Modeling recent positive selection in Americans of European ancestry

Seth D. Temple,[1*] Ryan K. Waples,[2] Sharon R. Browning[2]

[1] Department of Statistics, University of Washington

[2] Department of Biostatistics, University of Washington

[*] Address for correspondence: Seth D. Temple (sdtemple@uw.edu)

Abstract

Recent positive selection can result in an excess of long identity-by-descent (IBD) haplotype segments. The statistical methods that we propose here address three major objectives in studying classical selective sweeps: scanning for regions of interest, identifying possible sweeping alleles, and estimating a selection coefficient s. First, we implement a selection scan to locate regions of excess IBD rate. Second, we develop a statistic to rank alleles in strong linkage disequilibrium with a putative sweeping allele. We aggregate these scores to estimate the allele frequency of the sweeping allele, even if it is not genotyped. Lastly, we propose an estimator for the selection coefficient and quantify uncertainty using the parametric bootstrap. Comparing against state-of-the-art methods in extensive simulations, we show that our methods are better at pinpointing alleles at low frequency and estimating $s \geq 0.015$. We apply these methods to study positive selection in inferred European ancestry samples from the TOPMed project. We find twelve loci where their IBD rates exceed four standard deviations above the population median. The excess IBD rate at LCT is thirty-five standard deviations above the population median, and our estimates of its selection coefficient imply strong selection within the past 120 to 163

generations. Overall, we present robust and accurate new approaches to study very recent adaptive evolution under mild assumptions.

Introduction

Positive natural selection favors alleles advantageous for survival or reproduction. The process of evolution via positive selection is often modeled via hard or soft selective sweeps. In hard sweeps, a new mutation increases rapidly in frequency because it provides a significant evolutionary benefit. In soft sweeps, an existing neutral allele at low frequency increases when it becomes newly beneficial in the context of environmental change[1–3]. Alleles on the same haplotype as a sweeping allele increase in frequency as well until crossover recombination separates them from the sweeping allele. Motivated by this observation of genetic "hitch-hiking", most methods to detect and model recent positive selection evaluate patterns of unusual linkage disequilibrium[4]. Selective sweep events in the past five hundred generations can leave profound footprints on genomic architecture around a selected locus. For instance, extended haplotype homozygosity about the LCT gene in northern Europeans spans multiple megabases (Mb) of genetic sequence[5]. Loci inferred to be under strong positive selection in humans reside in gene-dense regions, including gene complexes important to pigmentation, the innate immune system, and the adaptive immune system. Studying positive selection can be complementary to large association studies; for instance, inflammatory disorders show evidence of strong and weak selection[6]. The purpose, magnitude, and prevalence of such adaptive evolution remain open questions in population genetics, medical genetics, and conservation genetics.

The popular mathematical model to study a selective sweep from a single allele frames the problem in terms of the selection coefficient s. This parameter measures the relative

advantage that the sweeping allele has over other alleles. In the haploid model, the formula $p_s(t-1) = p_s(t) \times (1+s)$ describes how the allele frequency changes at time t generations ago; that is, the allele frequency increases by $(1+s)$ percent in the next generation. This interpretation highlights the impact that the selection coefficient imposes on the time series of allele frequency. The function may be inverted to express the allele frequency $p_s(t)$ one generation further back in time in terms of the allele frequency $p_s(t-1)$ and the selection coefficient s. Besides estimating s, further challenges include locating the selected allele and estimating its frequency. Together the selection coefficient and the frequency of the sweeping allele imply a distribution on the time of *de novo* mutation, as well as historical allele frequencies.

To scan for targets of recent positive selection, standard methods study decay patterns of extended haplotypes[5,7], excess recent shared ancestry[8–10], haplotype differences between ancestrally diverged populations[11,12], allele frequencies in ancient DNA samples[13], or excess ancestry proportions in admixed samples[14]. All these approaches compute statistics derived from population genetics, normalize over the genome, and catalog regions where these statistics differ from a genome-wide central tendency. Here we take an identity-by-descent (IBD) based approach. Modern algorithms can infer haplotype tracts shared from a common ancestor with high confidence if the ancestor of the sweeping allele is recent enough so that many of the shared tracts carrying that allele span a length of two or more centiMorgans (cM)[8,9,15–19]. Because IBD tracts of these lengths ≥ 2.0 cM mostly come from ancestors in the last five hundred generations, we define "recent" to apply to selection events within that time range. Our IBD methods examine the most recent partial sweeps; haplotype homozygosity methods study slightly older selection within a population; differences between populations relate to even older selection.

To identify possible sweeping alleles, the earliest method proposed is extended haplotype homozygosity (EHH). This statistic measures the proportion of shared haplotypes at increasing physical distances[5]. Current methods derive from this initial idea[7,12,20,21]. In contrast, Akbari et al. (2018) evaluate the difference between two estimators of derived haplotype frequency about a given marker[22]. They observe in simulations that these two estimators skew in different directions under positive selection, which is why their statistics should highlight regions undergoing a selective sweep. Their method, integrated Selection of Allele Favored by Evolution (iSAFE), performs favorably compared to other methods[20,23]. The goal of these methods is to rank alleles in terms of their evidence for being a sweeping allele. These methods do not supply a frequency estimate for the sweeping allele.

To estimate the selection coefficient, three different paradigms have been considered. Some approaches use ancient DNA to infer allele frequency changes over time[13,24], or mark-recapture phenotype frequencies in the case of wildlife[25]. These methods are limited to population studies where data is collected across multiple time points. Analyses of contemporary datasets rely on population genetics theory. The coalescent defines a model for how samples relate to each other in the context of an ancestral tree [26–28]. Recent methods enable inference of the coalescent process across neighboring markers in the entire genome[29–32]. Using an estimated coalescent tree, Stern, Wilton, and Nielsen (2019) optimize the selection coefficient in terms of a likelihood equation[33]. Their method is sensitive to the initial task of tree estimation and does not quantify uncertainty. Some tree inference algorithms apply ad-hoc rules to handle increasing sample sizes, but these methods do not express uncertainty over the space of possible tree sequences[29,32,34]. Alternative approaches express phylogenetic uncertainty by sampling from a Bayesian posterior distribution, but these approaches scale poorly with increasing biobank

sizes[30,35,36]. As opposed to statistical inference, machine learning methods simulate genetic data from selective sweeps[37–40] and then train neural networks. Features in these supervised learning approaches amount to encodings of the inferred ancestral tree[41,42], vectors of summary statistics[43], or image processing of genetic variants[44]. Predictive accuracy will depend on how the user simulates genetic data and trains the model, and simulating genetic data from large populations may require terabytes or more computer memory[41].

No existing method addresses all aspects of selective sweep inference. Another concern in evolutionary biology is that alternative mechanisms could explain the patterns of haplotype structure that these methods leverage. For instance, the major histocompatibility complex (MHC) appears in selection scans in many European populations, but authors repeatedly suggest this signal is an example of balancing selection rather than a sweep[6,45]. Further confounding factors are population structure, assortative mating, and unaccounted for familial relatedness, but these issues can be partially addressed by preliminary data analysis[46–48]. Some existing methods require ancient or archaic samples that are available for Eurasian populations but limited for other populations or studies on non-model organisms. A complete methodology that is motivated by statistical and population genetics theory, makes as few assumptions as possible, and is easily interpretable would broadly contribute to the research community.

Here we offer a unifying framework for recent and strong selective sweeps that connects shared haplotypes to a statistical model for recent ancestry. At a biallelic locus, haplotypes carrying and not carrying the sweeping allele bifurcate into two coalescent trees[49,50]. The subtree with the adaptive allele coalesces much earlier than expected given genome-wide relatedness. This imbalance is more profound the greater the selection coefficient. We view IBD tracts of shared genetic material in modern sequences as a summary of the unobserved coalescent tree.

The basic idea is that we expect to observe more IBD tracts at a selected locus due to sharing among haplotypes descendant from the adaptive haplotype. This perspective motivates our methodology.

In this paper, we propose a suite of new methods iSWEEP (**i**ncomplete **S**elective **s**weep **W**ith **E**xtended haplotypes **E**stimation **P**rocedure) to study all three aspects of a selective sweep analysis from a sample of whole genome sequences. This inference framework examines the effects of recent and strong selection on identity-by-descent within a population. In simulations, we recover simulated selection coefficients and allele frequencies in various demographic scenarios and in the face of model misspecification. We also find that our methods to identify the sweeping allele and estimate the selection coefficient perform favorably or competitively compared to state-of-the-art approaches. We apply our estimation procedure to investigate positive selection in cohorts of European ancestry in the TOPMed project[51], recapitulating results in previous studies for twelve loci[8–10,24]. We give 95% confidence intervals for selection coefficients at five of these loci, and these estimates are supported by statistical theory and simulations to contain the selection coefficient 95% of the time. Annotations of genes discovered in our selection scan highlight pigmentation, skin tissue, hair cells, and adaptive immunity as possible targets of selection in Europeans. Overall, we suggest that selective sweeps with coefficients $s \geq 0.015$ may be a domain in which reliable inference is possible using identity by descent tracts, and that such selection within the past five hundred generations may have left profound footprints on the genomic architecture of Europeans.

Results

Brief overview of our methods

Selective sweeps driven by alleles with selection coefficients $s \geq 0.015$ can create extreme local IBD rates compared to average genome-wide relatedness and large clusters of IBD haplotypes. Like EHH-based methods, we survey the genome for loci with an excess of inferred IBD tracts. Then, we calculate scores for those alleles with frequencies that are maximally differentiated between an inferred haplotype group of excess IBD and the rest of the sample. These scores serve as the basis for ranking alleles as possible targets of selection. Using these scores, we produce average estimates for the location and frequency of the sweeping allele. Next, with the allele frequency, we propose a conditional estimator for the selection coefficient that relates the amount of IBD to the coalescent under selection. Finally, we perform a fast parametric bootstrap procedure to express uncertainty in our inference.

Robust statistical inference of the selection coefficient

We perform ten thousand experiments to validate that our selection coefficient estimator is accurate, and its 95% confidence intervals contain the true selection coefficient 95% of the time. Our default simulation settings are selection coefficient $s = 0.03$, allele frequency $p(0) = 0.50$, effective sizes $N_e(t)$ at coalescent time t from a population bottleneck (BN), and an ongoing hard sweep from *de novo* mutation. For this simulation study, we simulate true IBD tract lengths from a coalescent under selection. The procedure is to estimate $\hat{s}$ with the known simulated model and IBD lengths greater than 3.0 cM. We explore perturbations from these defaults settings in terms of $s = 0.02$ or 0.04, allele frequency $p(0) = 0.25$ or 0.75, and effectives sizes from super-exponential growth (SE) or constant size 25,000 diploids (C25). Our

procedure gives accurate point estimates and well-calibrated uncertainty for these examples (Table 1).

Next, we assess the extent to which our estimator $\hat{s}$ is robust to model misspecifications, such as bias in allele frequency estimates, incorrect inferred $\widehat{N}_e(t)$, or sweeps from neutral variation at low frequency. Here the estimates are slightly biased but their confidence intervals still cover the true value nearly 95% of the time (Supplementary Table 1). These biased estimates are directly interpretable in terms of how the model misspecification affects the amount of IBD. For example, to fit IBD counts to a misspecified model with elevated effective sizes, a greater selection coefficient is required, and vice versa for lower effective sizes. Interestingly, the estimator is robust to sweeps from standing variation because IBD tracts ≥ 3.0 cM mainly represent relatedness within the last 50-100 generations. The simplicity of our method and our explicit assumptions permit these experiments to examine average behavior and robustness of our statistical inference.

Identifying the sweeping allele and estimating its frequency

The performance of our estimation procedure is impacted by the accuracy of IBD detection methods. To investigate this concern, we conduct a second study where we make inferences based on IBD tracts inferred from sequence data. We simulate evolution within 8 cM regions in which there is an adaptive mutation of $s = 0.030$ introduced within the last five hundred generations. We use established methods to infer IBD tracts in these simulated datasets. Then, we apply our methods using the inferred IBD tracts to identify variants strongly correlated with the sweeping allele, estimate its allele frequency, and infer its selection coefficient.

As in the first study, we explore different selection coefficients, allele frequencies, and effective sizes. We simulate genetic data for five thousand diploids from large populations of constant effective size $N_e = 25{,}000$ or effective sizes like those of Northern Europeans or European Americans[52,53]. We refer to these examples as constant size (C25), super-exponential growth (SE), and population bottleneck (BN). For the bottleneck example, we also study selected alleles with coefficients $0.015 \leq s \leq 0.035$. A model violation to the population bottleneck example that we consider is instead two subpopulations that randomly mix with proportion 40% starting five hundred generations ago.

We evaluate how well iSWEEP and comparable methods iSAFE[22], iHS[7], and nSL[21] identify the selected allele or variants strongly correlated with it. Given that there may be tens of thousands of variants in the region of interest, this task is far from trivial. The rank estimate is the SNP ordering of the true selected allele that is given by each method, where one is the best possible rank. To see if the ordering prefers SNPs in strong linkage disequilibrium with the selected allele, we look at mean absolute deviations (MAD) in frequency and location for the ten SNPs of lowest rank. This metric discerns if a method is favoring alleles on the same haplotypes as the selected allele, in the case that the selected allele is not genotyped.

Two methods, iSWEEP and iSAFE, show improved ability to pinpoint the causal mutation compared to the older methods iHS and nSL (Supplementary Figure 1). They assign high scores to SNPs that have frequencies within 10% and physical locations within 250 kb of the adaptive allele and are robust to different demographic models (Supplementary Figure 1). Next, we juxtapose iSWEEP and the state-of-the-art method iSAFE in terms of their strengths and weaknesses (Figure 1). Both methods perform well in all simulations, so the following commentary concerns small differences in accuracy. iSAFE is better at the task it is designed for:

ordering variants. For smaller selections coefficients ($s = 0.015, 0.020$) and allele frequencies between 10-30%, iSWEEP is as good if not better than iSAFE across all three tasks. The accuracy of iSAFE increases with the allele frequency, and vice versa for iSWEEP. This observation emphasizes that iSWEEP is designed for more recent positive selection.

The statistics in iSWEEP have clear interpretations and require few assumptions. The method's performance is driven by its main data input: abundant and accurately inferred IBD segments. For the population bottleneck and constant N_e examples, there are more IBD tracts to learn from compared to the super-exponential growth example (Supplementary Figure 1). The scores are simple differences in proportions. They also do not require knowledge of which allele is favored, but instead infer the favored allele as the one at higher frequency in an outgroup of excess IBD. Haplotype clusters in the inferred outgroup may themselves be of interest. We suggest an information entropy statistic from the cluster proportions of excess IBD as another diagnostic to support the selective sweep hypothesis. This value measures cluster agglomeration, where zero indicates one large outgroup. We observe that entropy greater than 0.50 may indicate weak selection (Supplementary Figure 3). Another explanation for entropy greater than 0.50 could be that the locus has multiple adaptive alleles, violating the sweep model. We incorporate the IBD entropy statistic into our real data analyses to discern if selection signals are better explained by other evolutionary processes.

