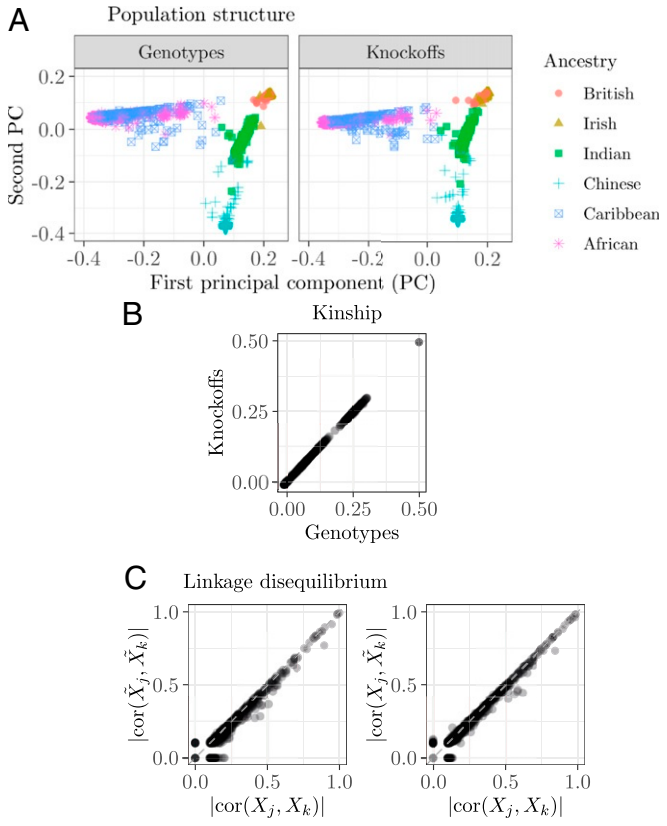

Fig. 1. Exchangeability of knockoffs and UK Biobank genotypes. (*A*) PCA for 10,000 individuals, separately for genotypes and knockoffs. (*B*) Kinship between 2,000 pairs of related individuals, computed separately on genotypes and knockoffs. Kinship is measured by means of kinship coefficients estimated by the KING software (42) so that a value of 0.5 indicates monozygotic twins and 0 indicates no relatedness. (*C*) Pairwise correlations between nearby variants on chromosome 22 (minor allele frequency $\geq$0.01) for the individuals in *A*, with (*Left*) or without (*Right*) swapping genotypes (X) and knockoffs ($\tilde{X}$).

estimated kinship of any two related individuals is the same regardless of whether it is based on $\mathbf{X}$ or $\tilde{\mathbf{X}}$; this was also not guaranteed by earlier methods. Finally, Fig. 1*C* demonstrates our knockoffs preserve short-range LD, similar to that in refs. 30 and 39. These plots reveal the pairwise correlations between genotypes on nearby variants (within 100 kb of each other) are unchanged when one or both variables are replaced by their knockoffs. In *SI Appendix*, Figs. S6–S8 show our knockoffs preserve longer-range LD. We emphasize these results are not easily achieved; for example, it would not suffice to let $\tilde{\mathbf{X}}$ be an independent and identically distributed sample of $\mathbf{X}$ from the same population because that would violate the exchangeability in Fig. 1*C*, *Right*, as well as other symmetries not visualized here for lack of space. Of course, trivially valid knockoffs could be obtained by making identical copies of the genotypes, $\tilde{X} = X$, but that cannot yield any discoveries (36), whereas the results in *Application to the UK Biobank Data* will demonstrate our method is powerful.

Modeling Genotypes and Constructing Knockoffs. We explain here the main ideas of our knockoff construction, while the details are in *Materials and Methods*. First, we assume all haplotypes have been phased and we denote by $H^{(i,a)}, H^{(i,b)} \in \{0,1\}^p$ those inherited by individual i from each parent, so that $X^{(i)} = H^{(i,a)} + H^{(i,b)}$. As in ref. 30, we model the haplotypes and leverage them to construct phased knockoffs, namely $\tilde{H}^{(i,a)}, \tilde{H}^{(i,b)}$, which can then be simply combined to obtain valid knockoff genotypes: $\tilde{X}^{(i)} = \tilde{H}^{(i,a)} + \tilde{H}^{(i,b)}$.

The haplotype distribution is approximated by an HMM in the style of SHAPEIT (43–45). This overcomes the main limitation of the fastPHASE HMM (46), which was used to construct knockoffs (30, 39) for homogeneous populations (30) but cannot describe both LD and population structure (47). The SHAPEIT HMM describes the haplotypes as a mosaic of K reference motifs corresponding to the haplotypes of other individuals in the dataset, where K is fixed (e.g., $K = 100$). Different individuals may have different sets of motifs, chosen based on their similarity (*Materials and Methods*). The intuition is that the haplotypes of someone from England should be well approximated by a mosaic of haplotypes from other English samples. However, the two copies of the same chromosome for one individual (phased haplotypes) are allowed to have different sets of motifs, reflecting the idea that the parents may have different ancestries. Further, the reference motifs can vary across chromosomes and even locally across wide windows within each chromosome, enabling the description of possible admixtures (48) (*Materials and Methods*). Conditional on the references, the identity of the motif copied at each position is described by a Markov chain with transition probabilities proportional to the genetic distances between neighboring sites; different chromosomes are treated as independent. Conditional on the Markov chain, the motifs are copied imperfectly, as relatively rare mutations can independently occur at any site. After inferring the unobserved Markov chain in this HMM, we carefully perturb it to construct knockoffs. Fig. 2*A* presents a schematic visualization of this method.

The above model ignores familial relatedness because it describes all haplotypes as conditionally independent given the reference motifs, which cannot explain long identical-by-descent (IBD) segments (49, 50). To handle this, we model jointly haplotypes in the same family. First, we detect long IBD segments in the data (51–54). If the pedigrees are known, the IBD search can be focused within the given families; otherwise, there exists software to approximately reconstruct families (55). The results define a relatedness graph in which two haplotypes are connected if they share an IBD segment; we refer to the connected components of this graph as the IBD-sharing families. Second, we define a larger HMM jointly describing all haplotypes in each IBD-sharing family, conditional on the location of the segments. Marginally, each haplotype is modeled by the SHAPEIT HMM; however, the Markov chains for related individuals are forced to match along the IBD segments. This coupling will be preserved in the knockoffs, which will contain exchangeable IBD segments in the same locations, although they may not always have the original alleles (Fig. 2*B* and *SI Appendix*, Fig. S9).

The Knockoff Filter. Although knockoffs are constructed to be statistically indistinguishable from the genotypes, it may be possible to tell them apart by looking also at the phenotype, since $\tilde{X} \perp\!\!\!\perp Y \mid X$ but $X \not\perp\!\!\!\perp Y \mid \tilde{X}$. Loosely speaking, this implies differences in the comparisons of Y with either X_G or $\tilde{X}_G$ can provide evidence against the null hypothesis $X_G \perp\!\!\!\perp Y \mid X_{-G}$. Such property is leveraged by estimating importance measures, T_G and $\tilde{T}_G$, for each group of SNPs and knockoffs, respectively. The importance measures are combined into a test statistic for each group $G \in \mathcal{G}$; i.e., $W_G = T_G - \tilde{T}_G$. This is designed such that a large value of $W_G > 0$ is evidence against the null hypothesis, while the sign of statistics for null groups is independent coin flips (36). The knockoff filter computes an adaptive significance threshold for these statistics, provably controlling the FDR if the knockoffs are correctly exchangeable. See *SI Appendix*, Fig.

