Controlling the false discovery rate in GWAS with population structure

Matteo Sesia

Department of Data Sciences and Operations, University of Southern California

Stephen Bates

Department of Statistics, Stanford University

Emmanuel Candès

Departments of Statistics and of Mathematics, Stanford University

Jonathan Marchini

Genetics Center, Regeneron Pharmaceuticals

Chiara Sabatti

Departments of Statistics and of Biomedical Data Sciences, Stanford University

Abstract

This paper proposes a novel statistical method to address population structure in genome-wide association studies while controlling the false discovery rate, which overcomes some limitations of existing approaches. Our solution accounts for linkage disequilibrium and diverse ancestries by combining conditional testing via knockoffs with hidden Markov models from state-of-the-art phasing methods. Furthermore, we account for familial relatedness by describing the joint distribution of haplotypes sharing long identical-by-descent segments with a generalized hidden Markov model. Extensive simulations affirm the validity of this method, while applications to UK Biobank phenotypes yield many more discoveries compared to BOLT-LMM, most of which are confirmed by the Japan Biobank and FinnGen data.

INTRODUCTION

Genome-wide association studies (GWAS) measure hundreds of thousands of single-nucleotide polymorphisms (SNP) in thousands of individuals to identify variants affecting a phenotype of interest. The objective is to reliably determine which genotype-phenotype associations are likely to be important and which are spurious; several challenges make this difficult, particularly for polygenic traits which may be influenced by thousands of variants. First, spurious associations may arise from multiple comparisons; i.e., strong correlations will occur by chance as numerous variables are tested simultaneously.[1] The typical solution is to apply a stringent significance threshold designed to control the family-wise error rate (FWER)—the probability of a single false positive. However, this is too conservative for polygenic phenotypes because numerous discoveries are expected.[2–5] Indeed, SNPs with effect sizes large enough to pass the FWER threshold do not fully explain the heritability of complex traits.[6] A second source of spurious associations is linkage disequilibrium (LD): the stochastic dependence of alleles on the same chromosome,[7,8] which tends to be stronger for those physically closer to each other. Thus, SNPs with no effect on the phenotype may be *marginally* associated with it simply because they are in LD with a causal variant.[5,9] Standard GWAS methods cannot account for LD,[5] so their findings do not precisely localize causal variants and may be difficult to recognize as distinct when multiple nearby loci are discovered.[9] Such ambiguity may help explain why large studies are reporting associations densely across the genome.[10] Lastly, population structure (heterogeneous degrees of similarity between different individuals, due to diverse ancestries or familial relatedness) may also analogously lead to spurious associations and has long been of concern in genetic analyses.[11–13] Population structure not only induces dependence between distant loci, it can even create spurious associations when there are no causal variants at all. For example, if two populations differ in the distribution of the trait solely due to their environments, then any SNP whose allele frequency varies across populations will be associated with the trait. Hence, several methods were developed to account for population structure. An early remedy built on principal component analysis (PCA),[14] although linear mixed models (LMM) have subsequently become predominant.[15–17] While these corrections mitigate the impact of population structure, they are limited to *marginal* testing (i.e., they ignore LD) with FWER control.

KnockoffZoom[9] was recently proposed to address the limitations of marginal testing and FWER control. This method accounts for LD through *conditional* (on nearby loci) rather than marginal testing, so that its discoveries are clearly distinct and indicative of the presence of causal variants.[5,9] The inferences are valid for both quantitative (e.g., body measurement) and qualitative (e.g., disease status) phenotypes, regardless of their genetic architecture, in contrast to traditional methods, which rely on linear models whose correctness may be difficult to verify. Practically, KnockoffZoom partitions the genome into disjoint blocks and tests a conditional association hypothesis for each of them, which, if rejected, suggests the presence of causal effects.[9] While this requires testing the importance of fixed groups of SNPs (determined at will, but before looking at the phenotype), it can be easily applied at increasing levels of resolution, with finer and finer genome partitions, in order to localize causal variants as precisely as possible. KnockoffZoom does not analyze imputed variants[18] because, without additional assumptions, it is theoretically impossible to test conditional associations beyond the resolution of the SNP array (imputation cannot introduce additional information about the phenotype compared to that contained in the typed SNPs).[9] Nonetheless, its findings achieve resolution comparable to that of fine-mapping methods[19–21] applied to typed variants.[9] Furthermore, KnockoffZoom is more powerful than LMM approaches because it controls the false discovery rate[22] (FDR)—the expected proportion of false discoveries—instead of the FWER. The main limitation of KnockoffZoom is that, as originally implemented, it is only applicable to relatively homogeneous samples because it does not fully account for population structure. To address this missing component, we now present an extension that accounts for population structure while retaining the advantages of conditional testing and FDR control.

Before presenting our results, we review the mechanics of KnockoffZoom. The method establishes statistical significance through careful data augmentation: it constructs imperfect copies (*knockoffs*)[5,23] of the genotypes for each individual that are in LD with the real variants and also have the same allele frequencies, in such a way that replacing a group of genotypes with the corresponding knockoffs would keep the modified data set statistically indistinguishable from the original one, except possibly for some reduced association with the phenotype. Knockoffs serve as negative controls:[23,24] they are *blindly* analyzed jointly with the genotypes (the algorithm ignores which variables are knockoffs until the very end), through any procedure of choice (e.g., an LMM, a sparse regression model, or any machine learning tool). The significance of each genetic segment is determined by contrasting the estimated importance of its genotypes to that of the corresponding knockoffs. This is a fair comparison because knockoffs behave as the real non-causal variants. In order to define precisely, and achieve practically, this exchangeability, the distribution of genotypes is approximated with a hidden Markov model (HMM).[5] This assumption is well grounded: HMMs have already been widely and successfully employed to describe LD,[8] for the purposes either of ancestry inference,[25,26] of identifying population-specific haplotype blocks,[27] or of phasing and imputation.[28–32]

However, KnockoffZoom has so far relied practically on the fastPHASE[29] HMM, which is not designed to account for population structure. In fact, this model cannot simultaneously describe association between different loci due to *mixture*[33] (individuals with different ancestries), *admixture*[34] (individuals of mixed ancestry), and *linkage* (allele dependencies between nearby loci)—we demonstrate this limitation empirically in Supplementary Notes A–C and Supplementary Figures 1–13. Therefore, KnockoffZoom was previously applied only to homogeneous and unrelated samples.[5,9] In this paper, we build upon the more flexible SHAPEIT[35–37] HMM and expand the applicability of KnockoffZoom to populations with diverse and possibly admixed ancestries. We also further extend this method to account for familial relatedness, by jointly describing the distribution of haplotypes from multiple individuals sharing long *identical-by-descent* (IBD) genetic segments,[38] and then developing a corresponding construction of knockoffs. Crucially, our method is applicable even if the population structure and the familial relatedness are cryptic, and remains completely model-free regarding the genetic architecture of the trait.

After testing our method through careful simulations with real genotypes and simulated phenotypes, we shall apply it to study height, body mass index, platelet count, systolic blood pressure, cardiovascular disease, respiratory disease, hypothyroidism, and diabetes in the UK Biobank data set,[39] including virtually all samples therein. We will show this yields more numerous (between 25%, for height, and 320%, for cardiovascular disease) distinct discoveries compared to BOLT-LMM.[40] Furthermore, we will verify that most additional discoveries (between 37.3%, for platelet count, and 88.5%, for diabetes) are validated by the GWAS Catalog,[52] the Japan Biobank Project[41], or the FinnGen resource,[42] while many of the remaining ones have known associations to related traits. Finally, we highlight a novel discovery for cardiovascular disease, as one of many promising findings that may be worthy of further validation. Our full results, which involve thousands of discoveries, are available from `https://msesia.github.io/knockoffzoom-v2/`, along with an efficient software implementation of our method.

RESULTS

Knockoffs preserving population structure

We assume all haplotypes have been phased[9] and we approximate their distribution with an HMM similar to that of SHAPEIT.[35–37] This model describes each haplotype sequence as a mosaic of K reference motifs corresponding to the haplotypes of other individuals in the data set, where

K is fixed (e.g., $K = 100$); critically, different haplotypes may use different sets of motifs. The references are chosen based on haplotype similarity; see Methods. The idea is that the ancestry of an individual should be approximately reflected by the choice of references; e.g., the haplotypes of someone from England should be well-approximated by a mosaic of haplotypes primarily belonging to other English individuals. 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: relatively rare mutations can independently occur at any site. This model effectively accounts for population structure as well as LD in the context of phasing,[35–37] and our simulations will demonstrate its usefulness for conditional testing.

Having defined an HMM for each haplotype sequence, knockoffs are generated by repurposing the algorithm in KnockoffZoom v1,[9] which was originally based on the fastPHASE model, but is sufficiently general to apply here. Our software implementation takes as input phased haplotypes in standard binary format, builds the data-adaptive SHAPEIT HMM, and returns as a knockoff-augmented data set that can be conveniently used by KnockoffZoom for conditional testing at the desired resolution. The knockoff generation procedure is explained in the Methods.

