Probabilistic estimation of identity by descent segment endpoints and detection of recent selection

Sharon R. Browning[1*] and Brian L. Browning[2]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.

[2]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA 98195, USA.

*Correspondence: sguy@uw.edu

Abstract

Most methods for fast detection of identity by descent (IBD) segments report identity by state segments without any quantification of the uncertainty in the endpoints and lengths of the IBD segments. We present a method for determining the posterior probability distribution of IBD segment endpoints. Our approach accounts for genotype errors, recent mutations, and gene conversions which disrupt DNA sequence identity within IBD segments. We find that our method's estimates of uncertainty are well calibrated for homogeneous samples. We quantify endpoint uncertainty for 7.7 billion IBD segments from 408,883 individuals of White British ancestry in the UK Biobank, and use these IBD segments to find regions showing evidence of recent natural selection. We show that many spurious selection signals are eliminated by the use of unbiased estimates of IBD segment endpoints and a pedigree-based genetic map. Nine of the top ten regions with the greatest evidence for recent selection in our scan have been identified as selected in previous analyses using different approaches. Our computationally efficient method for quantifying IBD segment endpoint uncertainty is implemented in the open source ibd-ends software package.

Introduction

Pairs of individuals within a population can share one or more long segments of their genomes identical by descent due to inheritance from common ancestors. Identity by descent (IBD) segments are used in many applications, such as estimation of kinship,[1-3] recent demography,[4-8] mutation rates,[9-12] and recombination rates,[13] and detection of recent selection.[14-16]

The three main types of test for recent positive selection are based on population differentiation, admixture proportions, and haplotype structure. The first type looks for variants that differ markedly in frequency between populations.[17; 18] The second type looks for regions in which the sample ancestry proportions in admixed individuals differ from those elsewhere in the genome.[19; 20] One subtype of this test involves archaic admixture, such as introgression from Neanderthals into modern humans, and searches for regions in which the frequency of the archaic haplotype in a modern population is unusually high.[21] The third type looks for high-frequency haplotypes that are unusually long.[22; 23] IBD-based selection scans fall into this category.[14] IBD scans look for genomic regions that have a significantly higher than average number of IBD segments. If the genome were completely neutral, and there are no biases in detecting IBD segments or estimating their centiMorgan (cM) lengths, the expected number of IBD segments exceeding some cM length threshold would be constant across the genome. In contrast, if certain haplotypes in a genomic region have a selective advantage, the effective size of the population is reduced in that region, which leads to a higher than expected number of IBD segments. IBD-based tests can also detect the effects of negative selection and balancing selection, since any type of selection will tend to decrease the effective population size within the genomic region.

An IBD segment for a pair of haplotypes is a segment of DNA inherited from a single common ancestor, with no crossovers occurring within the segment in the lineages of the two haplotypes since the common ancestor.[4; 6] Within a shared IBD segment, sequence identity can be disrupted by mutation and gene

conversion. In addition, genotype error can cause two haplotypes to appear to be discordant at a position. At such positions, two "identical by descent" haplotypes are in fact not identical. This non-identity needs to be considered when detecting IBD segments.

In the human genome, de novo single nucleotide mutations occur at an average rate of around 1.3×10^{-8} per base pair per meiosis,[12] which is similar to the average rate of crossing over per base pair per meiosis. Thus, regardless of the number of generations since the most recent common ancestor, an average of approximately one mutation is expected in the lineage of an IBD segment.

For a pair of haplotypes drawn at random from a population, most of the genome is comprised of very short segments of IBD, with a very large number of generations to the most recent common ancestor. Since each short segment contains an average of approximately one discordance caused by mutation in addition to discordances caused by gene conversion, a series of closely-spaced discordances is a clear indication that the genomic interval containing the discordances is comprised of a sequence of short IBD segments. In contrast, when one observes a long segment without discordances, it is usually (depending on the population's demographic history) highly probable that this segment is primarily comprised of a single long IBD segment resulting from recent common ancestry.

There are three primary paradigms for IBD segment detection. The first paradigm considers a pair of haplotypes to be either "IBD" or "not IBD" at each position in the genome. A hidden Markov model, with pre-determined IBD proportion and rates of transition between the IBD and non-IBD states, may be used to obtain posterior probabilities of IBD and non-IBD at each position.[24-29] This paradigm developed out of the analysis of pedigree data, and is very natural in that setting.[30] However, for population data with unknown relationships, the dichotomy into IBD and non-IBD is artificial and ignores the fact that each pair of haplotypes has a common ancestor at each position in the genome, although that ancestor may have lived a long time ago.

The second paradigm considers the length of shared segments. A segment is identical by descent if it is inherited from a common ancestor and exceeds a length threshold. In practice, if identity by state (IBS) sharing extends beyond some threshold, the segment is reported as identical by descent.[31-33] This paradigm recognizes the potential existence of IBD segments that are shorter than the threshold, but does not try to find them. The threshold is typically chosen to be a length above which the accuracy of the reported segments is high.[31; 32]

The third paradigm considers two haplotypes to be identical by descent if their time to most common ancestor (TMRCA) is less than some specified number of generations.[16; 34]

In this work we take a different perspective. We recognize that a pair of haplotypes is, strictly speaking, identical by descent at every point in the genome. However, for any given point in the genome, the endpoints of the IBD segment containing that point are unknown (Figure 1A). It is possible that one or more discordances at the end of the segment are actually contained within the long IBD segment (Figure 1B). It is also possible that IBD ends before IBS ends, so that the end of the IBS segment is not part of the long IBD segment, but instead contains one or more neighboring short IBD segments (Figure 1C). In some cases, two or more long IBD segments in a region can be mistaken for a single long IBD segment (Figure 1D).[35] Our approach quantifies this uncertainty. In Results, we show that our quantification is well-calibrated, and we apply our method to perform an IBD-based selection scan in the UK Biobank.

Methods

Overview of method

The input data for our method are phased genotypes and candidate IBD segments. Highly-accurate phased genotypes can be obtained from statistical phasing in large cohorts of accurately-genotyped individuals.[36] The candidate IBD segments may be obtained using a length-based IBD detection method such as hap-

ibd.[33] Our method estimates the posterior probability distributions of the endpoints of the candidate IBD segments and outputs quantiles and samples from these posterior distributions.

Before estimating segment endpoints, we apply a minor allele frequency (MAF) filter to the phased genotypes that excludes variants with frequency less than 0.1%. In Results we show that these rare variants are not modelled as well as the more common variants and that including these rare variants negatively impacts the accuracy of the endpoint estimates.

We model allele discordance within an IBD segment using a user-specified error rate. Analysis results are not overly sensitive to the exact choice of error rate (see Results). Discordances within a segment are assumed to occur independently except when two or more closely-spaced discordances could have originated from the same gene conversion event.

We also model IBS extending beyond the end of an IBD segment. IBS segments can be comprised of multiple IBD segments. We do not try to directly model each of these IBD segments individually, but instead model the distribution of IBS segments found in the data. Short regions of IBS are modelled using the local context, because the IBS length distribution varies across the genome due to factors such as mutation rate and selection. Longer segments of IBS are modelled using chromosome-wide data because there is limited information about longer IBS segments from the local context.

We estimate the probability of the observed discordance data as a function of the IBD endpoints. We then use Bayes' rule to obtain the probability distribution of each IBD endpoint. We work from a focal position within an IBS segment (Figure 1A) and estimate the probability distributions for the positions of the left and right endpoints of the IBD segment that covers the focal position.

Notation

We wish to estimate the endpoints of the IBD segment covering a position x_0 for a given pair of haplotypes, H_1 and H_2. All positions are measured in terms of genetic distance in Morgans, and haplotype phase is assumed to be known. In this description we are only concerned with the estimation of the right endpoint of the IBD segment covering x_0. The estimation of left endpoints is similar. Index the markers to the right of x_0 by $1, 2, 3, \ldots, M$, where M is the last marker on the chromosome. Let the positions of these markers be $x_1, x_2, x_3, \ldots, x_M$.

The IBS data, D, is the observed IBS status (identical or discordant) for the alleles on haplotypes H_1 and H_2 at the markers to the right of x_0. Let $D[a, b]$ denote the IBS status at markers with indices $a \leq i \leq b$.

Let ϵ be the average proportion of discordant markers within IBD segments (the "error rate" mentioned above). We approximate $(1 - \epsilon)$ with 1 and thus omit terms of $(1 - \epsilon)$ in our calculations.

Modelling the IBS data for the IBD segment

We model the IBS data, D, from the focal point x_0 rightwards as being generated by two processes. The first, up to the IBD endpoint R, requires that alleles should be identical except at a small number of discordances due to mutation, gene conversion, or genotype error. If discordances in the IBD segment are independent and the right endpoint is in the interval (x_i, x_{i+1}), the probability of the data in the part of the IBD segment to the right of the focal point (i.e. $P(D[1, i])$) is ϵ^{n_i}, where n_i is the number of discordances between the first marker after the focal point and the i-th marker (inclusive) and factors of $(1 - \epsilon)$ are approximated by 1.

We use Bayes rule to obtain the posterior distribution of R, the position of the right endpoint ($R > x_0$). For each inter-marker interval (x_i, x_{i+1}), $i = 0, 1, \ldots, M - 1$, the probability that the right endpoint is contained in the open interval (x_i, x_{i+1}) satisfies:

$$P(R \in (x_i, x_{i+1})|D) \propto P(D \mid R \in (x_i, x_{i+1}))P\big(R \in (x_i, x_{i+1})\big)$$
$$= \epsilon^{n_i} P(D[i+1, M] \mid R \in (x_i, x_{i+1}))\, P\big(R \in (x_i, x_{i+1})\big)$$

Equation 1

where "∝" denotes proportionality. Normalizing the probabilities in Equation 1 to sum to 1 over the i gives the posterior probability that the endpoint occurs in each interval (x_i, x_{i+1}).

The probability $P\big(R \in (x_i, x_{i+1})\big)$ is a prior probability for the position of the right endpoint, given the position L_0 of the left endpoint (see "Iterative updating of endpoints and focal point" below). We model the population as having constant effective size 10,000 to obtain these probabilities. Details are given in Appendix 1.

The remaining component in Equation 1 is the probability $P(D[i+1, M] \mid R \in (x_i, x_{i+1}))$, which is the probability of the IBS data to the right of the right IBD endpoint. We model this by considering the points of discordance as being the points of renewal in a renewal process. That is, we obtain probabilities of the length of each segment that contains all of the non-discordant positions up to and including the next discordance, with each such segment being treated as being independent. The probabilities of each of these sequences is obtained empirically from the observed data. Details are given in Appendix 2.

Appendix 3 describes how to obtain the posterior cumulative distribution function for the endpoint from the interval probabilities given in Equation 1.

Iterative updating of endpoints and focal point

In the preceding section, we assumed that the left endpoint L_0 was known, however it also needs to be estimated, and we do this estimation iteratively. We start by using the left endpoint of the input candidate IBD segment as the value of L_0. After estimating the posterior distribution of the right endpoint, we use this distribution to obtain a new "right endpoint" R_0 that is set equal to the 5th percentile of this

distribution. Percentiles are referenced by distance from the focal point x_0. Thus, small percentiles are located closer to the focal point than larger percentiles. This choice of percentile is conservative (it reduces the estimated length of the IBD segment for the purpose of these calculations) and thus is likely to speed the convergence of the iterative approach. After estimating the right endpoint distribution, we estimate the left endpoint using the newly calculated value of R_0 and obtain a new value of L_0 as the 5th percentile of the left endpoint distribution. Then if $L_0 - x_0$ is significantly altered (>10% change in length) from the previous value, we use the new value of L_0 to re-estimate the right endpoint distribution and obtain a new value of R_0. If R_0 is significantly altered from the previous value, we use the new value of R_0 to re-estimate the left endpoint, and so on. Whenever we change the value of L_0 or R_0 we update the focal point x_0 which is located half-way between L_0 and R_0 in base coordinates. We perform a maximum of 10 updates of each endpoint. In order to prevent the focal point from moving outside the input candidate IBD segment, we constrain L_0 and R_0 to stay within the input candidate IBD segment.