To model a selective sweep, we require an allele frequency estimate, not the identity of the sweeping allele. Our approach aggregates information over the lowest rank SNPs to estimate the allele frequency $\hat{p}(0)$. As illustrated in Supplementary Figure 2, our allele frequency estimates are often within 5-10% of the true value in panmictic populations. We highlight that

our method of estimating the allele frequency does not necessitate that the allele itself is genotyped; it borrows information from neighboring correlated markers.

Estimating selection coefficients in simulated data

Given an allele frequency estimate and accurate IBD detection, we expect inference of $\hat{s}$ to be in line with those of the idealized coalescent simulation study. We compare our estimation procedure with the deep learning method ImaGene[44] and a tree inference method CLUES[33]. Our method has higher accuracy with larger values of s and performs well in the constant, super-exponential growth, and population bottleneck examples. Figure 2 shows that the 0.25 to 0.75 percentiles of our estimates contain the true selection coefficient for $s \geq 0.025$, and medians of our estimates are closer to the true values than in the other two methods. This result holds for the constant and super-exponential growth demographic models as well. In general, our inference maintains limited bias on average and reasonable confidence intervals coverages when inferred IBD tracts are used (Supplementary Table 2). Overestimation when $s < 0.025$ corresponds mostly to replicates that do not pass the genome-wide significance threshold derived from our real data analysis (Supplementary Figure 3).

Estimates $\hat{s}$ from CLUES are around or below 0.01 for all simulations (Figure 2). The authors of this method acknowledge that their method is limited to $s \leq 0.01$. An analysis of CLUES from separate authors observes results like ours for $s \geq 0.010$. On the other hand, they report that CLUES can estimate $0.010 \leq s \leq 0.020$ when given the true simulated tree[41]. This finding may concern the fundamental algorithmic challenge in inferring high-dimensional ancestral recombination graphs. We leverage data from an upstream inference as well, but numerous papers demonstrate that detecting IBD tracts longer than 3.0 cM is very accurate[15–17].

We train two neural networks with the ImaGene software package to compare against our method. As training and validation data, we generate genetic sequences with similar parameter settings to this simulation study. The first neural network is trained on 100,000 simulated datasets, and we find that it consistently underestimates selection coefficients. The second neural network is trained on two million simulated datasets, and we find that it consistently overestimates selection coefficients (Figure 2). The simulated genetic data is just shy of 100 Gb memory, and the larger model fits in 165 hours wall clock time using 24 computer nodes. Training the supervised learning model may explain differences in the results between ImaGene and our unsupervised method. We do not compare against ImaGene for the constant and super-exponential growth examples because we would have to fit custom neural networks for each. We also do not compare against another machine learning method that takes coalescent trees as input due to high memory costs[41].

Previous simulation studies using CLUES and ImaGene evaluate performance on small selection coefficients[33,42]. Overall, we contrast our method as suited for recent strong positive selection $s \geq 0.015$ whereas current alternatives address weak selection $s \leq 0.015$ [33,44,54]. An estimator should be able to infer a wide range of parameter values so as to not erroneously confuse strong selection as weak. An IBD-based selection scan[8] can subset loci beforehand to just putatively strong sweeps, which is the domain that our estimator is designed for.

Selection scan in TOPMed samples of European ancestry

Prior selection scans in European ancestry populations have implicated many regions of interest. In this paper, we focus on statistical inference of selective sweeps parameters by analyzing cohorts from the Trans-Omics for Precision Medicine (TOPMed) consortium project. We start with 38,387 whole genome sequences from haplotype phasing performed with Beagle

5.2 as previously reported[55]. Using principal components 1-4 and global ancestry inference, we define a group of European ancestry (abbreviated EUR). We estimate kinship coefficients and remove samples with kinship coefficients greater than 0.125 due to close familial relatedness[56]. After filtering, the EUR group consists of 13,778 sample individuals, respectively.

Our scan for recent positive selection proceeds as follows. First, we apply hap-ibd and ibd-ends to find candidate IBD segments and refine their endpoints across all 22 autosomes[8,15]. Using the 2.0 cM threshold from the ibd-ends output, we count IBD tracts overlapping a focal position every 20 kb. Then, we calculate a median and standard deviation for the rate of IBD. We look for locations where the IBD rate exceeds the median plus four times the standard deviation[8]. This threshold represents controlling the family wise error rate of 0.05, adjusting for hypothesis tests roughly every 2 Mb. We further insist that regions of excess IBD span at least 1 cM in length. Our IBD-based selection scan gives results consistent with analyses on white British individuals in the UK Biobank[8–10] and pan-European analyses using ancient DNA[13,24]. Figure 3 shows the IBD rates along the autosomes, the median, and the threshold for the EUR group. Table 2 reports the loci that have excess IBD rates extending for more than 1 cM. To estimate $\hat{s}$ for these loci, we infer the recent population effective sizes $\widehat{N}_e(t)$ using 2.0 cM segments as input to IBDNe[52]. The excess IBD rates at LCT, MHC, and TRPM1 are 35, 26, and 10 standard deviations above the population median, respectively. We do not claim that specific phenotypes are targets of selection, but we do annotate genes that have been explored by other authors.

Improved modeling of LCT selection in European ancestry populations

Lactase persistence is widely believed to be subject to a selective sweep in Europe. We view it as a positive control to evaluate our method. The putative selected allele is a regulatory mutation upstream of LCT in intron 13 of MCM6. This allele enhances the promoter of LCT,

allowing an alternative path for gene expression. Four different alleles in this intron increase lactase production *in vitro* among East African populations[57], but we concentrate in this study on the single mutation -13.910C>T that is common in Northern Europeans. For LCT we estimate $\hat{s} = 0.0330$ with 95% confidence interval $(0.02877, 0.0382)$, and $\hat{p}(0) = 0.675$ at location 135,735,797 (Table 3). Our result is close to current estimates using ancient DNA[24]. We also estimate selection coefficients separately within each moderately sized cohort BioMe, MLOF, VTE, VUAF, and WHI (Supplementary Table 3). The individual and aggregate analyses show nearly perfect concordance. The IBD entropy statistics are all equal to zero as well, meaning that there is a single haplotype cluster of excess IBD. All analyzes identify the -13.910C>T mutation at position 135,851,076 as the maximally differentiated SNP between the inferred outgroup and the rest of the sample.

The sweep model with these estimates would place the allele at 2% frequency around 120 to 163 generations ago (3.5 – 5 thousand years ago (kya); 95% confidence intervals). We prefer to report time estimates when the model reaches low frequencies because genetic drift can play a greater role than selection when the allele is rare. Figure 4 shows that the putative sweep at LCT may have progressed more rapidly than any other putative sweeps in Europeans. Our model assumes a constant selection coefficient. Klunk et al. (2022) look at ancient DNA samples from Britian and suggest instead a dramatically different model of extreme and periodic soft selective sweeps[58]. This hypothesis does align with Evershed et al. (2022) who used milk residues in archaeological pottery to argue against the hypothesis of coevolution with dairying practices[59], but the Klunk et al. data analysis has been criticized for technical and conceptual errors[60].

Possible selective sweeps around pigmentation genes

Protection from ultraviolet light exposure, vitamin absorption, or light sensitivity at new latitudes may explain why some regions of interest contain pigmentation genes. In our analysis, we make note of the same chromosome 15 region containing OCA, HERC2, and TRPM1 as in a previous report[8]. These genes all express in retinal cells and melanocytes and have been linked to eye and skin color phenotypes[61,62]. TRPM1 passes our selection scan in all six cohort analyses, and it is the third largest signal overall. Each cohort analysis ranks first the SNP at basepair 31,437,832. This SNP frequency and our estimates $\hat{p}(0)$ range from 12-15% (Supplementary Table 3). Because entropy is slightly above 0.50 for each cohort and greater than 1.0 for the whole EUR group, we suggest that the highlighted SNP may play a role in selection but that there is not a preponderance of evidence in support of the sweep model.

The neighboring genes OCA/HERC2 pass our selection scans in all but the BioMe cohort. We compute IBD entropy equal to 0.32 resulting from three large IBD groups that include 73% of the haplotypes in the EUR sample. Entropy is zero for VTE, VUAF, and WHI cohorts since there is one large outlier group. We estimate $\hat{s} = 0.0190$ with 95% confidence interval $(0.0168, 0.0212)$ and $\hat{p}(0) = 0.75$ (Table 3). The sweep model in Figure 4 implies that the allele would be at frequency 2% between 221 to 311 generations ago (~ 6.5 to 9.5 kya; 95% confidence interval). The best-ranked variant in the EUR group, rs1129038, is known to inhibit gene regulation and affect the blue eye color phenotype[63]. Three cohorts instead rank first the ~35% frequency SNP at position 28,075,818. This region contains ~200 kb of low marker coverage upstream of rs112903. Low marker density as well as small sample size can impact the precision in our methods, instead emphasizing hitchhiking variants further away the selected allele. On the other hand, this SNP may be functionally important to a selected haplotype.

Possible selective sweeps on introgressed haplotypes

Studies on archaic introgression have identified a Neanderthal haplotype about OAS genes 1, 2, and 3 as a putative target of positive selection[64,65]. For this region of interest, entropy is slightly above 0.50 for the aggregate EUR group but below 0.50 for 4 of 5 cohort studies (Table 3, Supplementary Table 3). These genes transcribe antiviral proteins and may thus be another example of possible selection for immune response. The introgressed haplotype harbors a splice variant that upregulates expression of a specific isoform[66]. Prior studies estimate this haplotype in the range of 40% frequency among modern Europeans. Applying our method, we find the same introgressed SNPs in our best evidence SNP set as in prior studies[64–66]. Haplotype structure in the OAS region differs from our simulations of a single causal allele. This figure highlights that the causal allele in our model is not necessarily a SNP variant but could be a haplotype with high linkage disequilibrium (Supplementary Figure 4). Again, all but one cohort-specific analysis and the aggregate EUR analysis estimate $\hat{p}(0)$ between 40-50%. Our estimate $\hat{s} = 0.0183$ $(0.0158, 0.0208$; 95% confidence interval) indicates the time to allele frequency 2% to be around 179 to 257 generations ago (~ 5 to 7.5 kya; 95% confidence interval) (Table 2). Our statistical method can estimate selection coefficients in such a soft sweep, but this *post hoc* -interpretation of adaptive introgression is out of its scope.

Keratin genes are expressed in epithelial cells like hair follicles. There are tens of genes in this family on chromosome 12 [67]. The KRT71 gene may be an example of adaptive introgression[65]. The aggregate EUR group and the largest cohort WHI rank first the SNP at basepair 53,083,566. We estimate $\hat{s} = 0.0207$ with 95% confidence interval $(0.0186, 0.0228)$, $\hat{p}(0) = 0.725$, and time to the frequency of 2% to be between 202 to 276 generations ago (~ 6 to 8.5 kya; 95% confidence interval) (Table 3). Figure 4 shows that putative sweeps around KRT71

and OCA/HERC2 loci may have progressed in a similar fashion. Entropy is below 0.50 for 4 of 5 analyses, and the estimates $\hat{p}(0)$ vary greatly. This observation may reflect the small sample size of some cohorts, or it may indicate selection more complex than the sweep model.

Limited evidence for the strong sweep model at many human loci

MHC and TLR1-6-10 involve families of genes relevant to immune response. Human leukocyte antigen genes encode MHC molecules which bind to antigens and mediate the presentation of these to the surface of immune cells. These proteins help cells to recognize self from foreign, thereby influencing donor compatibility for organ transplant. Toll-like cell receptors activate the innate immune system by recognizing pathogens. Prior studies suggest the selection signal in TLR1-TLR6-TLR10 may be due to an introgressed Neanderthal haplotype[66,68]. For these two regions of interest, we find cohesive IBD groups in the cohort-specific analyses, but entropy is greater than 0.50 for the aggregate EUR group (Supplementary Table 3). This insight suggests that there may be geographic differentiation[8,69] or multiple selected haplotypes or in these gene complexes. It is tempting to speculate on these loci given their functional roles, but the interpretation of a selection coefficients only makes sense in the context of the sweep model. Balancing selection is a clear violation of our model, as are multiple independent sweeps.

We give the opinion that estimating a selection coefficient involves strong belief in the sweep model. We do not estimate selection coefficients for any other loci (Table 2) because they do not have as clear a signal in our analysis. This list includes genes PRDM1, SLC45A2, and HFNB1 which have all been discussed before in selection scans with respect to their phenotypic function[8,13,24]. Interestingly, this list includes a non-annotated region at 205 Mb in chromosome 1 that we find in 5 of 6 analyses. For all these loci, we calculate IBD entropy lower than 0.50 for

some cohort-specific analyses, but entropy greater than 1.0 for the aggregate EUR analysis. We also fail to come to a consensus on allele frequency estimates between the cohort-specific analyses. This pattern mirrors our results for the MHC and TLR1-6-10 gene complexes. We acknowledge that these loci could be examples of older sweeps, sweeps with small selection coefficients, balancing selection, or other possible evolutionary processes.

Discussion

We applied new methods to detect and model very recent selective sweeps in Americans of inferred European ancestry from the TOPMed consortium project. These methods use accurately inferred identity-by-descent (IBD) tracts which can be informative about a positively selected locus. They address three major aspects of such inference: pinpointing selection signals to a resolution of 250 kb, estimating frequencies for unknown sweeping alleles, and inferring selection coefficients with uncertainty quantification. The selection coefficient estimator we propose is a function of the IBD rate, such that we can predict how it changes with different IBD rates or under different model assumptions. It is also robust to standing neutral variation at low frequency, such that we can estimate selection coefficients from soft sweeps or adaptive introgression. Across a broad range of simulation studies, we verify the statistical behaviors of our methods and juxtapose them to be favorable or competitive alongside alternative methods.

In our selection scan on European Americans, we replicate the results of previous reports from haplotype-based methods. We refine estimates of the selection coefficient for signals at the LCT, OAS1-2-3, and OCA/HERC2 genes. For instance, at the LCT gene we estimate a 95% confidence interval for the selection coefficient that is of width less than 0.01. This small uncertainty quantification is uniquely supported by robust inference and extensive testing. We

comprehensively explore multiple IBD-based statistics to gather evidence for or against the hard sweep model at other loci with excess IBD rate in our selection scan. We find that there may not be enough evidence in our dataset to justify modeling a selective sweep at MHC, TLR1-6-10, PRDM1, SLC45A2, and HFNB1, and that there is weak evidence for sweeps at TRPM1 and KRT71. We will make publicly available the variant scores in these regions of interest. Further studies could explore the extent to which balancing selection and population structure confound statistical inference of selective sweeps. Another striking observation is that we conduct IBD-based analyses on less than 2,000 samples, and we find their results to be broadly consistent with our analysis on a larger consortium dataset of 13,778 samples. The sample size necessary to have a certain level of statistical power depends on effective sizes. Population samples in the Americas may have good statistical power in our methods because population bottlenecks result in more IBD tracts. If the selection coefficient $s \geq 0.015$ for a sweep, our methods may apply to studies with small sample size.