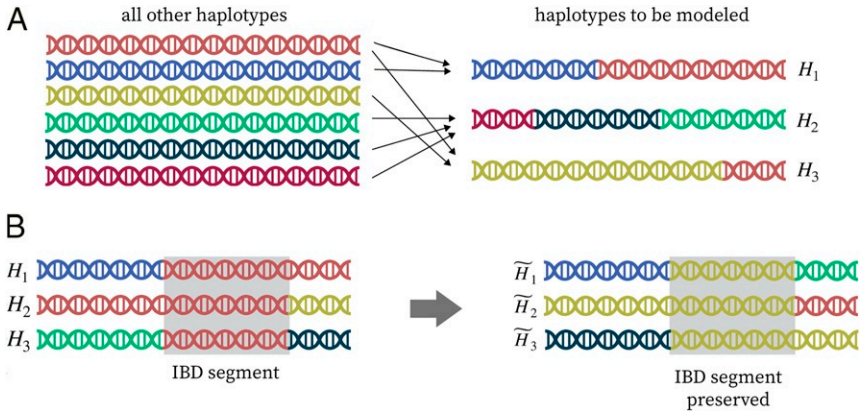

Fig. 2. Visualization of the haplotype HMM and knockoff construction. (*A*) Each haplotype sequence, $H^{(i)}$, is described as a mosaic of motifs from a subset of other haplotypes (in different colors). (*B*) Haplotypes and knockoffs of closely related individuals share IBD segments where their alleles match exactly (shaded segments).

S10, for a full schematic. In addition to controlling the FDR, we can assess the significance of individual findings, either through a local estimate of the false discovery proportion (FDP) (30, 56), which requires a sufficiently large number of discoveries, or through a q value (57), which is defined as the smallest nominal FDR level at which that discovery could have been reported.

The FDR guarantee holds regardless of the unknown relation between X and Y and with any importance statistics. The statistics can be computed by virtually any method and easily incorporate prior knowledge (37). As in earlier works (30, 37, 39), we utilize a sparse generalized linear model (Lasso) (12), although our inference never assumes its validity. This is practical with large data (58, 59), interpretable (4), and powerful compared to linear mixed models (LMMs). Concretely, we fit a sparse regression of Y on $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, after standardizing the columns to have unit variance. The importance measures are $T_G = \sum_{j \in G} |\hat{\beta}_j(\lambda_{\mathrm{CV}})|$ and $\tilde{T}_G = \sum_{j \in G} |\hat{\beta}_{j+p}(\lambda_{\mathrm{CV}})|$, where $\hat{\beta}_j(\lambda_{\mathrm{CV}})$ [resp. $\hat{\beta}_{j+p}(\lambda_{\mathrm{CV}})$] is the Lasso coefficient for X_j (resp. $\tilde{X}_j$) at a value of the regularization parameter tuned by cross-validation to achieve a low prediction error.

To localize causal variants as precisely as possible, we apply the knockoff filter at multiple resolutions; partitions with larger groups yield more power, but at the cost of less informative findings. The FDR is then controlled separately for each level of resolution. Further, it is possible to coordinate the results at different resolutions to ensure they are consistent with one another (30), provably retaining the same FDR guarantees. This requires a variation of the multilayer knockoff filter first proposed in ref. 40, which we do not apply here in the interest of simplicity. Note that naively aggregating the results obtained at different resolutions may not control the false discovery rate (60), which is why we report them separately. For example, reporting only the highest-resolution finding in each locus would not be theoretically valid, although it sometimes performs quite well in practice (30).

Imputed Variants. Imputed variants—those not directly measured but instead predicted with a model based on nearby observed variants—are commonly included in fine-mapping analyses (41). However, correctly utilizing imputed variants requires care because they are not as informative as the measured ones. In particular, imputed variants are conditionally independent of the trait given the observed genotypes, as they are constructed based only on the latter. As such, it is impossible to attribute distinct signals to imputed variants without stronger assumptions; by definition, any association observed between them and the trait could be equally well explained as a (possibly nonlinear) association with the observed variants. Two forms of modeling assumptions could be used to help attribute signals to imputed variants. First, functional annotations could suggest the imputed variant is more plausibly responsible for an observed association. Second, one may posit a linear model and search for variants that individually explain as much of the response as possible. The latter is the most typical approach (28).

We do not include imputed variants in our analysis due to the fundamental issue above: one could never conclude that an unseen variant is causal based on the data alone. Our method identifies relations supported by the available evidence without the need of modeling assumptions, such as sparsity and linearity, and does not make further claims. In particular, we never single out an imputed variant as causal, but instead we identify promising regions of the genome containing at least one measured variant. This is an impartial reporting of the evidence at hand, isolating statistical associations only to the resolution achievable with the data. While it may sometimes be desirable to conduct fine mapping at the resolution of imputed variants, we keep this task distinct from our analysis. Further, unmeasured variants would remain a possible source of confounding (as illustrated in *SI Appendix*, Fig. S1) regardless of whether imputed SNPs are included in the analysis, although such confounding can be expected to be small when interpreting our findings at low resolution (e.g., hundreds of kilobases).

Leveraging Covariates. The flexibility of our method allows one to easily leverage additional information to compute even more powerful statistics. In our application, we include relevant measured covariates U^m (such as sex, age, and squared age, as well the top five genetic principal components) in the above regression model, replacing $[\mathbf{X}, \tilde{\mathbf{X}}]$ with $[\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{U}^m]$. These covariates explain some of the phenotypic variation that cannot be captured by a sparse linear combination of genotypes, thus reducing the noise in the model. The coefficients for U^m are not regularized, and they do not directly enter the calculated statistics. The validity of the knockoff filter requires the knockoffs to be exchangeable with the genotypes conditional on any covariates utilized in the computation of the test statistics (Eq. **2**). This is the case in our analysis, since sex and age can be safely assumed to be independent of the genotypes (we do not analyze the sex chromosomes), and the principal components are related to the genotypes through the population structure, for which our knockoffs already account. In general, one can thus explicitly analyze any subset U^m of the covariates U that

are already implicitly taken into account by our conditional hypotheses.

Application to the UK Biobank Data

Preprocessing of the UK Biobank Data. We test our method through simulations with the phased haplotypes of 489,000 individuals (592,000 SNPs, chromosomes 1 to 22). After some preprocessing (*Materials and Methods*), we partition each autosome into contiguous groups at seven levels of resolution, ranging from that of single SNPs to that of 425-kb-wide groups (*SI Appendix*, Table S2). The partitions are obtained chromosome by chromosome through complete-linkage hierarchical clustering, based on the dissimilarity measures defined below, and cutting the resulting dendrogram at different heights to obtain groups at different levels of resolution, similar to that in ref. 30. This procedure guarantees the partitions are nested: Each group is contained in exactly one larger group at the resolution immediately below. The dissimilarity between two SNPs is defined as their genetic distance measured in centimorgans, as previously estimated in a European population (61). This makes the groups contiguous and homogeneous in terms of LD at each resolution, and it may result in heterogeneous numbers of SNPs and physical widths because the recombination rate varies across the genome.

The individuals have diverse ancestries, although most are British (430,000), Irish (13,000), or other Europeans (16,000). There are 136,818 samples with reported close relatives, divided into 57,164 families (*Materials and Methods*). We apply RaPID (55) to detect IBD segments longer than 3 cM, chromosome by chromosome, ignoring (for simplicity) those shared by individuals who are not in the same family; this gives us 7,087,643 segments over the entire genome. Then, we generate knockoffs preserving both population structure and IBD segments.

Analysis of Simulated Phenotypes. We simulate continuous phenotypes conditional on the real genotypes, from a homoscedastic linear model with 4,000 causal variants paired in 2,000 (100-kb-wide) loci placed uniformly across the genome, so that each locus contains 2 causal variants. The total heritability is varied as a control parameter. This gives us a controlled but realistic testing environment. We take BOLT-LMM (25) as a benchmark, applying it on the same data with standard parameters. However, the comparison requires some care since LMMs are designed to test marginal associations, accounting for population structure (25) but not LD (30), and controlling the FWER instead of the FDR.