Estimation of error rate

After running the endpoint estimation algorithm, our ibd-ends software estimates the error parameter ϵ by measuring the rate of discordant alleles within inferred IBD segments. The procedure is as follows. For each IBD segment that has been analyzed with the endpoint estimation algorithm, take the interval bounded by the posterior 5th percentile of the left and right endpoint distributions, and use these to obtain a cM length. If this length is < 2 cM, ignore the segment. Within the genomic region bounded by these endpoints, examine the alleles on the two IBD haplotypes. Count the number of mismatches, and the total number of positions examined. Across all segments, report the total number of mismatches, divided by the sum of the number of positions examined in each segment. If the estimated error rate does not differ significantly from the error rate used in the analysis (e.g. less than a three-fold difference), it is not necessary to re-run the analysis with the new value (see Results). For a large study, a pilot analysis on a small chromosome can be used to determine the error rate that should be used in the full analysis.

Modified error rate to account for gene conversion

Gene conversions copy material from one haplotype to the other during meiosis and can thus result in discordant alleles between IBD haplotypes. The typical length of a gene conversion tract is around 300 base pairs.[39] Changes will only occur at positions at which the individual in whom the gene conversion occurred was heterozygous. Thus, many gene conversions have no effect on allele discordance, but some gene conversions can result in more than one allele discordance occurring in proximity. Since these discordances are not independent events, we do not include an error term ϵ for each one, since that would be overly harsh and tend to result in premature truncation of the IBD segment. Instead, when more than one discordance occurs within 1 kb, we apply the error rate ϵ for the first discordance, and a less severe gene conversion error rate of ϵ' for each successive discordance within 1 kb of the first discordance (by default, $\epsilon = 0.0005$ and $\epsilon' = 0.1$).

Analysis pipeline

Our software, ibd-ends, requires the input of candidate segments for which endpoints will be evaluated. In this work, we use hap-ibd[33] to find the candidate segments. For many applications, one wishes to assess endpoint uncertainty for all IBD segments that exceed a length threshold. In that case, the key consideration is to avoid false negatives when detecting candidate IBD segments. If a potential IBD segment is not included in the input data to ibd-ends, it will not be included in the results. False positives (candidates for which the true IBD segment is actually shorter than the threshold) are less serious – they increase compute times but will be shown to be unlikely to be true long IBD segments when the segment endpoints are estimated. Thus, one should try to cast a wide net when identifying candidate IBD segments.

When analyzing sequence data, the high density of variants and the presence of genotype error can cause a high rate of discordances between IBD haplotypes. The hap-ibd method permits some discordances in a segment. It does so by finding seed IBS segments that exceed a certain length, and then extending these

segments if there is another IBS segment that exceeds a minimum extension length and that is separated from the seed segment by a short non-IBS gap. One way to apply hap-ibd to sequence data is to reduce the seed and extension lengths, which effectively increases the permitted density of discordances, but this can significantly increase computation time. Here we take a different approach. We reduce the marker density of the sequence data for the hap-ibd analysis (but not for the ibd-ends analysis). We choose a minimum MAF for hap-ibd that reduces the marker density to approximately that of a 600k SNP array, and we apply this threshold using hap-ibd's "min-mac" parameter. We retain the highest MAF variants because these are the most informative for detecting the candidate segments. This approach greatly reduces the density of variants, and thus reduces the number of IBD segments that would otherwise go undetected due to genotype error.

Except as otherwise noted, genotype data were phased using Beagle 5.1,[40] input IBD segments for ibd-ends were obtained using hap-ibd,[33] and default parameters were used for all programs.

Simulation overview

We generated three sets of simulated data to investigate three conditions under which estimation of endpoints could be challenging: gaps in marker coverage, non-constant population size, and heterogeneous samples.

We added genotype error to the simulated data for each marker at a rate equal to the minimum of 0.02% and one-half the MAF for the marker. This error rate produces a discordance rate of 0.04% in markers with MAF > 0.04%, which matches the discordance rate seen in the TOPMed data[41] and is six times higher than the 0.0067% discordance rate seen in the UK Biobank data.[42] We also wanted to confirm that the method produces accurate results with higher error rates, so we added error to one data set at a rate equal to the minimum of 0.1% and one-half the MAF for the marker.

Simulation of constant size population with variable marker density

We generated 60 Mb of data for 2000 individuals from a population with constant size of 10,000 diploid individuals. The recombination rate is 1×10^{-8} per base pair per meiosis. During the most recent 5000 generations, gene conversion initiations occurred at a rate of 2×10^{-8} and gene conversion tracts had a mean length of 300 base pairs. We used SLiM v3.3 to simulate the past 5000 generations,[43] and msprime to add mutations and simulate the more distant past.[44; 45] The mutation rate varied along the simulated region, with a new mutation rate each 100 kb that was uniformly distributed between 0 and 3×10^{-8} per bp per meiosis. In addition, we made a 1.5 Mb gap by removing genetic markers between positions 20.0 and 21.5 Mb, to represent a centromeric region. We added genotype error at a rate of 0.02% as described above. We used the simulated ancestral recombination graph to determine the true endpoints of IBD segments of length > 1 cM for all pairs of individuals within a subset of 500 individuals so that we could evaluate the accuracy of the inferred IBD segment endpoints.

We used the true (simulated) haplotypes, including any alleles changed by the addition of genotype error, for all analyses of these data. When detecting candidate IBD segments with hap-ibd, we used a minor allele count threshold of 1700 (minor allele frequency of 0.425; see Analysis pipeline), resulting in 10,760 markers after excluding markers in the 1.5 MB gap region, which corresponds to a mean density of one marker per 5.4 kb in the remaining 58.5 Mb. All markers with MAF > 0.1% (241,010 markers) were used in the ibd-ends analysis.

Simulation of non-constant size population

We generated 60 Mb of data for 50,000 individuals from a UK-like population. These simulated data have been described previously.[33] The demographic model has a population size of 24,000 in the distant past, a reduction to 3,000 occurring 5,000 generations ago, growth at rate 1.4% per generation starting 300 generations ago, and growth at rate 25% beginning 10 generations ago. The mutation rate is 1.3×10^{-8}

per base pair per meiosis, while the recombination rate is 1×10^{-8} per base pair per meiosis. During the most recent 5000 generations gene conversion initiations occurred at a rate of 2×10^{-8} and gene conversion tracts had a mean length of 300 base pairs. We used SLiM v3.3 to simulate the most recent 5000 generations[43] and msprime to add mutations and simulate the more distant past.[44; 45] We generated two copies of the data: one with 0.02% added genotype error and one with 0.1% added genotype error as described above. We also created a SNP-array version of the data with 0.02% genotype error and 10,000 randomly selected markers with minor allele frequency > 5% (1 marker per 6 kb on average, corresponding to approximately 500k markers genome-wide). We used the simulated ancestral recombination graph to determine the true endpoints of IBD segments of length > 1 cM for all pairs of individuals within a subset of 1000 individuals so that we could evaluate the accuracy of the inferred IBD segment endpoints.

When applying hap-ibd to the UK-like sequence data we applied a minor allele count threshold of 4500 (minor allele frequency threshold of 0.45; see Analysis pipeline), resulting in 11,524 markers across the 60 Mb with a mean density of one marker per 5.2 kb. All markers with MAF > 0.1% (198,566 markers) were used in the ibd-ends analysis of the sequence data. When applying hap-ibd to the SNP-array data, we used the default minor allele count threshold (min-mac=2), and we analyzed all variants with frequency > 0.1% in the ibd-ends analysis of the SNP-array data.

Simulation of a heterogeneous population

We simulated 10 Mb of data for 500 individuals of African-like ancestry and 500 individuals of European-like ancestry. The demographic history is the two-population model of Tennessen et al.,[46; 47] implemented in stdpopsim.[48] The combined sample represents an ancestrally heterogeneous population, which violates an assumption of our modelling of endpoint uncertainty. The recombination rate and mutation rate are both 1×10^{-8} per base pair per meiosis. We did not include gene conversion in the simulation. We

simulated the data with msprime,[44] and we added genotype error at a rate of 0.02% as described above. We used the simulated ancestral recombination graph to determine the true endpoints of IBD segments of length > 0.5 cM for all pairs of individuals so that we could evaluate the accuracy of the inferred IBD segment endpoints.

When applying hap-ibd to these data we applied a minor allele count threshold of 700 (MAF threshold of 0.35; see Analysis pipeline), resulting in 2215 markers across the 10 Mb with a mean density of one marker per 4.5 kb. We set the hap-IBD min-seed and min-output parameters to 1 cM since there are not many IBD segments of length > 2 cM in these data. All markers with MAF > 0.1% were used in the ibd-ends analyses (48,074 markers).

UK Biobank data

We phased QC-filtered UK Biobank data (487,373 individuals) using Beagle 5.1,[40] and then used hap-ibd to find candidate IBD segments among 408,883 White British individuals identified by the UK Biobank.[42] We ran ibd-ends with default settings on the candidate IBD segments from the White British individuals to estimate the uncertainty in the endpoints of these IBD segments. We used Bherer et al.'s European genetic map which is based on family data from Iceland and other European populations.[49] Variants located outside the bounds of the map are excluded from the ibd-ends analyses because extrapolated cM positions for markers outside the map can differ significantly from their true cM positions, leading to substantial under- or over- estimation of IBD segment lengths.

Results

Simulated data with variable marker density

Figure 2 shows the results of ibd-ends analysis on the simulated sequence data with constant population size, variable marker density, and a 1.5 Mb gap in marker coverage (see Methods for details). Even with

the large gap and uneven marker density, the endpoints uncertainty is well-calibrated (Figure 2A), and coverage of sampled IBD segments across the simulated region is even (Figure 2B). 58% of sampled endpoints are located within 5 kb of the true value.

Our method currently does not give special treatment to chromosome ends. A significant number of IBD segments terminate at the chromosome end, but our method does not assign any positive probability to an IBD segment ending exactly at the end of the chromosome. For analyzed segments with true right endpoint at the right end of the chromosome in these data, 25% have sampled right endpoint within 2 kb of the chromosome end, and 67% have sampled right endpoint within 100 kb of the chromosome end, but none have sampled right endpoint exactly at the chromosome end.

The ibd-ends analyses of these data used the default error rate of 0.0005. The error rate estimated by ibd-ends was 0.00039.

UK-like simulated data

Figure 3 shows results for the simulated UK-like data with a 0.02% genotype error rate. The estimated uncertainty is well-calibrated (upper row of Figure 3), even when using inferred haplotype phase and when using data thinned to represent a 500k SNP array. As expected, endpoint uncertainty, as measured by the difference between the endpoint sampled from the uncertainty distribution and the true endpoint, is much higher when analyzing SNP array data rather than full sequence data (lower row of Figure 3, right vs left and middle columns). Results are also accurate with smaller sample sizes (200 or 1000 individuals; Figure S1). When we removed markers to form a 1.5 cM gap (at 20-21.5 Mb), we found no noticeable change in calibration, and no excess rate of IBD in the region of the gap (Figure S2). We performed further analyses to evaluate endpoint estimation accuracy with a higher genotype error rate (Figure S3) and with different MAF thresholds (Figure S4). The results are well-calibrated with the higher genotype error rate (0.1%) when using true haplotypes. When using inferred haplotypes, the higher genotype error rate

results in phase errors that reduce accuracy. The results are not particularly sensitive to the choice of MAF, but some miscalibration is observed when very rare variants are included (0.01% MAF).