Prior analyses of LCT selection offer either no statistical support for their confidence intervals, wide confidence intervals, or no confidence intervals at all. No confidence intervals for the selection coefficient have demonstrated 95% coverage experimentally. The 95% confidence interval in Bersaglieri et al. (2004)[70] has width greater than 0.10 whereas the 95% confidence interval in Hejase et al. (2022)[41] has width less than 0.001 versus. There is a distinction between parameter uncertainty and model uncertainty. The dropout technique in Hejase et al. (2022)[41] expresses model uncertainty, meaning that we cannot interpret their confidence intervals in their strict probabilistic formulation. Likewise, the high density posterior intervals in Torada et al. (2019)[44] express model uncertainty in that its Monte Carlo sampling from a fitted posterior distribution is a *post-hoc* model interpretation. The Bayesian approach in Stern, Wilton, and

Nielsen (2019) [33] could offer parameter uncertainty via a 95% credible interval from the posterior distribution, but the method is too slow. In contrast, our statistically valid confidence intervals support our refined results on LCT selection.

Our methods are not a panacea for all selection studies in all populations and species. The limitations of our inference go hand in hand with those of algorithms to infer IBD tracts. The method relies on accurate detection of 2 to 3 cM shared haplotype segments. Longer segments are too few to provide adequate information; inferring shorter segments may lead to many false negatives and positives. Additionally, we use these IBD tracts to infer effective population sizes, which are conditioned on to estimate the selection coefficient. These estimates of effective population size can be precise for some human populations[52,53,71]. To what extent the threshold in our selection scan controls the false positive rate we do not know. IBD rates across sliding windows can contain the same IBD segments, making derivation of genome-wide significance complicated. The Bonferroni correction we use is conservative in our case of correlated hypothesis testing.

All the results are obtained from automated bioinformatics pipelines that run the statistical methods presented here. These methods are implemented in the freely available software package iSWEEP (**i**ncomplete **S**elective **S**weep **W**ith **E**xtended haplotypes **E**stimation **P**rocedure). The most computationally intensive aspects of our methods involve the use of highly efficient hap-ibd and ibd-ends algorithms to infer IBD tracts[8,15]. Using a 24-core Intel Xeon Silver 4214 2.2 GHz compute node, wall clock computing time to detect ≥1.0 cM IBD tracts on chromosome 2 is 4.5 hours for hap-ibd and 12.5 hours for ibd-ends. These algorithms scale to the analysis of large sequence datasets which far exceed the sample sizes of existing approaches. The tool ibd-ends is especially important to adjust for genotyping errors, gene conversion, and

low marker density in sequence data[8]. Our methods do not assume causal phenotypes, nor do they require that the sweeping allele be genotyped or derived from a distant ancestor. The input is high coverage whole genome sequences as opposed to other methods based on ancient or archaic DNA. These minimal assumptions and data requirements may facilitate analysis in non-European populations or large samples of non-model organisms.

Materials and Methods

Method Details

The identity-by-descent model with selection

Let $p_s(t)$ and $N_e(t)$ be the allele frequency of the sweeping allele and effective population time at coalescent time $t > 0$ (in generations ago). The formulas define two subpopulations $2N_e(t) \times p_s(t)$ and $2N_e(t) \times (1 - p_s(t))$ that have or do not have the sweeping allele. Standard models in literature exist for $p_s(t)$ given additive, multiplicative, dominance, recessive, and heterozygote selection. We invert $p_s^{-1}(t)$ with the quadratic formula and choose the solution in the interval [0,1]. These formulas concern the mean behavior in the Wright Fisher process with selection. Namely, the random count variable $X(t) \sim Binomial(N_e(t), p_s(t-1))$ be the number of haploids with the sweeping allele. For estimating s, we use the deterministic $p_s(t)$ from the backward formulas. For bootstrap uncertainty, we use estimates $\hat{p}_s(t) = X(t)/[2N_e(t)]$ when simulating back in time. Note that the sample mean estimate $\hat{p}_s(t)$ is a maximum likelihood estimate and should be close to $p_s(t)$ for large $N_e(t)$.

Let coalescent time T and identity-by-descent (IBD) tract length L be random variables. Let c be a user-defined genetic distance threshold. Indices i, j refer to haplotypes that share IBD tracts from a common ancestor. For two randomly sampled haplotypes,

$$P(T = \tau) = \prod_{\tau=1}^{t-1} \left(1 - \frac{1}{\left(2N_e(\tau)p(\tau)\right)}\right) \frac{1}{(2N_e(t)p(t))}.$$

Given coalescent time $T = t$, the IBD tract length L follows a Gamma distribution parameterized by a function of t [52,72]. We use the function $f(t) = \frac{t}{100}$ to measure genetic distance in

centiMorgans (cM). Recombination endpoints to the right of a focal position are exponentially distributed with rate equal to $f(t)$. The shared IBD to the right of the focal position from two independent, identically distributed recombination endpoints is the minimum length. The entire IBD tract length $L \mid T = t \sim Gamma(2, 2f(t))$ is the sum of shared IBD to the left and right of the focal position. The probability $P(L \geq c) = \int \int P(T = t)P(L \geq c \mid T = t)$ is calculated using the law of total probability. In the case of a neutral biallelic marker, the probability mass for the coalescent time is

$$P(T = t) = p(0)^2 P(T_{11} = t) + \left(1 - p(0)\right)^2 P(T_{00} = t),$$

where T_{11}, T_{00} are the coalescent times for randomly sampled haplotypes with alleles 1 and 0. The tail probability $P_s(L \geq c)$ is not necessarily strictly monotone in s if $P_s(T = t)$ is not strictly monotone in s for all t. For IBD tract lengths from haplotypes in the allele 1 subpopulation with $s > 0$, their tail probability $P_s(L_{11} \geq c)$ is strictly monotone. The opposite is true in the allele 0 subpopulation: the tail probability $P_s(L_{00} \geq c)$.For large values $N_e(t)$ and length threshold c, we argue that $P_s(L \geq c)$ to be strictly monotone in s because IBD is rare and s increasing leads to more IBD in the allele 1 subpopulation than is lost in the allele 0 subpopulations. One can verify this condition numerically given fixed $p(0)$.

Selection coefficient estimation

Let $L_{i,j}$ be the IBD tract length shared by haplotypes i and j in a sample of n, and $Y_{i,j} = I(L_{i,j} \geq c)$ is the binary variable denoting if $L_{i,j}$ is longer than threshold c. The empirical tail probability $\hat{P}_s(L \geq c) = \binom{n}{2}^{-1} \sum_{i<j} Y_{i,j}$ is always an unbiased estimator of $P_s(L \geq c)$. We want even stronger statistical properties for estimation. Consider the composite likelihood assumption that IBD tracts $(Y_1, Y_2, \dots, Y_M)$ are independent, identically distribution Bernoulli random

variables with probability parameter $P_s(L \geq c)$. The sample mean $\hat{P}_s(L \geq c) = \bar{Y}_{1:M}$ is then the maximum likelihood estimate (MLE) for the binomial experiment. We highlight three statistical properties of MLEs. First, consistency means that the estimator gets arbitrarily close to the parameter as sample size M increases. Second, sufficiency means that $\bar{Y}_{1:M}$ provides all relevant information to estimate $P_s(L \geq c)$. Third, the MLE is normally distributed in large samples. If $\hat{\theta}$ is the MLE of parameter θ, then $f(\hat{\theta})$ is the MLE of parameter $f(\theta)$[73]. These statistical properties we discuss as they relate to our selection coefficient estimator which is a function of the IBD rate.

We propose the value $\hat{s}$ that solves the equation $P_{\hat{s}}(L \geq c) = \bar{Y}_{1:M}$ as our estimator. Because inverting $P_s(L \geq c)$ is analytically intractable, we optimize using scipy[74] the expression

$$\hat{s} = \arg\min s \left(\binom{n}{2}^{-1} \sum_{i<j} I\left(L_{i,j} \geq c\right) - P_s(L \geq c) \right)^2 .$$

Our estimator is approximate in the sense that IBD tracts are not independent binary variables with probability mass $P_s(L \geq c)$. In simulation study, we find that $\hat{s}$ is close to the true value and appears to be normally distributed (Supplementary Figure 6). We also observe that the inferred counts of IBD tracts greater than the 2.0 cM threshold in the EUR data analysis appear to be normally distributed (Supplementary Figure 6). These arguments suggest that our estimator behaves like a maximum likelihood estimator; that is, it should be consistent, sufficient, and normally distributed as sample size gets large. Deep learning methods say that, instead of low-dimensional statistics, we require high-dimensional data to infer selection coefficients [41,44,75]. This assertion is not true for sufficient statistics. Given $\widehat{N}_e(t)$ and $\hat{p}(0)$, under the composite likelihood the IBD rate is a 1-dimensional sufficient statistic.

We provide a parametric bootstrap approach to give 95% confidence intervals[76,77]. The parametric bootstrap draws realizations from a probability model conditional on a parameter estimate, and it then re-estimates the parameter with the bootstrap sample. Simulating tracts independently undersells variance in the IBD tract length distribution (Supplementary Figure 6), affecting the width of confidence intervals. We present an algorithm to generate IBD segments from the coalescent given parameter estimates for the hard sweep.

Algorithm 1.

1. Initialize all nodes i to have recombination endpoints $l_i, r_i = \infty$.
2. Simulate a conditional coalescent tree given $s, p(0), N_e(t)$, and a genetic model.
3. At each coalescence to a common ancestor:
    a. Draw left and right recombination endpoints $l_I, r_I \sim Exp((t_k - t_I)/100)$ where t_I, t_K is the time ago for some incoming interior node and the common ancestor.
    b. Update $t_I = t_K$, $r_i = \min(r_i, r_I)$, $l_i = \min(l_i, l_I)$.
    c. Compute $r = \min(r_i, r_j)$ and $l = \min(l_i, l_j)$ for sample nodes i, j under interior nodes I, J.
    d. If $r + l \geq c$ for cM threshold c, call an IBD segment between sample nodes i, j.
4. Stop once there is a single most recent common ancestor.

We simulate the conditional coalescent tree using a discrete time Wright Fisher process with selection. This exact solution as opposed to a coalescent approximation is essential given that those sample nodes with the *de novo* mutation coalescence to a single founder in recent history.

We implement modifications to this algorithm to improve run time. Most sample nodes share very short haplotypes with their common ancestors after tens of generations, so we stop

comparing against a node i once $r_i + l_i < c$. If the lengths $r_i = r_j$ and $l_i = l_j$ for sample nodes i, j, merge these nodes and compared the merged node against other sample nodes in the preceding steps. This subroutine avoids redundant calculations. We observe in simulations that these implementations lead to approximately linear instead of quadratic run time. This algorithm makes it feasible to form bootstrap intervals from two hundred draws for sample size as large as 128,000 diploids.

We use a sample standard deviation $\hat{\sigma}_K$ from K parametric bootstraps to form a confidence interval $\hat{s} \pm z_{1-\alpha/2} \times \hat{\sigma}_K$, where $z_{1-\alpha/2}$ is the normal quantile. We do not implement bootstrap bias-correction[78] because the optimization procedure can be unstable for small s. This instability is due to few additional IBD tracts being observed between small s and the neutral model. This scenario is circumvented by implementing our selection scan beforehand. The apparent normal distribution in Supplementary Figure 6 would explain why the 95% confidence intervals in this our work maintain adequate coverage over the selection coefficient s.

Pinpointing the focal position of the sweeping allele

Genomic regions passing our selection scan encompass many Mb of genetic sequence. The IBD outgroup algorithm is sensitive to the initial location of the IBD detection. First, we narrow in on smaller regions of limited sequence variation. Field et al. (2016) [79] study regions where there is less distance between singleton variants compared to a genome-wide average. They motivate their approach with the observation that the coalescent with strong positive selection has long terminal branches which pick up probabilistically more *de novo* mutations. Instead of modeling rare sequence variation, our approach integrates over common sequence variation. Our motivation is that variants hitchhiking on the causal allele should segregate at similar frequency, suppressing common variation locally about the causal position

(Supplementary Figure 5). We take a sliding window approach to localize the signal. Consider $Q_w = \sum_{k=1}^{10} p_{(k)}$ where is $p_{(k)}$ is the kth percentile of allele frequencies in the window w. High Q_w values correspond to windows where common variation is greater than that of an allele frequency spectrum at neutrality. We use window sizes and step sizes of 250 kb and 50 kb, respectively. To address low marker density regions, we do not consider windows where there is a gap of 25 kb with no SNPs. The window with the highest Q_w is a refined focal location for further steps in the estimation procedure.

Inferring an outgroup

We develop a procedure to infer an outgroup A with excess IBD sharing to contrast with the rest of the sample B. First, we detect IBD segments with length 1.0 cM or greater that overlap a focal location. Let haplotypes be nodes and IBD segments be edges in the context of a network graph. Because extended IBD is infrequent in a population sample, this graph should contain many disconnected haplotype clusters[80]. Sizes of these haplotype clusters $C_1, C_2, \ldots$ amount to an empirical distribution for IBD cluster size. We compute a sample mean $\bar{C}$ and sample standard deviation $\hat{\sigma}_C$ for cluster size. Then, we identify clusters whose size exceeds $\bar{C} + 3 \times \hat{\sigma}_C$. Let these inferred clusters $\hat{a}$ form the outgroup $\hat{A}$. The rest of the sample we denote as $\hat{B}$. The rationale for this subroutine is that haplotypes with the causal allele descend from the mutant founder haplotype. If a selective sweep is recent and strong, IBD tracts longer than 1.0 cM may be preserved, leading to much larger cluster sizes.

Estimating the sweep frequency

For each variant j, we compute the scaled difference in proportions. Let $p_j \coloneqq p_j(0)$ be the current allele frequency in the entire sample, and $p_{j,\hat{A}}$ is the allele frequency in the inferred outgroup. The scaled difference in proportions in the main text is:

$$t_j = \frac{p_{j,\hat{A}} - p_{j,\hat{B}}}{\sqrt{p_j(1-p_j)}}$$

Except for a constant scalar, these scores represent the Student t-statistic in the two-sample difference in proportions test. They are similar to the DAF metric in Grossman et al. (2010) [23]. However, as opposed to assuming a derived allele, we determine *post-hoc* which allele is more frequent in the inferred outgroup $\hat{A}$. One can sort SNPs based on our t_k scores. This ordering ranks SNPs according to their evidence as selected or in strong linkage with a selected allele.

We aggregate evidence across variants to estimate a haplotype frequency. We take a sliding window approach in terms of physical distance and allele frequency. We consider rectangles of 10% length and 500 kb width, and we move them by increments of 2.5% and 50 kb. For each rectangle, we average t_k variant scores contained in it. To address sparse marker coverage, we insist that rectangles include at least five markers. The rectangle with the highest t_k average is our best subset of candidate SNPs. Its midpoints serve as our final estimates for allele frequency $\hat{p}(0)$ and location $\hat{l}$.