It is standard to combine (or clump) marginally significant LMM discoveries from nearby SNPs, e.g., using the standard PLINK (26) algorithm as in ref. 25. Ideally, this should allow each clump to be interpreted as indicating a distinct discovery; however, the solution is imperfect because even relatively far-apart SNPs may not be completely independent of each other if they are on the same chromosome. This issue is particularly important at the Biobank scale, as larger sample sizes make even weak correlations statistically significant, complicating the interpretation of marginal hypotheses. We address this difficulty by further consolidating the clumps computed by PLINK if they are within 100 kb of each other, as in ref. 30. In our simulations, this strategy ensures the final discoveries are distinct because the true causal loci are well spaced, but that may not always be the case in practice. Therefore, it is unclear how to best clump marginal associations in general.

Fig. 3 visualizes our discoveries within a locus containing two causal variants, although the method operates genome-wide (*SI Appendix*, Fig. S11). Fig. 3 also shows the nearby LMM findings, clumped by PLINK but unconsolidated. Here, our method localizes the causal variants precisely, while the LMM findings span a long region including many spurious associations. *SI Appendix*, Fig. S12 describes how the results obtained with each method change as the signal strength is varied.

Regarding the discrepancy in the target error rates, it is unfortunately difficult to use LMMs to control the FDR. The issue is this: LMMs test marginal hypotheses (and the test statistics are not independent of each other), while we ultimately need to report distinct (conditional) discoveries. This issue was discussed in ref. 30, and it is consistent with the general difficulty of controlling the FDR after applying any sort of postprocessing to the output of a testing procedure (60). We consider two possible solutions to make the error rates more comparable, which are informative within our simulations but would not work in practice. The naive approach is to apply the Benjamini–Hochberg (BH) correction (19) to the marginal *P* values before clumping.

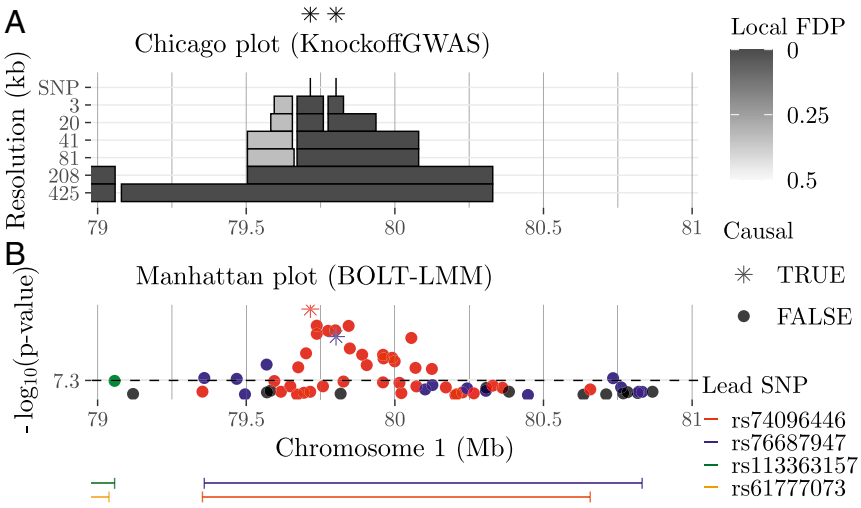

Fig. 3. KnockoffGWAS and BOLT-LMM discoveries for a simulated trait based on the genotypes of 489,000 UK Biobank individuals with population structure. (*A*) The shaded rectangles represent our discoveries at different resolutions. The FDR level is 10%. Darker rectangles have lower estimated local FDP. The lighter rectangles are false discoveries because they do not contain causal variants (whose positions are marked by asterisks on top). (*B*) BOLT-LMM *P* values from the same data and PLINK clumps (segments below) at the genome-wide significance level (5×10^{-8}), utilizing different colors for different clumps. The colors match those of the corresponding *P* values.

We shall see this inflates the type-I errors. The second approach involves an imaginary oracle that knows the identities of the causal variants and exploits them to automatically adjust the significance level for the LMM P values, such that the proportion of false discoveries (after clumping) equals the target FDR exactly. Obviously, this oracle does not exist for real phenotypes.

Fig. 4 compares the performances of KnockoffGWAS and BOLT-LMM. The latter targets the FWER (5×10^{-8} genome-wide significance level) or the FDR with either of the two aforementioned strategies. Performance is assessed in terms of power, false discoveries, and resolution. Fig. 4*A* focuses on KnockoffGWAS at low resolution (genome partition with median group size equal to 208 kb) and BOLT-LMM with strong clumping (consolidating clumps within 100 kb of each other). Here, power is measured as the proportion of the 2,000 causal loci encompassed by at least one discovery, while the FDP is defined as the fraction of findings that do not contain causal variants. The resolution is measured as the median width of the reported genetic segments. The results show our method is almost as powerful as the imaginary LMM oracle, but is slightly more conservative and reports narrower (more informative) discoveries. By contrast, BOLT-LMM makes fewer discoveries when targeting the FWER and reports too many false positives when heuristically targeting the FDR.

Fig. 4*B* summarizes the KnockoffGWAS discoveries at different resolutions, reporting only the resolution with the most findings, for simplicity. Since the reported resolution may be different for different heritability values, we measure power in terms of the total number of true discoveries rather than as a fraction. Here, the goal is to detect as many distinct associations as possible, ideally also distinguishing between multiple causal variants in the same locus, so the LMM findings are clumped but unconsolidated. KnockoffGWAS always controls the proportion of false discoveries below the target FDR, and it is either comparable to or more powerful than the oracle. Furthermore, we localize causal variants more precisely as the heritability increases, while the LMM clumps become wider and increasingly polluted by spurious associations, as visualized in *SI Appendix*, Fig. S12. Additional simulations on subsets of the UK Biobank samples (*SI Appendix*, section S2.a) demonstrate our method is robust and powerful even when the individuals have extremely diverse ancestries (*SI Appendix*, Table S1 and Figs. S13 and S14) or very strong familial relatedness (*SI Appendix*, Table S3 and Fig. S15), in which cases the FDR would have been much larger than desired had we not taken the population structure into account.

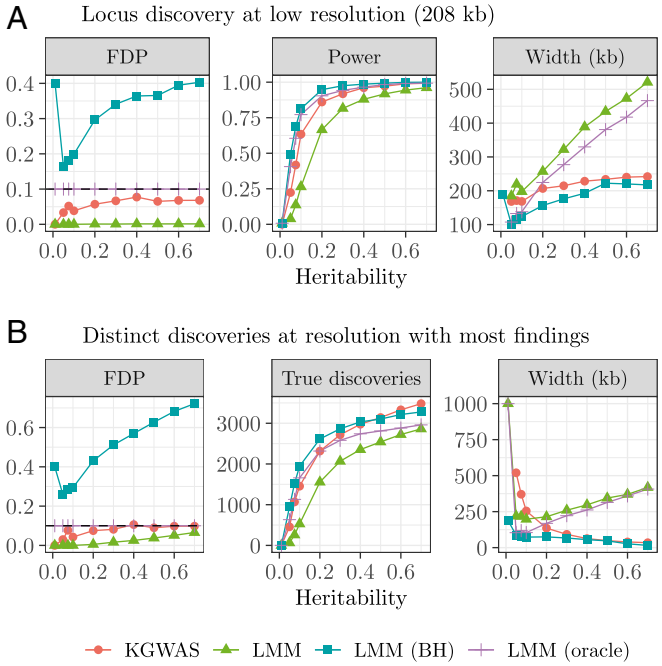

Fig. 4. Performance on real genotypes and synthetic phenotypes from a model with 4,000 causal variants. The results obtained with either KnockoffGWAS (nominal FDR 0.1) or BOLT-LMM (5×10^{-8}, heuristic FDR, and oracle FDP calibration) are shown as a function of the total heritability. (*A*) Low-resolution KnockoffGWAS discoveries and strongly clumped LMM findings. (*B*) Multiresolution KnockoffGWAS discoveries (reported only at the resolution with the most findings for each value of heritability) and weakly clumped BOLT-LMM findings. Other details are as in Fig. 3.