We found that using an analyses error rate that is up to three times higher or lower than the estimated error rate gave accurate results (Figure S5 and Table S1). When using an analysis error rate that is outside this range, the estimated error rates produced by ibd-ends are within the range of error rates that will provide good results in a subsequent analysis (Table S1). If the estimated error rate differs from the analysis error rate by more than three-fold, we recommend redoing the analysis using the estimated rate. Pilot results on a small chromosome can be used to determine whether the analysis error rate needs to be changed from the default value.

Compute times for ibd-ends analyses with default settings were 1.4 hours for the full UK-like data with 50,000 individuals (17 million IBD segments), and 0.5 hours for 1000 individuals (7000 IBD segments) using a 24-core compute node with 24 Intel Xeon Silver 4214 2.2 GHz processors and 382 GB of memory.

Heterogeneous simulated data

In the heterogeneous simulation, half of the simulated individuals are from a population with an African demographic history, while the other half are from a population with a European demographic history (see Methods). Analyzing these data together violates the assumptions of the endpoints modelling, however the results are not excessively mis-calibrated (Figure S6A). For example, 18% of the true endpoints are closer to the center of the IBD segment than the nominal 10^{th} percentile. When analyzing the African individuals separately (Figure S6B) or the European individuals separately (Figure S6C), the results are well-calibrated.

UK Biobank data

We sampled one left endpoint and one right endpoint from the endpoint uncertainty distributions of 77.7 billion candidate autosomal IBD segments of length 2 cM or larger that hap-IBD detected in the 408,883 White British UK Biobank individuals. We refer to the segment defined by the sampled end points as the "sampled IBD segment". Analysis was parallelized by chromosome. Total wall clock computing time across all chromosomes was 7.5 hours for hap-ibd and 150 hours for ibd-end using a 24-core Intel Xeon Silver 4214 2.2 GHz compute node. The estimated error rates from each chromosome varied from 0.00027 to 0.00034 when using the default analysis error rate of 0.0005.

There were 51.7 billion sampled IBD segments with length > 2 cM. Every 10 kb along each chromosome we computed the number of sampled IBD segments with length > 2 cM covering the position (Figure 4). The IBD rate is the number of IBD segments covering a position divided by the number of haplotype pairs. Each individual contributes two haplotypes, and all haplotype pairs are considered except those pairs within the same individual. A high rate of IBD at a position is a signal of possible recent strong natural selection.[14-16]

The median IBD rate is 0.0126%. There are ten regions with an IBD rate higher than 0.024% (Table 1). Nine of these are genomic regions known to be undergoing significant levels of selection, indicating the success of this approach in finding real signals of selection. Four of the regions have been shown to have adaptive introgression from Neanderthals (*OAS*, *CCR9/CXCR6*, *TLR1/6/10*, Type II Keratins). Five of the regions of selection play a role in immunity (*MHC*, *OAS*, *CCR9/CXCR6*, *TLR1/6/10*, *PRDM1*). Other regions are involved in nutrition (*LCT*), skin and hair traits (Type II Keratins and *TRMP1*), and fertility (*MAPT* inversion).

The highest selection signal, with an IBD rate of 0.14%, comes from a chromosome 2 region containing the *LCT* gene which has a variant selected for lactose tolerance in Europeans.[50] The selected variant is thought to have arisen, or at least begun to increase in frequency, around the time of the advent of cattle

farming in Europe, around 7500 years ago,[51] and selection has been so strong that the selected variant allele frequency is now around 75% in individuals of British descent.[52] Since IBD is affected by recent selection, it is not surprising that this signal is most prominent.

In contrast, the immunity-related *HLA* genes in the multi histocompatibility complex (*MHC*) on chromosome 6 have been under selection over a much longer time period,[53; 54] and this region has a much lower peak IBD rate (0.028%) than the *LCT* region. Various sub-regions within the *MHC* appear to have been subject to adaptive introgression from Neanderthals and Denisovans, but it is difficult to be certain because long-term balancing selection across the region can result in signals that look like adaptive introgression.[55]

The second-highest signal, with an IBD rate of 0.051%, comes from a chromosome 15q13 region containing *TRPM1*, a pigmentation gene that has been shown to have been subject to selection in non-Africans.[56; 57]

The high IBD rate region on chromosome 12q24 (0.029% IBD rate) encompasses the *OAS* locus which is involved in immunity.[58] This locus has a Neanderthal haplotype present at high frequency in non-Africans that has been subject to positive selection.[59]

The high IBD rate region on chromosome 17q21 (0.026% IBD rate) encompasses the 17q21.31 *MAPT* inversion, for which the H2 form has undergone positive selection in Europe and is associated with increased fertility and higher recombination rates in females in Iceland.[60]

The high IBD rate region on chromosome 6q21 (0.026% IBD rate) contains the *PRDM1* and *ATG5* genes. This pair of genes is associated with autoimmune diseases and cancer[61-63] and shows a signal of recent selection in HapMap data.[64]

The high IBD rate region on chromosome 3p21 (0.025% IBD rate) is a region of adaptive introgression from Neanderthals.[65; 66] It contains chemokine receptor genes, including CCR9 and CXCR6, that are involved in immunity.[67; 68]

The high IBD rate region on chromosome 4p14 (0.025% IBD rate) contains several toll-like receptor genes (*TLR1, TLR6,* and *TLR10*) that are involved in immunity, and this region has experienced adaptive introgression from Neanderthals.[66] As well as the adaptive introgression, this region shows other signals of selection, including geographic differentiation within the UK[69] and signs of recent positive selection among non-Africans.[70]

The high-IBD rate region on chromosome 12q13 (0.024% IBD rate) contains the Type II Keratin genes, which code filament proteins that provide a major structural role in epithelial cells.[71] This region has experienced adaptive introgression from Neanderthals.[66]

The remaining locus on chromosome 22q11 (0.025% IBD rate) has not previously been highlighted as being under selection, to the best of our knowledge. This locus includes *UBE2L3* which is associated with multiple auto-immune diseases.[72] These associations make this locus a strong candidate for natural selection.[73]

We also investigated the next ten regions with highest IBD rates (Table 2). These regions include the pigmentation gene *SLC45A2* on chromosome 5p13 (0.024% IBD rate) and the epidermal differentiation complex locus on chromosome 1q21.3 (0.020% IBD rate). Both of these loci are known to have undergone recent positive selection.[74; 75]

The variability of IBD rate across the genome is much lower for ibd-ends than for hap-ibd (Figure S7). The standard deviation of the ibd-ends IBD rate is 0.0056%, whereas the standard deviation for hap-ibd is an order of magnitude higher at 0.060%. Ibd-ends is a much better tool for investigating regions of potential selection than length-based methods because length-based methods can report artifactually high rates of

IBD in some regions, such as regions containing large gaps in marker coverage. For example, IBD segment detection with four length-based IBD detection methods produce a 40-3000 fold higher IBD rate at the chromosome 1 centromere in UK Biobank data compared to the background rate.[33]

Misspecification of the genetic map can also lead to spurious regions of high IBD rate in IBD selection scans, because overestimation of genetic length will result in a larger number of IBD segments that pass the length threshold. To investigate the impact of map choice on the IBD selection analysis, we analyzed a subset of 50,000 White British individuals with three maps. We observed considerable differences in the regions of highest IBD rate when using different maps (Figure S8). With the HapMap linkage-disequilibrium (LD) based map[76] there are 20 regions with IBD rate > 0.05% (compared to two with Bherer et al.'s family-based map),[49] with three of these occurring at the locations of centromeres. Furthermore, the IBD rate has higher variability when using the HapMap map, with a standard deviation of 0.018% (compared to 0.0057% with Bherer et al.'s map in this 50,000-individual subset). LD-based maps are known to be biased in regions of selection, with genetic distances underestimated in these regions,[77] which would lead to an apparent decrease in IBD rate in these regions. Thus, increases in apparent IBD rate with the LD-based HapMap map compared to Bherer et al.'s pedigree-based map must be due to other effects, such as unmodelled features of the data. The IBDrecomb map is designed to obtain uniform IBD rate across the genome.[13] Thus regions of likely selection such as *LCT* do not have high IBD rates when using this map. The analysis with IBDrecomb still has some regions of high IBD rate (19 regions with IBD rate > 0.05%), which are concentrated at chromosome ends (where inflation of map distances is known to occur[13]) and at centromeric gaps. These regions are likely to be artifacts.

The MHC region has a much higher selection signal from analysis with the HapMap map than with Bherer et al.'s map. The genetic lengths of the MHC are 3.29 cM for the HapMap map, 2.20 cM for Bherer et al.'s map and 1.44 cM for the IBDrecomb map. Three previous IBD-based selection analyses have used the HapMap map,[14-16] and two of these analyses also had very strong MHC signals.[14; 16]

Discussion

We presented a method for calculating the posterior probability distributions of IBD segment endpoints. We showed that the method can be applied to large data sets, such as the UK Biobank SNP array data on 408,883 White British individuals and simulated sequence data on 50,000 individuals. In the UK Biobank data, we analyzed 77.7 billion candidate IBD segments, and found 51.7 billion IBD segments for which the length based on sampled endpoints is greater than 2 cM.

In addition to quantifying endpoint uncertainty, a major advantage of our method is that it handles genotype errors and other discordances within IBD segments in a principled way, in contrast to many other methods for IBD segment detection which use ad hoc approaches. Our method does not directly account for haplotype phase uncertainty, however statistical phasing of non-rare variants in large array-typed cohorts is now extremely accurate,[36] and technologies for generating highly-accurate, phase-resolved sequence data are becoming available.[37]

The sampled lengths and endpoints of the segments are unbiased and have low variance. IBD segments defined by the sampled endpoints can be used in downstream analyses. In addition, the estimates of endpoint uncertainty can be used to include more data in analyses. For example, when estimating genome-wide mutation rates from IBD segments, it is important to be confident that one does not count mutations that actually lie outside the IBD segment. In the past, this has been achieved by trimming 0.5 cM from the putative IBD endpoints,[10; 12] but trimming using a small quantile of the uncertainty distribution would be expected to result in less trim (and hence more data), while maintaining accuracy. Alternatively, methods could be developed that directly account for endpoint uncertainty without trimming.

Another analysis that could incorporate estimated IBD segment end point uncertainty is IBD-based estimation of recombination maps. Endpoints of IBD segments are points of past recombination which

are used to estimate the map. Consequently, misspecification of IBD endpoints adds noise and bias to the resulting map. One could improve recombination map estimation by incorporating endpoint uncertainty into an iterative procedure that uses the current estimated recombination map to refine the estimates of IBD segment endpoints.

IBD segment endpoint uncertainty could also be used to improve IBD-based estimation of recent demographic history. For this application, it is not the actual IBD endpoints that matter, but rather the distribution of IBD segment lengths, which one could estimate by sampling endpoints from the posterior distribution. A threshold of 2 or 3 cM on IBD segment length is usually applied with current methods since lengths of shorter segments have higher relative uncertainty.[4-8] With sampled endpoints, it may be possible to use a lower length threshold and thus estimate demographic history further into the past.

We found that the IBD segments obtained from sampled endpoints provide excellent input data for an IBD-based selection scan. Nine of the ten top regions in our UK Biobank analysis are regions of known selection, and the remaining region is a good candidate for selection. Two particular features of our IBD-based selection analysis contribute to its success. The first is unbiased estimation of IBD segment lengths, even in the presence of gaps in marker coverage, which eliminates many spurious signals. In contrast, length-based IBD detection methods tend to have regions with inflated IBD rates,[33] which would cause spurious signals if used in a selection scan. The second is the choice of genetic map. We found that for IBD-based selection scans, pedigree-based maps based on actual observations of recombination produce more accurate results than maps based on LD or on IBD sharing, because effects of selection and other unmodelled features in local genomic regions can bias LD-based and IBD-based maps, and this bias results in spurious signals of selection.

Two limitations of the current study could be addressed in future work. First, our modelling assumes a homogeneous population with the same distribution of IBS segment lengths for all pairs of individuals.