An information entropy criterion supporting the hard sweep model

Loci that pass a selection scan may not be the result of a selective sweep. Our method and others look for patterns of extended haplotype sharing. However, other processes may give rise to such an observation. For instance, under balancing selection there could be multiple fit haplotypes that maintain appreciable frequency. In the recent and strong sweep model, we expect one predominant group of extended haplotype homozygosity. Recall $\hat{a}$ denotes IBD clusters in the inferred outgroup $\hat{A}$. Let $q_{\hat{a}}$ be the proportion of haplotypes in cluster $\hat{a}$ out of $\hat{A}$. The information entropy statistic referred to in the main text is then:

$$IBD\ entropy := \sum_{\hat{a}} -q_{\hat{a}} \log_{10} q_{\hat{a}} \ .$$

If $\hat{A}$ consists of one cluster, the entropy is 0.00. If $\hat{A}$ consists of two equally sized clusters, the entropy is 0.30. If $\hat{A}$ consists of three equally sized clusters, the entropy is 0.52. In general, entropy increases with more clusters. We use the criterion entropy less than 0.50 to give supporting evidence for the hard sweep model. We observe that most replicates in our simulation study have entropy less than 0.50 (Supplementary Figure 3). The outgroup $\hat{A}$ should be of appreciable size as well in a hard selective sweep. Thus, we consider the outgroup $\hat{A}$ being at least 5% as further support for the hard sweep model. Human loci that do not pass these thresholds have weak support for the hard sweep model and may be the result of another biological process.

Implementation of all methods

1. Infer outgroups $\hat{A}_1$ and the rest of the sample $\hat{B}_1$ about the location of max IBD rate.
2. Calculate Q_w for sliding windows of size 250 kb and step 50kb.
3. Infer outgroups $\hat{A}_2$ and the rest of the sample $\hat{B}_2$ about the location of max Q_w.
    a. Compute IBD entropy given $\hat{A}_2$ and $\hat{B}_2$.
4. Compute t_j for markers j based on $\hat{A}_2$ and $\hat{B}_2$.
5. Average t_j scores across sliding windows of allele frequency and physical distance.
6. Estimate the selection coefficient conditional on the allele frequency estimate.
7. Form 95% confidence intervals from the parametric bootstrap.

Statistical Analyses

Demographic Scenarios

We consider three demographic models in our simulation studies. We refer to these demographic scenarios as bottleneck (BN), super-exponential growth (SE), and constant size of 25,000 diploids (C25). These parameters refer to the effective population size $N_e(t)$ which is the number of haploids in an idealized population. The bottleneck demography model has an ancestral size 10,000 diploids, 2% exponential growth beginning three hundred generations ago, an instantaneous drop at 20 generations ago to effective size one million, and then continued 2% exponential growth. The super-exponential growth model has an ancestral effective size 5,000 diploids and 1%, 7% and 15% exponential growth beginning 300, 60, and 10 generations ago. We design these $N_e(t)$ to resemble estimates of European-American and White British effective sizes[52,53].

Simulation study on IBD from the coalescent

We argue *in silico* that our selection coefficient estimator $\hat{s}$ is approximately unbiased and consistent, and that the 95% confidence interval has proper coverage as the number of parametric bootstraps increases. These properties depend on correct model specification and perfect detection of IBD segments. First, we define a general model with the following parameters: $s = 0.030$, $p(0) = 0.50$, $N_e(t)$ is the population bottleneck demographic model, the multiplicative genetic model is $1{:}\,(1+s){:}\,(1+s)^2$, and the selective sweep is ongoing and from *de novo* mutation. All subsequent models derive from this model. Unless otherwise specified, we use IBD segments ≥ 3.0 cM and estimate 95% confidence intervals from fifty bootstraps. Bias is the estimate subtracted by the true parameter. Width is the (absolute) difference between the left

and right endpoints of the confidence interval. Coverage is if the bootstrap interval contains the true s. These summary statistics are averaging over 10,000 replicates.

This analysis involves simulating true IBD segments from a coalescent tree and estimating the selection coefficient conditional on model specification. We perturb one parameter at a time from the general model and evaluate its impact on our inference method. We consider the following choices: $s \in [0.02, 0.03, 0.04]$; $p(0) \in [0.25, 0.50, 0.75]$; bottleneck (BN), super-exponential growth (SE), and constant size (C25) demographic scenarios. For these simulations, we estimate conditional on correct model specification. Table 1 shows empirical unbiasedness, consistency, and 95% coverage for these settings.

Next, we evaluate how model misspecification impacts our inference method. Here, we consider the following perturbations: mis-estimated allele frequency or $N_e(t)$, and standing variation from 1%, 2%, and 5%; a partial sweep that ended 20 or 50 generations ago. We simulate from these models but estimate conditional on the general model. For mis-specified demography, 1.1x and 0.9x denote uniform scaling of $N_e(t)$. The first and second columns of Supplementary Table 1 show the true perturbed parameter and the assumed value for conditional estimation. Biased estimation of $\hat{p}(0)$ or $\widehat{N}_e(t)$ can skew the estimate $\hat{s}$ in a predictable direction. These two parameters reflect the structure of the coalescent with selection just as the selection coefficient does. For instance, smaller $\widehat{N}_e(t)$ will underestimate $\hat{s}$ to fit the hard sweep model to the fixed IBD rate, and vice versa for large $\widehat{N}_e(t)$. The estimator is robust to low levels of standing variation (1%, 2%, or 5%) because it fits to a model of IBD in the last 100 or so generations, whereas some approaches estimate standing variation parameters[81]. We observe unbiased estimates and adequate 95% coverage in this misspecification. If the sweep ended recently (20 generations), there can be marginal effects to bias and coverage. The fact that the

IBD rate is a sufficient statistic means that we can use its simplicity instead of a discretized length distribution[52,72]. Because $\hat{s}$ is a function of the IBD rate, its behavior can be interpreted directly in terms of how model misspecification affects the coalescent and the IBD length distribution.

Simulation study on inferred IBD tracts

We generate sequence data for five thousand diploid individuals using SLiM[37,39] and msprime[38]. Parameters common to all simulations are constant mutation and recombination rates of 1e-8, gene conversion rate 2e-8, mean gene conversion tract length 300 bp, and 0.0002 genotyping rate. At the coalescent time specified in Supplementary Table 2, we place a *de novo* sweep mutation at the center of an 8 cM genomic region. We focus on partial sweeps, so we re-run simulations if the *de novo* mutation is not segregating between 10 and 90% allele frequency in the present day. For each experiment, we report averages and coverage proportions from one hundred replicate simulations.

We use two existing software to infer IBD tracts: hap-ibd[15] and ibd-ends[8]. To detect IBD segments ≥ 3.0 cM, we find candidate segments with hap-ibd and then refine their endpoints with ibd-ends. For hap-ibd, we use min-seed=0.5, min-ext=0.2, min-out=1.0, min-maf=0.4; for ibd-ends, we use min-maf=0.001, err=2e-4, and the posterior median endpoints. These settings aim to minimize false positive and negative IBD calls. To detect IBD segments ≥ 1.0 cM, we use hap-ibd with min-seed=0.5, min-ext=0.2, min-out=1.0, min-maf=0.1. These settings find more IBD calls at the cost of accuracy. VCF files have the true phase from the simulated data. These simulations and their analysis are automated in a bioinformatics pipeline[82].

Methods for identifying selected alleles

Our methods estimate the frequency and location of a sweeping allele and rank SNPs based on their evidence for being the selected allele. The first feature is unique in that no existing software makes these estimates without fixing an allele as the causal allele. As a result, we cannot provide a comparison against any current methods. One reason this works is because we have high coverage sequence data. Numerous methods exist to rank plausible sweep variants. We compare against iSAFE, iHS, and nSL using their default settings and as input a phased VCF with 1.2 Mb to the left and right of the sweeping allele. These three methods are implemented with their default settings in selscan[12]. This comparison is unfair to iSWEEP because the method must initially infer the focal region of the sweeping allele. For estimating the aggregate allele frequency and location, it is important that iSWEEP scores high variants in strong linkage disequilibrium with the sweeping allele.

Methods for estimating selection coefficients

Few methods exist to estimate the selection coefficient with only modern genetic sequences. We compare against CLUES[33] and ImaGene[44] which optimize a likelihood for the ancestral recombination graph (ARG) and train a neural network, respectively. To apply these methods, we downsample to five hundred of the five thousand diploids due to limitations in computer memory and runtime. CLUES uses importance sampling from a distribution on ARGs. To sample the posterior distribution, we run Relate[29] with 500 of the 5000 diploids, a buffer of 100 kb to the left and right of the sweep allele, the true population effective sizes, and 500 MCMC samples for the branch lengths. We choose these parameters to provide as much data to CLUES while keeping compute times within 1 day for each replicate. We run CLUES using default parameters, MCMC samples of the inferred ARG, and the exact allele frequency of the sweep mutation. We simulate data for ImaGene using the msms software[83]. We simulate using

the exact same genetic and demographic settings. We considered selection coefficients between 0.00 and 0.05 discretized every 0.0005. In total, we simulated two million replicates uniformly proportioned among the different selection coefficients. These simulations required 98 Gb of computer storage in msms format. We split the samples into 80% training and 20% validation for the deep learning convolutional neural network. Otherwise, all details are as in the ImaGene paper[44]. We do not compare against method Selection Inference using the Ancestral recombination graph (SIA) because of limitations in computer memory. They train a neural network model on one million simulated ARGs, which amount to 1.6 terabytes of storage[34]. The compressed tree objects we simulate amount to roughly thirty megabases memory, which sums to thirty terabytes of compressed storage if done two million times.

Real Data Analysis

The hard sweep model we assume has no population structure or familial relatedness. We subset the original 38,387 whole genome sequences to address these model misspecfications. We conduct a prinicpal components analysis utilizing the GENESIS and SNPRelate software programs[84,85]. Based on data visualization, we determine bounds on PCs 1-4 to define the EUR and non-EUR ancestry groups. We approach clustering this way because there is a European outgroup and many admixed samples in the TOPMed dataset. We use the 1000 Genomes and Human Genome Diversity Panel (HGDP) datasets to perform supervised global ancestry inference with ADMIXTURE[46]. The CEU, CHB, GIH, and YRI samples are used from 1000 Genomes and 61 indigenous American samples are used from HGDP. For the EUR group, global ancestry with respect to the CEU reference is minimum 73%, interquartile range (90%, 97%), and mean 93%. Other groups of common continental ancestry do not have enough samples for the IBD-based analyses explored here. Lastly, we remove samples with kinship coefficients ϕ

greater than 0.125 due to close familial relatedness[56]. How we do this is by forming a graph edge if haplotypes i, j have $\phi_{i,j} \geq 0.125$ and randomly keeping one haplotype from each connected component.

After applying these filters, the EUR ancestry group consists of 13,778 sample individuals. There are 4172 White, 9583 Other, and 23 non-White samples from self-report surveys. From cohort studies there are 7782 WHI, 2157 MLOF, 1217 VTE, 1108 BioMe, 976 VUAF, 345 FHS, 287 CCAF, and 1 HyperGen individuals in the EUR group. Since some of the cohort studies in the EUR group have sample sizes around one thousand or more, we conduct individual analyses to further support the selection scan results.

In IBD calling, we use the same algorithm parameters as in our simulation study to call segments longer than 1.0 and 3.0 cM. We call 2.0 cM segments with ibd-ends as well to have more power in the selection scan. We use the 2019 recombination map from deCODE Genetics in Iceland to convert physical distance to genetic distance[86]. To infer the nonparametric effective size, we use these 2.0 cM segments as input to IBDNe with default settings[52]. To estimate the selection coefficient, we use this $\widehat{N}_e(t)$ which better infers recent selection as opposed to sequentially Markovian coalescent approaches[87,88]. We check that there are similar inferred effective population sizes for the European American cohorts.

For our selection scan, before the procedure described in the main text, we remove IBD data where counts are three standard deviations below the median. This quality check takes out data near telomeres and centromeres where IBD calling is difficult. We use the median instead of mean because it is a measure of central tendency less affected by outliers. Next, we do two sanity checks. First, we look if any significant loci are in regions that are known to be difficult to

sequence, because bioinformatics issues could affect IBD calling. To our knowledge, none of the loci we discuss fall into this category. Second, in inferring outgroups $\hat{A}$, we verify that they contain samples from multiple cohort studies, as opposed to being effects of cohort-specific ascertainment. To estimate selection coefficients, we assume the additive genetic model $1:(1+s):(1+2s)$. To display allele frequencies in Figure 4, we draw from a random walk parameterized Binomial($\widehat{N}_e(t), p_{\hat{s}_k}(t)$), where $\hat{s}_k$ is an estimate from the kth bootstrap.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG005701. S.D.T. acknowledges funding support from the National Defense Science and Engineering Graduate Fellowship and National Institute of Health T32 GM081062 Pre-doctoral Training Grant in Statistical Genetics.

Molecular data for the Trans- Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). Core support including centralized genomic-read mapping and genotype calling along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL- 117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample- identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I).

We gratefully acknowledge the studies and participants who provided biological samples and data for the TOPMed project. Funding for the Barbados Asthma Genetics Study was provided by National Institutes of Health (NIH) R01HL104608, R01HL087699, and HL104608 S1. The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by funds from the NHLBI and the National Human Genome Research Institute (NHGRI) (U01HG00638001; U01HG007417; X01HL134588); genome sequencing was funded by contract HHSN268201600037I. The Cleveland Clinic Atrial Fibrillation study was supported by NIH grants R01 HL 090620 and R01 HL 111314, the NIH National Center for Research Resources for Case Western Reserve University and Cleveland Clinic Clinical and

Translational Science Award UL1-RR024989, the Cleveland Clinic Department of Cardiovascular Medicine philanthropy research funds, and the Tomsich Atrial Fibrillation Research Fund; genome sequencing was supported by R01HL092577. The Framingham Heart Study was supported by contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the NHLBI and grant supplement R01 HL092577-06S1; genome sequencing was funded by HHSN268201600034I and U54HG003067. The Hypertension Genetic Epidemiology Network Study is part of the NHLBI Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472, U01 HL054473, U01 HL054495, and U01 HL054509; genome sequencing was funded by R01HL055673. The Jackson Heart Study is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from NHLBI and the National Institute for Minority Health and Health Disparities (NIMHD); genome sequencing was funded by HHSN268201100037C. The My Life, Our Future samples and data are made possible through the partnership of Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, and Bioverativ; genome sequencing was funded by HHSN268201600033I and HHSN268201500016C. The Venous Thromboembolism project was funded in part by grants from the NIH, NHLBI (HL66216 and HL83141) and the NHGRI (HG04735). The Vanderbilt Genetic Basis of Atrial Fibrillation study was supported by grants from the American Heart Association (EIA 0940116N), and grants from the National Institutes of Health (HL092217, U19 HL65962, and UL1 RR024975), and by CTSA award (UL1TR000445) from the National Center for Advancing Translational Sciences; genome

sequencing was funded by R01HL092577. The Women's Health Initiative program is funded by NHLBI through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005; genome sequencing was funded by HHSN268201500014C.