Knockoffs preserving familial relatedness

The above model is not directly applicable to closely related samples because it describes them independently (conditional on the reference motifs), whereas these share long IBD segments.[38,43] Recall that knockoffs must follow the genotype distribution; therefore, processing related samples independently would yield knockoffs breaking the relatedness structure, invalidating our inferences (see Methods for a full explanation). We fix this by jointly modeling haplotypes in the same family.

We begin by detecting long IBD segments in the data.[44–47] If the pedigree is known in advance, one can restrict the IBD search within the given families; otherwise, there exists efficient software to approximately reconstruct families from data at the UK Biobank scale.[48] The results thus obtained define a relatedness graph, where two haplotype sequences are connected if they share an IBD segment; we refer to the connected components of this graph as the *IBD-sharing families*.

Conditional on the location of the IBD segments, we define a larger HMM jointly describing the distribution of all haplotypes in each IBD-sharing family. Marginally, each haplotype is modeled by the SHAPEIT HMM; however, different haplotypes are coupled along the IBD segments (and we avoid using haplotypes in the same family as references for one another). This model is explicitly described in the Methods, where we explain how to generate knockoffs for it. The details are technical, since the algorithm in Sesia et al.[9] is no longer feasible due to the much larger state space of this new HMM. We shall demonstrate empirically that our method generates knockoffs sharing the same IBD segments as the real haplotypes, and also simultaneously preserves LD.

Numerical experiments with genetic data

Setup

We test the proposed methods via simulations based on subsets of phased haplotypes from the UK Biobank data, chosen as to have strong population structure either due to diverse ancestries or familial relatedness. After some pre-processing (see Methods), we partition each of the 22 autosomes into contiguous groups of SNPs at 7 different levels of resolution, ranging from that of single SNPs to that of 425 kb-wide groups; see Supplementary Table 1. These partitions are obtained by applying complete-linkage hierarchical clustering to the SNPs (genetic distances are used as similarity measures) and cutting the resulting dendrogram at different heights. For each such partition, we generate knockoffs with two alternative methods. In one case, we fit fastPHASE using $K = 50$ latent HMM motifs, even though this is not designed for data with population structure, and then apply KnockoffZoom v1.[9] In the other case, we apply our new method to generate knockoffs, and then we proceed from there to select important groups of SNPs as in KnockoffZoom v1; we refer to this approach as KnockoffZoom v2.

Knockoffs preserving population structure

We generate knockoffs for 10,000 unrelated individuals with one of 6 different self-reported ancestries (Supplementary Table 2). We will perform a diagnostic check of these knockoffs by verifying their exchangeability with the genotypes through a PCA and a covariance analysis; then, we will assess the performance of KnockoffZoom for simulated phenotypes.

A property of valid knockoffs is that their distribution is the same as that of the genotypes.[5,23] (This follows from the stronger exchangeability requirement; see Methods). Thus, the proportion of knockoff variance explained by their top principal components should be close to the corresponding quantity computed on the genotypes. The PCA in Supplementary Figures 14–15 demonstrates that our knockoffs preserve population structure quite accurately in this sense, unlike those based on the fastPHASE HMM. This holds even at low resolution, where the sensitivity to model misspecification is highest.[9] The covariance analysis[9] in Supplementary Figures 16–22 confirms that our knockoffs are approximately exchangeable with the genotypes and should have power comparable to that of knockoffs based on the fastPHASE HMM.

Next, we simulate continuous phenotypes conditional on the true genotypes, from a homoscedastic linear model with 500 causal variants distributed uniformly across the genome; the total heritability is varied as a control parameter.[9] We apply KnockoffZoom v1 and v2 on these data, using lasso-based statistics.[49] Supplementary Figure 23 shows the histogram of test statistics, which should be symmetric around zero for null groups (i.e., those without causal variants). The statistics obtained with the new knockoffs satisfy this property, while the fastPHASE model leads to a rightward bias, which may result in an excess of false positives. By contrast, both methods yield similarly distributed statistics for causal groups. The power[9] and FDR are compared in Figure 1: KnockoffZoom v2 has slightly lower power, but always controls the FDR.

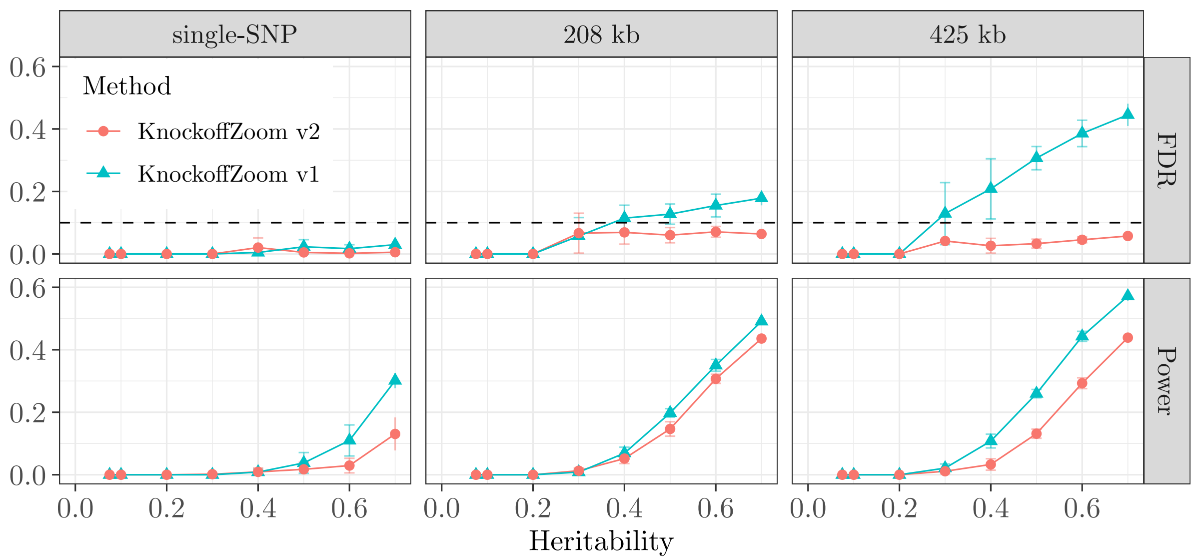

FIG. 1. **Power and FDR in simulations involving samples with diverse ancestries**. KnockoffZoom performance at different resolutions on artificial phenotypes and real genotypes of 10,000 samples with very diverse ancestries, using alternative knockoff constructions. The nominal FDR is 10%. The results are averaged over 10 experiments with independent phenotypes; the vertical bars indicate standard errors.

Supplementary Figure 24 presents analogous results in simulations with sparser signals, while Supplementary Figure 25 summarizes findings at different resolutions by counting only the most specific ones.[9] Supplementary Figure 26 reports on simulations where SNP importance is estimated by BOLT-LMM instead of the lasso;[50] the LMM is less powerful, but it makes fastPHASE knockoffs even more susceptible to population structure, while our new method remains valid.

Knockoffs preserving familial relatedness

We test our method on 10,000 British individuals in 4,900 self-reported families; see Supplementary Table 3 and Supplementary Figure 27 for details. We use RaPID[48] to detect IBD segments wider than 3 cM, chromosome-by-chromosome, adopting the recommended parameters. After discarding, for simplicity, segments shared by individuals who do not belong to the same self-reported family, we are left with 723,454 of them. Their mean width is 19.6 Mb, or 26.1 cM, and each contains 4238 SNPs on average (Supplementary Figure 28). We then generate knockoffs preserving these IBD segments, and compare the results with those obtained disregarding relatedness.

Supplementary Figure 29 shows that knockoffs would not preserve IBD segments if we did not explicitly enforce such constraint, especially at low resolution. The diagnostics in Supplementary Figure 30 confirm that our method correctly preserves LD, and Supplementary Figure 31 demonstrates that accounting for relatedness does not decrease power; to the contrary, it can increase it by ensuring that closely related haplotypes are not used as references for one another, which would reduce the desired contrast between genotypes and knockoffs.

We simulate binary phenotypes from a liability threshold (probit) model with 100 uniformly distributed causal variants; the numbers of cases and controls are balanced. (We consider binary phenotypes, as opposed to continuous phenotypes as in the previous section, simply to highlight the flexibility of our method, which is equally valid regardless of the distribution of the trait). We include in this model an additive random term for each family, mimicking shared environmental effects, whose strength is smoothly controlled by a parameter $\gamma \in [0, 1]$ (Methods). The phenotypes of different individuals in the same family are conditionally independent given the genotypes if $\gamma = 0$, while identical twins will always have the same phenotype if $\gamma = 1$. In theory, environmental effects may introduce spurious associations, unless the knockoffs account for familial relatedness; this point is demonstrated empirically below and explained rigorously in the Methods.

Figure 2 reports FDR and power at low-resolution, with and without preserving relatedness. This shows that preserving IBD segments enables FDR control even with extreme environmental factors ($\gamma = 1$), with virtually no power loss. However, KnockoffZoom v2 is reasonably robust even if relatedness is ignored, especially at higher resolution (Supplementary Figure 32). This partly depends on the multivariate importance statistics used here (i.e., sparse logistic regression); in fact, marginal statistics are more vulnerable to confounding, as illustrated in Supplementary Figure 33. Finally, Supplementary Figures 34–35 confirm that the test statistics for null groups of SNPs are symmetrically distributed if relatedness is preserved.