Our analysis of a simulated combined African and European ancestry sample indicates that violation of this assumption introduces a little mis-calibration in the posterior endpoint probabilities. One solution would be to adjust for ancestry of the samples, or for local ancestry in the case of admixed samples. Second, we did not give any special treatment to the ends of chromosomes. Consequently, IBD segments that truly extend to the end of the chromosome are estimated to have endpoints occurring near, rather than at, the chromosome ends. One work-around would be to use the input candidate segment endpoint in preference to the ibd-ends endpoint when the input candidate segment endpoint is located at the chromosome end. A more sophisticated solution would be to use the estimated distribution of IBD segment lengths as a prior distribution when analyzing IBD segments that may extend to the end of the chromosome.

Appendices

Appendix 1: Prior for IBD length distribution

In this appendix we derive a formula for the prior probabilities $P\big(R \in (x_i, x_{i+1})\big)$ in Equation 1. Let $Y = R - L_0$ be the length (measured in Morgans) from the left endpoint, L_0, of the segment to the right endpoint R. Here we assume that the left endpoint is known, but in practice we iteratively update its estimated value as described in the main text. We write $F(y) = P(Y \leq y)$ for the prior probability distribution on Y. We model the population size as constant diploid size N, with $N = 10{,}000$, which reflects the approximate average historical size of out-of-Africa populations.

Summing over possible values for G, the number of generations to the common ancestor of the IBD segment, we obtain

$$
\begin{aligned}
F(y) &= P(Y \le y) \\
&= \sum_{g=1}^{\infty} P(Y \le y | G = g) P(G = g) \\
&= \sum_{g=1}^{\infty} (1 - e^{-2gy}) \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N} \\
&= 1 - (2Ne^{2y} - 2N + 1)^{-1}
\end{aligned}
$$

In these calculations, we use the fact that the IBD segment length conditional on the number of generations g to the common ancestor is exponentially distributed with rate parameter $2g$,[7] and the fact that the number of generations to the most recent common ancestor in a population of constant size N is geometrically distributed.[4] The final expression for $F(y)$ in the calculation is obtained by applying the formula for a geometric series.

The prior distribution for the right endpoint is:

$P(R \le x) = P(Y \le x - L_0) = F(x - L_0)$ for $L_0 < x \le x_M$

and the probability in Equation 1 that R falls in a certain interval can be calculated as:

$$P\big(R \in (x_i, x_{i+1})\big) = F(x_{i+1} - L_0) - F(x_i - L_0).$$

In Appendix 3, we will require the inverse of F, which is $F^{-1}(p) = \frac{1}{2}\log\left(\frac{p+2N(1-p)}{2N(1-p)}\right)$.

Appendix 2: Modelling the IBS data beyond the IBD segment

In this appendix we describe how to calculate the conditional probability $P(D[i+1, M] \,|R \in (x_i, x_{i+1}))$ in Equation 1. This is the probability of the IBS data to the right of the right IBD endpoint.

Let $m_i(j)$ $(j \ge 1)$ be the ordered indices of the discordant markers to the right of x_i, plus the last marker on the chromosome if that marker is not included in the list of discordant positions. We approximate the IBS process to the right of the IBD endpoint as a renewal process with a renewal every time there is a

discordant marker. Then the probability of the IBS data to the right of x_i given that the IBD endpoint occurred in the interval (x_i, x_{i+1}) is:

$$P(D[i+1, M] \,|R \in (x_i, x_{i+1})) \approx P(D[i+1, m_i(1)]) \prod_{j>1} P(D[m_i(j-1)+1, m_i(j)])$$

Each interval of IBS data in the preceding equation has the form $D[a, b]$, where a and b are marker indices, and has the property that the alleles on the two haplotypes H_1 and H_2 are identical at all positions i satisfying $a \leq i < b$, but are discordant at position b (except in some cases where b is the last marker on the chromosome). Our approach to estimating the probabilities of these IBS interval data is to estimate the probability $p_{a,b}$ that two randomly chosen haplotypes have identical alleles over the interval $[a, b]$ (i.e. for all marker indicies i with $a \leq i \leq b$). If haplotypes H_1 and H_2 have identical alleles at marker b, then we set $P(D[a, b]) = p_{a,b}$, and if the two haplotypes have discordant alleles at marker b, we set $P(D[a, b]) = p_{a,(b-1)} - p_{a,b}$. We estimate $p_{a,b}$ empirically from the data.

Our method for estimating $p_{a,b}$ depends on the length of the interval $[a, b]$. For short intervals, we use data in the interval so that the estimated probability incorporates the local genomic context, such as high or low marker density, high or low heterozygosity, and high or low levels of LD. For long intervals, we will use chromosome-wide data. As the interval length becomes longer, $p_{a,b}$ will tend to decrease because the probability of observing a long IBS interval for a random pair of haplotypes is small. Small probabilities are more difficult to estimate than long ones, so we need to bring in data from the rest of the genome. Furthermore, long intervals represent long IBS (alleles at all markers in the interval are identical except for the final marker), and long IBS is likely to be the result of a long IBD segment. Excluding the effects of selection, the distribution of IBD lengths should be uniform across the genome, so estimating the frequency of such long IBS segments from data across the genome is appropriate.

For short intervals, we estimate $p_{a,b}$ by the proportion of pairs of haplotypes that have identical by state alleles for markers in the interval $[a, b]$. We consider an interval $[a, b]$ to be a short interval if the estimated $\hat{p}_{a,b}$ satisfies $\hat{p}_{a,b} \geq 0.001$ when it is estimated from the haplotypes in the interval. We use 10,000 randomly sampled haplotypes (or all haplotypes if the sample size is 5000 or fewer individuals) to estimate the short interval IBS probabilities.

For longer intervals we estimate $p_{a,b}$ using the global distribution of sampled one-sided IBS lengths. A one-sided IBS length is the cM distance from a random starting position to the first non-IBS marker in the direction toward the center of the chromosome for a random pair of haplotypes. We estimate $p_{a,b}$ by the proportion of sampled one-sided IBS lengths that are greater than $(x_b - x_a)$. When estimating the global IBS length, we randomly select 1000 positions and randomly sample 2000 one-sided IBS lengths for each position. We then exclude the sampled lengths at positions for which the IBS length are significantly longer than average. We do this by calculating the 90th percentile of the 2000 segment lengths for each position. If the 90th percentile at a position is more than 3 times the median 90th percentile from all 1000 positions, we discard the position. This filtering protects against selecting positions near gaps in marker coverage.

Appendix 3: Obtaining the posterior cumulative distribution function for the endpoint

In this appendix we describe how to calculate the cumulative distribution function of the posterior distribution for the right endpoint of the IBD segment, as well as how to sample from this posterior distribution.

We assume that the data D are not informative about the location of the endpoint within the inter-marker interval given that the endpoint occurs within the interval, i.e. we assume that $P(x_i < R \leq x | x_i < R < x_{i+1}, D) = P(x_i < R \leq x | x_i < R < x_{i+1})$ for $x_i < x < x_{i+1}$, since there are no data within the interval. Write $p_i = P(R \leq x_i | D)$, which can be estimated using Equation 1 with the

procedures described above. As in Appendix 1, we write $Y = R - L_0$ for the length of the segment (in Morgans), and $F(y) = P(Y \leq y)$ for the prior on IBD lengths. For x satisfying $x_i < x < x_{i+1}$:

$$\begin{aligned}
P(R \leq x|D) &= P(R \leq x_i|D) + P(x_i < R \leq x|D) \\
&= P(R \leq x_i|D) + P(x_i < R \leq x|x_i < R \leq x_{i+1}, D)P(x_i < R \leq x_{i+1}|D) \\
&= P(R \leq x_i|D) + P(x_i < R \leq x|x_i < R \leq x_{i+1})P(x_i < R \leq x_{i+1}|D) \\
&= P(R \leq x_i|D) + \frac{P(x_i < R \leq x)}{P(x_i < R \leq x_{i+1})}P(x_i < R \leq x_{i+1}|D) \\
&= p_i + \frac{P(x_i - L_0 < Y \leq x - L_0)}{P(x_i - L_0 < Y \leq x_{i+1} - L_0)}(p_{i+1} - p_i) \\
&= p_i + \frac{F(x - L_0) - F(x_i - L_0)}{F(x_{i+1} - L_0) - F(x_i - L_0)}(p_{i+1} - p_i)
\end{aligned}$$

Note that $p_i = P(R \leq x_i|D)$ is conditional on the data, whereas $F(x_i - L_0) = P(Y \leq x_i - L_0) = P(R \leq x_i)$ is a prior probability and is not conditional on the data.

Then to find the p-th quantile we want to find $x^{(p)}$ such that $P\left(R < x^{(p)}\middle|D\right) = p$. Solving the above equation for $x = x^{(p)}$we obtain

$$x^{(p)} = L_0 + F^{-1}\left(\left(F(x_{i+1} - L_0) - F(x_i - L_0)\right)\frac{p - p_i}{p_{i+1} - p_i} + F(x_i - L_0)\right).$$

The formula for $F^{-1}(p)$ can be found in Appendix 1.

In order to obtain a sampled value from the posterior probability of the endpoint, we first generate a realization u from the Uniform(0,1) distribution, and we then obtain $x^{(u)}$ using the above formula with u in place of p, which is then a realization from the desired distribution. This is an example of using the inverse transform principle to sample from a distribution.[78]

Supplemental Data

Supplemental data consist of 8 figures and 1 table.

Declaration of Interests

The authors declare no competing interests

Acknowledgments

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award numbers HG005701 and HG008359. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research has been conducted using the UK Biobank Resource under Application Number 19934.

Web Resources

ibd-ends: https://github.com/browning-lab/ibd-ends

Bherer et al.'s refined European map:

https://github.com/cbherer/Bherer_etal_SexualDimorphismRecombination/blob/master/Refined_EUR_genetic_map_b37.tar.gz

HapMap map: ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/genetic_map_HapMapII_GRCh37.tar.gz

IBDrecomb European-American map:

https://github.com/YingZhou001/IBDrecomb/tree/master/maps.b37/FHS.map.tar.gz

References

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

2. Huff, C.D., Witherspoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., Zhang, Y., Tuohy, T.M., Neklason, D.W., Burt, R.W., Guthery, S.L., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Research 21, 768-774.

3. Ramstetter, M.D., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Mezey, J.G., and Williams, A.L. (2017). Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. Genetics 207, 75-82.

4. Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. American Journal of Human Genetics 91, 809-822.

5. Palamara, P.F., and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. Bioinformatics 29, i180-188.

6. Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol 11, e1001555.

7. 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.

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

9. Campbell, C.D., Chong, J.X., Malig, M., Ko, A., Dumont, B.L., Han, L., Vives, L., O'Roak, B.J., Sudmant, P.H., Shendure, J., et al. (2012). Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44, 1277-1281.

10. Palamara, P.F., Francioli, L.C., Wilton, P.R., Genovese, G., Gusev, A., Finucane, H.K., Sankararaman, S., Genome of the Netherlands Consortium, Sunyaev, S.R., de Bakker, P.I.W., et al. (2015). Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. American Journal of Human Genetics 97, 775–789.

11. Narasimhan, V.M., Rahbari, R., Scally, A., Wuster, A., Mason, D., Xue, Y., Wright, J., Trembath, R.C., Maher, E.R., van Heel, D.A., et al. (2017). Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat Commun 8, 303.

12. Tian, X., Browning, B.L., and Browning, S.R. (2019). Estimating the Genome-wide Mutation Rate with Three-Way Identity by Descent. Am J Hum Genet 105, 883-893.

13. Zhou, Y., Browning, B.L., and Browning, S.R. (2020). Population-specific recombination maps from segments of identity by descent. The American Journal of Human Genetics.

14. Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. Genetics 186, 295-308.

15. Han, L., and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. European Journal of Human Genetics 21, 205-211.

16. 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.

17. Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175-195.

18. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome research 12, 1805-1814.