Software Resources

Name	Description	Citation	URL
ADMIXTURE	Global ancestry inference	[46]	http://dalexander.github.io/admixture/
bcftools	Utilities for VCF files and other genetic data	[89]	https://samtools.github.io/bcftools/
Beagle	Haplotype phasing	[55,90,91]	https://faculty.washington.edu/browning/beagle/beagle.html
CLUES	Method to estimate selection coefficient	[33]	https://github.com/standard-aaron/clues
GENESIS	Principal components analysis and GWAS	[84]	https://www.bioconductor.org/packages/release/bioc/html/GENESIS.html
hap-ibd	Fast detection of IBD in genetic sequences	[15]	https://github.com/browning-lab/hap-ibd
ibd-ends	Refines IBD segment endpoints in genetic sequences	[8]	https://github.com/browning-lab/ibd-ends
IBDkin	Relatedness estimation using IBD segments	[56]	https://github.com/YingZhou001/IBDkin
IBDNe	Estimation of nonparametric effective sizes using IBD segments	[52]	https://faculty.washington.edu/browning/ibdne.html
ImaGene	Deep learning for selection	[44]	https://github.com/mfumagalli/ImaGene
iSAFE	Method to rank SNPs in sweep study	[22]	https://github.com/alek0991/iSAFE
matplotlib	Plotting in Python		https://matplotlib.org/
msms	Coalescent simulations	[83]	https://www.mabs.at/publications/software-msms/
msprime	Coalescent simulations	[38]	https://github.com/tskit-dev/msprime
networkx	Tools for graphs		https://networkx.org/
NumPy	Scientific utilities in Python		https://numpy.org/
pandas	Tabular data utilities in Python		https://pandas.pydata.org/
pyslim	Utilities for tree sequences		https://github.com/tskit-dev/pyslim
Relate	Infer ancestral recombination graph	[29]	https://myersgroup.github.io/relate/
scipy	Scientific utilities in Python	[74]	https://scipy.org/
seaborn	Plotting in Python		https://seaborn.pydata.org/
selscan	Analyzing selection	[12]	https://github.com/szpiech/selscan
SLiM	Simulating genetic data	[39]	https://messerlab.org/slim/
snakemake	Bioinformatics pipelines	[82]	https://snakemake.github.io/
SNPRelate	Genetics toolkit for relatedness and PCA	[85]	https://github.com/zhengxwen/SNPRelate
tskit	Utilities from tree sequences	[34]	https://github.com/tskit-dev/tskit
tszip	Data compression for tree sequences		https://pypi.org/project/tszip/
USCS Genome Browser	Looking up genomic regions		https://genome.ucsc.edu/

The iSWEEP software is freely available under the open source CC0 1.0 Universal License (https://github.com/sdtemple/isweep). This includes a package and multiple automated bioinformatics pipelines written in Python and snakemake.

Author Contributions

S.D.T. planned the study, developed the methods, wrote the software, conducted the analyses, and wrote the manuscript. S.R.B. proposed the study and advised on all aspects of the project, including the methodology and analysis. R.K.W. helped with software development and provided advice for the direction of the study. All authors contributed to editing the manuscript.

Declaration of Interests

The authors declare no competing interests.

Figures

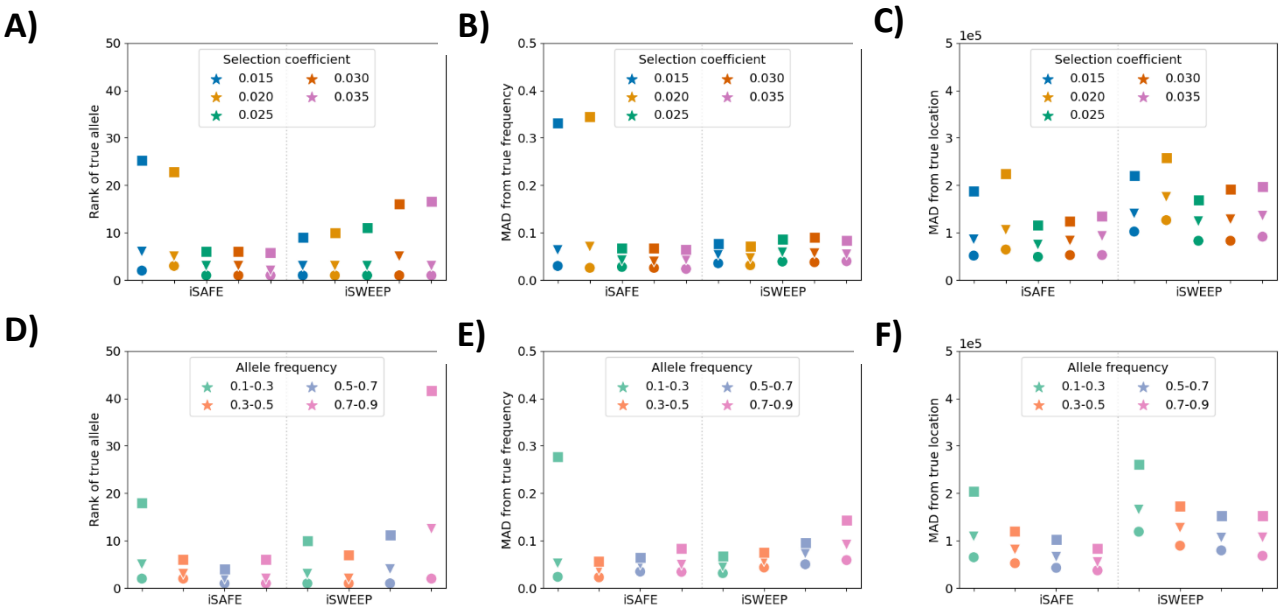


Figure 1. Identifying the sweeping allele and those alleles correlated with it

Comparison among haplotype-based methods iSWEEP and iSAFE. (A-C) Performance for varying selection coefficients and (D-E) performance for varying allele frequencies. (A, D) Ranking of the true simulated allele, where one is the best rank. (B, E) Mean absolute deviation (MAD) in frequency over lowest ten sorted alleles. (B, E) MAD in location over lowest ten sorted alleles. All results aggregated over two hundred simulations of five thousand diploids from bottleneck effective sizes. Circle is 20th percentile, triangle is 50th percentile, and square is 80th percentile.

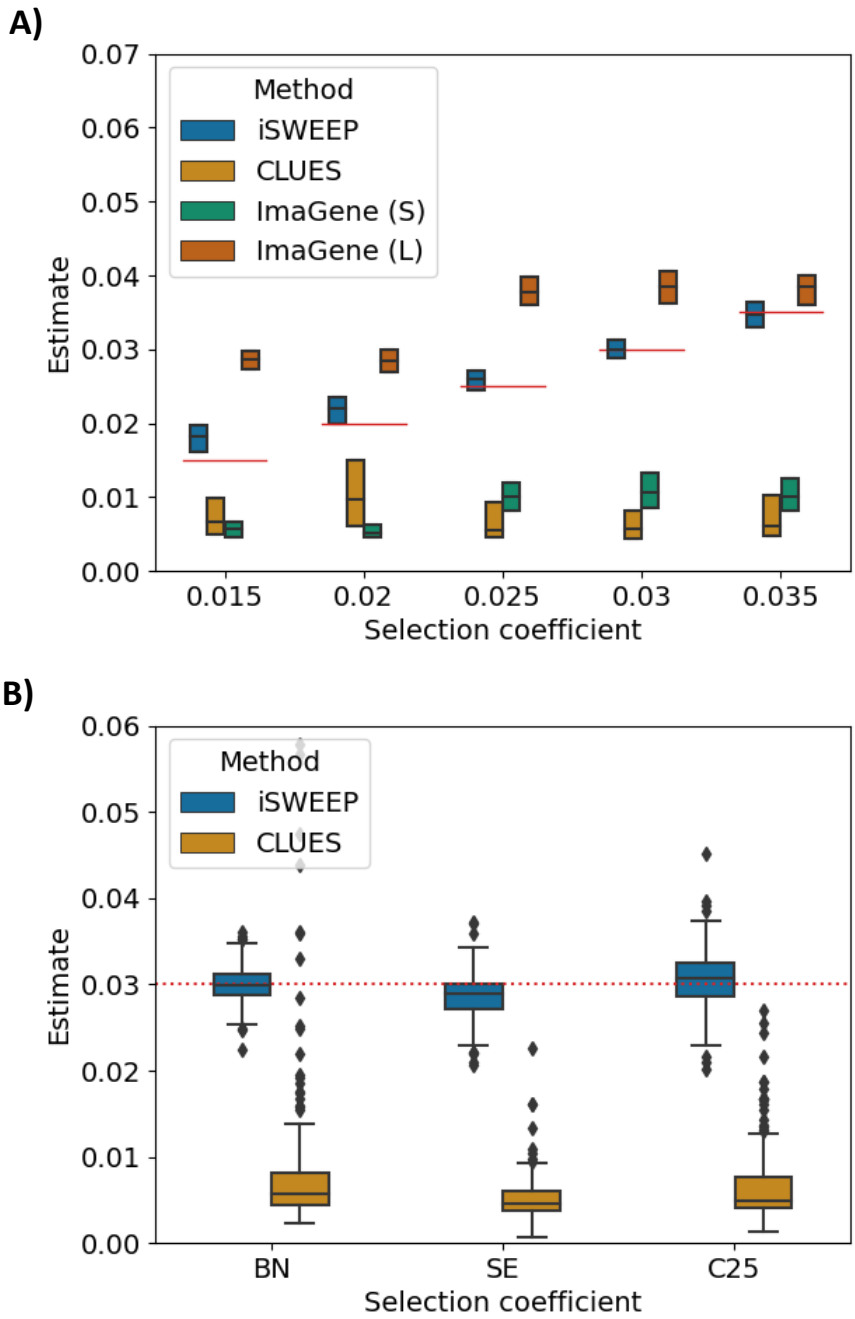


Figure 2. Robust estimates of the selection coefficient from inferred IBD tracts

(A) 25-75 percentile boxes for selection coefficient estimates of iSWEEP, CLUES, and ImaGene trained on 100,000 (S) or two million (L) simulated datasets. (B) Boxplots for selection coefficient estimates of iSWEEP and CLUES given population bottleneck (BN), super-exponential growth (SE), constant size $N_e = 25{,}000$ (C25) effective sizes. Horizontal red lines correspond to the true selection coefficient.

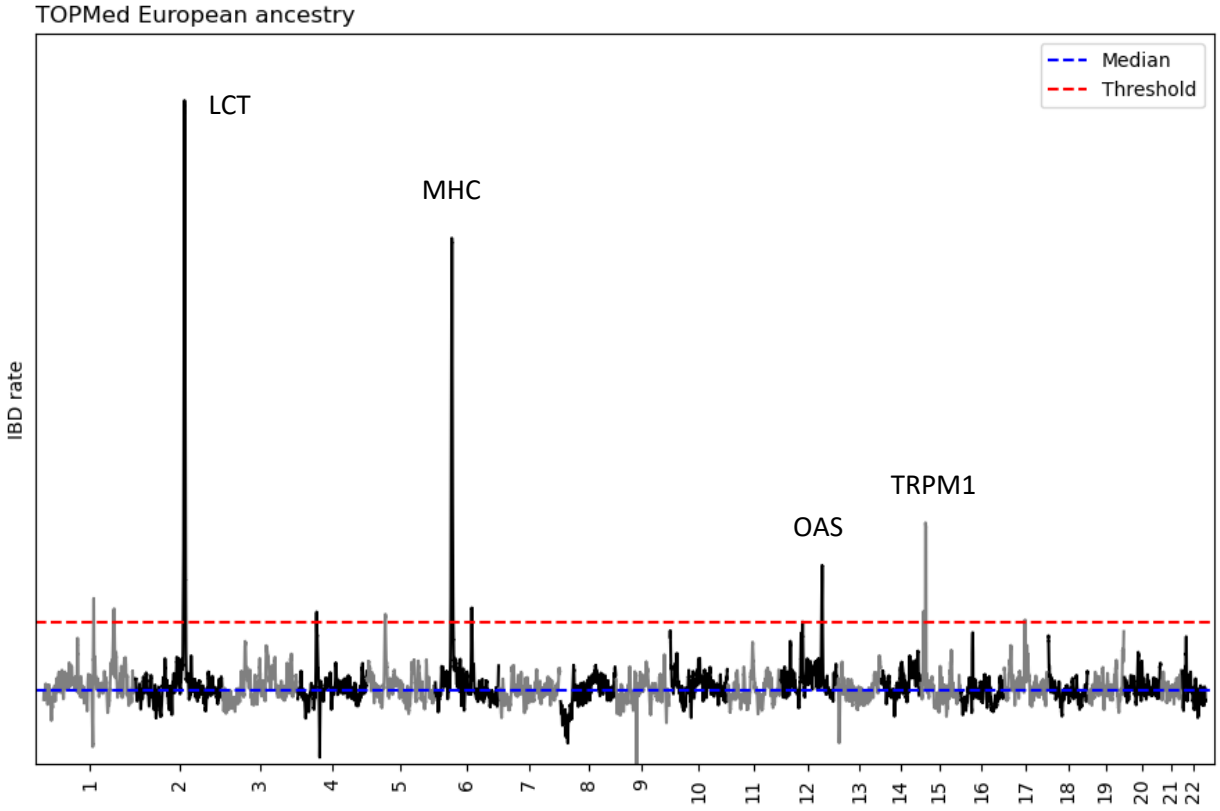


Figure 3. Selection scan in TOPMed European Americans (EUR)

Identity-by-descent (IBD) segments longer than 2.0 cM were inferred using hap-ibd[15] and ibd-ends[8]. The rate of IBD in the population sample is computed every 20 kb. The median (blue) plus four times the standard error (red) are presented as horizontal lines. LCT, MHC, OAS, and TRPM1 loci are annotated.

