A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data

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

Segments of identity by descent (IBD) are used in many genetic analyses. We present a method for detecting identical-by-descent haplotype segments in phased genotype data. Our method, called hap-IBD, combines a compressed representation of haplotype data, the positional Burrows-Wheeler transform, and multi-threaded execution to produce very fast analysis times. An attractive feature of hap-IBD is its simplicity: the input parameters clearly and precisely define the IBD segments that are reported, so that program correctness can be confirmed by users. We evaluate hap-IBD and four state-of-the-art IBD segment detection methods (GERMLINE, iLASH, RaPID, and TRUFFLE) using UK Biobank chromosome 20 data and simulated sequence data. We show that hap-IBD detects IBD segments faster and more accurately than competing methods, and that hap-IBD is the only method that can rapidly and accurately detect short 2–4 centiMorgan (cM) IBD segments in the full UK Biobank data. Analysis of 485,346 UK Biobank samples through the use of hap-IBD with 12 computational threads detects 231.5 billion autosomal IBD segments with length ≥ 2 cM in 24.4 h.

Introduction

Segments of identity by descent (IBD) are genomic regions over which a pair of individuals share a haplotype due to inheritance from a recent common ancestor. IBD segments are useful in a wide variety of applications because they capture information about genetic relationships between individuals. Correlation between pairwise IBD and phenotypic similarity can be used to detect genomic regions harboring trait-affecting variants[1–6] and to estimate heritability.[7–10] IBD segments are also used to estimate kinship coefficients,[11] detect close relationships,[12–14] and identify fine-scale population structure.[15–20]

Recent demographic history can be inferred from IBD segments.[15,16,21–24] Populations with smaller effective population size have more IBD sharing because individuals are more closely related on average. Short segments have a larger time to the most recent common ancestor (TMRCA), and thus are informative about less recent effective size, while long segments have a smaller TMRCA and are informative about very recent effective size. Similarly, IBD segments shared between populations are informative about migration rates. Approximately the past 100 generations of demographic history can be inferred from IBD segments.[23]

IBD segments are also useful for estimating population genetic parameters, including mutation rates[25–28] and recombination rates,[29] and for detecting regions undergoing recent selection.[10,20,30–32] The mutation rate is estimated from the observed discordance rate in IBD haplotypes. Recombination rate maps can be estimated using the rate of IBD segment endpoints. Selection is detected by looking for genomic regions with higher rates of IBD sharing.

There are several classes of methods for detecting IBD segments. The first class of methods are probabilistic. These methods include PLINK,[2] Beagle IBD,[33] and others.[4,10,34–39] For these methods, the unobserved IBD status for a pair (or set) of individuals at a locus takes two (IBD/non-IBD) or more possible states. Typically, a hidden Markov model is used to infer the IBD state at each marker. In the context of a pedigree, with shared haplotypes inherited only through pedigree founders, this IBD-state approach makes sense. However, in population samples, the concept of "IBD state" is ill defined. Two haplotypes are identical by descent if they are descended from a common ancestor, which is trivially true for all pairs of haplotypes at each position in the genome.

The second class of methods, which includes all the methods presented in this paper, look for long segments of identical-by-state allele sharing either in phased or in unphased genotype data. These identity-by-state (IBS) methods include GERMLINE[40] and others.[41–44] In contrast to most of the probabilistic methods, these methods do not dichotomize pairwise haplotypic sharing into "IBD" and "non-IBD," but instead dichotomize it into "long-IBD" and "not-long-IBD," which better fit the realities of population-based IBD sharing. Ideally, reported IBD segments should primarily represent IBD from a single common ancestor, rather than a conflation of segments from multiple ancestors, and this is achieved when the length threshold is relatively long.[45] A drawback to these methods is that the handling of allelic discordances within IBD segments tends to be *ad hoc*.

For IBS methods, the requirement that two individuals share a haplotype is more stringent than the requirement that the two individuals share at least one allele in their genotypes across a given region. Thus, haplotype-based

[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: browning@uw.edu

https://doi.org/10.1016/j.ajhg.2020.02.010.

methods can detect short IBD segments (e.g 2–10 centiMorgan [cM] in length) with much greater accuracy than genotype-based methods can. However, haplotype-based methods can break up a long-IBD segment into a sequence of shorter IBD segments if there are haplotype phasing errors in the long-IBD segment. For some downstream applications, after detecting IBD segments, it is necessary to perform a merging step in order to recover the original long-IBD haplotype. On the other hand, genotype-based methods do not require accurately phased genotype data, and they can detect long segments (≥ 15 cM) with high accuracy, which is sufficient for highly accurate detection of first- and second-degree relatives.[13]

A third class of methods are those that combine aspects of probabilistic modeling and length-based thresholding on IBS. Typically, these methods detect candidate long shared segments and then form a likelihood ratio for IBD versus non-IBD.[11,46–48] These methods tend to be more computationally efficient than the full probabilistic methods, but unlike some of the purely length-based IBS methods, they cannot analyze biobank-scale datasets.