19. Workman, P., Blumberg, B., and Cooper, A. (1963). Selection, gene migration and polymorphic stability in a US White and Negro population. American journal of human genetics 15, 429.

20. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., and Risch, N.J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81, 626-633.

21. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science 328, 710-722.

22. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., 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.

23. 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.

24. Leutenegger, A.L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E.A. (2003). Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics 73, 516-523.

25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559-575.

26. Browning, S.R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178, 2123-2132.

27. Thompson, E.A. (2008). The IBD process along four chromosomes. Theoretical Population Biology 73, 369-373.

28. Browning, S.R., and Browning, B.L. (2010). High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86, 526-539.

29. Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F.C., and Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genetic Epidemiology 33, 266-274.

30. Han, L., and Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype data. Genetic Epidemiology 35, 557-567.

31. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics 40, 1068-1075.

32. 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.

33. 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. The American Journal of Human Genetics.

34. Thompson, E.A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194, 301-326.

35. Chiang, C.W., Ralph, P., and Novembre, J. (2016). Conflation of Short Identity-by-Descent Segments Bias Their Inferred Length Distribution. G3 (Bethesda) 6, 1287-1296.

36. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat Commun 10, 1-10.

37. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., and Olson, N.D. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature biotechnology 37, 1155-1162.

38. Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nature reviews Genetics 12, 703-714.

39. Williams, A.L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S.R., Curran, J.E., Duggirala, R., et al. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4.

40. 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.

41. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., and Kang, H.M. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv, 563866.

42. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209.

43. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Molecular biology and evolution 36, 632-637.

44. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol 12, e1004842.

45. 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.

46. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., and Jun, G. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. science 337, 64-69.

47. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.

48. Jeffrey, R.A., Christopher, B.C., Noah, D., Jared, G.G., Ariella, L.G., Christopher, C.K., Aaron, P.R., Georgia, T., Franz, B., and Jedidiah, C. A Community-Maintained Standard Library of Population Genetic Models. eLife 9, e54967.

49. Bhérer, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat Commun 8, 1-9.

50. 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. The American Journal of Human Genetics 74, 1111-1120.

51. Itan, Y., Powell, A., Beaumont, M.A., Burger, J., and Thomas, M.G. (2009). The origins of lactase persistence in Europe. PLoS Comput Biol 5, e1000491.

52. Ho, M., Povey, S., and Swallow, D. (1982). Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples. American journal of human genetics 34, 650.

53. Hughes, A.L., and Yeager, M. (1998). Natural selection at major histocompatibility complex loci of vertebrates. Annual review of genetics 32, 415-435.

54. Meyer, D., Single, R.M., Mack, S.J., Erlich, H.A., and Thomson, G. (2006). Signatures of demographic history and natural selection in the human major histocompatibility complex loci. Genetics 173, 2121-2142.

55. Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. Nature Reviews Genetics 16, 359-371.

56. Storz, J.F., Payseur, B.A., and Nachman, M.W. (2004). Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Molecular biology and evolution 21, 1800-1811.

57. Hider, J.L., Gittelman, R.M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J.M., and Parra, E.J. (2013). Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. BMC evolutionary biology 13, 150.

58. Hu, J., Wang, X., Xing, Y., Rong, E., Ning, M., Smith, J., and Huang, Y. (2018). Origin and development of oligoadenylate synthetase immune system. BMC evolutionary biology 18, 1-13.

59. 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, 1-15.

60. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., and Gudnadottir, V.G. (2005). A common inversion under selection in Europeans. Nature genetics 37, 129-137.

61. Raychaudhuri, S., Thomson, B.P., Remmers, E.F., Eyre, S., Hinks, A., Guiducci, C., Catanese, J.J., Xie, G., Stahl, E.A., and Chen, R. (2009). Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nature genetics 41, 1313-1318.

62. Zhou, X.-j., Lu, X.-l., Lv, J.-c., Yang, H.-z., Qin, L.-x., Zhao, M.-h., Su, Y., Li, Z.-g., and Zhang, H. (2011). Genetic association of PRDM1-ATG5 intergenic region and autophagy with systemic lupus erythematosus in a Chinese population. Ann Rheum Dis 70, 1330-1337.

63. Best, T., Li, D., Skol, A.D., Kirchhoff, T., Jackson, S.A., Yasui, Y., Bhatia, S., Strong, L.C., Domchek, S.M., and Nathanson, K.L. (2011). Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin's lymphoma. Nature medicine 17, 941-943.

64. Ramos, P.S. (2017). Population genetics and natural selection in rheumatic disease. Rheumatic Disease Clinics 43, 313-326.

65. Ding, Q., Hu, Y., Xu, S., Wang, J., and Jin, L. (2014). Neanderthal introgression at chromosome 3p21. 31 was under positive natural selection in East Asians. Molecular biology and evolution 31, 683-695.

66. 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 e59.

67. Paust, S., Gill, H.S., Wang, B.Z., Flynn, M.P., Moseman, E.A., Senman, B., Szczepanik, M., Telenti, A., Askenase, P.W., Compans, R.W., et al. (2010). Critical role for the chemokine receptor CXCR6 in NK cell-mediated antigen-specific memory of haptens and viruses. Nat Immunol 11, 1127-1135.

68. Papadakis, K.A., Prehn, J., Nelson, V., Cheng, L., Binder, S.W., Ponath, P.D., Andrew, D.P., and Targan, S.R. (2000). The role of thymus-expressed chemokine and its receptor CCR9 on lymphocytes in the regional specialization of the mucosal immune system. J Immunol 165, 5069-5076.

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

71. Rogers, M.A., Edler, L., Winter, H., Langbein, L., Beckmann, I., and Schweizer, J. (2005). Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13. 13. Journal of investigative dermatology 124, 536-544.

72. Lewis, M.J., Vyse, S., Shields, A.M., Boeltz, S., Gordon, P.A., Spector, T.D., Lehner, P.J., Walczak, H., and Vyse, T.J. (2015). UBE2L3 polymorphism amplifies NF-κB activation and promotes plasma cell

development, linking linear ubiquitination to multiple autoimmune diseases. The American Journal of Human Genetics 96, 221-234.

73. Ramos, P.S., Shedlock, A.M., and Langefeld, C.D. (2015). Genetics of autoimmune diseases: insights from population genetics. Journal of human genetics 60, 657-664.

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

75. Mathyer, M.E., Brettmann, E.A., Schmidt, A.D., Goodwin, Z.A., Quiggle, A.M., Oh, I.Y., Tycksen, E., Zhou, L., Estrada, Y.D., and Wong, X.C.C. (2019). An enhancer: involucrin regulatory module impacts human skin barrier adaptation out-of-Africa and modifies atopic dermatitis risk. bioRxiv, 816520.

76. The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.

77. O'Reilly, P.F., Birney, E., and Balding, D.J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome research 18, 1304-1313.

78. Devroye, L. (1986). Non-uniform random variate generation.(New York: Springer-Verlag).

79. Zody, M.C., Jiang, Z., Fung, H.-C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., and Abouelleil, A. (2008). Evolutionary toggling of the MAPT 17q21. 31 inversion region. Nature genetics 40, 1076-1083.

80. Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G.P., Bresolin, N., Clerici, M., and Sironi, M. (2010). Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet 6, e1000849.

81. Mischke, D., Korge, B.P., Marenholz, I., Volz, A., and Ziegler, A. (1996). Genes encoding structural proteins of epidermal cornification and S100 calcium-binding proteins form a gene complex ("

epidermal differentiation complex") on human chromosome 1q21. Journal of Investigative Dermatology 106.

82. Mikkelsen, T., Hillier, L., Eichler, E., Zody, M., Jaffe, D., Yang, S.-P., Enard, W., Hellmann, I., Lindblad-Toh, K., and Altheide, T. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69-87.

83. Freathy, R.M., Mook-Kanamori, D.O., Sovio, U., Prokopenko, I., Timpson, N.J., Berry, D.J., Warrington, N.M., Widen, E., Hottenga, J.J., and Kaakinen, M. (2010). Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. Nature genetics 42, 430-435.

84. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., and Gloyn, A.L. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature genetics 42, 105-116.

85. Bitarello, B.D., de Filippo, C., Teixeira, J.C., Schmidt, J.M., Kleinert, P., Meyer, D., and Andrés, A.M. (2018). Signatures of long-term balancing selection in human genomes. Genome Biol Evol 10, 939-955.

86. Graf, J., Hodgson, R., and Van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Human mutation 25, 278-284.

87. Van Der Lee, R., Wiel, L., Van Dam, T.J., and Huynen, M.A. (2017). Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. Nucleic Acids Res 45, 10634-10648.

88. Horikawa, Y., Iwasaki, N., Hara, M., Furuta, H., Hinokio, Y., Cockburn, B.N., Lindner, T., Yamagata, K., Ogata, M., and Tomonaga, O. (1997). Mutation in hepatocyte nuclear factor–1β gene (TCF2) associated with MODY. Nature genetics 17, 384-385.

89. Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J.T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B.A., and Baker, A. (2007). Two variants on chromosome

17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nature genetics 39, 977-983.

Figures and Tables

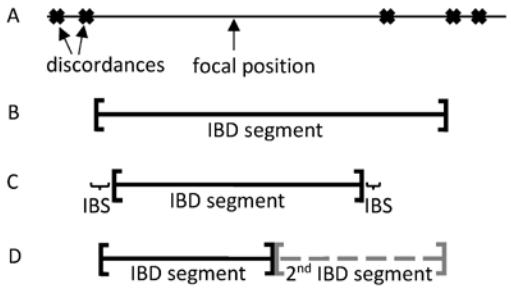

Figure 1: Uncertainty in IBD endpoints. A: Allele discordances between two haplotypes are represented as crosses. We wish to estimate the endpoints of the IBD segment that covers the focal position in the middle of the longest identity by state (IBS) interval. B-D: Three of the possibilities for the shared IBD segment that covers the focal position. B: The IBD segment contains the first discordance to the right of the focal position. C: The IBD segment does not extend all the way to the discordances and has short flanking segments of IBS. D: Two moderately long IBD segments are adjacent. In this case, the second IBD segment is not of direct interest because it does not cover the focal position.

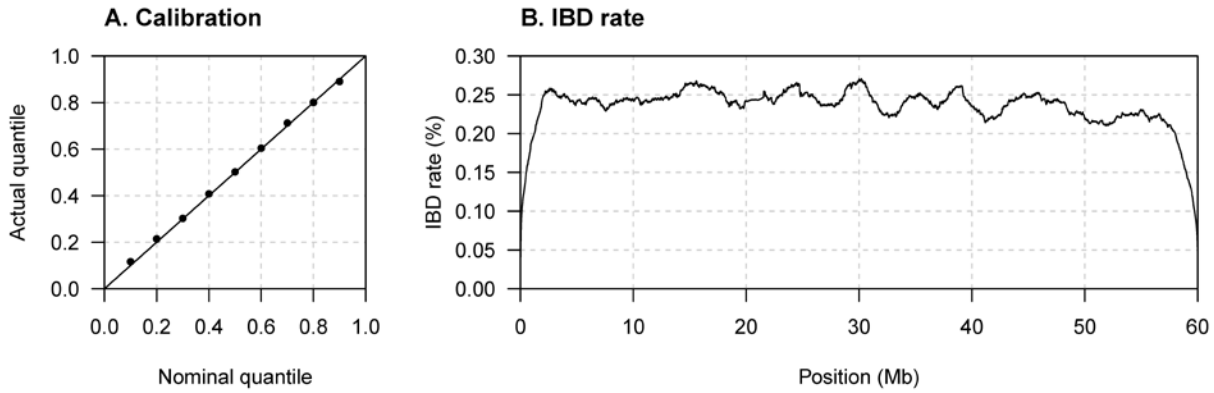

Figure 2: Method performance with uneven marker density. Sequence data on 2000 individuals were simulated under a constant effective population size. Markers located between 20 and 21.5 Mb were removed, and marker density varies every 100 kb (see Methods). The true haplotype phase is used in the analysis. (A) Quantile-quantile plot assessing the calibration of the estimated endpoint uncertainty. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint (points on the plot). The $y = x$ line is shown for comparison. Results for the left endpoints are similar but are not shown. (B) The y-axis is the IBD rate, which is the percentage of pairs of haplotypes for which the position on the chromosome is covered by a sampled IBD segment with length > 2 cM for the haplotype pair. The IBD rate is calculated at 10 kb intervals.