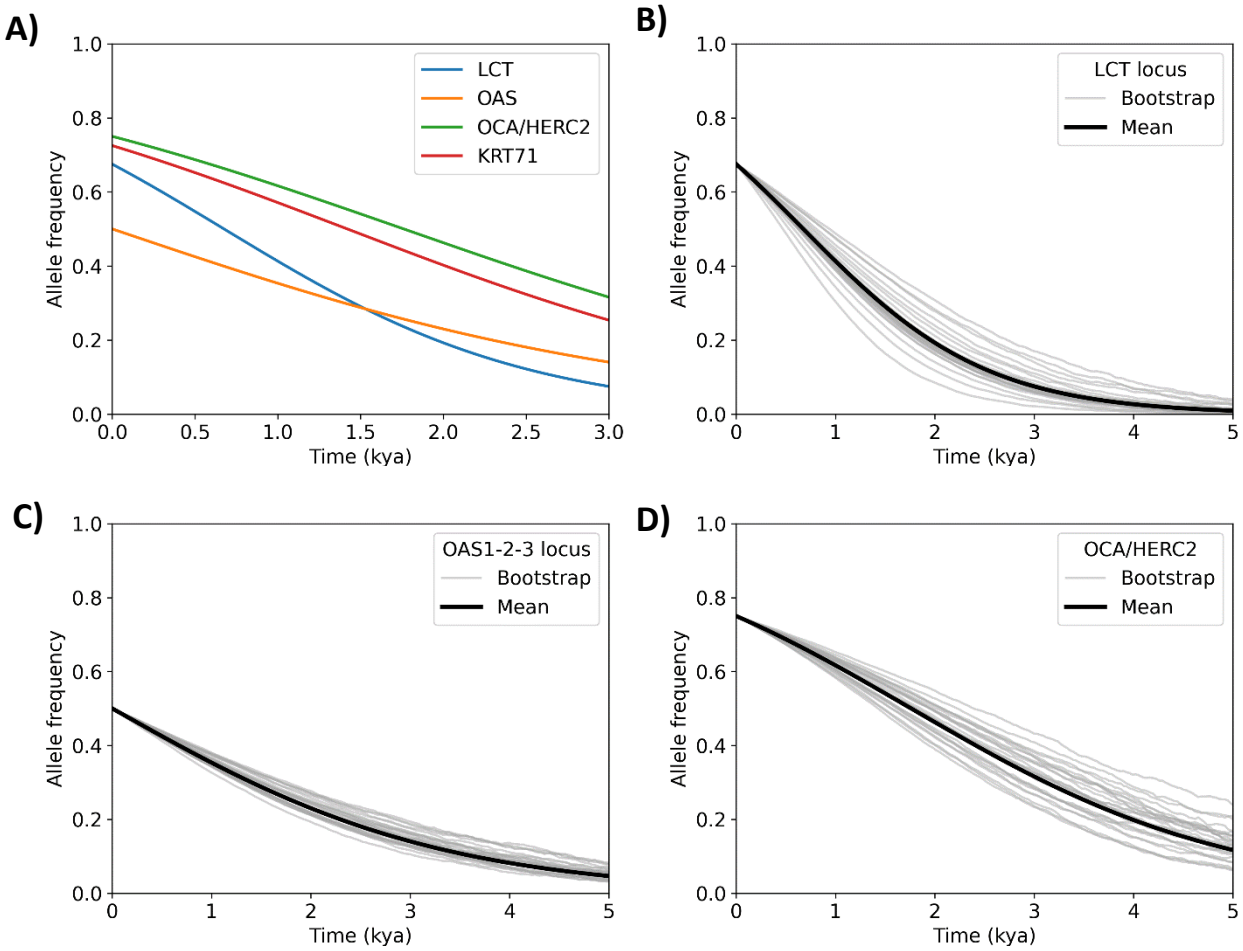
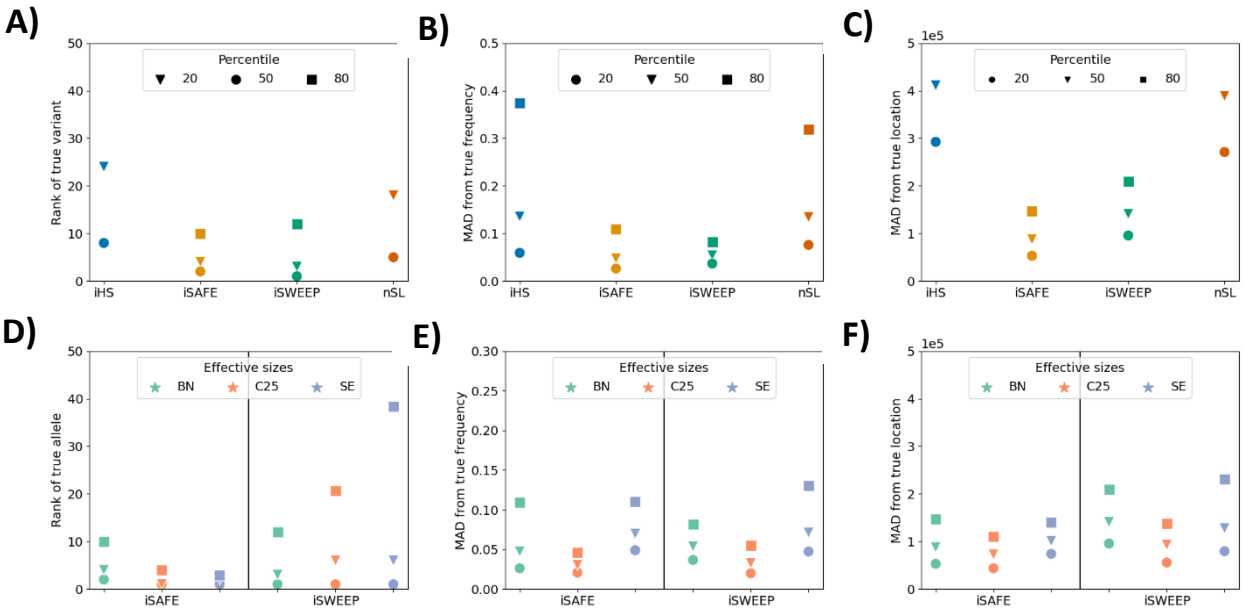


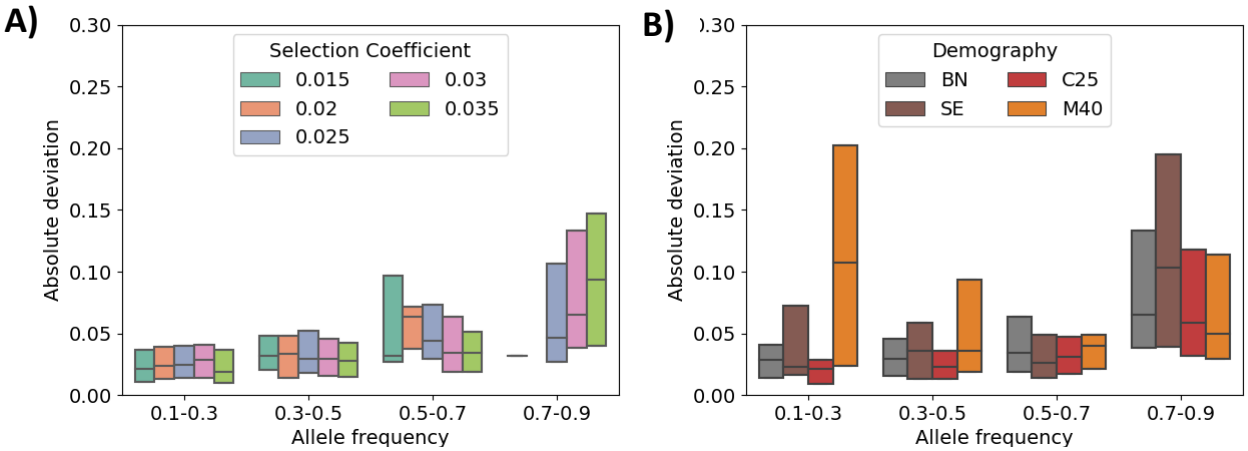
Figure 4. Historical allele frequencies in last 100 generations for human loci under selection

(A) Historical allele frequency estimates at LCT, OAS1-2-3, OCA/HERC2, and KRT71 for the EUR ancestry analysis. Parametric bootstraps shown (B) LCT, (C) OAS1-2-3, and (D) OCA/HERC2 loci. One hundred generations (100 generations = 3 kya) correspond to the resolution of 3.0 cM IBD tracts. We assume the selective sweep model and extrapolate results the past five thousand years (5 kya).



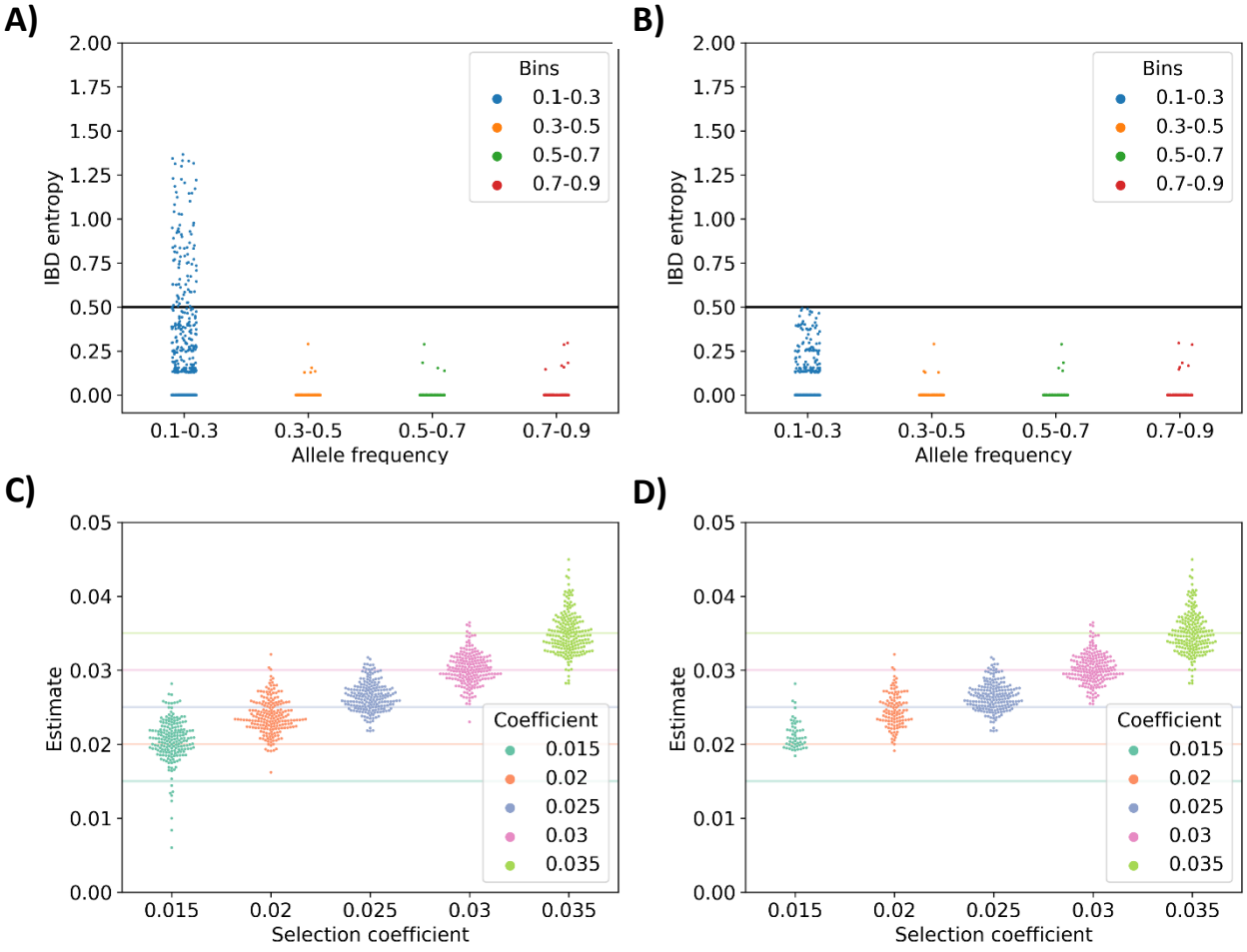
Supplementary Figure 1. Identifying the sweeping allele and those correlated with it

(A-C) Performance between four haplotype-based methods and (D-E) performance for effective sizes bottleneck (BN), super-exponential (SE), and constant $N_e = 25{,}000$ (C25). (A, D) Ranking of the true simulated allele, where one is the best rank. (B, E) Mean absolute deviation (MAD) in frequency over lowest ten sorted alleles. (B, E) MAD in location over lowest ten sorted alleles. All results aggregated over two hundred simulations of five thousand diploids. Circle is 20th percentile, triangle is 50th percentile, and square is 80th percentile.



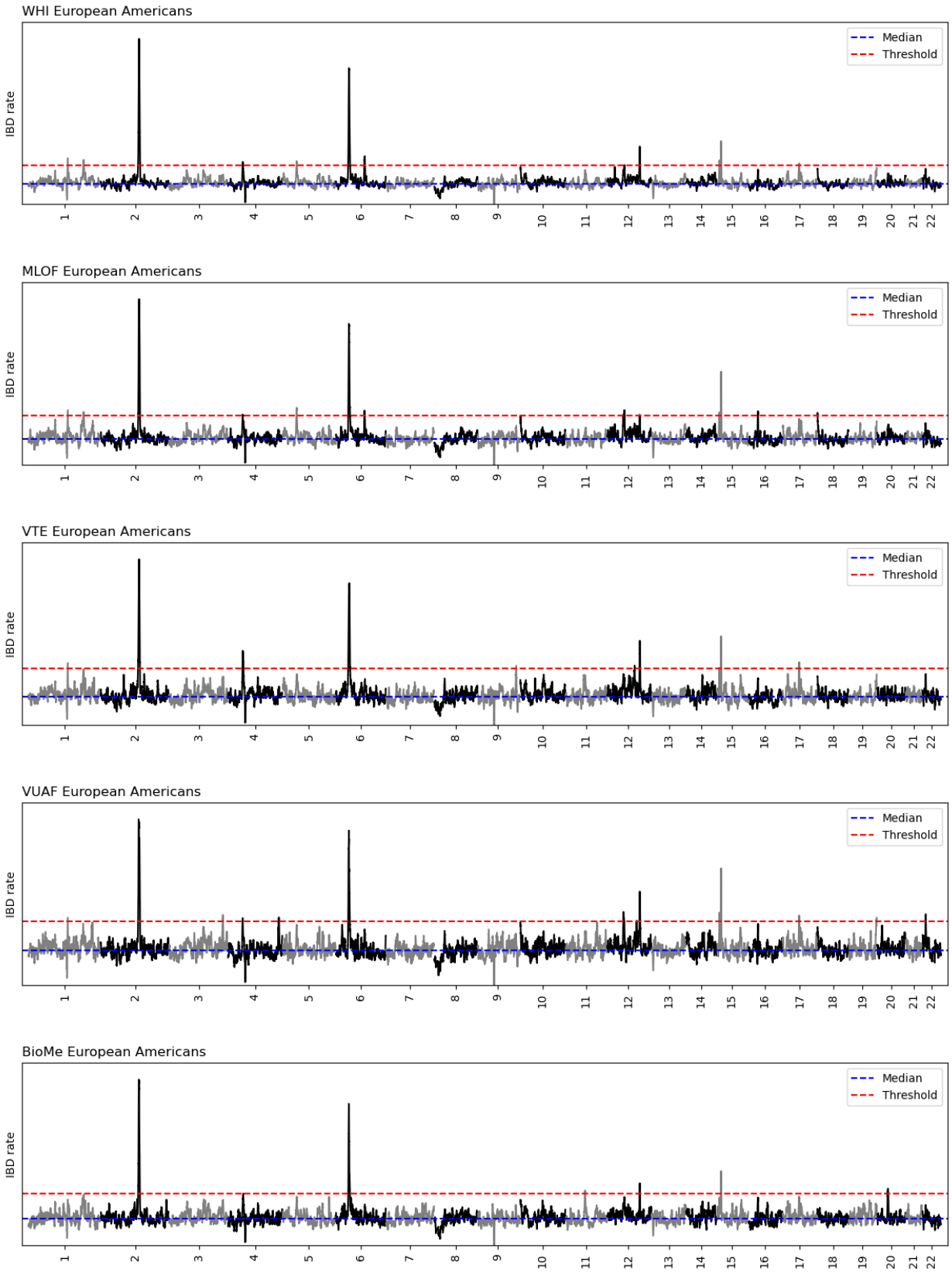
Supplementary Figure 2. Robust estimates of allele frequency from inferred IBD tracts

Percentile 25-75 boxes for allele frequency estimates in terms of varying selection coefficients (A) or effective sizes (B). Effective sizes are population bottleneck (BN), super-exponential growth (SE), constant size 25k (C25), and two subpopulations with a migration rate of forty percent (M40). Results in (A) are based on the BN effective sizes, five thousand diploid samples, and one thousand simulations. Results in (B) are based on five thousand diploid samples and two hundred simulations for each effective sizes example.