Although "identity by descent" implies allelic identity, in fact, positions of discordance will be observed. Causes of this discordance include mutation or gene conversion since the common ancestor, and genotype error. Probabilistic methods allow for these discordances via an error term in the modeling, whereas length-based methods allow for short, infrequent gaps in allele sharing.

Genotype error rates vary greatly across datasets. Data from two recent studies give genotype error rate estimates of 0.008 per Mb per individual in a large SNP array study[49] and 25 per Mb per individual for single-nucleotide variants in a large sequencing study[50] (with error rates estimated as half the discordance between duplicate samples after quality control filtering, multiplied by the average number of called and/or assayed variants per Mb). Exclusion of rare variants can decrease the genotype error rate,[51] particularly for sequence data.[50]

With increasingly large datasets, computational issues become significant. The detection of sets of shared haplotypes can be reduced to linear computational complexity by means of hashing[40,44] or by use of the positional Burrows-Wheeler transform (PBWT).[43] However, the generation of pairwise IBD segments from these sets scales quadratically with sample size, because the number of pairs of individuals grows quadratically with sample size. Consequently, detecting IBD segments in biobank-scale data is challenging. As well as computation time being an issue, some algorithms require unfeasibly large amounts of computer memory to analyze such datasets.

In this work, we present hap-IBD, a method which scales to biobank-sized data, has greater accuracy than competing methods, and is notable for the simplicity of its algorithm and tuning parameters. Hap-IBD utilizes the PBWT[52] and parallel computation to reduce computing time, and it uses data compression to reduce memory requirements.[53] It addresses the issue of allele discordance in IBD segments by requiring that a reported segment have a central core (the "seed") that is free of discordance, while allowing extension beyond the seed after a short gap containing discordance. The key parameters for hap-IBD are the minimum seed length, the minimum extension length, the maximum gap length, and the minimum length of reported IBD segments. These parameters directly control which IBD segments are detected and reported. The hap-IBD program is open-source and freely available for academic and commercial use.

Material and Methods

The Hap-IBD Algorithm

The hap-IBD method employs a simple seed-and-extend algorithm. A seed is an IBS segment with genetic length (i.e., length in cM) greater than a specified minimum length (2 cM by default). The hap-IBD algorithm finds all seed segments, and it extends each seed if possible (Figure 1). A seed segment is extended if there is another long-IBS segment for the same pair of haplotypes that is separated from the seed segment by a short non-IBS gap. The maximum number of base pairs between the first and last markers in the non-IBS gap and the minimum cM length of the extension IBS segment can be specified by the user, and these are 1,000 base pairs and 1 cM, respectively, by default. A segment may be extended multiple times. When it is no longer possible to extend the segment, the segment is written to the output file if its centiMorgan length is greater than a specified minimum output length (2 cM by default).

Allowing short non-IBS gaps provides robustness to three sources of discordant alleles in IBD segments: genotype error, gene conversion, and mutation since the most recent common ancestor. Genotype error and mutation will typically introduce a single discordant allele in an IBD segment. Gene conversion will generally produce a very short interval containing one or a few discordant alleles in an IBD segment. When the phasing of the surrounding alleles is correct, the mismatching alleles on the pair of IBD haplotypes result in two IBS segments for the same pair of haplotypes, separated by a single marker, or at most a few markers in the case of gene conversion. Our method allows these breaks in IBS sharing to be detected and the IBS segments on each side of the break to be included in the same reported IBD segment. The maximum length is specified in base pairs rather than centiMorgans, because breaks due to mutation or genotype error typically involve a single marker, and breaks due to gene conversion are typically less than 1,000 base pairs.[54]

Two or more distinct IBS seed segments can result in the same IBD haplotype after each seed is extended. If an IBS segment that extends the seed segment to the left is itself a valid seed segment, we stop the extension process and discard the seed segment that is being extended, because the same IBD haplotype will be generated by a seed segment that occurs earlier on the chromosome.