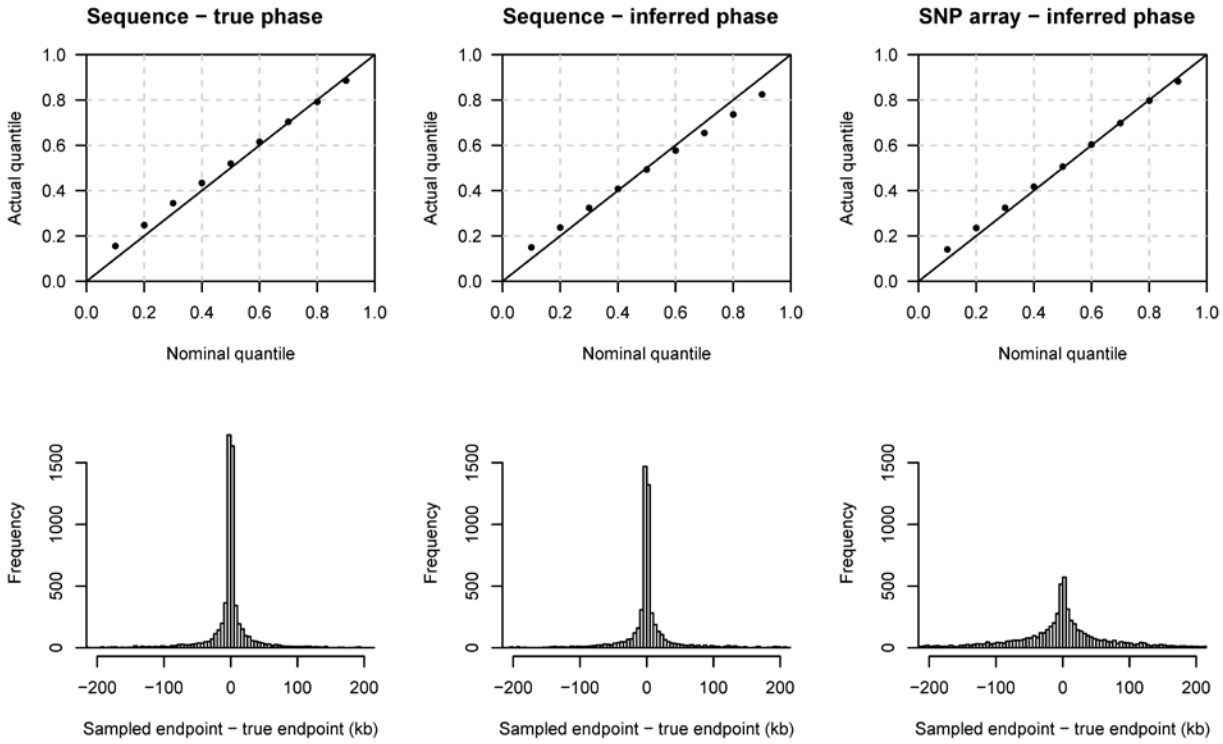

Figure 3: Method performance on UK-like simulated sequence and SNP array data. The data comprise 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. True IBD segment endpoints were determined for 1000 individuals, and these individuals were used to generate the results in this figure. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The left column is for analysis using the true haplotype phase. The middle column is for analysis using haplotype phase inferred using Beagle 5.1. The right column is for

data thinned to match a SNP array with 500,000 markers genome-wide (10,000 markers in the simulated 60 Mb interval), and with haplotype phase inferred using Beagle 5.1.

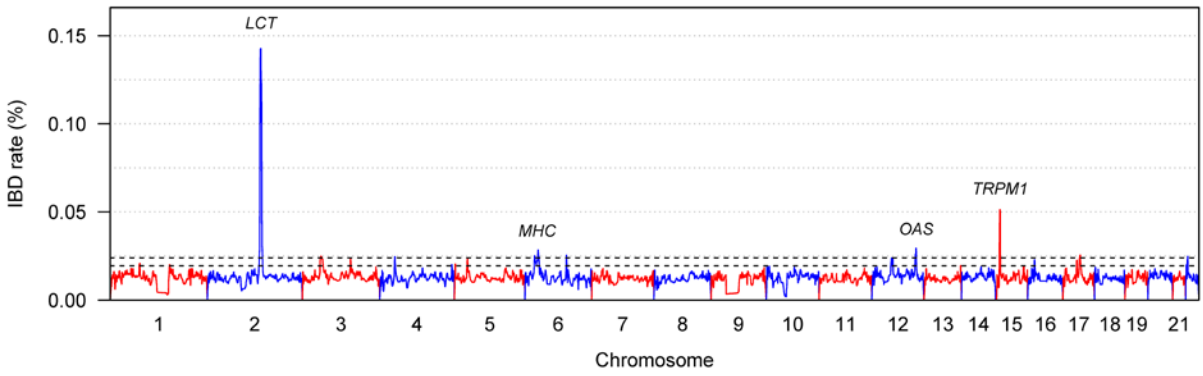

Figure 4: Rate of IBD segments along the autosomes in UK Biobank White British data. The x-axis shows position along each chromosome. Chromosomes alternate in color. Notable genes and regions (*LCT*, *MHC*, *OAS*, and *TRPM1*) located within the four highest peak regions are labelled. The y-axis is the IBD rate, which is the percentage of pairs of haplotypes for which the position on the chromosome is covered by a sampled IBD segment with length > 2 cM for the haplotype pair. The IBD rate is calculated at 10 kb intervals. The black dashed lines show the thresholds of 0.024% and 0.0196% used for the results in Tables 1 and 2 respectively.

Table 1: Regions of highest IBD rate in UK Biobank White British analysis. Regions in which the maximum IBD percentage is at least 0.024% are shown. Positions are in GRCh37 coordinates.

Chromosome	Peak position (Mb)	Region (Mb)[a]	Max IBD %	Notes[b]
2	136.65	134.75-138.47 (2q21-22)	0.1426	*LCT* (136.55-136.59) is known to be subject to recent positive selection.[50]
3	47.62	45.18-53.22 (3p21)	0.0249	Chemokine receptor genes including *CCR9* (45.93-45.94) and *CXCR6* (45.98-45.99). This region shows adaptive intogression from Neanderthals.[65]
4	38.84	38.12-40.21 (4p14)	0.0246	Toll-like receptor genes including *TLR1, TLR6,* and *TLR 10* (38.77-38.86). This locus is known to be under positive selection.[69]
6	24.87 33.97	24.03-36.63 32.43-36.27 (6p21-22)	0.0252 0.0284	MHC locus containing *HLA* genes (28.48-33.45) is known to be subject to strong pressures of natural selection.[54]
6	106.33	105.98-107.31 (6q21)	0.0255	*PRDM1* (106.53-106.56). This locus shows a signal of recent selection.[64]
12	53.46	49.17-54.75 (12q13)	0.0240	Type II Keratins (*KRT1-8*; 52.63-53.32). This locus has experienced adaptive introgression from Neanderthals.[66]

12	113.44	111.64-114.11 (12q24)	0.0294	*OAS* locus (*OAS*1-3) (113.34-113.45). This locus has experienced adaptive introgression from Neanderthals.[59]
15	31.61	30.90-32.46 (15q13)	0.0513	*TRPM1* (31.29-31.45). A pigmentation gene under selection in non-Africans.[56]
17	44.78	42.01-45.89 (17q21)	0.0256	17q21.31 *MAPT* inversion (43.44-44.85).[c] H2 form has been positively selected in Europe.[60]
22	22.56	21.36-23.05 (22q11)	0.0248	*UBE2L3* (21.90-21.98) is associated with multiple auto-immune diseases.[72]

[a] Region in which the IBD percentage is at least 80% of the value at the peak.

[b] See the main text for further notes and references.

[c] Positions from Zody et al.[79] lifted over from build 36 (chromosome 17: 40.80-42.20 Mb) to build 37.

Table 2: Secondary regions from UK Biobank selection scan. Regions with maximum IBD rate between 0.0195% and 0.024% are shown. Positions (in Mb) are in GRCh37 coordinates.

Chromosome	Peak position (Mb)	Region (Mb)[a]	Max IBD %	Notes
1	76.49	74.93-77.33 (1p31.1)	0.0209	*ST6GALNAC3* (76.54-77.10) is subject to virus-driven selective pressure.[80]
1	152.78	151.58-153.75 (1q21.3)	0.0201	Epidermal differentiation complex locus (151.96-153.60).[81] The locus with the greatest differentiation between humans and chimpanzees.[82] Recent positive selection in humans.[75]
3	123.41	122.36-124.70 (3q21)	0.0233	*ADCY5* (123.00-123.17) is associated with birth weight[83] and type II diabetes,[84] and is under long term balancing selection.[85]
4	184.59	184.30-185.48 (4q35.1)	0.0201	
5	2.68	2.32-2.99 (5p15.33)	0.0205	
5	33.84	32.56-34.63 (5p13)	0.0239	*SLC45A2* (33.94-33.98, previously known as *MATP*) is a pigmentation gene that has undergone recent positive selection in Europe.[74; 86]

13	112.87	112.54-113.55 (13q34)	0.0196	*SPACA7* (113.03-113.09) is a reproduction-related gene under positive selection in primates.[87]
16	18.08	17.35-18.74 (16p12.3)	0.0229	
17	36.03	35.17-36.51 (17q12)	0.0226	*HNF1B* (36.05-36.10, previously known as *TCF2*) is associated with diabetes.[88; 89]
22	18.89	18.57-19.60 (22q11.21)	0.0199	

[a] Region in which the IBD percentage is at least 80% of the value at the peak.

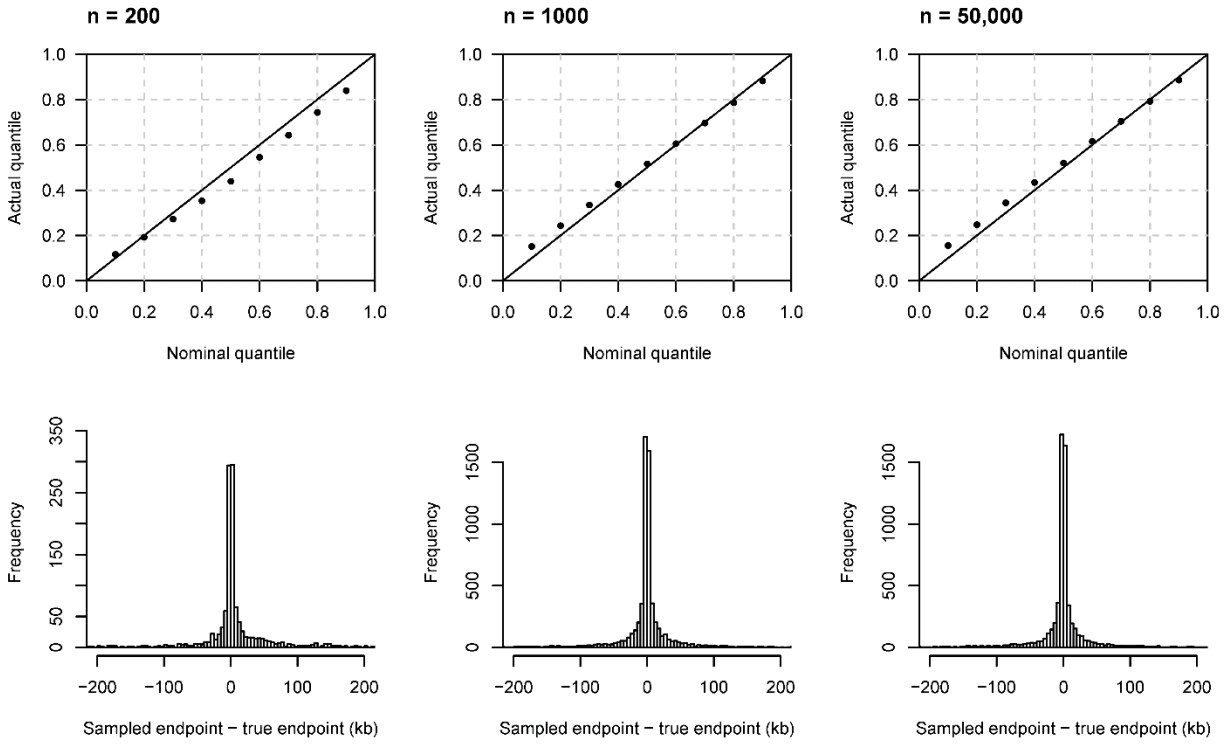

Figure S1: Performance on UK-like simulated data with varying sample sizes. The input data comprise sequence data on individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%, and true haplotype phase. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The left column is for five analyses each using 200 individuals. The middle column is for analysis using 1000 individuals. The right column is for analysis using 50,000 individuals with results were assessed on 1000 individuals for which true IBD endpoints were determined; these results are the same as the left column in Figure 3. The five n=200 analyses (left column) altogether

involve approximately 20% as many haplotype pairs as the other two analyses, and hence the results are subject to more statistical variation. Since there are 20% fewer data points than the n=1000 and n=50,000 analyses, the y-axis of the n=200 histogram (lower left) has been scaled proportionally, in order to make it comparable to the other two analyses.