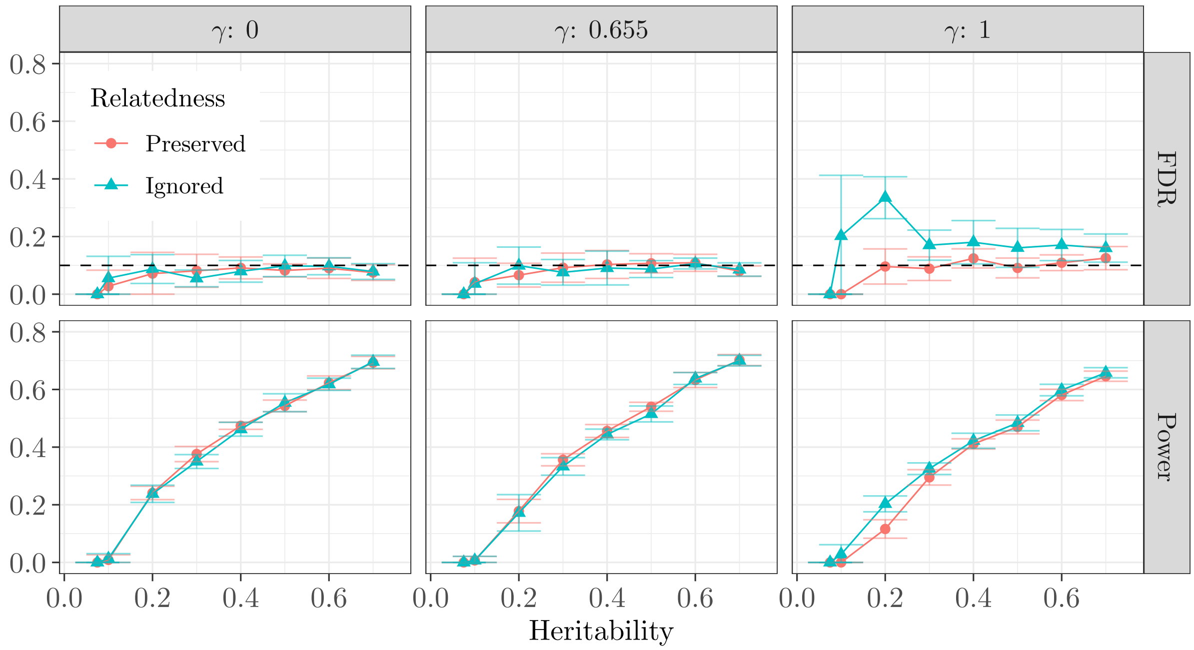

FIG. 2. **Power and FDR in simulations with familial relatedness**. KnockoffZoom v2 performance on artificial phenotypes and real genotypes of 10,000 related samples. Our method is applied with and without preserving IBD segments. Results for phenotypes with different strengths of environmental effects γ are in separate columns ($\gamma = 0$: no environmental effects, $\gamma = 1$: strongest environmental effects; see Methods for more information about γ). Knockoff resolution equal to 425 kb. Other details are as in Figure 1.

.

Analysis of the UK Biobank phenotypes

Setup

We analyze 486,975 individuals both genotyped and phased in the UK Biobank, 136,818 of which have close relatives; see Methods for details on quality control. Most samples are British (429,934), Irish (12,702), or other Europeans (16,292). By running RaPID[48] within the 57,164 families, as in the simulations, we detect 7,087,643 long IBD segments across the 22 autosomes. We then apply KnockoffZoom v2 at different resolutions (as in the simulations), aiming to control the FDR below 10%, using knockoffs preserving both population structure and familial relatedness.

Discoveries with different subsets of individuals

We study 4 continuous traits (height, body mass index, platelet count, systolic blood pressure) and 4 diseases (cardiovascular disease, respiratory disease, hyperthyroidism, diabetes), using different subsets of samples to compare performance. The phenotypes are defined in Supplementary Table 4. In order to increase power, we include in the KnockoffZoom predictive model, along with the genotypes and the knockoffs, a few relevant covariates that help explain some variation in the phenotype,[9] as explained in the Methods. These covariates include the top principal components of the genetic matrix,[14] although it is worth emphasizing that the validity of our method does not depend on them, since we account for population structure through the knockoffs. The numbers of low-resolution (208 kb) discoveries are in Figure 3; this demonstrates that including relatives yields many more discoveries, while including different ancestries tends to have a smaller impact (unsurprisingly, since we have relatively few non-British samples). Table I summarizes the gains in the numbers of discoveries at different resolutions allowed by KnockoffZoom v2, either by leveraging related samples, or by including non-British individuals. This shows an increase in power, except at the single-SNP resolution; this exception may be partially explained by the fact that single-SNP discoveries are fewer and thus more affected by the random variability in the knockoffs.[9] Supplementary Tables 5–6 report the numbers of discoveries at other resolutions, as well as those obtained from the analysis of European non-British samples only. The full results are available at `https://msesia.github.io/knockoffzoom-v2/`, along with an interactive visualization tool.

Table II demonstrates that KnockoffZoom v2 is much more powerful than BOLT-LMM, when the latter is applied to the 459k European samples;[40] in fact, we discover almost all findings reported by BOLT-LMM and many new ones. (See Supplementary Tables 7–8 for more detailed results at other levels of resolution.) This is consistent with KnockoffZoom v1,[9] although our method is even more powerful (except for the single-SNP resolution), as shown in Supplementary Tables 9–10.

Validation of novel discoveries

We begin to validate our findings by comparing them with those in the GWAS Catalog,[52] in the Japan Biobank Project,[41] and in the FinnGen resource[42] (we use the standard 5×10^{-8} threshold for the p-values reported by the latter two). For simplicity, hereafter we focus on our findings obtained including both related and non-British individuals. Supplementary Tables 11–12 show that most of our high-resolution discoveries correspond to SNPs previously known to be

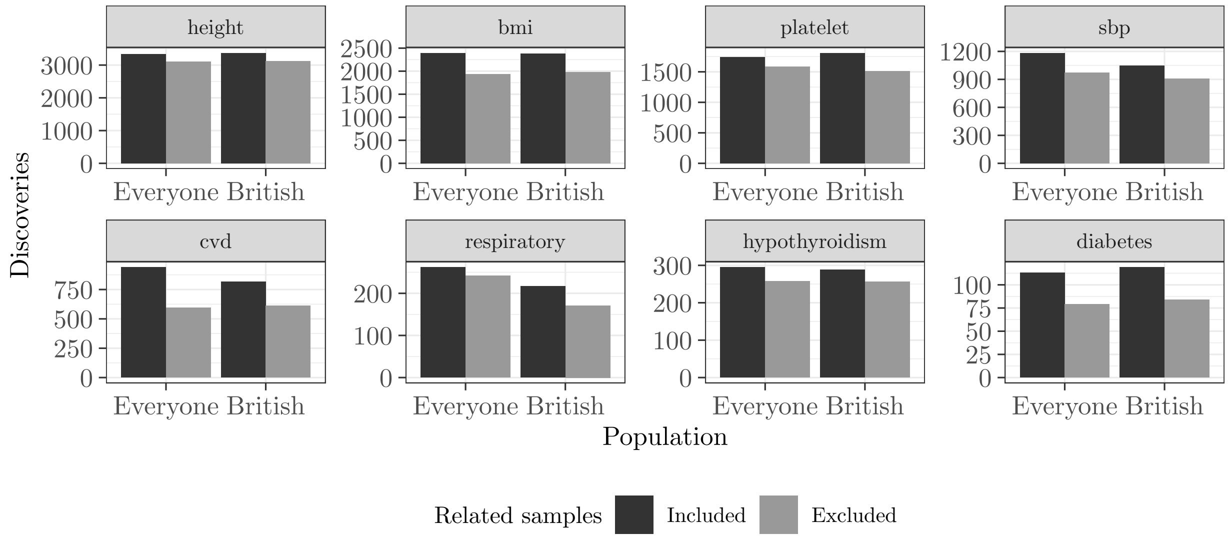

FIG. 3. **Numbers of KnockoffZoom discoveries for UK Biobank phenotypes**. Low-resolution (208 kb) discoveries using data from subsets of individuals with different self-reported ancestries.

	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	2814–3527	25.3	2808–3355	19.5	3355–3527	5.1	2808–2814	0.2
41 kb	4419–5867	32.8	4353–5354	23.0	5354–5867	9.6	4353–4419	1.5
81 kb	6784–8031	18.4	6676–7781	16.6	7781–8031	3.2	6676–6784	1.6
208 kb	8776–10270	17.0	8635–10049	16.4	10049–10270	2.2	8635–8776	1.6
425 kb	9401–10730	14.1	9028–10297	14.1	10297–10730	4.2	9028–9401	4.1
Sample size	408k–487k	19.4	356k–430k	20.8	430k–487k	13.0	356k–408k	14.6

TABLE I. **Effect of sample size increases on numbers of KnockoffZoom v2 discoveries**. Cumulative numbers of discoveries for all UK Biobank phenotypes at different resolutions, with different subsets of the samples. For example, including related individuals increases by 16.4% the number of discoveries obtained from the British samples at the 208 kb resolution (from 8635 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 10049 to 10,270).