Analysis of UK Biobank Phenotypes. We study four continuous traits (height, body mass index, platelet count, systolic blood pressure) and four diseases (cardiovascular disease, respiratory disease, hyperthyroidism, diabetes), as defined in *SI Appendix*, Table S4. To increase power, we include a few covariates, as explained in *The Knockoff Filter*. Table 1 reports the numbers of low-resolution (208 kb) discoveries (target FDR 10%) and compares them to those obtained by BOLT-LMM, which are clumped with PLINK (5×10^{-8} significance threshold) but unconsolidated, consistent with ref. 25. The agreement between these discoveries is summarized as in ref. 30: Two findings are said to overlap if they indicate nondisjoint genomic regions. As explained in *Analysis of Simulated Phenotypes*, these two methods test different hypotheses and target different error rates, so we do not expect them to yield the same numbers of findings. Nonetheless, the comparison is informative because the findings of the LMM are commonly attributed the same meaning as ours: pointing to genomic regions likely to contain distinct causal variants. However, to achieve this interpretation, the LMM results first need to be clumped. It is important to note that, because LMMs test marginal hypotheses, applying standard FDR controlling procedures to their P values would not result in valid FDR control for the clumped discoveries, as demonstrated in Fig. 4 and discussed earlier in ref. 30.

BOLT-LMM is applied on 459,000 European samples (25) for all phenotypes except diabetes and respiratory disease, for which it is applied on 350,000 unrelated British samples (30) for the sake of consistency in phenotype definitions. These results suggest KnockoffGWAS is more powerful: It discovers almost all findings reported by the LMM and many other ones. The findings at other resolutions are summarized in *SI Appendix*, Table S5. Note that the model assumed by BOLT-LMM assumes continuous-valued phenotypes, and other LMM-based methods have been specifically developed for case–control studies (62); however, BOLT-LMM is still a standard benchmark here because the ratio of cases and controls is not very small (25) (*SI Appendix*, Table S4). *SI Appendix*, Fig. S16 visualizes our discoveries for cardiovascular disease in the form of Manhattan plots, using q values (57) to measure the individual significance of each finding. The full list of our discoveries is available online at https://msesia.github.io/knockoffgwas/, along with an interactive visualization tool.

SI Appendix, Table S6 confirms our findings are consistent with those of ref. 30, although our method is more powerful because it leverages a larger sample; see *SI Appendix*, Table S7 for more details. The only exception is at the single-SNP resolution, which may be partially explained by these discoveries being fewer and thus more susceptible to the random variability of knockoffs. Table 2 summarizes the increase in discoveries at each resolution directly resulting from the inclusion of samples with close relatives or non-British ancestry. The inclusion of related individuals yields many more discoveries, while it is unsurprising that

Table 1. KnockoffGWAS discoveries (208-kb resolution, 10% FDR), using all 487,000 UK Biobank samples, and corresponding BOLT-LMM findings (5×10^{-8})

	Knockoff GWAS discoveries		BOLT-LMM discoveries	
Phenotype	Total	Overlap with LMM	Total	Overlap with KZ
bmi	2,395	898 (37.5%)	697	689 (98.9%)
cvd	940	274 (29.1%)	257	249 (96.9%)
diabetes	113	52 (46.0%)	62	55 (88.7%)
height	3,339	2,228 (66.7%)	2,464	2,430 (98.6%)
hypothyroidism	295	129 (43.7%)	143	142 (99.3%)
platelet	1,743	1,057 (60.6%)	1,204	1,183 (98.3%)
respiratory	262	82 (31.3%)	94	92 (97.9%)
sbp	1,183	561 (47.4%)	568	530 (93.3%)

For example, we report 940 distinct discoveries for cardiovascular disease, 274 of which contain significant LMM associations. The LMM reports 257 discoveries for this phenotype, 96.9% of which overlap with at least one of our discoveries

diverse ancestries bring smaller gains since there are relatively few non-British samples.

Validation of Discoveries. We begin to validate our findings by comparing them to the GWAS Catalog (3), the Japan Biobank Project (63), and the FinnGen resource (64) (standard 5×10^{-8} threshold for the P values reported by the latter two). *SI Appendix*, Table S8 indicates most high-resolution discoveries correspond to SNPs with a known association to the phenotype. This is especially true for findings also detected by BOLT-LMM, although many of our additional ones are confirmed (*SI Appendix*, Table S9). For example, we report 1,089 discoveries for cardiovascular disease at the 425-kb resolution, only 255 of which are detected by BOLT-LMM; however, 85.6% of our additional 834 discoveries are confirmed in at least one of the aforementioned resources. Further, *SI Appendix*, Table S10 suggests most relevant associations in the Catalog (above 70%) are confirmed by our findings, which is again indicative of high power. The relative power (proportion of known associations that we discover) seems above 90% for quantitative traits, but below 50% for all diseases except hypothyroidism, probably due to the relatively small number of cases in the UK Biobank dataset compared to more targeted case–control studies.

The 5×10^{-8} threshold for the Japan Biobank Project and the FinnGen resource is too conservative if the goal is to confirm selected discoveries. Therefore, we next carry out an enrichment analysis. The idea is to compare the distribution of the external statistics within our selected loci to those from the rest of the genome; see *SI Appendix*, section S2.c. This approach estimates the number of replicated discoveries but cannot tell exactly which ones are confirmed; therefore, we will consider alternative validations later. (A more precise analysis is possible, but has low power; see *SI Appendix*, section S2.c). *SI Appendix*, Table S11 shows many additional discoveries can thus be validated, especially at high resolution. See *SI Appendix*, Table S12 for more details about enrichment. Table 3 summarizes the confirmatory results. Respiratory disease is excluded here because the FinnGen resource divides it among several fields, so it is unclear how to best obtain a single P value. Regardless, the GWAS Catalog and the FinnGen resource already directly validate 90% of those findings.

We continue by cross-referencing with the literature the discoveries missed by BOLT-LMM and unconfirmed by the above studies, focusing for simplicity on the 20-kb resolution. *SI Appendix*, Table S13 shows most of these discoveries point to SNPs with known associations to phenotypes closely related to the one of interest; see *SI Appendix*, Table S14 for details. Finally, *SI Appendix*, Table S15 shows that many lead SNPs (those with the largest importance measure in each group) within our unconfirmed discoveries have known consequences in the protein-coding sequence.

Fig. 5 shows a discovery for cardiovascular disease, which seems unlikely to be false based on the estimated local FDP. The highest-resolution finding here spans four genes, but we could not find previously reported associations with cardiovascular disease within this locus. However, one gene (SH3TC2) is associated with blood pressure (65) and another (ABLIM3) with body mass index (66). *SI Appendix*, Fig. S17 visualizes the same discovery within a wider portion of the chromosome.

Discussion

We developed a method for constructing knockoffs that preserve population structure and familial relatedness, as well as LD, thereby obtaining a fully operational conditional testing strategy for the analysis of GWAS data. In particular, we can now analyze Biobank-scale datasets both efficiently, leveraging the available prior knowledge and the power of virtually any machine-learning tools, and agnostically, without parametric assumptions about the phenotype. While we cannot identify causal variants exactly due to possible unaccounted confounders such as missing variants, we push farther in that direction compared to the traditional pipeline.

The inclusion of related and ethnically diverse individuals is crucial for several reasons (67). First, large studies are sampling

Table 2. Cumulative numbers of discoveries for all UK Biobank phenotypes at different resolutions, utilizing different subsets of the samples.