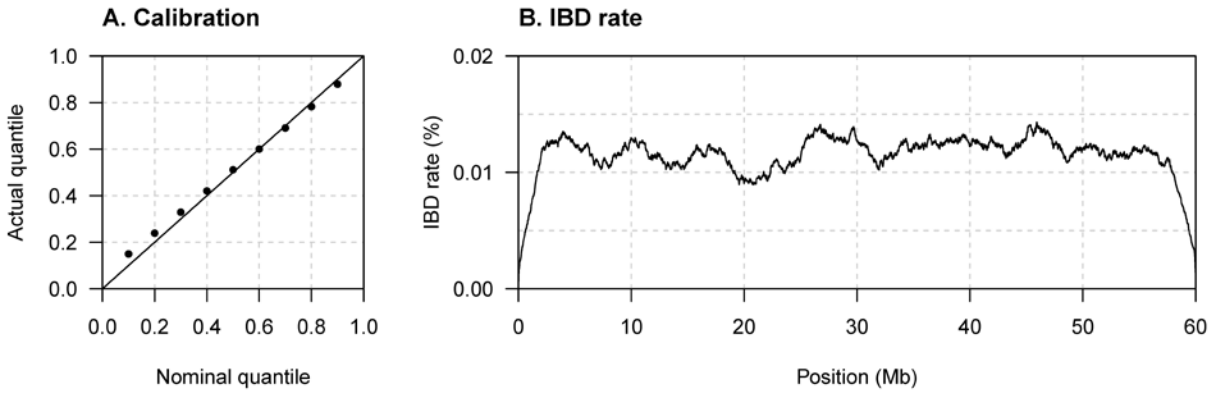

Figure S2: Method performance on simulated UK-like data with 1.5 Mb marker gap. 1000 individuals from the UK-like simulated sequence data with true phase were analyzed. Markers located between 20 and 21.5 Mb were removed. (A) Quantile-quantile plot assessing the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. Results for the left endpoints are similar but are not shown. (B) The y-axis is the IBD rate, which is the percentage of pairs of haplotypes for which the position on the chromosome is covered by a sampled IBD segment with length > 2 cM. The IBD rate is calculated at 10 kb intervals.

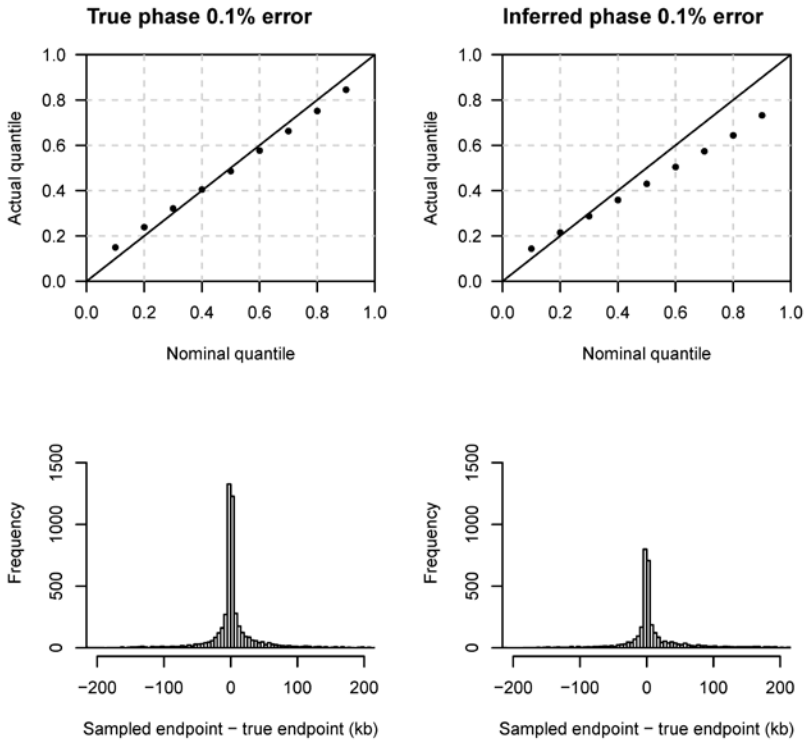

Figure S3: Performance on UK-like simulated data with a 0.1% rate of genotype error. The data comprise sequence data on 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.1%. Results were calculated using true IBD segment endpoints for all pairs of individuals within a subset of 1000 individuals. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The left column is for analysis using the true haplotype phase. The right column is for analysis using haplotype phase inferred using Beagle 5.1. There are 26% fewer

segments in the inferred phase analysis, because the higher genotype error rate results in some undetected IBD segments at the hap-ibd detection step.

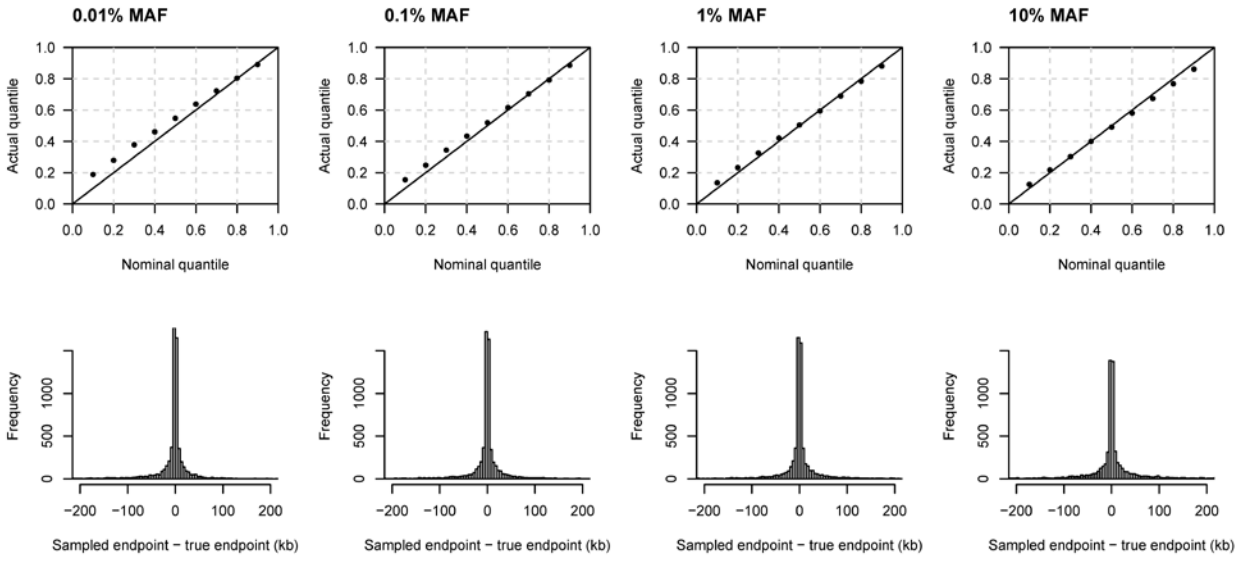

Figure S4: Performance on UK-like simulated data with varying minor allele frequency filters. The data comprise sequence data on 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The true haplotype phase is used in all analyses. The minimum minor allele frequency threshold applied in the ibd-ends analysis is shown above each column. A MAF threshold of 0.01% corresponds to at least 10 copies of the minor allele in these data. Compute times were 4.6 hours (0.01% MAF), 1.4 hours (0.1% MAF), 60 minutes (1% MAF), and 51 minutes (10% MAF). The default MAF for ibd-ends is 0.1%, so the second column shows the same results as the left column of Figure 3.

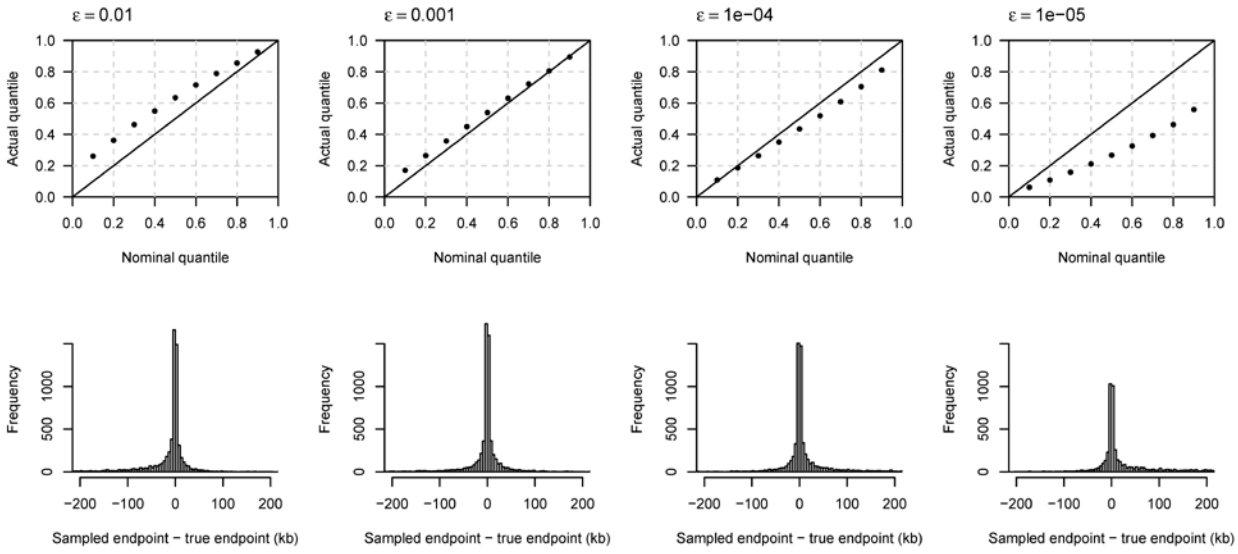

Figure S5: Performance on UK-like simulated data with varying analysis error rates. The data comprise sequence data on 1000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The true haplotype phase is used in all analyses. The analysis error rate, ϵ, is shown above each column. The estimated error rates can be found in Table S1.