	KnockoffZoom v2		BOLT-LMM	
Phenotype	Discoveries	Overlap with LMM	Discoveries	Overlap with KZ
bmi	2395	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	3339	2228 (66.7%)	2464	2430 (98.6%)
hypothyroidism	295	129 (43.7%)	143	142 (99.3%)
platelet	1743	1057 (60.6%)	1204	1183 (98.3%)
respiratory	262	82 (31.3%)	94	92 (97.9%)
sbp	1183	561 (47.4%)	568	530 (93.3%)

TABLE II. **Comparison of KnockoffZoom v2 at low resolution with BOLT-LMM**. KnockoffZoom discoveries (208 kb resolution, 10% FDR) using all 487k UK Biobank samples for different phenotypes, and corresponding BOLT-LMM genome-wide significant discoveries (5×10^8); the latter is applied on 459k European samples[40] for all phenotypes except diabetes and respiratory disease, for which it is applied on 350k unrelated British samples,[9] for the sake of consistency in phenotype definitions. For example, we report 940 distinct discoveries for cardiovascular disease, 274 of which contain significant associations according to BOLT-LMM. The latter reports a total of 257 discoveries (clumped[9] with the standard PLINK[51] algorithm) for this phenotype, 96.9% of which overlap with one of our discoveries.

associated with the phenotype of interest. This is particularly true for those findings that are also detected by BOLT-LMM, although many of our additional discoveries are also confirmed; see Supplementary Tables 13–14. For example, our method reports 1089 findings for cardiovascular disease at the 425 kb resolution, only 255 of which can be detected by BOLT-LMM; however, 85.6% of our additional 834 discoveries are confirmed in at least one of the aforementioned resources. Furthermore, Supplementary Table 15 suggests that most relevant associations in the Catalog (above 70%) are confirmed by our findings, which is again indicative of high power. The relative power of our method (i.e., the proportion of previously reported associations that we discover) seems to be above 90% for quantitative traits, but lower than 50% for all diseases except hypothyroidism, probably due to the relatively small number of cases in the UK Biobank data set compared to more targeted case-control studies.

The 5×10^{-8} genome-wide threshold for the Japan Biobank Project and the FinnGen resource is overly conservative given that our goal is to confirm selected discoveries. Therefore, we next utilize these independent summary statistics for an enrichment analysis. The idea is to compare the distribution of the external statistics corresponding to our selected loci to that of loci from the rest

	Total			Not found by BOLT-LMM		
		Confirmed			Confirmed	
Phenotype	Discoveries	Other	Other or Enrich.	Discoveries	Other	Other or Enrich.
bmi	2395	1076 (44.9%)	1620 (67.6%)	1497	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	3339	1886 (56.5%)	2493 (74.7%)	1111	164 (14.8%)	556 (50.0%)
hypothyroidism	295	156 (52.9%)	226 (76.6%)	166	43 (25.9%)	101 (60.8%)
platelet	1743	453 (26.0%)	1017 (58.3%)	686	29 (4.2%)	256 (37.3%)
respiratory	262	241 (92.0%)	NA	180	159 (88.3%)	NA
sbp	1183	643 (54.4%)	885 (74.8%)	622	154 (24.8%)	358 (57.6%)

TABLE III. **Validation of findings through comparisons with other studies and enrichment**. 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. 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).

of the genome, as explained in Supplementary Note D. This approach can estimate the number of replicated discoveries but it has the limitation that it cannot tell exactly which ones are confirmed; therefore, we will consider alternative validation methods later. (A more precise analysis is possible here in theory, but has low power; see Supplementary Note D). Supplementary Tables 16–17 show that many additional discoveries can thus be validated, especially at high resolution. (See Supplementary Tables 18–19 for more details about enrichment.) Table III summarizes these confirmatory results. Respiratory disease is excluded from the enrichment analysis because the FinnGen resource divides it among several fields, so it is unclear how to best obtain a single p-value. In any case, the GWAS Catalog and the FinnGen resource already directly validate 90% of our new findings for this phenotype.

We continue the validation by inspecting the novel discoveries (i.e., those missed by BOLT-LMM and unconfirmed by the above studies) and cross-referencing them with the genetics literature, focusing for simplicity on the 20 kb resolution. Supplementary Table 20 shows that almost all discoveries contain genes, and most have known associations to phenotypes closely related to that of interest (Supplementary Table 21). Furthermore, most lead SNPs (those with the largest importance measure in each group[9]) have functional annotations (Supplementary Table 22).

Figure 4 showcases one of our novel discoveries for cardiovascular disease. The finest finding here spans 4 genes, but we could not find previously reported associations with cardiovascular disease within this locus. However, one of these genes (SH3TC2) is known to be associated with blood pressure,[53] while another (ABLIM3) is known to be associated with body mass index.[54]

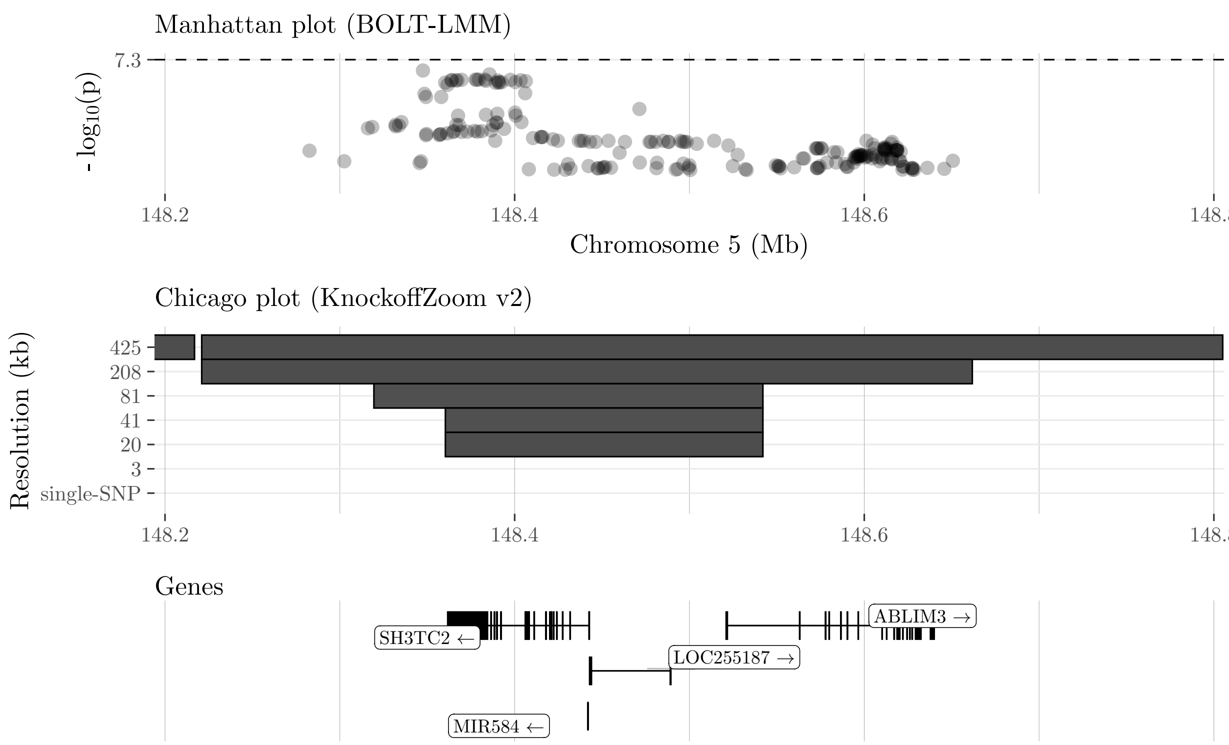

FIG. 4. **Novel discovery for cardiovascular disease on the UK Biobank data**. Center: the shaded rectangles indicate the genetic segments detected by our method at different resolutions. Bottom: genes in the locus spanned by our finest discovery. Top: BOLT-LMM marginal p-values computed on UK Biobank samples with European ancestry,[40] for genotyped and imputed variants within this locus. All BOLT-LMM p-values within this locus are larger than 5×10^{-8}; those larger than 10^{-3} are hidden. Note that KnockoffZoom does not analyze imputed variants, since these do not carry any additional information about the phenotype.[9]

DISCUSSION

This paper has developed a new algorithm for constructing genetic knockoffs, inspired by the SHAPEIT phasing method, that extends the applicability of KnockoffZoom to the analysis of data with population structure. In particular, we can include related and ethnically diverse individuals while controlling the FDR. This extension is crucial for several reasons. Firstly, very large studies are sampling entire populations quite densely,[42,55] which yields many closely related samples. It would be wasteful to discard this information, and potentially dangerous not to carefully account for relatedness. Secondly, the historical lack of diversity in GWAS (which mostly involve individuals of European ancestry) is a well-recognized problem,[56,57] which biases our scientific knowledge and disadvantages the health of under-represented populations. While this issue goes beyond the statistical difficulty of analyzing diverse GWAS data, our work should at least help remove a technical barrier. Firstly, by allowing the simultaneous analysis of diverse populations, we can borrow strength from one another and increase power, as the different patterns of LD in different populations may uncover causal variants more effectively.[58] Secondly, discoveries obtained from the analysis of diverse data may improve our ability to explain phenotypic variation in the minority populations.[59,60] Since the UK Biobank mostly comprises of British individuals, the increase in power resulting from the analysis of diverse samples can only be relatively small. Nonetheless, we observe some gains when we include non-British individuals. This is promising, especially since simulations demonstrate that our inferences are valid even when the population is extremely heterogeneous. Therefore, in the near future, we would like to apply our method to more diverse data sets, such as that collected by the Million Veteran Program,[61] for example.