	Including related samples				Including non-British samples			
	Everyone		British		Related		Unrelated	
Resolution	Total	Change (%)	Total	Change (%)	Total	Change (%)	Total	Change (%)
Single SNP	138–167	21.0	155–125	−19.4	125–167	33.6	155–138	−11.0
3 kb	921–992	7.7	655–971	48.2	971–992	2.2	655–921	40.6
20 kb	2,814–3,527	25.3	2,808–3,355	19.5	3,355–3,527	5.1	2,808–2,814	0.2
41 kb	4,419–5,867	32.8	4,353–5,354	23.0	5,354–5,867	9.6	4,353–4,419	1.5
81 kb	6,784–8,031	18.4	6,676–7,781	16.6	7,781–8,031	3.2	6,676–6,784	1.6
208 kb	8,776–10,270	17.0	8,635–10,049	16.4	10,049–10,270	2.2	8,635–8,776	1.6
425 kb	9,401–10,730	14.1	9,028–10,297	14.1	10,297–10,730	4.2	9,028–9,401	4.1
Sample size	408k–487k	19.4	356k–430k	20.8	430k–487k	13.0	356k–408k	14.6

For example, including related individuals increases by 16.4% the number of discoveries obtained from the British samples at the 208-kb resolution (from 8,635 to 10,049). As another example, adding non-British individuals (including related ones) increases by 2.2% the number of discoveries obtained from the British samples (including related ones) at the 208-kb resolution (from 10,049 to 10,270).

Table 3. Numbers of low-resolution (208-kb) discoveries obtained with our method and confirmed by other studies or by an enrichment analysis carried out on external summary statistics.

	Total discoveries			Discoveries not found by BOLT-LMM		
	No.	Confirmed		No.	Confirmed	
Phenotype		Other (%)	Other or enrichment (%)		Other (%)	Other or enrichment (%)
bmi	2,395	1,076 (44.9)	1,620 (67.6)	1,497	335 (22.4)	806 (53.8)
cvd	940	738 (78.5)	764 (81.3)	666	472 (70.9)	493 (74.0)
diabetes	113	97 (85.8)	106 (93.8)	61	46 (75.4)	54 (88.5)
height	3,339	1,886 (56.5)	2,493 (74.7)	1,111	164 (14.8)	556 (50.0)
hypothyroidism	295	156 (52.9)	226 (76.6)	166	43 (25.9)	101 (60.8)
platelet	1,743	453 (26.0)	1,017 (58.3)	686	29 (4.2)	256 (37.3)
respiratory	262	241 (92.0)	NA	180	159 (88.3)	NA
sbp	1,183	643 (54.4)	885 (74.8)	622	154 (24.8)	358 (57.6)

For example, 81.3% of our 940 discoveries for cardiovascular disease are confirmed either by the results of other studies or by the enrichment analysis. The results are stratified based on whether our findings can be detected by BOLT-LMM using the UK Biobank data (excluding non-European individuals). bmi, body mass index; cvd, cardiovascular disease; sbp, systolic blood pressure.

entire populations densely (64, 68), which yields many close relatives. It would be wasteful to discard this information and potentially dangerous not to account for relatedness. Second, the historical lack of diversity in GWAS (which mostly involve European ancestries) is a well-recognized problem (69, 70) that biases our scientific knowledge and disadvantages the health of underrepresented populations. While this issue goes beyond the difficulty of analyzing diverse GWAS data, our work at least helps remove a technical barrier.

GWAS data from different populations are typically analyzed separately, and only later may the results be combined through meta-analyses (71, 72), partly out of concern for population structure. However, our method makes such sample splitting unnecessary. By allowing a simultaneous analysis, we can increase power because different LD patterns uncover causal variants more effectively (48). Our discoveries may also be useful to better explain phenotypic variation in minority populations (73, 74). Since the UK Biobank mostly comprises British individuals, the increase in power resulting from the analysis of other samples can only be relatively small. Nonetheless, we observe some gains when we include non-British individuals. Simulations demonstrate our inferences are valid even when the population is very heterogeneous, suggesting our approach might be particularly suitable for the analysis of more diverse data, such as those collected by the Million Veteran Program (75), for example.

Finally, including individuals with diverse ancestries opens additional research opportunities. For example, it would be interesting to understand which discoveries are consistently found in different populations (76), as this may help further weed out false positives, explain observed variations in certain phenotypes, and possibly shed more light on the underlying biological pathways. Further avenues for future research may focus on the analysis of whole-genome sequencing data including rare variants. The analysis of rare variants involves additional challenges, both computational (our method scales linearly in the number of variables, but even that cost may still be high) and statistical (signals on rare variants are more difficult to detect because they result in smaller effective sample sizes). Further, it may be more difficult to accurately model the distribution of rare variants and construct valid knockoffs. However, the SHAPEIT HMM with individual-specific reference haplotypes is known to be relatively accurate even for rare variants (45, 77)

Materials and Methods

Software Availability. Our methods are implemented in an open-source software package available from https://msesia.github.io/knockoffgwas/. This includes a standalone program written in C++, which takes as input phased haplotypes in BGEN format (78) and outputs genotype knockoffs at the desired resolution in the PLINK (26) BED format. The package also includes R scripts to partition the genome into contiguous groups of SNPs at different resolutions, compute Lasso-based test statistics, apply the knockoff filter, and visualize the discoveries interactively. Furthermore, the repository contains Bash scripts to connect the different modules of this pipeline and carry out a complete GWAS analysis from beginning to end, an example of which can be conveniently run on a small toy dataset provided with the package. Our software is specifically designed for the analysis of large datasets, as it is multithreaded and memory efficient. Furthermore, knockoffs for different chromosomes can be generated in parallel. For reference, it took us ∼4 d using 10 cores and 80 GB of memory to generate knockoffs on chromosome 1 for the UK Biobank data (∼1 million haplotype sequences, 600,000 SNPs, and 600,000 IBD segments).

The SHAPEIT HMM. We say a sequence of phased haplotypes $H = (H_1, \ldots, H_p) \in \{0, 1\}^p$ is distributed as an HMM with K hidden states if there exists a vector of latent random variables $Z = (Z_1, \ldots, Z_p) \in \{1, \ldots, K\}^p$ such that

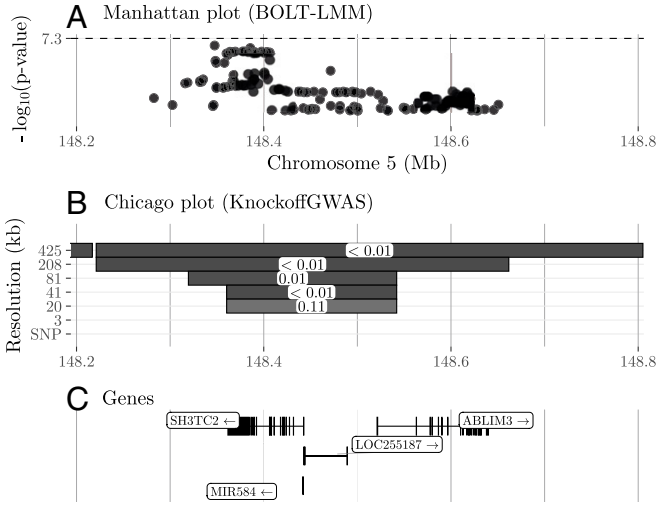

Fig. 5. Results of the analysis of UK Biobank data on cardiovascular disease, within a small portion of chromosome 5. (*A*) Marginal P values computed by BOLT-LMM on the subset of samples with European ancestry (25), for genotyped and imputed variants within this locus. All P values here are larger than 5×10^{-8}. (*B*) Findings reported by KnockoffGWAS. The shaded rectangles indicate the genetic segments discovered at different resolutions; the darker ones are more statistically significant, i.e., with a lower estimated local FDP (white labels). (*C*) Location of genes in the locus spanned by our highest-resolution discovery.

$$\begin{cases} Z \sim \mathrm{MC}\,(Q), & \text{(latent Markov chain)}, \\ H_j \mid Z \overset{\text{ind.}}{\sim} f_j(H_j \mid Z_j), & \text{(emission distribution)}. \end{cases} \quad [3]$$

Above, $\mathrm{MC}\,(Q)$ is a Markov chain with initial probabilities Q_1 and transition matrices $(Q_2, \ldots, Q_p)$.