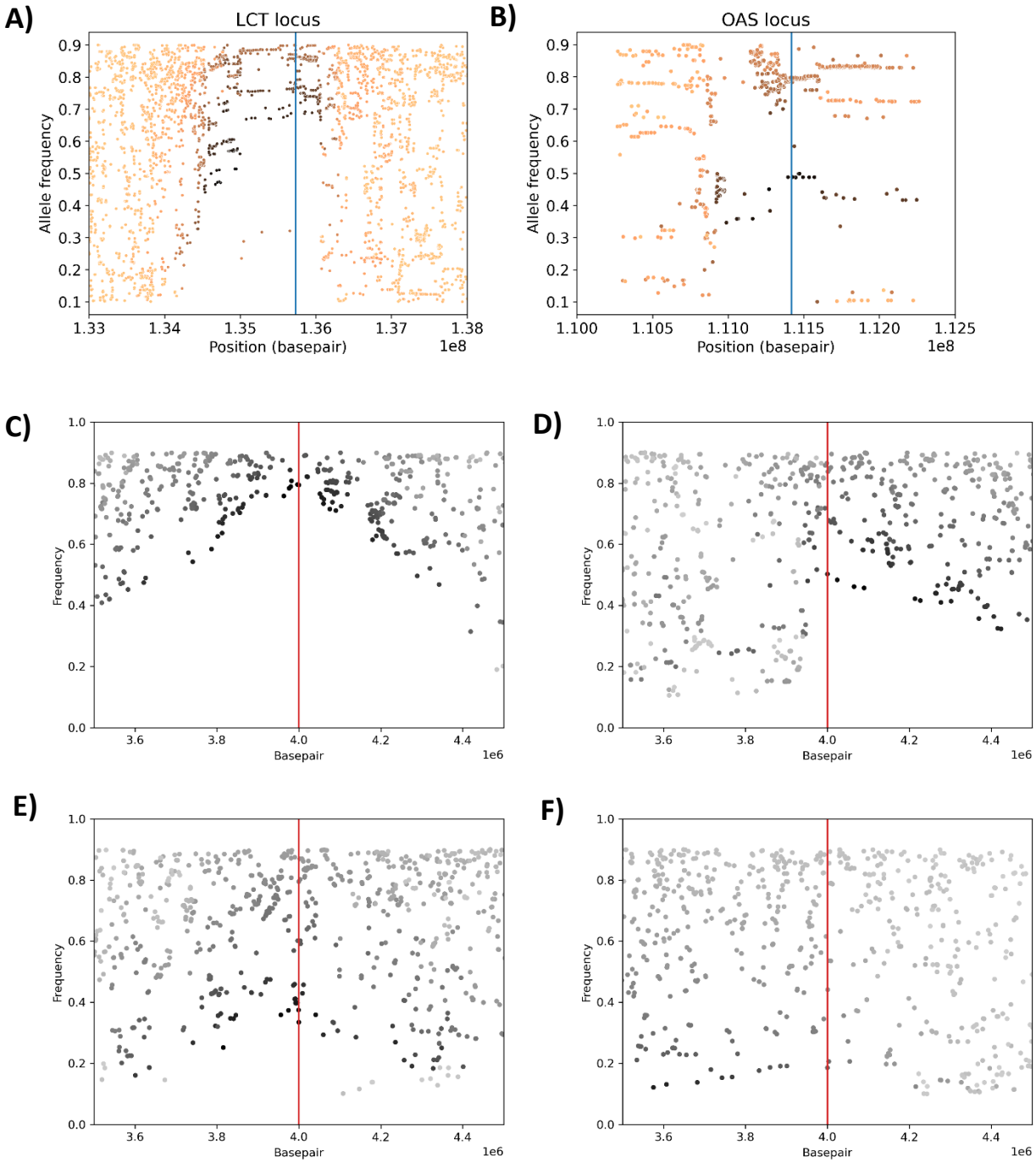


Supplementary Figure 3. Partial sweeps at low frequency and from $s \leq 0.020$ can have low information entropy

(A) The IBD entropy statistic calculated at different allele frequencies. (B) Of the partial sweeps at frequency 10-30% and where the IBD rate is less than four times the standard error plus population median, 95% of those have IBD entropy less than 0.50. (C, D) Those simulations with high IBD entropy correspond to partial sweeps at frequency 10-30% and with $s = 0.015$ or $s = 0.020$. Results aggregated over one thousand simulations of five thousand diploids in the population bottleneck example.

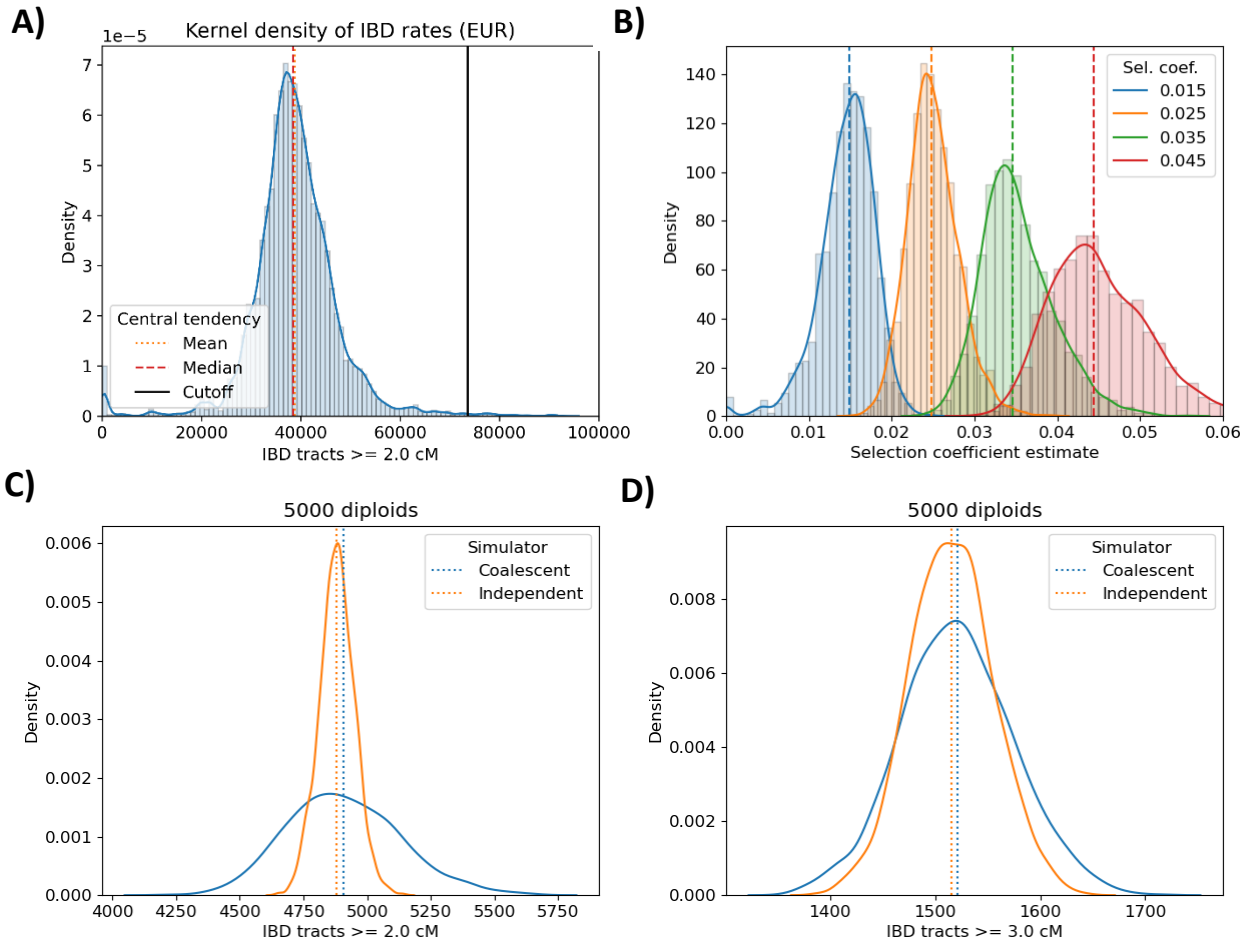


Supplementary Figure 4. Cohort-specific selection scans of TOPMed European Americans



Supplementary Figure 5. Selective sweeps show patterns of depleted common variation

For each marker, physical distance by allele frequency is plotted for different simulations or real genetic loci. Color gradient is based on scaled difference in allele frequency between inferred excess IBD group and the rest of the sample. More differentiated alleles are shaded darker. (A, B) Results for LCT and OAS1-2-3 in EUR ancestry analysis. Vertical blue lines are location estimates. (C-F) Simulation examples where the sweeping allele frequency is in the ranges 70-90%, 50-70%, 30-50%, and 10-20%. The vertical red line is the location of the selected allele.



Supplementary Figure 6. Identity-by-descent rates and selection coefficient estimates look to be normally distributed. The parametric bootstrap reflects the true variance in the coalescent process

(A) Histograms and fitted curve for inferred IBD rates along the genome in EUR ancestry samples. (B) Histograms and fitted curve for selection coefficient estimates in simulations. (C, D) Kernel density estimates over 2,500 replicate simulations using coalescent simulator for IBD tracts (blue) as opposed to a simulator for independent IBD tracts (orange).

Tables

Table 1. Selection coefficient estimation based on true IBD tracts

Each row represents results of 10,000 replicates. Summary statistics are the mean estimate, the mean bias, the width of confidence interval, and the percentage of intervals containing the true selection coefficient. Default settings are $s = 0.03$, $p(0) = 0.50$, a population bottleneck demography, the multiplicative genetic model, an ongoing sweep from a new beneficial mutation. Effective sizes considered are population bottleneck (BN), super-exponential growth (SE), and constant N_e =25000 (C25). Ninety-five percent confidence intervals are based on fifty bootstraps. Sample size is five thousand diploids.

Scenario	True s	Estimate $\hat{s}$	Width	Coverage
$s = 0.020$	0.0200	0.0202	0.0085	97.67%
$s = 0.030$	0.0300	0.0300	0.0085	95.21%
$s = 0.040$	0.0400	0.0401	0.0099	95.11%
$p(0) = 0.25$	0.0300	0.0300	0.0087	95.31%
$p(0) = 0.50$	0.0300	0.0300	0.0066	94.89%
$p(0) = 0.75$	0.0300	0.0300	0.0058	95.19%
$N_e(t)$ = BN	0.0300	0.0300	0.0066	95.26%
$N_e(t)$ = SE	0.0300	0.0300	0.0062	95.01%
$N_e(t)$ = C25	0.0300	0.0300	0.0066	95.79%

Table 2. Top twelve regions identified in selection scan on TOPMed European Americans

Populations are defined based on principal components analysis and validated with self-reported Race and global ancestry inference. Physical and genetic positions are shown in megabases (Mb) and centiMorgans (cM) for the location of maximum IBD rate. The interval is the left and right endpoints where IBD rate exceeds the four standard deviations threshold. The interval must span at least 1.0 cM. Genes of interest are annotated for regions discussed in the main text. Locations are aligned to build hg38.

Population	Chr	Max IBD Rate	Mb	cM	Genes
EUR	2	0.003635	141.94 (130.52-141.94)	141.79	LCT
EUR	6	0.002881	38.12 (22.12-38.12)	49.35	MHC
EUR	15	0.001323	34.16 (28.34-34.16)	15.42	TRPM1
EUR	12	0.001088	115.64 (108.9-115.64)	120.54	OAS
EUR	1	0.000909	154.56 (149.5-154.56)	146.25	Epidermal
EUR	6	0.000857	108.46 (103.76-108.46)	107.11	PRDM1
EUR	1	0.000852	209 (203.5-209)	205.05	.
EUR	15	0.000839	30.84 (25.94-30.84)	10.19	OCA2
EUR	4	0.000834	40.96 (36.3-40.96)	56.17	TLR
EUR	5	0.000822	35.98 (31-35.98)	53.34	SLC45A2
EUR	17	0.000791	39.74 (35.66-39.74)	62.14	HFNB1
EUR	12	0.000780	55.06 (50.94-55.06)	64.04	KRT71

Table 3. Estimates of location, allele frequency, and selection coefficient for human loci suspected to be under positive selection

Statistical inference at putatively selected loci in European Americans (EUR). Common names are given for genes in the relevant literature or in the UCSB genome browser. Low values of IBD entropy indicate a few large, inferred clusters of excess IBD. Locations are aligned to build hg38.

Population	**Gene**	**Chr**	**Location**	$\hat{\boldsymbol{p}}(\mathbf{0})$	$\hat{\boldsymbol{s}}$	**Entropy**	**Freq** $\widehat{\boldsymbol{A}}$
EUR	LCT	2	135,735,797	0.675	$0.0330 \pm 0.0.0052$	0.00	0.65
EUR	OAS	12	111,417,995	0.500	0.0183 ± 0.0025	0.58	0.21
EUR	OCA HERC2	15	28,045,464	0.750	0.0189 ± 0.0021	0.32	0.73
EUR	KRT71	12	52,915,123	0.725	0.0207 ± 0.0021	0.10	0.38

Supplementary Table 1. Selection coefficient estimation based on true IBD segments and model misspecification

Each row represents results of 10,000 replicates. Summary statistics are the mean estimate, the mean bias, the width of confidence interval, and the percentage of intervals containing the true selection coefficient. Default settings are $s = 0.03$, $p(0) = 0.50$, a population bottleneck demography, the multiplicative genetic model, an ongoing sweep from a new beneficial mutation. Ninety-five percent Confidence intervals are based on fifty bootstraps. Sample size is five thousand diploids. The Assumed column is the value used in estimating s. Effective sizes considered are population bottleneck (BN), super-exponential growth (SE), and constant N_e =25000 (C25). Abbreviations 0.9x and 1.1x mean that the $N_e(t)$ is uniformly scaled by that value. SV concerns the allele frequency of standing neutral variation.

Parameter			$\boldsymbol{s}$				
Type	**True**	**Assumed**	**True**	**Estimate**	**Bias**	**Width**	**Coverage**
$p(0)$	0.40	0.50	0.0300	0.0305	-0.0005	0.0074	95.75%
$p(0)$	0.60	0.50	0.0300	0.0301	-0.0001	0.0063	94.61%
$p(0)$	0.20	0.25	0.0300	0.0316	-0.016	0.0100	93.55%
$p(0)$	0.30	0.25	0.0300	0.0289	0.0011	0.0080	90.33%
$N_e(t)$	BN	0.9x	0.0300	0.0282	0.0018	0.0070	75.90%
$N_e(t)$	BN	1.1x	0.0300	0.0314	0.0014	0.0064	92.08%
SV	0.05	0.00	0.0300	0.0300	0.0000	0.0093	94.59%
SV	0.02	0.00	0.0300	0.0300	0.0000	0.0073	94.88%
SV	0.01	0.00	0.0300	0.0300	0.0000	0.0064	94.92%

Supplementary Table 2. Selection coefficient estimation based on inferred IBD segments

For each row, estimates $\hat{s}$, mean absolute deviation (MAD), bias, and coverage are averaged over one hundred simulations. Demographic scenarios in the first column abbreviated as bottleneck (BN), super-exponential growth (SE), constant size $N_e = 25{,}000$ (C25), and random mixing with proportion $m = 0.40$ between two subpopulations. Time refers to generations ago when the *de novo* mutation arose.