Our method accounts for LD, population structure, and cryptic relatedness, which are the major sources of confounding in GWAS data, so our discoveries are directly interpretable and may be portrayed in a causal sense relatively safely;[5,9] indeed, a closely related approach yields formal causal claims in the special case of parent-child trio data.[62] Furthermore, our inferences require no assumptions about the genetic architecture of the phenotype, which makes KnockoffZoom very versatile. It is worth stressing that our simulations are based on genetic data, do not provide any additional information to our method about the relatedness or population structure other than that already available to the analyst, and do not exploit any knowledge of the architecture of the simulated traits. Therefore, the results are informative about the general validity of our method.[5,9]

Confirmatory analyses demonstrate that a large proportion of our discoveries are either consistent with the findings of BOLT-LMM on the same data, or are supported by other studies. This is of interest, especially since we do not have access to studies larger than the UK Biobank, so we cannot expect to already replicate all novel discoveries. Furthermore, many of our unconfirmed findings are associated to related phenotypes or contain protein-coding genes that are over-expressed in the relevant tissues. Even though our method does not perform fine-mapping in the traditional sense—we do not work with imputed variants—it is more flexible and appreciably more powerful than the existing genome-wide alternatives, such as BOLT-LMM. Furthermore, our discoveries are more tightly localized than those based on marginal testing, and hence immediately more interpretable, which will facilitate any follow-up analysis.

The possibility to include individuals of diverse ancestries opens alluring research opportunities. For example, we would like to understand which discoveries are consistent across populations and which are more specific, as this may help further weed out false positives, explain observed variations in phenotypes, and possibly shed more light onto the underlying biology.

SOFTWARE AVAILABILITY

We have implemented our methods in a standalone software written in C++, which is available from `https://msesia.github.io/knockoffzoom-v2/`. This takes as input phased haplotypes in BGEN format[63] and outputs genotype knockoffs at the desired resolution in the PLINK[51] BED format. Our software is designed for the analysis of large data: it is multi-threaded and memory efficient. Furthermore, if sufficient computational resources are available, knockoffs for different chromosomes can be constructed in parallel. For reference, it took us approximately 4 days using 10 cores and 80GB of memory to generate knockoffs for the UK Biobank data on chromosome 1 (approximately 1M haplotype sequences, 600k SNPs, and 600k IBD segments).

DATA AVAILABILITY

Data from the UK Biobank Resource (application 27837); see `https://www.ukbiobank.ac.uk/`.

ACKNOWLEDGEMENTS

M. S. was in the Department of Statistics 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 useful computer code. The authors acknowledge the participants and investigators of the UK Biobank project, the FinnGen study, and the Japan Biobank Project.

---

1 Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).

[2] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A* **100**, 9440–9445 (2003).

[3] Sabatti, C., Service, S. & Freimer, N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 829–833 (2003).

[4] Brzyski, D. *et al.* Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75 (2017).

[5] Sesia, M., Sabatti, C. & Candès, E. Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18 (2019).

[6] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

[7] Hill, W. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).

[8] Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

[9] Sesia, M., Katsevich, E., Bates, S., Candès, E. & Sabatti, C. Multi-resolution localization of causal variants across the genome. *Nat. Comm.* **11**, 1093 (2020).

[10] Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

[11] Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

[12] Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).

[13] Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).

[14] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

[15] Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).

[16] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

[17] Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

[18] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

[19] Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

[20] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B.* (2020).

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

[22] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* **57**, 289–300 (1995).

[23] Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. B.* **80**, 551–577 (2018).

[24] Barber, R. F. & Candès, E. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).

[25] Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

[26] Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).

[27] Koivisto, M. *et al.* An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. In *Biocomputing 2003*, 502–513 (World Scientific, 2002).

[28] Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703 (2011).

[29] Scheet, P. & Stephens, M. 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).

[30] Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906 (2007).

[31] Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).

[32] Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).

[33] Weir, B. S. *Genetic Data Analysis* (Sinauer, Sunderland, Massachusetts, 1990).

[34] Stephens, J. C., Briscoe, D. & O'Brien, S. J. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.* **55**, 809 (1994).

[35] Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179 (2012).

[36] Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5 (2013).

[37] O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817 (2016).

[38] Thompson, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).

[39] Bycroft, C. *et al.* The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[40] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

[41] Japan, B. Biobank Japan Project (2020). URL `http://jenger.riken.jp/en/`.

[42] FinnGen. FinnGen documentation of r3 release (2020). URL `https://finngen.gitbook.io/documentation/`.

[43] Sved, J. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2**, 125–141 (1971).

[44] Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068 (2008).

[45] Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).

[46] Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).

[47] Bjelland, D. W., Lingala, U., Patel, P. S., Jones, M. & Keller, M. C. A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. *Eur. J. Hum. Genet.* **25**, 617–624 (2017).

[48] Naseri, A., Liu, X., Tang, K., Zhang, S. & Zhi, D. Rapid: ultra-fast, powerful, and accurate detection of segments identical by descent (ibd) in biobank-scale cohorts. *Genome Biol.* **20**, 143–143 (2019).

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

[50] Marchini, J. L. Discussion of gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 27–28 (2019).

[51] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

[52] Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005–D1012 (2019).

[53] Hoffmann, T. J. *et al.* Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature Genet.* **49**, 54 (2017).

[54] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

[55] deCODE genetics. https://www.decode.com/ (2019). Accessed: 2019-12-06.

[56] Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature News* **538**, 161 (2016).

[57] Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

[58] Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).

[59] Bitarello, B. D. & Mathieson, I. Polygenic scores for height in admixed populations. *bioRxiv preprint* (2020).

[60] Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *bioRxiv preprint* (2020).

[61] Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

[62] Bates, S., Sesia, M., Sabatti, C. & Candès, E. J. Causal inference in genetic trio studies. *arXiv preprint* 2002.09644 (2020).

[63] Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *BioRxiv* 308296 (2018).

[64] Sesia, M., Sabatti, C. & Candès, E. Rejoinder: Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 35–45 (2019).

[65] Sabatti, C. *Multivariate Linear Models for GWAS*, 188–207 (Cambridge University Press, 2013).

[66] Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R, packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).

[67] Consortium, I. H. . *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).

[68] Kinderman, R. & Snell, S. *Markov random fields and their applications* (American Mathematical Society, Providence, RI, USA, 1980).

[69] Yedidia, J., Freeman, W. & Weiss, Y. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millenium*, vol. 8, 239–269 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003).

[70] Bates, S., Candès, E., Janson, L. & Wang, W. Metropolized knockoff sampling. *J. Am. Stat. Assoc.* 1–25 (2020).

METHODS

Formal definition of the inferential objective

We begin by formally stating our objective. Let $Y \in \mathbb{R}^n$ be a phenotype of interest measured for n subjects, and let $X \in \mathbb{R}^{n \times p}$ be the corresponding genotypes at p sites, which are assumed to be random variables from an HMM, as we will discuss shortly. Our goal is to detect genomic regions containing distinct associations with the phenotype, as precisely as possible. Formally, we seek this objective by testing *conditional* null hypotheses[23] in the form of

$$\mathcal{H}_{0,g} : Y \perp\!\!\!\perp X_{G_g} \mid X_{-G_g}, \tag{1}$$

where $\mathcal{G} = (G_1, \ldots, G_L)$ is a fixed partition of $\{1, \ldots, p\}$, X_{G_g} denotes the variants in group G_g, and X_{-G_g} denotes those outside it. This framing is different from that employed by classical GWAS techniques, where one is only able to test marginal independence between the phenotype and a single variant X_j—a scientifically less interesting hypothesis.[5,64] In particular, conditioning on the remainder of the genome (X_{-G_g} in the above notation) ensures that a rejection of the null hypothesis implies the presence of a unique association of Y with the variants in the region G_g, rather than a spurious correlation caused by LD. Moreover, such tests are naturally robust to the confounding of population structure,[64] and even enable formal causal inferences in some cases.[62] While we do not have the space for a full account here, previous work has already discussed at length the relative advantages of conditional testing over marginal testing.[5,9,62,64] The only caveat is that correctly implementing these tests using GWAS data is technically challenging in the presence of population structure[9]—a difficulty that we address here by building upon KnockoffZoom.[9]

Review of KnockoffZoom

KnockoffZoom is a recently-introduced statistical technique for genome-wide association studies that simultaneously tests the conditional null hypotheses defined above for all groups of variants in any given partition of the genome, controlling the FDR (the expected number of false discoveries) below a desired target level (e.g., 10%). By applying this procedure multiple times with increasingly refined partitions, one can localize interesting conditional associations at different levels of resolutions, ranging from a few hundred kilo-bases to single-SNP precision. Unlike those of other approaches, the statistical guarantees of KnockoffZoom do not rely on any assumption about the genetic architecture of the trait, such as linearity and Gaussian errors. Instead, KnockoffZoom only needs to model the distribution of the genotypes with an HMM, consistently with the standard approaches for phasing and imputation.[5,9] To test the hypotheses in (1), KnockoffZoom leverages a statistical technology called knockoffs,[5,23,24] which we describe informally below.