Taking inspiration from SHAPEIT (43–45), we assume the ith haplotype sequence can be approximated as an imperfect mosaic of K other haplotypes in the dataset, indexed by $\{\sigma_{i1}, \ldots, \sigma_{iK}\} \subseteq \{1, \ldots, 2n\} \setminus \{i\}$. (Note the slight overload of notation: i denotes hereafter a phased haplotype sequence, two of which are available for each individual). We will discuss later how the references are determined; for now, we take them as fixed and describe the other aspects of the model. Mathematically, the mosaic is described by an HMM in the form of Eq. **3**, where the ith Markov chain has

$$Q_1^{(i)}(k) = \alpha_k^{(i)},$$
$$Q_j^{(i)}(k' \mid k) = \begin{cases} \left(1 - e^{-\rho d_j}\right)\alpha_{k'}^{(i)} + e^{-\rho d_j}, & \text{if } k' = k, \\ \left(1 - e^{-\rho d_j}\right)\alpha_{k'}^{(i)}, & \text{if } k' \neq k. \end{cases} \quad [4]$$

Above, d_j indicates the genetic distance between loci j and $j-1$, which is assumed to be fixed and known [in practice, we will use distances previously estimated in a European population (61), although our method could easily accommodate different distances for different populations]. The parameter $\rho > 0$ controls the rate of recombination and can be estimated with an expectation-maximization (EM) technique (*SI Appendix*). However, we have observed it works well with our data to simply set $\rho = 1$; this is consistent with the approach of SHAPEIT (43–45), which also uses fixed parameters. The positive α weights are normalized so that their sum equals one and they can be interpreted as characterizing the ancestry of the ith individual. In this paper, we simply set all αs to be equal to $1/K$, although these parameters could also be estimated by EM (*SI Appendix*). Conditional on Z, each element of H follows an independent Bernoulli distribution:

$$f_j^{(i)}(H_j^{(i)} \mid k) = \begin{cases} 1 - \lambda_j, & \text{if } H_j^{(i)} = H_j^{(\sigma_{ik})}, \\ \lambda_j, & \text{if } H_j^{(i)} \neq H_j^{(\sigma_{ik})}. \end{cases} \quad [5]$$

Above, λ_j is a site-specific mutation rate, which makes the mosaic imperfect. Earlier works that first proposed this model suggested analytic formulas for determining ρ and $\lambda = (\lambda_1, \ldots, \lambda_p)$ in terms of physical distances between SNPs and other population genetics quantities (79). However, we choose to estimate λ by EM (*SI Appendix*) since our dataset is large. We noticed it works well to also explicitly prevent λ from taking extreme values (e.g., $10^{-6} \leq \lambda_j \leq 10^{-3}$).

To save computations and mitigate the risk of overfitting, K should not be too large; here, we take $K = 100$. Larger values improve the goodness of fit relatively little, while reducing power by increasing the similarity between variables and knockoffs. Having thus fixed K, the identities of the reference haplotypes for each i, $\{\sigma_{i1}, \ldots, \sigma_{iK}\}$, are chosen in a data-adaptive fashion as those whose ancestry is most likely to be similar to that of $H^{(i)}$. Concretely, we can apply *Algorithm 1* using the Hamming distance to define haplotype similarities, chromosome by chromosome. Instead of computing pairwise distances between all haplotypes, which would be computationally unfeasible, we first divide them into clusters of size N, with $K \ll N \ll 2n$ (i.e., $N \approx 5{,}000$), through recursive 2-means clustering, and then we compute only distances within clusters, following in the footsteps of SHAPEIT v3 (45).

In practice, it is preferable to apply a more sophisticated variation of *Algorithm 1*, which utilizes a different set of local references in different parts of the chromosome. We will describe this extension later, after discussing the knockoff generation algorithm.

Algorithm 1. Choosing the HMM reference haplotypes.

Input: haplotypes $H \in \{0, 1\}^{2n \times p}$, parameter K;
Input: hyperparameters N_1, N_2, s.t. $K \ll N_1 < N_2 \ll n$.
Input: a distance measure ξ between haplotypes.
Divide $\{1, \ldots, 2n\}$ into M sets C_c s.t. $N_1 \leq |C_c| \leq N_2$.
for $c = 1, \ldots, M$ **do**
  Compute a distance matrix $D \in \mathbb{R}^{|C_c| \times |C_c|}$ using ξ.
  for i in C_c **do**
    Define set $R(i)$ of K nearest neighbors of H_i in C_c.
  $--$
Output: a set $R(i)$ of K references for each haplotype $H^{(i)}$.

Algorithm 2. Knockoffs preserving population structure.

Input: haplotypes $H \in \{0, 1\}^{2n \times p}$, genetic map $\rho \in \mathbb{R}^{p-1}$,
  partition $\mathcal{G}$ of $\{1, \ldots, p\}$; parameter K.
for $i = 1, \ldots, 2n$ **do**
  Assign references $R(i) = \{\sigma_{i1}, \ldots, \sigma_{iK}\}$ (*Algorithm 1*).
  Initialize $\alpha_k^{(i)} \leftarrow \frac{1}{K}$, for each $k \in \{1, \ldots, K\}$.
Estimate $\lambda = (\lambda_1, \ldots, \lambda_p)$ by EM (*SI Appendix*).
Initialize $\rho \leftarrow 1$.
for $i = 1, \ldots, 2n$ **do**
  Define the HMM $\{R(i), \rho, \lambda\}$.
  Sample $Z^{(i)} = (Z_1^{(i)}, \ldots, Z_p^{(i)})$ from $\mathbb{P}[Z^{(i)} \mid H^{(i)}]$.
  Sample a knockoff copy $\tilde{Z}^{(i)}$ of $Z^{(i)}$ with respect to $\mathcal{G}$.
  Sample $\tilde{H}^{(i)}$ from $\mathbb{P}[H^{(i)} \mid Z^{(i)} = \tilde{Z}^{(i)}]$.
Output: knockoff haplotypes $\tilde{H} \in \{0, 1\}^{2n \times p}$.

Generating Knockoffs Preserving Population Structure. Above, we have described each $H^{(i)}$ as an HMM conditional on the references, $\{\sigma_{i1}, \ldots, \sigma_{iK}\}$. Therefore, it suffices to apply the general algorithm from ref. 30 on each customized HMM to generate knockoffs, conditioning on the individual sets of haplotype motifs. *Algorithm 2* outlines the solution; concretely, $Z^{(i)}$ is sampled from $\mathbb{P}[Z^{(i)} \mid H^{(i)}]$ with step I of algorithm 2 in ref. 30, $\tilde{Z}^{(i)}$ is obtained from step II of the same algorithm, and $\tilde{H}^{(i)}$ from step III.

Knockoffs with Local Reference Motifs Based on Hold-Out Distances. Relatedness is not necessarily homogeneous across the genome. This is particularly evident in the case of admixture, which may cause an individual to share haplotypes with a certain population only in part of a chromosome. Therefore, we extend *Algorithms 1* and *2* to accommodate different local references within the same chromosome.

First, we divide each chromosome into relatively wide (e.g., 10 Mb) genetic windows; then, we choose the references separately within each, based on their similarities outside the window of interest. Similarities are computed looking only at the two neighboring windows, to make the references as locally adaptive as possible. This approach is inspired by SHAPEIT v3 (45), although the latter does not hold out the SNPs in the current window to determine local similarity. Our approach is better suited for knockoff generation because it reduces overfitting—knockoffs too similar to the original variables—and consequently increases power. Having assigned the local references, we apply *Algorithm 2* window by window. To avoid discontinuities at the boundaries, we consider overlapping windows (expanded by 10 Mb on each side). More precisely, we condition on all SNPs within 10 Mb when sampling the latent Markov chain but then we generate knockoffs only within the window of interest.