$N_e(t)$	Time	s	$\hat{s}$	Bias	MAD	Coverage
BN	325	0.015	0.0168	0.0018	0.0046	83
BN	375	0.015	0.0175	0.0025	0.0038	82
BN	250	0.020	0.0220	0.0020	0.0032	90
BN	300	0.020	0.0216	0.0016	0.0023	95
BN	275	0.025	0.0264	0.0014	0.0022	94
BN	325	0.025	0.0254	0.0004	0.0013	94
BN	250	0.030	0.0304	0.0004	0.0018	98
BN	300	0.030	0.0295	-0.0005	0.0016	86
BN	225	0.035	0.0361	0.0011	0.0024	99
BN	275	0.035	0.0336	-0.0014	0.0020	86
C25	250	0.030	0.0311	0.0011	0.0029	97
C25	300	0.030	0.0300	-0.0000	0.0023	95
SE	250	0.030	0.0292	-0.008	0.0019	91
SE	300	0.030	0.0279	-0.0021	0.0026	66
$m = 0.4$	250	0.030	0.0294	-0.0006	0.0024	79
$m = 0.4$	300	0.030	0.0291	-0.0009	0.0016	87

Supplementary Table 3. Cohort-specific estimation for TOPMed European Americans

Same description as in Table 4, expect the location, $\hat{p}(0)$, and $\hat{s}$ are based on the lowest (best) ranked SNP. If multiple cohort-specific analyses find the same SNP, this may provide supporting evidence for that SNP being in strong linkage disequilibrium with a sweeping allele. Low values of IBD entropy indicate a few large, inferred clusters of excess IBD. We do not estimate selection coefficients when IBD entropy is less than 0.5. We do not estimate at MHC and TLR1-6-10 which may subject to other forms of selection.

Cohort	Name	Chr	Location	$\hat{p}(0)$	$\hat{s}$	Entropy	Freq. $\hat{A}$
BioMe	LCT	2	135,851,076	0.6769	0.0291	0	0.58
MLOF	LCT	2	135,851,076	0.6599	0.0294	0	0.57
VTE	LCT	2	135,851,076	0.7067	0.0319	0	0.61
VUAF	LCT	2	135,851,076	0.6834	0.0300	0	0.47
WHI	LCT	2	135,851,076	0.6863	0.0317	0	0.64
BioMe	***OAS***	***12***	***111,270,654***	***0.4607***	***0.0115***	***0.42***	***0.12***
MLOF	***OAS***	***12***	***112,781,063***	***0.2881***	***0.0157***	***0.39***	***0.19***
VTE	***OAS***	***12***	***111,598,263***	***0.1138***	***N/A***	***0.60***	***0.08***
VUAF	***OAS***	***12***	***111,270,654***	***0.4360***	***0.0196***	***0.42***	***0.13***
WHI	***OAS***	***12***	***111,279,654***	***0.4475***	***0.0175***	***0.47***	***0.19***
BioMe	OCA/HERC2	15	N/A	N/A	N/A	N/A	N/A
MLOF	OCA/HERC2	15	28,075,818	0.3417	0.0174	0	0.25
VTE	OCA/HERC2	15	28,632,448	0.8266	0.0196	0.23	0.63
VUAF	OCA/HERC2	15	28,075,818	0.3653	0.0199	0	0.29
WHI	OCA/HERC2	15	28,141,480	0.4170	0.0164	0	0.69
BioMe	***TRPM1***	***15***	***31,437,832***	***0.1223***	***N/A***	***0.57***	***0.06***
MLOF	***TRPM1***	***15***	***31,437,832***	***0.1249***	***N/A***	***0.53***	***0.09***
VTE	***TRPM1***	***15***	***31,437,832***	***0.1163***	***N/A***	***0.59***	***0.05***
VUAF	***TRPM1***	***15***	***31,437,832***	***0.1322***	***N/A***	***0.66***	***0.08***
WHI	***TRPM1***	***15***	***31,437,832***	***0.1209***	***N/A***	***1.18***	***0.19***
BioMe	KRT71	12	N/A	N/A	N/A	N/A	N/A
MLOF	KRT71	12	52,808,399	0.3797	N/A	0.93	0.11
VTE	KRT71	12	N/A	N/A	N/A	N/A	N/A
VUAF	KRT71	12	50,334,028	0.3607	N/A	0.45	0.09
WHI	KRT71	12	53,083,566	0.6873	0.0188	0.15	0.33
BioMe	***MHC***	***6***	***28,737,297***	***0.1002***	***N/A***	***0.29***	***0.13***
MLOF	***MHC***	***6***	***28,865,011***	***0.1014***	***N/A***	***0.30***	***0.15***
VTE	***MHC***	***6***	***27,442,643***	***0.1031***	***N/A***	***0.29***	***0.13***
VUAF	***MHC***	***6***	***27,442,643***	***0.1025***	***N/A***	***0.30***	***0.12***
WHI	***MHC***	***6***	***27,230,939***	***0.5040***	***N/A***	***0.75***	***0.28***
BioMe	TLR1-6-10	4	39,078,952	0.1692	N/A	0.26	0.11
MLOF	TLR1-6-10	4	38,717,111	0.4205	N/A	0.20	0.26
VTE	TLR1-6-10	4	38,764,099	0.6385	N/A	0.26	0.36
VUAF	TLR1-6-10	4	39,078,952	0.1885	N/A	0.28	0.07
WHI	TLR1-6-10	4	38,748,564	0.6919	N/A	0.17	0.47

References

1. Hermisson, J., and Pennings, P.S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics *169*, 2335–2352.

2. Pennings, P.S., and Hermisson, J. (2006). Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol. Biol. Evol. *23*, 1076–1084.

3. Pennings, P.S., and Hermisson, J. (2006). Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. *2*, e186.

4. Smith, J.M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. *23*, 23–35.

5. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

6. Pankratov, V., Yunusbaeva, M., Ryakhovsky, S., Zarodniuk, M., Estonian Biobank Research Team, and Yunusbayev, B. (2022). Prioritizing autoimmunity risk variants for functional analyses by fine-mapping mutations under natural selection. Nat. Commun. *13*, 7069.

7. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol. *4*, e72.

8. Browning, S.R., and Browning, B.L. (2020). Probabilistic Estimation of Identity by Descent Segment Endpoints and Detection of Recent Selection. Am. J. Hum. Genet. *107*, 895–910.

9. Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., and Palamara, P.F. (2020). Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. *11*, 1–15.

10. Palamara, P.F., Terhorst, J., Song, Y.S., and Price, A.L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nat. Genet. *50*, 1311–1317.

11. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

12. Szpiech, Z.A., and Hernandez, R.D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. *31*, 2824–2827.

13. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.

14. Salter-Townshend, M., and Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics *212*, 869–889.

15. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. Am. J. Hum. Genet. *106*, 426–437.

16. Shemirani, R., Belbin, G.M., Avery, C.L., Kenny, E.E., Gignoux, C.R., and Ambite, J.L. (2021). Rapid detection of identity-by-descent tracts for mega-scale datasets. Nat. Commun. *12*, 3546.

17. Freyman, W.A., Mcmanus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., and Auton, A. (2021). Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. Mol. Biol. Evol. *38*, 2131–2151.

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

19. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'Er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.

20. Wang, M., Huang, X., Li, R., Xu, H., Jin, L., and He, Y. (2014). Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. Mol. Biol. Evol. *31*, 3068–3080.

21. Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. *31*, 1275–1291.

22. Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. Nat. Methods *15*, 279–282.

23. Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.

24. Mathieson, I., and Terhorst, J. (2022). Direct detection of natural selection in Bronze Age Britain. Genome Res. *32*, 2057–2067.

25. Mathieson, I., and McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics *193*, 973–984.

26. Kingman, J.F.C. (1982). The coalescent. Stochastic Process. Appl. *13*, 235–248.

27. Kingman, J.F.C. (1982). On the genealogy of large populations. J. Appl. Probab. *19*, 27–43.

28. Wakeley, J. (2020). Developments in coalescent theory from single loci to chromosomes. Theor. Popul. Biol. *133*, 56–64.

29. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. *51*, 1321–1329.

30. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS Genet. *10*, e1004342.

31. Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. Nat. Genet. *51*, 1330–1338.

32. Zhang, B.C., Biddanda, A., Gunnarsson, Á.F., Cooper, F., and Palamara, P.F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat. Genet. 10.1038/s41588-023-01379-x.

33. Stern, A.J., Wilton, P.R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. *15*, e1008384.

34. Kelleher, J., Thornton, K.R., Ashander, J., and Ralph, P.L. (2018). Efficient pedigree recording for fast population genetics simulation. PLoS Comput. Biol. *14*, e1006581.

35. Kuhner, M.K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics *22*, 768–770.

36. Kuhner, M.K., Yamato, J., and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. Genetics *156*, 1393–1401.

37. Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W., and Ralph, P.L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol. Ecol. Resour. *19*, 552–566.

38. Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A.P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E.C., Galloway, J.G., et al. (2022). Efficient ancestry and mutation simulation with msprime 1.0. Genetics *220*. 10.1093/genetics/iyab229.

39. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol. Biol. Evol. *36*, 632–637.

40. Adrion, J.R., Cole, C.B., Dukler, N., Galloway, J.G., Gladstein, A.L., Gower, G., Kyriazis, C.C., Ragsdale, A.P., Tsambos, G., Baumdicker, F., et al. (2020). A community-maintained standard library of population genetic models. Elife *9*. 10.7554/eLife.54967.

41. Hejase, H.A., Mo, Z., Campagna, L., and Siepel, A. (2022). A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. Mol. Biol. Evol. *39*. 10.1093/molbev/msab332.

42. Hejase, H.A., Mo, Z., Campagna, L., and Siepel, A. (2021). Sia: Selection inference using the ancestral recombination graph. BioRxiv.

43. Riley, R., Mathieson, I., and Mathieson, S. (2023). Interpreting Generative Adversarial Networks to Infer Natural Selection from Genetic Data. bioRxiv, 2023.03.07.531546. 10.1101/2023.03.07.531546.

44. Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., and Fumagalli, M. (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data. BMC Bioinformatics *20*, 337.

45. Garrigan, D., and Hedrick, P.W. (2003). Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution *57*, 1707–1722.

46. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

47. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics *155*, 945–959.

48. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

49. Wiuf, C., and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. Theor. Popul. Biol. *56*, 183–201.

50. Hein, J., Schierup, M., and Wiuf, C. (2004). Gene Genealogies, Variation and Evolution: A primer in coalescent theory (Oxford University Press, USA).

51. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.

52. Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418.

53. Cai, R., Browning, B.L., and Browning, S.R. (2022). IBD-based estimation of X chromosome effective population size with application to sex-specific demographic history. bioRxiv, 2022.07.06.499007. 10.1101/2022.07.06.499007.

54. Kern, A.D., and Schrider, D.R. (2018). diploS/HIC: an updated approach to classifying selective sweeps. G3: Genes, Genomes, Genetics.

55. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890.

56. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). IBDkin: fast estimation of kinship coefficients from identity by descent segments. Bioinformatics *36*, 4519–4520.

57. Ségurel, L., and Bon, C. (2017). On the Evolution of Lactase Persistence in Humans. Annu. Rev. Genomics Hum. Genet. *18*, 297–319.

58. Klunk, J., Vilgalys, T.P., Demeure, C.E., Cheng, X., Shiratori, M., Madej, J., Beau, R., Elli, D., Patino, M.I., Redfern, R., et al. (2022). Evolution of immune genes is associated with the Black Death. Nature *611*, 312–319.

59. Evershed, R.P., Davey Smith, G., Roffet-Salque, M., Timpson, A., Diekmann, Y., Lyon, M.S., Cramp, L.J.E., Casanova, E., Smyth, J., Whelton, H.L., et al. (2022). Dairying, diseases and the evolution of lactase persistence in Europe. Nature, 1–10.

60. Barton, A.R., Santander, C.G., Skoglund, P., Moltke, I., Reich, D., and Mathieson, I. (2023). Insufficient evidence for natural selection associated with the Black Death. bioRxiv, 2023.03.14.532615. 10.1101/2023.03.14.532615.

61. Bajpai, V.K., Swigut, T., Mohammed, J., Naqvi, S., Arreola, M., Tycko, J., Kim, T.C., Pritchard, J.K., Bassik, M.C., and Wysocka, J. (2023). A genome-wide genetic screen uncovers determinants of human pigmentation. Science *381*, eade6289.

62. Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G., and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am. J. Hum. Genet. *82*, 424–431.

63. Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K.W., and Hansen, L. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. Hum. Genet. *123*, 177–187.

64. Sams, A.J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J.E., Messer, P.W., and Barreiro, L.B. (2016). Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome Biol. *17*, 246.

65. Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell *173*, 53-61.e9.

66. Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akey, J.M. (2016). Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Curr. Biol. *26*, 3375–3382.

67. Moll, R., Divo, M., and Langbein, L. (2008). The human keratins: biology and pathology. Histochem. Cell Biol. *129*, 705–733.

68. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. *5*, e1000562.

69. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

70. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am. J. Hum. Genet. *74*, 1111–1120.

71. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLoS Genet. *14*, e1007385.

72. Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. Am. J. Hum. Genet. *91*, 809–822.

73. Casella, G., and Berger, R.L. (1993). Statistical Inference. Biometrics *49*, 320.

74. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

75. Korfmann, K., Gaggiotti, O.E., and Fumagalli, M. (2023). Deep Learning in Population Genetics. Genome Biol. Evol. *15*. 10.1093/gbe/evad008.

76. Efron, B. (1982). The jackknife, the bootstrap and other resampling plans.

77. Efron, B. (1987). Better Bootstrap Confidence Intervals. J. Am. Stat. Assoc. *82*, 171–185.

78. Manly, B.F.J. (2018). Randomization, Bootstrap and Monte Carlo Methods in Biology: Texts in Statistical Science (chapman and hall/CRC).

79. Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I., et al. (2016). Detection of human adaptation during the past 2000 years. Science *354*, 760.

80. Shemirani, R., Belbin, G.M., Burghardt, K., Lerman, K., Avery, C.L., Kenny, E.E., Gignoux, C.R., and Ambite, J.L. (2022). Selecting Clustering Algorithms for Identity-By-Descent Mapping. In Biocomputing 2023 (WORLD SCIENTIFIC), pp. 121–132.

81. Yair, S., Lee, K.M., and Coop, G. (2021). The timing of human adaptation from Neanderthal introgression. Genetics *218*. 10.1093/GENETICS/IYAB052.

82. Kö Ster, J., and Rahmann, S. (2012). Genome analysis Snakemake-a scalable bioinformatics workflow engine. BIOINFORMATICS APPLICATION NOTE *28*, 2520–2522.

83. Ewing, G., and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics *26*, 2064–2065.

84. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics *35*, 5346–5348.

85. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

86. Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science *363*. 10.1126/science.aau1043.

87. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nature Genetics 2014 46:8 *46*, 919–925.

88. Terhorst, J., Kamm, J.A., and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. *49*, 303–309.

89. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. 10.1093/gigascience/giab008.

90. Browning, S.R. (2006). Multilocus association mapping using variable-length Markov chains. Am. J. Hum. Genet. *78*, 903–913.

91. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.