For the genotypes $X^{(i)}$ of any individual i, we construct *in silico* a synthetic "knockoff" version $\tilde{X}^{(i)} \in \mathbb{R}^p$, in such a way that X and $\tilde{X}$ are statistically exchangeable at the population level, except for the fact that $\tilde{X}$ has no conditional associations with Y. In other words, for any group G_g in the given genome partition, the distribution of X_{G_g} and $\tilde{X}_{G_g}$ are indistinguishable *unless there is a conditional association between X_{G_g} and Y*. This property, which will be formally stated later, allows the knockoff genotypes $\tilde{X}$ to serve as negative controls for X.[23] The intuition is that, by construction of the knockoffs, any detectable difference between X_{G_g} and $\tilde{X}_{G_g}$ implies that the region G_g must contain a distinct association with Y. Practically, one can control the FDR for the conditional hypotheses in (1) by computing a test statistics for each group of variants (i.e., as the contrast between any empirical association measures between Y and X_{G_g}, $\tilde{X}_{G_g}$, respectively) and then rejecting the null hypotheses whose statistics are greater than the data-adaptive significance threshold computed by the knockoff filter.[24] To maximize the number of discoveries, one should use powerful association (or, *importance*) measures; the typical solution is to compute these starting from a multivariate predictive model, as explained next.[23]

KnockoffZoom[9] fits a sparse generalized linear regression model of Y given the augmented data $[X, \tilde{X}] \in \mathbb{R}^{n \times 2p}$, after standardizing X and $\tilde{X}$ to have unit variance. Then, letting $\hat{\beta}_j(\lambda_{\text{CV}})$ and $\hat{\beta}_{j+p}(\lambda_{\text{CV}})$ indicate the estimated coefficients for X_j and $\tilde{X}_j$, respectively, it defines the importance measures $T_g = \sum_{j \in G_g} |\hat{\beta}_j(\lambda_{\text{CV}})|$ and $\tilde{T}_g = \sum_{j \in G_g} |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|$ for each group of variants G_g. (The regularization parameter λ_{CV} is tuned by cross-validation). We adopt these statistics because they have the advantage of being powerful,[9] interpretable for GWAS,[65] and computationally feasible even with very large data.[66] However, the knockoffs framework can theoretically accommodate virtually any statistics.[23] The importance measures are combined into a test statistic for each group of variants, i.e., $W_g = T_g - \tilde{T}_g$, which is finally provided as input to the knockoff filter[24] to select significant associations. In particular, the knockoff filter selects groups of SNPs whose W is sufficiently large (i.e., here large values of W imply that the corresponding real genotypes have larger regression coefficients compared to the knockoffs). See Figure 5 for a schematic summary of the entire procedure.

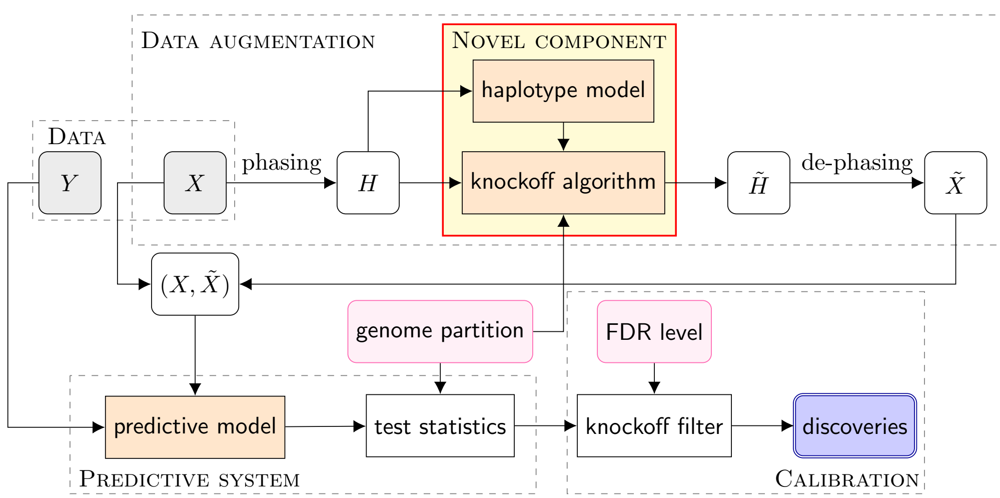

FIG. 5. **KnockoffZoom workflow**. The novelty consists of an HMM for the distribution of haplotypes, H, that can account for population structure and familial relatedness as well as LD, and of the associated algorithm for generating knockoffs. For computational reasons, the genotypes are phased prior to generating knockoffs, and the knockoff haplotypes are then de-phased to obtain knockoff genotypes.[9]

KnockoffZoom model setup

KnockoffZoom v1[5] assumes all samples to be independent and identically distributed (i.i.d.):

$$(X^{(i)}, Y^{(i)}) \overset{\text{i.i.d.}}{\sim} P_{X,Y} = P_X \cdot P_{Y|X},$$

where the distribution of the genotypes, P_X, is approximated by the fastPHASE HMM,[29] and $P_{Y|X}$ is left completely unspecified.[23] The fastPHASE HMM restricts the applicability of KnockoffZoom to homogeneous samples[9] (see also Supplementary Notes A– C), while the i.i.d. assumption is naturally ill-suited to described closely related samples that, in addition to sharing long nearly identical portions of DNA, may also be exposed to the same environmental factors affecting their phenotypes. Our present work shows how to overcome these limitations. We generalize the above setup by assuming a fixed family structure and by grouping together individuals in the same family, which are no longer treated independently. Furthermore, the distribution of the genotypes is allowed to vary across families, depending on the population from which the family descended. Formally, let $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_{|\mathcal{F}|}\}$ indicate a fixed partition of $\{1, \ldots, n\}$, where $\mathcal{F}_f \subseteq \{1, \ldots, n\}$ for all $f \in \{1, \ldots, |\mathcal{F}|\}$. Then, we assume that genotypes and phenotypes, along with a shared family effect E, are sampled independently for individuals in different families, from some joint distribution:

$$(X^{(\mathcal{F}_f)}, Y^{(\mathcal{F}_f)}, E^f) \sim P^f_{X,Y,E} = P^f_E \cdot P^f_X \cdot P^f_{Y|X,E}. \tag{2}$$

Above, the random matrix $X^{(\mathcal{F}_f)} \in \mathbb{R}^{|\mathcal{F}_f| \times p}$ contains the genotypes (i.e., the rows of X) for all individuals in the fth family and is sampled jointly from P^f_X, the random vector $Y^{(\mathcal{F}_f)} \in \mathbb{R}^{|\mathcal{F}_f|}$ describes the corresponding phenotypes, and E^f is the environmental factor shared by all family members. Crucially, these distributions are allowed to vary across families, which encodes the fact that the distribution of genotypes and phenotypes varies across populations. Importantly, the distributions P^f_E and $P^f_{Y|X,E}$ are not assumed to be known; only P^f_X must be specified to carry out the procedure. Conditional on X^f and E^f, the phenotypes Y^f are assumed to be sampled independently for each individual:

$$p(Y^{(\mathcal{F}_f)} \mid X^{(\mathcal{F}f)}, E^f) = \prod_{i \in \mathcal{F}_f} p^f(Y^{(i)} \mid X^{(i)}, E^f),$$

for some distribution function p^f, which may also vary across families. Lastly, note that $X \perp\!\!\!\perp E$, which means E is assumed to be an environmental factor unrelated to the genetic effects.

Before proceeding to the mechanics of the test, we pause to explain why testing the hypotheses in (1) properly already implicitly accounts for the family effects E. Since our motivation is to help build a better genetic understanding the phenotype, the most intuitive goal within the above setup is to test the null hypotheses

$$\mathcal{H}_{0,g} : Y \perp\!\!\!\perp X_{G_g} \mid X_{-G_g}, E,$$

for all $g \in \{1, \dots, L\}$. However, these hypotheses are equivalent to those in (1) because we have assumed $X \perp\!\!\!\perp E$. Therefore, tests of (1) already account for family effects, as long as we can carry them out correctly, which is the problem we address here.