A Generalized HMM with IBD Segments. We jointly model all haplotypes within an IBD-sharing family (defined in *Modeling Genotypes and Constructing Knockoffs*) $F = \{1, \ldots, m\}$, namely $H^{(F)} = (H^{(i(1))}, \ldots, H^{(m)})$, as an HMM with a K^m-dimensional Markov chain. We write the latter as $Z^{(F)} = (Z^{(1)}, \ldots, Z^{(m)})$, where $Z^{(i)} \in \{1, \ldots, K\}^p$. Conditional on $Z^{(i)}$, each element of $H^{(i)}$ is independent and follows the same emission distribution as in Eqs. **3**– **5**:

$$\mathbb{P}\left[H_j^{(i)} = 1 \mid Z_j^{(i)} = k\right] = f_j^{(i)}(1 \mid k) = \begin{cases} 1 - \lambda, & \text{if } H_j^{(\sigma_{ik})} = 1, \\ \lambda, & \text{if } H_j^{(\sigma_{ik})} = 0. \end{cases} \quad [6]$$

This would reduce to m HMMs as in Eqs. **3**–**5** if the $Z^{(i)}$ s were independent of each other. However, we couple the $Z^{(i)}$ s along the (a priori) fixed IBD segments to account for relatedness.

Define $\partial(i, j) \subset \{1, \ldots, m\}$ as the set of haplotype indexes that share an IBD segment with $H^{(i)}$ at position j, and $\eta_{i,j} = 1/(1 + |\partial(i, j)|) \in (0, 1]$. Then, we model $Z^{(F)}$ as follows: For $1 < j \leq p$,

$$\mathbb{P}\left[Z_j^{(F)} = z_j^{(F)} \mid Z_{j-1}^{(F)} = z_{j-1}^{(F)}\right] = \prod_{i=1}^{m} \left(Q_j^{(i)}(z_j^{(i)} \mid z_{j-1}^{(i)})\right)^{\eta_{i,j}} \prod_{i' \in \partial(i,j)} \mathbb{1}[z_j^{(i)} = z_j^{(i')}], \quad [7]$$

Algorithm 3. Knockoffs preserving population structure and relatedness.

Input: $H \in \{0,1\}^{2n \times p}$, $d \in \mathbb{R}^{p-1}$, $\mathcal{G}$, and K as in *Algorithm 2*, list of IBD segments $\mathcal{I}$.
Define the set $\mathcal{U} \subseteq \{1,\dots,n\}$ of haplotype indexes that do not share any IBD segments in $\mathcal{I}$.
Divide the remaining haplotypes into L distinct families $F_f \subseteq \{1,\dots,n\}$, for $f \in \{1,\dots,L\}$.
for $f \in 1,\dots,L$ **do**
  Assign references $R(f) = \{\sigma_{f1},\dots,\sigma_{fK}\}$ (*SI Appendix*).
  Initialize $\alpha_k^{(f)} \leftarrow \frac{1}{K}$, for each $k \in \{1,\dots,K\}$.
Estimate $\lambda = (\lambda_1,\dots,\lambda_p)$ by EM (*SI Appendix*).
Initialize $\rho \leftarrow 1$.
for $f \in 1,\dots,L$ **do**
  Define the HMM $\{R(f), \rho, \lambda\}$.
  Sample $(Z^{(i)})_{i \in F_f}$ from $\mathbb{P}[(Z^{(i)})_{i \in F_f} \mid (H^{(i)})_{i \in F_f}]$.
  Generate knockoffs $(\tilde{Z}^{(i)})_{i \in F_f}$ of $(Z^{(i)})_{i \in F_f}$ given $\mathcal{G}$.
  Sample $\tilde{H}^{(i)}$ from $\mathbb{P}[H^{(i)} \mid Z^{(i)} = \tilde{Z}^{(i)}]$, for all $i \in F_f$.
for $i \in \mathcal{U}$ **do**
  Generate knockoffs $\tilde{H}^{(i)}$ given $\mathcal{G}$ as in *Algorithm 2*.
Output: knockoff haplotypes $\tilde{H} \in \{0,1\}^{2n \times p}$.

where the transition matrices $Q_j^{(i)}$ are defined as in Eq. **4**, while

$$\mathbb{P}\left[Z_1^{(F)} = (k^{(1)},\dots,k^{(m)})\right] = \prod_{i=1}^{m} \left(\alpha_{k^{(i)}}^{(i)}\right)^{\eta_{i,j}} \prod_{i' \in \partial(i,j)} \mathbb{1}[k^{(i)} = k^{(i')}]. \quad [8]$$

The first term on the right-hand side of Eq. **7** describes the transitions in the Markov chain, while the second term constrains the haplotypes to match along the IBD segments. The purpose of the $\eta_{i,j}$ exponent is to make the marginal distribution of each sequence as consistent as possible with the model for unrelated haplotypes. (If $\eta_{i,j} = 1$, Markov chain transitions may occur with significantly different frequency inside and outside IBD segments.) In the trivial cases of size-one families, $\partial((1,j)) = \emptyset$ and $\eta_{1,j} = 1$, for all $j \in \{1,\dots,p\}$, so Eqs. **7** and **8** reduce to the model in Eq. **4**. In general, the latent states for different haplotypes in the same family will be identical along all IBD segments. See *SI Appendix*, Fig. S18*A* for a graphical representation of this model.

Generating Knockoffs Preserving IBD Segments. The knockoff generation algorithm from ref. 30 would have computational complexity $\mathcal{O}(npK^m)$ for the above model, which is unfeasible for large n unless $m = 1$. (Our model is an HMM with a K^m-dimensional latent Markov chain, where each vector-valued variable corresponds to the alleles at a specific site for all individuals in the family.) Fortunately, the joint distribution of $(Z^{(F)}, H^{(F)})$ can be equivalently seen as a more general Markov random field (80) with $2 \times m \times p$ variables, each taking values in $\{1,\dots,K\}$ or $\{0,1\}$, respectively. See *SI Appendix*, Fig. S18*B* for a graphical representation, where each node corresponds to one of the two haplotypes of an individual at a particular marker. The random field perspective opens the door to more efficient inference and posterior sampling based on message-passing algorithms (81). Leaving the technical details to *SI Appendix*, we outline the procedure in *Algorithm 3*.

In a nutshell, we follow the spirit of *Algorithm 2*, with the important difference that the HMM with a K-dimensional latent Markov chain of length p is replaced by a Markov random field with $2 \times m \times p$ variables; this requires some innovation.

- The K haplotype references in the model for each $H^{(i)}$ are shared by the entire family; see *SI Appendix, Algorithm S1*.
- $Z^{(F)} \mid H^{(F)}$ is sampled with *SI Appendix, Algorithm S2*, which replaces the forward–backward algorithm in refs. 30 and 39 with generalized belief propagation (81). This generally involves some degree of approximation, but it is exact for many family structures.
- $\tilde{Z}^{(F)}$ is generated with *SI Appendix, Algorithm S3*, which is a variation of that from ref. 30, circumventing the computational difficulties of the higher-dimensional model by breaking the couplings between different haplotypes through conditioning (82) upon the extremities of the IBD segments (*SI Appendix*, Fig. S18).

To clarify, conditioning on the extremities of the IBD segments means we make $\tilde{H}$ identical to H for a few sites in each family, which reduces power only slightly (we consider relatively long IBD segments, so there are few extremities), but greatly simplifies the computations (see *SI Appendix* for a full explanation). It is worth remarking that, for trivial size-one families, this is exactly equivalent to *Algorithm 2*. Finally, note that the extension to local references with holdout distances discussed earlier also applies seamlessly here. Our software implements this extension, but we do not explicitly write down the algorithms with local references for lack of space.