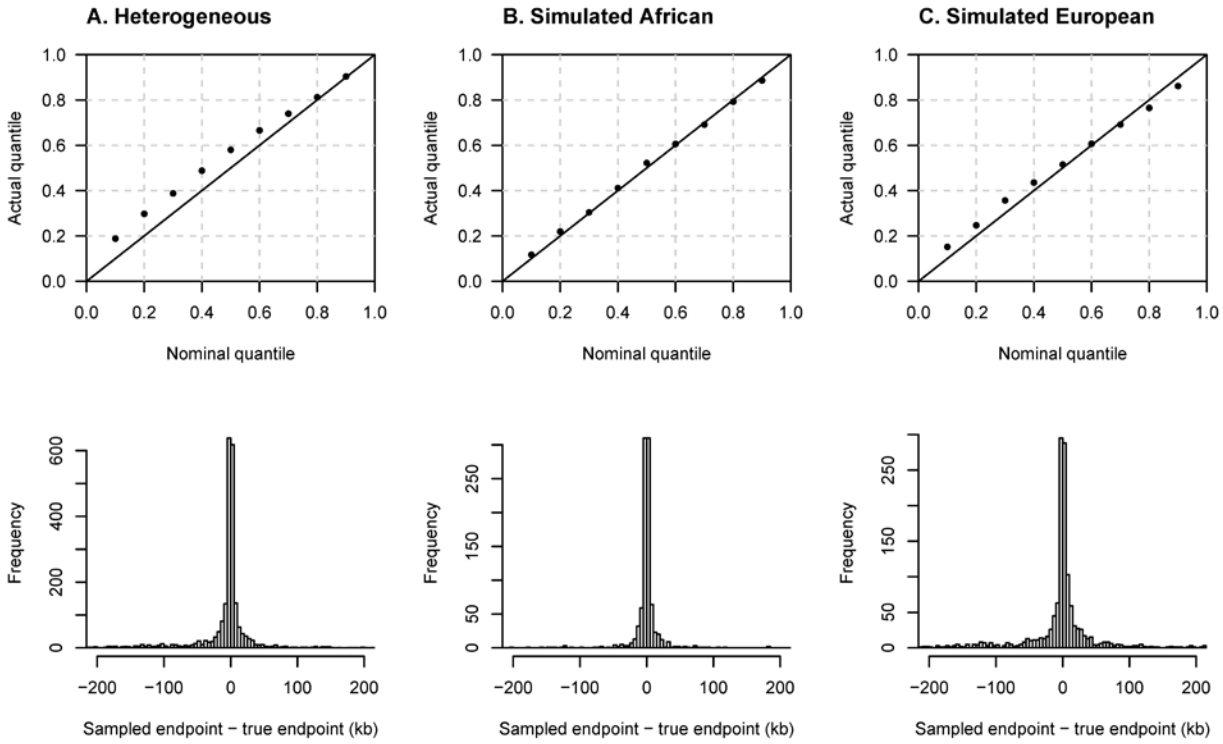

Figure S6: Performance in a simulated two-population model. The data comprise sequence data on 500 individuals each from simulated African and European demographic histories. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line is shown for comparison. The actual quantile (y-axis) corresponding to a given nominal quantile (x-axis) is the proportion of segments for which the reported nominal quantile of the right endpoint is greater than the true right endpoint. The bottom row shows histograms of the right endpoint sampled from the estimated posterior distribution minus the true right endpoint of the underlying segment. Histogram bin widths are 5 kb. Results for the left endpoints are similar but are not shown. The true haplotype phase is used in all analyses. (A) Combined analysis of individuals of simulated African and simulated European ancestry. (B) Analysis of individuals of simulated African ancestry only. (C) Analysis of individuals of simulated European ancestry only.

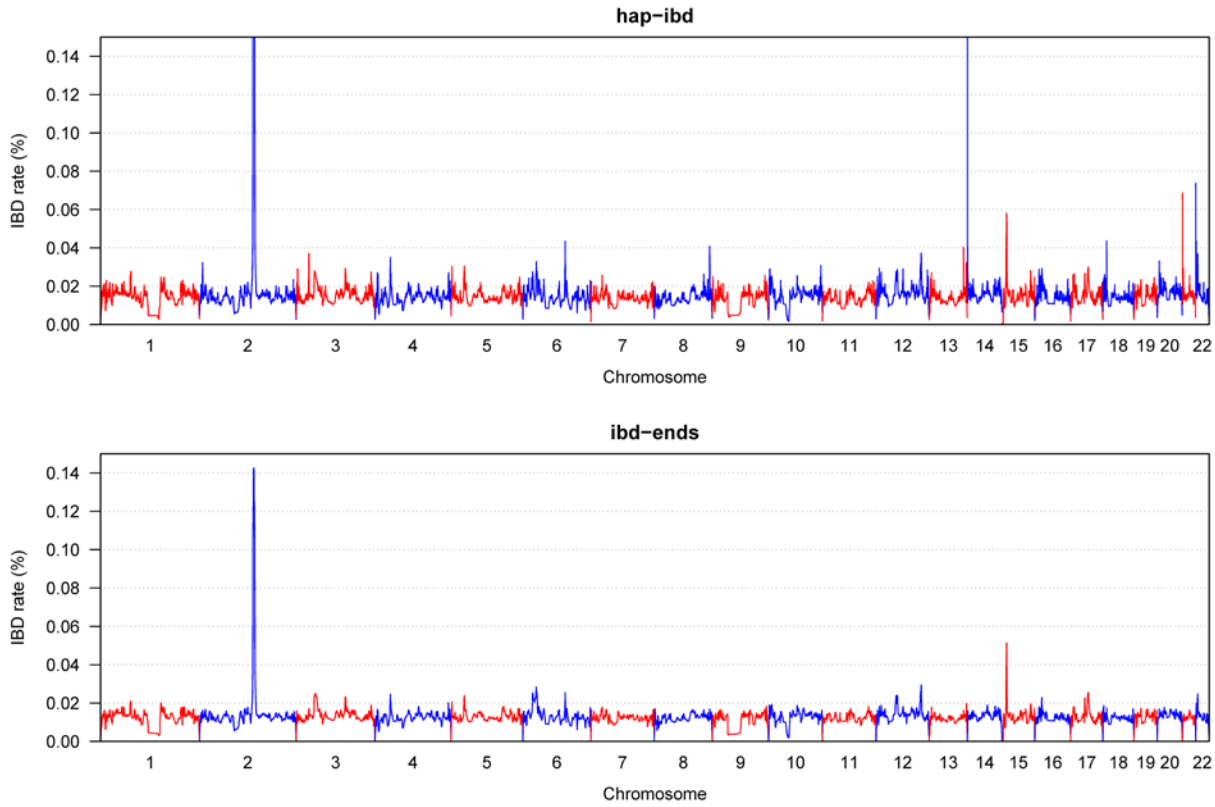

Figure S7: Comparison of rates of IBD detected by hap-ibd and ibd-ends in the UK Biobank White British data. The x-axis shows position along each chromosome. Chromosomes alternate in color. The y-axis is the IBD rate, which is the percentage of pairs of haplotypes for which the position on the chromosome is covered by a sampled IBD segment with length > 2 cM for the haplotype pair. The IBD rate is calculated at 10 kb intervals. The peak heights on chromosomes 2 and 14 for hap-ibd are 0.54% and 2.79% respectively. The lower panel is the same data as in Figure 4.

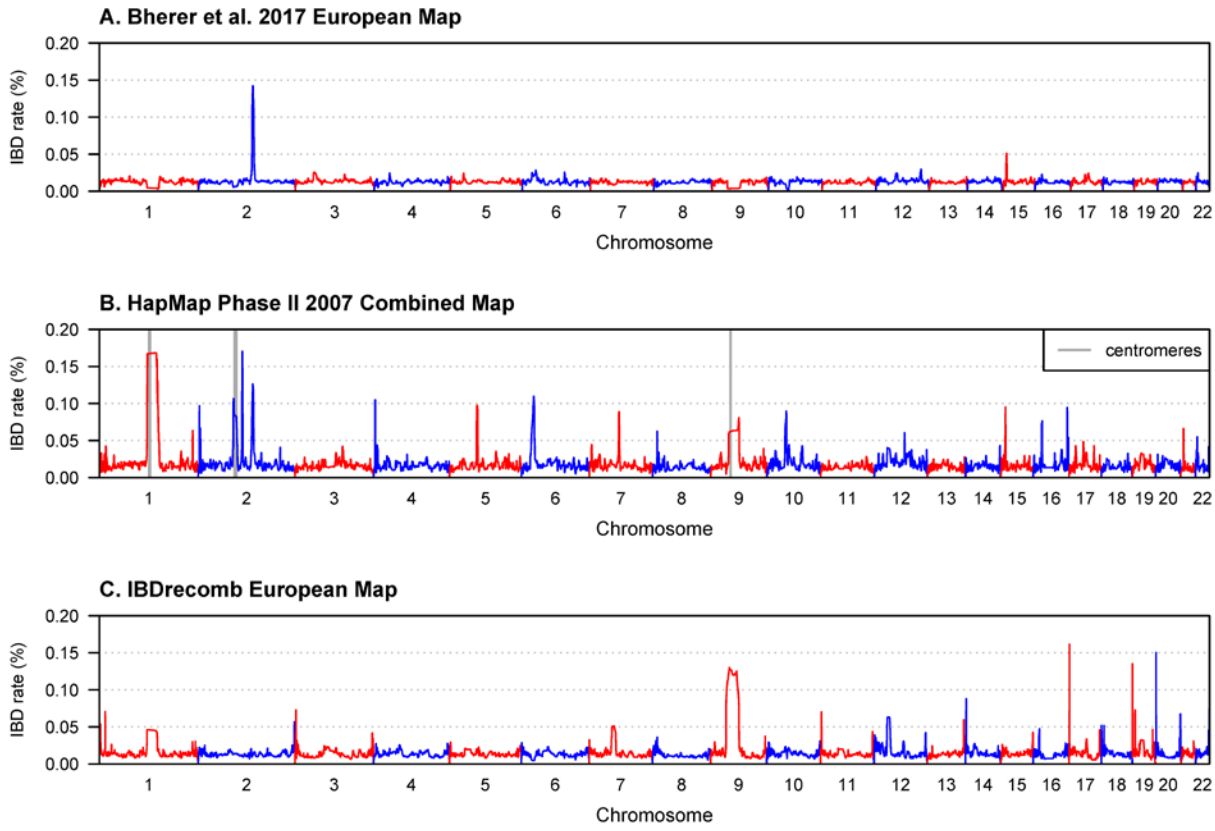

Figure S8: Effect of genetic map on inferred IBD rate. We estimated endpoint uncertainty for 50,000 White British individuals from the UK Biobank using three different maps. A) Bherer et al.'s refined European map.[49] B) The HapMap Phase II combined ancestry map,[76] with the centromeres for chromosomes 1, 2, and 9 notated. C) The IBDrecomb European map [13] calculated at a 1 kb scale. The y-axis is the IBD rate, which is the percentage of pairs of haplotypes for which the position on the chromosome is covered by a sampled IBD segment with length > 2 cM for the haplotype pair. The IBD rate is calculated at 10 kb intervals.

Table S1: Estimated error rates for UK-like simulated data

Genotype error rate	Sample size	Sequence or SNP data	True/inferred phase	Analysis error rate	Estimated error rate
0.02%	50,000	Sequence	True	0.0005	0.00037
0.02%	50,000	Sequence	Inferred	0.0005	0.00032
0.02%	50,000	SNP	Inferred	0.0005	0.00012
0.02%	1000	Sequence	True	0.0005	0.00054
0.02%	200	Sequence	True	0.0005	0.0013
0.1%	50,000	Sequence	True	0.0005	0.0011
0.1%	50,000	Sequence	Inferred	0.0005	0.00096
0.02%	1000	Sequence	True	0.01	0.00065
0.02%	1000	Sequence	True	0.001	0.00057
0.02%	1000	Sequence	True	0.0001	0.00047
0.02%	1000	Sequence	True	0.00001	0.00031
0.02%	1000	Sequence	True	0.000001	0.00012