The last preliminary step is to formally define the knockoffs. This requires a small extension of the existing framework,[23] because there are now dependent groups of observations (families) that follow different distributions. Nonetheless, the existing theory can be easily adapted for this setting. In particular, we can still test the hypotheses in (1) with the knockoff filter,[24] as long as the KnockoffZoom test statistics satisfy the *flip-sign* property in Lemma 3.3 of Candès et al.[23] To ensure this property holds in the presence of families, the knockoff exchangeability property[23] must be defined in the following way: for any fixed partition $\mathcal{G}$ of the variants, we require that swapping any group of SNPs with the corresponding knockoffs, *simultaneously* for all family members, would keep the joint distribution of genotypes and knockoffs statistically indistinguishable. Formally, this can be written as:

$$\left(X^{(\mathcal{F}_f)}, \tilde{X}^{(\mathcal{F}_f)}\right)_{\text{swap}(S;\mathcal{G})} \stackrel{d}{=} \left(X^{(\mathcal{F}_f)}, \tilde{X}^{(\mathcal{F}_f)}\right) \qquad \forall S \subseteq \mathcal{G}, \tag{3}$$

where $\mathcal{F}_f$ indicates a particular family, and $\text{swap}(S;\mathcal{G})$ is the operator that swaps all columns of $X^{(\mathcal{F}_f)}$ indexed by S with the corresponding columns of $\tilde{X}^{(\mathcal{F}_f)}$. In the following, we will develop a novel algorithm, based on a joint model for the genotypes in each family, to generate knockoff genotypes satisfying (3).

Modeling haplotypes with population structure

We begin by recalling some useful notation for HMMs.[9] We say that a sequence of phased haplotypes $H = (H_1, \dots, H_p)$, with $H_j \in \{0, 1\}$, is distributed as an HMM with K hidden states if there exists a vector of latent random variables $Z = (Z_1, \dots, Z_p)$, with $Z_j \in \{1, \dots, K\}$, such that:

$$\begin{cases} Z \sim \text{MC}\left(Q\right), & \text{(latent discrete Markov chain)}, \\ H_j \mid Z \sim H_j \mid Z_j \stackrel{\text{ind.}}{\sim} f_j(H_j \mid Z_j), & \text{(emission distribution)}. \end{cases} \tag{4}$$

Above, the Markov chain $\text{MC}\left(Q\right)$ has initial probabilities Q_1 and transition matrices $(Q_2, \dots, Q_p)$.

Taking inspiration from SHAPEIT,[35–37] we assume the i-th haplotype sequence can be approximated as an imperfect mosaic of K other haplotypes in the data set, 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 (4), where the distribution of the latent Markov chain is:

$$\mathbb{P}[Z_1^{(i)} = k] = \alpha_k^{(i)},$$

$$\mathbb{P}[Z_j^{(i)} = k' \mid Z_{j-1}^{(i)} = k] = Q_j^{(i)}(k' \mid k) = \begin{cases} \left(1 - e^{-\rho d_j}\right) \alpha_{k'}^{(i)} + e^{-\rho d_j}, & \text{if } k' = k, \\ \left(1 - e^{-\rho d_j}\right) \alpha_{k'}^{(i)}, & \text{if } k' \neq k. \end{cases} \tag{5}$$

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,[67] 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 (Supplementary Methods). However, we have observed it works well with our data to simply set $\rho = 1$; this is consistent with the approach of SHAPEIT,[35–37] which also uses fixed parameters. The positive weights $(\alpha_1^{(i)}, \ldots, \alpha_K^{(i)})$ are normalized so that their sum equals one and they can be interpreted as characterizing the ancestry of the i-th individual. In the applications discussed within this paper, we simply assume $\alpha_k^{(i)} = 1/K$ for all k, although these parameters could also be estimated by EM (Supplementary Methods). Conditional on the sequence Z, each element of H follows an independent Bernoulli distribution:

$$f_j^{(i)}(H_j^{(i)} \mid k) = \begin{cases} 1 - \lambda_j, & \text{if } H_j^{(i)} = H_j^{(\sigma_{ik})}, \\ \lambda_j, & \text{if } H_j^{(i)} \neq H_j^{(\sigma_{ik})}. \end{cases} \tag{6}$$

Above, the parameter λ_j is a site-specific mutation rate, which makes the HMM mosaic imperfect. Earlier works that first proposed this model also suggested formulae for determining suitable values of the parameters ρ and $\lambda = (\lambda_1, \ldots, \lambda_p)$ in terms of the physical distances between SNPs and other population genetics quantities.[8] However, given that we have access to large a data set, we choose instead to estimate λ by EM (Supplementary Methods). We have noticed that it works well to also explicitly prevent λ from being too large or too small (e.g., $10^{-6} \leq \lambda_j \leq 10^{-3}$).

To reduce the computational burden and mitigate the risk of overfitting, the value of K should not be too large; here, we adopt $K = 100$. We have observed that larger values improve the goodness-of-fit relatively little, while reducing power by increasing the similarity between variables and knockoffs.[9] 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, one may carry this out as outlined by Algorithm 1, separately chromosome-by-chromosome. Different options are available for defining similarities between haplotypes; for simplicity, we use the Hamming distance. Since computing the pairwise distances between all haplotypes would have quadratic complexity in the sample size, which is unfeasible for large data, we first divide the haplotypes into clusters of size N, with $K \ll N \ll 2n$ (i.e., $N \approx 5000$), through recursive 2-means clustering, and then we only compute distances within clusters, following in the footsteps of SHAPEIT v3.[37]

Algorithm 1 Choosing the HMM reference haplotypes

Input: haplotypes $H \in \{0,1\}^{2n \times p}$, parameter K;
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. $N_1 \leq |C_c| \leq N_2$, by recursive 2-means clustering.[37]
for $c = 1, \ldots, M$ **do**
  Compute a distance matrix $D \in \mathbb{R}^{|C_c| \times |C_c|}$ for all haplotypes in C_c, with ξ.
  for i in C_c **do**
    Define $R(i)$ as the set of K nearest neighbors of H_i in C_c.
Output: a set $R(i)$ of K references for each haplotype $H^{(i)}$.

While the approach in Algorithm 1 is the easiest to introduce first, in practice it is preferable for our purpose to utilize a different set of local references in different parts of the same chromosome. We will describe this extension later, after discussing the novel knockoff generation algorithm.

Generating knockoffs preserving population structure

Conditional on the reference indices, $\{\sigma_{i1}, \ldots, \sigma_{iK}\}$, the above setup describes each haplotype sequence $H^{(i)}$ as an HMM; a model for which we already know how to generate knockoffs in theory.[5,9] Therefore, generating knockoffs in our setting is straightforward: all we need to do is to define the customized HMM for each haplotype sequence, and then to apply the knockoff generation algorithm previously developed in KnockoffZoom v1.[9] This solution is outlined in Algorithm 2.

Algorithm 2 Knockoff haplotypes preserving population structure

Input: haplotypes $H \in \{0,1\}^{2n \times p}$, genetic map $\rho \in \mathbb{R}^{p-1}$, partition $\mathcal{G}$ of $\{1,\dots,p\}$; parameter K.

for $i = 1,\dots,2n$ **do**

Assign references $R(i) = \{\sigma_{i1},\dots,\sigma_{iK}\}$ with Algorithm 1.

Initialize $\alpha_k^{(i)} \leftarrow \frac{1}{K}$, for each $k \in \{1,\dots,K\}$.

Estimate $\lambda = (\lambda_1,\dots,\lambda_p)$ by EM (Supplementary Methods), initialize $\rho \leftarrow 1$.

for $i = 1,\dots,2n$ **do**

Define the HMM $\{R(i), \rho, \lambda\}$.

Sample $Z^{(i)} = (Z_1^{(i)},\dots,Z_p^{(i)})$ from $\mathbb{P}[Z^{(i)} \mid H^{(i)}]$, with step I of Algorithm 3 in Sesia et al.[9]

Sample a knockoff copy $\tilde{Z}^{(i)}$ of $Z^{(i)}$ with respect to $\mathcal{G}$, with step II of Algorithm 3 in Sesia et al.[9]

Sample $\tilde{H}^{(i)}$ from $\mathbb{P}[H^{(i)} \mid Z^{(i)} = \tilde{Z}^{(i)}]$, with step III of Algorithm 3 in Sesia et al.[9]

Output: knockoff haplotypes $\tilde{H} \in \{0,1\}^{2n \times p}$.

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, it is worth extending Algorithms 1–2 to accommodate different local references within the same chromosome. We implement this as follows.

First, we divide each chromosome into relatively wide genetic windows (10 Mb, for concreteness); then, we choose the reference haplotypes separately within each of them, based on their similarities outside the window of interest. In order to allow the choice of references to be as locally adaptive as possible, we only consider the alleles in the two neighboring windows when computing distances. This approach is similar to that of SHAPEIT v3,[37] 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. Once the local references for each haplotype have been defined, we can 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 (step I of Algorithm 3 in Sesia et al.[9]) but then we only generate knockoffs within the window of interest.

Knockoffs preserving familial relatedness

Within an IBD-sharing family of size m, we model the joint distribution of the haplotype sequences, namely $(H^{(1)}, \ldots, H^{(m)})$, as an HMM with a K^m-dimensional latent Markov chain. For simplicity, we write this as $Z^{(1:m)} = (Z^{(1)}, \ldots, Z^{(m)})$, 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 (4)–(6):

$$\mathbb{P}\left[H_j^{(i)} = 1 \mid Z_j^{(i)} = k\right] = f_j^{(i)}(1 \mid k) = \begin{cases} 1-\lambda, & \text{if } H_j^{(\sigma_{ik})} = 1, \\ \lambda, & \text{if } H_j^{(\sigma_{ik})} = 0. \end{cases} \tag{7}$$

If $Z^{(1)}, \ldots, Z^{(m)}$ were also independent of each other, this would reduce to a collection of m HMMs of the type in (4)–(6). However, to account for familial relatedness, we assume that different $Z^{(i)}$ are coupled at the IBD segments (which we have previously detected and hold fixed), as described below.