Quality Control and Data Preprocessing. We begin with 487,297 genotyped and phased subjects in the UK Biobank (application 27837). Among these, 147,669 have reported at least one close relative. We define families by clustering individuals with kinship greater than or equal to 0.05; then we discard 322 individuals (those with the most missing phenotypes) to ensure that no families have size larger than 10. The choice of a 0.05 kinship cutoff is motivated by the observation that values would result in larger families, significantly increasing the computational cost of generating knockoffs without providing any clear benefits in terms of type-I error control, given that our choice already accounts for the close relatedness of most concern. This leaves 136,818 related individuals divided into 57,164 families. The median family size is 2, and the mean is 2.4. The total number of individuals passing quality control (including those without relatives) is 486,975. We analyze only biallelic SNPs with minor allele frequency above 0.1% and in Hardy–Weinberg equilibrium (10^{-6}), among the subset of 350,119 unrelated British individuals analyzed in ref. 30. The final SNPs count is 591,513.

Data Availability Previously published data were used for this work. (See main text for the various genetic data repositories that we access.)

ACKNOWLEDGMENTS. M.S. and S.B. were advised by E.C. at Stanford University. M.S., S.B., E.C. and C.S. were supported by NSF Grant DMS 1712800. S.B. was also supported by a Ric Weiland fellowship. E.C. and C.S. were also supported by NSF Grant OAC 1934578 and by a Math+X grant (Simons Foundation). We thank Kevin Sharp (University of Oxford) for sharing computer code. We acknowledge the participants and investigators of the UK Biobank, the FinnGen, and the Japan Biobank projects.

1. N. Risch, K. Merikangas, The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
3. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. C. Sabatti, "Multivariate linear models for GWAS" in, *Advances in Statistical Bioinformatics*, K.-A. Do, Z. S. Qin, M. Vannucci, Eds. (Cambridge University Press, 2013), pp. 188–207.
5. L. Buzdugan *et al.*, Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* **32**, 1990–2000 (2016).
6. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).
7. C. Sabatti, S. Service, N. Freimer, False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 829–833 (2003).
8. R. A. Fisher, The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **53**, 399–433 (1918).
9. C. J. Hoggart, J. C. Whittaker, M. De Iorio, D. J. Balding, Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
10. T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, K. Lange, Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721 (2009).
11. M. Stephens, D. J. Balding, Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
12. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
13. J. Taylor, R. J. Tibshirani, Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7629–7634 (2015).
14. N. Meinshausen, P. Bühlmann, Stability selection. *J. R. Stat. Soc. B* **72**, 417–473 (2010).
15. J. Wu, B. Devlin, S. Ringquist, M. Trucco, K. Roeder, Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genet. Epidemiol.* **34**, 275–285 (2010).

16. B. A. Logsdon, G. E. Hoffman, J. G. Mezey, A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11**, 58 (2010).
17. Y. Guan, M. Stephens, Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
18. B. Devlin, N. Risch, A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
19. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
20. D. Brzyski *et al*., Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75 (2017).
21. A. L. Price *et al*., Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
22. H. M. Kang *et al*., Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
23. Z. Zhang *et al*., Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
24. J. Listgarten *et al*., Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
25. P. R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, A. L. Price, Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
26. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
27. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
28. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
29. F. Dudbridge, Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
30. M. Sesia, E. Katsevich, S. Bates, E. Candès, C. Sabatti, Multi-resolution localization of causal variants across the genome. *Nat. Commun.* **11**, 1093 (2020).
31. M. D. Gallagher, A. S. Chen-Plotkin, The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
32. D. J. Hunter, J. M. Drazen, Has the genome granted our wish yet? *N. Engl. J. Med.* **380**, 2391–2393 (2019).
33. V. Tam *et al*., Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
34. A. B. Popejoy *et al*.; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG), The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).
35. L. Duncan *et al*., Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
36. R. F. Barber, E. Candès, Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).
37. E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. B* **80**, 551–577 (2018).
38. S. Bates, M. Sesia, C. Sabatti, E. Candès, Causal inference in genetic trio studies. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24117–24126 (2020).
39. M. Sesia, C. Sabatti, E. J. Candès, Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18 (2019).
40. E. Katsevich, C. Sabatti, Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* **13**, 1–33 (2019).
41. J. Marchini, B. Howie, Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
42. A. Manichaikul *et al*., Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
43. O. Delaneau, J. Marchini, J. F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
44. O. Delaneau, J. F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
45. J. O'Connell *et al*., Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
46. P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
47. M. Sesia, "New methods for variable importance testing with applications to genetic studies," Ph.D. thesis, Stanford University, Stanford, CA (2020).
48. N. A. Rosenberg *et al*., Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
49. J. A. Sved, Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2**, 125–141 (1971).
50. E. A. Thompson, Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).
51. A. Kong *et al*., Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
52. A. Gusev *et al*., Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
53. B. L. Browning, S. R. Browning, A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
54. D. W. Bjelland, U. Lingala, P. S. Patel, M. Jones, M. C. Keller, A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. *Eur. J. Hum. Genet.* **25**, 617–624 (2017).
55. A. Naseri, X. Liu, K. Tang, S. Zhang, D. Zhi, RaPID: Ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biol.* **20**, 143 (2019).
56. B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, 2010).
57. J. D. Storey *et al*., The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
58. F. Privé, H. Aschard, A. Ziyatdinov, M. G. B. Blum, Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
59. F. Privé, H. Aschard, M. G. B. Blum, Efficient implementation of penalized regression for genetic risk prediction. *Genetics* **212**, 65–74 (2019).
60. E. Katsevich, C. Sabatti, M. Bogomolov, Filtering the rejection set while preserving false discovery rate control. *J. Am. Stat. Assoc.* **0**, 1–27 (2021).
61. D. M. Altshuler *et al*.; International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
62. W. Zhou *et al*., Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
63. Biobank Japan, Biobank Japan Project (2020). http://jenger.riken.jp/en/. Accessed 30 June 2020.
64. FinnGen, FinnGen documentation of r3 release (2020). https://finngen.gitbook.io/documentation/. Accessed 30 June 2020.
65. T. J. Hoffmann *et al*., Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* **49**, 54–64 (2017).
66. A. E. Locke *et al*.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; International Endogene Consortium, Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
67. L. A. Hindorff *et al*., Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
68. H Jónsson *et al*., Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
69. A. B. Popejoy, S. M. Fullerton, Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
70. G. Sirugo, S. M. Williams, S. A. Tishkoff, The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
71. M. B. Lanktree *et al*.; Hugh Watkins on behalf of PROCARDIS; Meena Kumari on behalf of the Whitehall II Study and the WHII 50K Group, Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height. *Am. J. Hum. Genet.* **88**, 6–18 (2011).
72. Y. R. Li, B. J. Keating, Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).
73. B. D. Bitarello, I. Mathieson, Polygenic scores for height in admixed populations. *G3 (Bethesda)* **10**, 4027–4036 (2020).
74. T. B. Cavazos, J. S. Witte, Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *HGG Adv.* **2**, 100017 (2021).
75. J. M. Gaziano *et al*., Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
76. S. Li, M. Sesia, Y. Romano, E. Candès, C. Sabatti, Searching for consistent associations with a multi-environment knockoff filter (2021).
77. S. Sariya *et al*., Rare variants imputation in admixed populations: Comparison across reference panels and bioinformatics tools. *Front. Genet.* **10**, 239 (2019).
78. G. Band, J. Marchini, BGEN: A binary file format for imputed genotype and haplotype data. arXiv [Preprint] (2018). https://doi.org/10.1101/308296. (Accessed 2 May 2018).
79. N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
80. R. Kinderman, S. Snell, *Markov Random Fields and Their Applications* (American Mathematical Society, Providence, RI, 1980).
81. J. Yedidia, W. Freeman, Y. Weiss, "Understanding belief propagation and its generalizations" in *Exploring Artificial Intelligence in the New Millennium*, G. Lakemeyer, B. Nebel, Eds. (Morgan Kaufmann Publishers Inc., San Francisco, CA, 2003), vol. 8, pp. 239–269.
82. S. Bates, E. Candès, L. Janson, W. Wang, Metropolized knockoff sampling. *J. Am. Stat. Assoc.* **0**, 1–15 (2020).