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

$$\mathbb{P}\left[Z_j^{(1:m)} = z_j^{(1:m)} \mid Z_{j-1}^{(1:m)} = z_{j-1}^{(1:m)}\right] = \prod_{i=1}^{m} \left(Q_j^{(i)}(z_j^{(i)} \mid z_{j-1}^{(i)})\right)^{\eta_{i,j}} \prod_{i' \in \partial(i,j)} \mathbb{1}[z_j^{(i)} = z_j^{(i')}], \tag{8}$$

where the transition matrices $Q_j^{(i)}$ are defined as in (5), while

$$\mathbb{P}\left[Z_1^{(1:m)} = (k^{(1)}, \ldots, k^{(m)})\right] = \prod_{i=1}^{m} \left(\alpha_{k^{(i)}}^{(i)}\right)^{\eta_{i,j}} \prod_{i' \in \partial(i,j)} \mathbb{1}[k^{(i)} = k^{(i')}]. \tag{9}$$

The first term on the right-hand-side of (8) describes the transitions in the Markov chain, while the second term is the constraint that makes the haplotypes 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 presented earlier for unrelated haplotypes. (If we set instead $\eta_{i,j} = 1$, transitions of the latent state may occur with significantly different frequency inside and outside IBD segments). In the trivial cases of families of size one, $\partial(1,j) = \emptyset$ and $\eta_{1,j} = 1$, for all $j \in \{1, \ldots, p\}$, so it is easy to see that (8)–(9) reduce to the original model for unrelated haplotypes in (5). By contrast, in the general case, the latent states for different haplotypes in the same family are constrained to be identical along all IBD segments. See Supplementary Figure 36 (a) for a graphical representation of this model.

Generating knockoffs with the existing algorithm for HMMs in this case would have computational complexity equal to $\mathcal{O}(npK^m)$, which is unfeasible for large data sets unless $m = 1$. To understand why this complexity is $\mathcal{O}(npK^m)$, note that the model is effectively an HMM with a K^m-dimensional latent Markov chain,[9] where each vector-valued variable corresponds to the alleles at a specific site for all individuals in the family. Fortunately, one can equivalently look at the joint distribution of $(Z^{(1:m)}, H^{(1:m)})$ as a more general Markov random field[68] with $2 \times m \times p$ variables, each taking values in $\{1, \ldots, K\}$ or $\{0, 1\}$, respectively. See Supplementary Figure 36 (b) for a graphical representation of this model, 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.[69] Leaving the technical details to the Supplementary Methods, we outline the new knockoff generation procedure in Algorithm 3.

In a nutshell, we follow the same spirit of the construction for unrelated haplotypes in Algorithm 2, with the important difference that the HMM with a K-dimensional latent Markov chain of length p is replaced by a latent Markov random field with $2 \times m \times p$ variables, which requires some innovations.

- First, the K haplotype references in the model for each $H^{(i)}$ are shared by all haplotypes within the same family; see Supplementary Algorithm 1.

- Second, the posterior sampling of $Z^{(1:m)} \mid H^{(1:m)}$ is carried out via generalized belief propagation[69] instead of the standard forward-backward procedure for HMMs;[5,9] see Supplementary Algorithm 2 and Supplementary Figure 37. (This generally involves some degree of approximation, but it is exact for many family structures).

- Third, the knockoff copies $\tilde{Z}^{(1:m)}$ of $Z^{(1:m)}$ are generated with a modified version of the existing algorithm for HMMs[9] that circumvents the computational difficulties of the higher-dimensional model by breaking the couplings between different haplotypes through conditioning[70] upon the extremities of the IBD segments; see Supplementary Algorithm 3 and Supplementary Figure 38. To clarify, conditioning on the extremities of the IBD segments means that we make the knockoffs identical to the true genotypes 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 reduces the computational burden (see Supplement for a full explanation).

It is worth remarking that, for trivial families of size one, this is exactly equivalent to Algorithm 2. Finally, note that the extension to local references with hold-out distances discussed earlier also applies seamlessly here. Even though our software implements this extension, we do not explicitly write down the algorithms with local references in this paper for the sake of clarity and space.

Algorithm 3 Knockoff haplotypes preserving population structure and familial relatedness

Input: $H \in \{0,1\}^{2n \times p}$, $d \in \mathbb{R}^{p-1}$, $\mathcal{G}$, and K as in Algorithm 2, collection of IBD segments $\mathcal{I}$.

Define the set $U \subseteq \{1, \ldots, n\}$ of haplotype indices that do not share any IBD segments in $\mathcal{I}$.

Divide the remaining haplotypes into L distinct families $F_f \subseteq \{1, \ldots, n\}$, for $f \in \{1, \ldots, L\}$.

for $f \in 1, \ldots, L$ **do**

  Assign references $R(f) = \{\sigma_{f1}, \ldots, \sigma_{fK}\}$ with Supplementary Algorithm 1.

  Initialize $\alpha_k^{(f)} \leftarrow \frac{1}{K}$, for each $k \in \{1, \ldots, K\}$.

Estimate $\lambda = (\lambda_1, \ldots, \lambda_p)$ by EM (Supplementary Methods), initialize $\rho \leftarrow 1$.

for $f \in 1, \ldots, L$ **do**

  Define the HMM $\{R(f), \rho, \lambda\}$.

  Sample $(Z^{(i)})_{i \in F_f} \in \{1, \ldots, K\}^{m \times n}$ from $\mathbb{P}[(Z^{(i)})_{i \in F_f} \mid (H^{(i)})_{i \in F_f}]$, with Supplementary Algorithm 2.

  Sample a knockoff copy $(\tilde{Z}^{(i)})_{i \in F_f}$ of $(Z^{(i)})_{i \in F_f}$ with respect to $\mathcal{G}$, with Supplementary Algorithm 3.

  Sample $\tilde{H}^{(i)}$ from $\mathbb{P}[H^{(i)} \mid Z^{(i)} = \tilde{Z}^{(i)}]$, coordinate by coordinate independently, for all $i \in F_f$.

for $i \in U$ **do**

  Generate a knockoff copy $\tilde{H}^{(i)}$ of $H^{(i)}$ as in Algorithm 2.

Output: knockoff haplotype matrix $\tilde{H} \in \{0,1\}^{2n \times p}$ that preserves the IBD segments.

Quality control and data pre-processing for the UK Biobank

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 in the data set. We define families by clustering together individuals with kinship greater or equal to 0.05; then we discard 322 individuals (chosen as those with the most missing phenotypes) to ensure that no families have size larger than 10. This leaves us with 136,818 individuals labeled as related, divided into 57,164 families. The median family size is 2, and the mean is 2.4). The total number of individuals that passed this quality control (included those who have no relatives) is 486,975. We only analyze 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 previous work.[9] The final SNPs count is 591,513.

Including additional covariates

We account for the sex, age and squared age of the subjects to increase power (squared age is not used for *height*), as in earlier work.[9,40] These covariates are leveraged in the KnockoffZoom predictive model, along with the top 5 genetic principal components. The inclusion of the principal components, which may capture population-wide shifts in the distribution of the phenotype, is useful to improve power because it increases the accuracy of the predictive model while keeping it sparse,[9] but it is unnecessary to avoid spurious discoveries due to population structure, since our method already accounts for that through the knockoffs. We thus fit a sparse regression model to predict Y given $[Z, X, \tilde{X}] \in \mathbb{R}^{n\times(r+2p)}$, where $Z, X, \tilde{X}$ are the r covariates, the p genotypes, and the p knockoffs, respectively. The model coefficients for Z are not regularized, and their estimates are not directly included in the test statistics.

Numerical experiments

Feature importance measures for each SNP are computed in three alternative ways: by fitting the lasso with cross-validation and taking the absolute value of the estimated regression coefficients;[9] by running BOLT-LMM[40] and taking the negative logarithm of the marginal p-values; and by performing univariate logistic regression (in the case of binary phenotypes) and taking the negative logarithm of the marginal p-values. These models are designed to predict Y given $[X, \tilde{X}]$; in the first two cases we also include the top 10 principal components of the genotype matrix (computed on the entire UK Biobank data set) as additional covariates. Then, the feature importance measures T_j and $\tilde{T}_j$, for the jth SNP and its corresponding knockoff, are combined in the usual way to define the knockoff test statistics for each group $G_g \subseteq \{1, \dots, p\}$ of variables: $W_g = \sum_{j\in G_g} T_j - \sum_{j\in G_g} \tilde{T}_j$.

In the simulations with related samples, the latent Gaussian variable for the ith individual in the probit model is given by:

$$L^{(i)} = \sum_{j=1}^{p} \beta_j X_j^{(i)} + \gamma E^{(f_i)} + \sqrt{1-\gamma^2}\epsilon^{(i)},$$

where $E^{(f_i)}$ and $\epsilon^{(i)}$ are independent standard normal random variables, $\gamma \in [0, 1]$, and f_i denotes the family to which the ith individual belongs. The magnitude of the nonzero genetic coefficients β is varied as a parameter, to control the total heritability of the trait.