The hap-IBD algorithm also has an optional min-markers parameter that requires seed IBS segments to have a minimum number of markers. The min-markers parameter can be useful for ensuring a minimum level of evidence for IBD in genomic regions having low marker density. When a min-markers parameter is specified, IBS segments that extend a seed are also required to have a minimum number of markers. We set the minimum

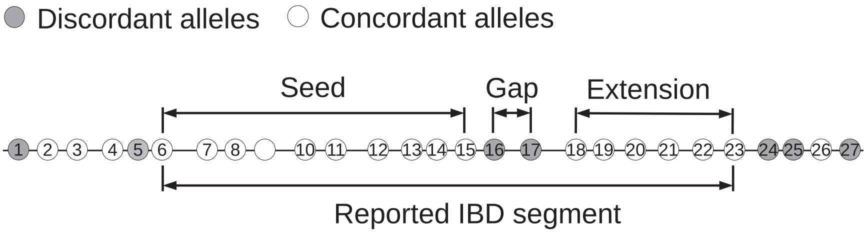

Figure 1. Overview of Hap-IBD Algorithm

The hap-IBD algorithm first identifies a seed identity-by-state (IBS) segment whose genetic length exceeds the minimum seed length. Here the seed segment is markers 6–15. The seed segment is extended to the right because it is followed on the right by a short non-IBS region (markers 16–17) with base-pair length that is less than the maximum gap length and another long IBS segment (markers 18–23) with genetic length that exceeds the minimum extension length. The seed segment is not extended to the left because the IBS segment to the left (markers 2–4) is shorter than the minimum extension length.

number of markers in an extension to be the product of the min-markers parameter and the ratio of the minimum extension length to the minimum seed length.

We first describe a single-threaded implementation of the preceding algorithm and then describe how the single-threaded implementation is modified to permit parallel computation.

Computationally Efficient Detection of Seed Segments

After the genotype data for a chromosome are read into memory, we apply the PBWT.[52] The PBWT sweeps through the markers in chromosome order, and at each marker, it sorts the reverse haplotype prefixes in lexicographic order (the reverse haplotype prefix at the m-th marker is the sequence of alleles at markers $m- \ 1$, $m- \ 2$, …). At marker m, we generate a “divergence” array that stores the first marker of the IBS segment containing marker $m-1$ for each pair of haplotypes that are adjacent after sorting.[52] The divergence array is used to efficiently identify all seed IBS segments that end at marker m (see Durbin’s Algorithm 3).[52] After a seed is identified, it is extended (if possible) by comparing the alleles on the two IBS haplotypes in the regions preceding and succeeding the seed segment as described above.

Memory-Efficient Computation

The hap-IBD program takes phased genotype data in variant call format (VCF) as input.[55] As the genotype data are read into memory, the data are immediately converted to binary reference format (version 3).[53] Binary reference format compresses low-frequency variants by storing only the indices of the haplotypes carrying non-major alleles. Higher-frequency variants are compressed by storing unique allele sequences in a region, along with a vector that maps haplotype indices to the allele sequence carried by the haplotype. We use binary reference format because it permits data for an entire chromosome to be stored compactly in memory, and it allows rapid queries of alleles carried by haplotypes at each marker.

The PBWT requires only two additional arrays of stored information, each with length equal to the number of haplotypes. Seed IBS segments are extended as soon as they are identified by the PBWT. After extension, segments that are longer than the minimum output segment length are immediately printed to an output buffer, which is flushed to disk when full. Consequently, only a limited number of IBD segments are stored in memory at any time.

Parallelization

The hap-IBD algorithm is parallelized by applying the PBWT concurrently in overlapping marker windows. If L is the genetic distance between the first and last markers on the chromosome, S is the minimum seed genetic length, and T is the number of computational threads, we sequentially define T overlapping marker windows $W_1, W_2, \ldots, W_T$ that each have length approximately equal to $((L-S)/T+S)$ cM, and that have approximately S cM overlap between adjacent windows. The first window W_1 begins at the first marker on the chromosome and ends at the first marker after genetic position $((L-S)/T+S)$ whose index is greater than the minimum number of markers required for an IBS seed segment. The first marker in W_{i+1} is the first marker in W_i that cannot be the start of a seed IBS segment contained within W_i because the number of markers or genetic distance separating the marker from the last marker in W_i is too small. The last marker in W_{i+1} is the first marker that is $\geq (L-S)/T$ cM away from the last marker in window W_i. With these definitions, every seed IBS segment will be detected in at least one of the overlapping windows.

We run the PBWT algorithm in each overlapping window in parallel. When a seed IBS segment is found, we ignore the window boundaries when we extend the segment, so that the extension process is the same as for the single-threaded case. If multiple seeds result in the same maximal IBD segment after extension, we keep the maximal IBD segment generated by the first seed, and we discard the duplicate IBD segments generated by later windows seeds.

Input and Output Data

The input data is a VCF file[55] with phased genotypes and no missing genotypes, and a PLINK-format genetic map.[2] Linear interpolation is used to estimate the genetic map positions for any marker whose position is not on the genetic map. Although the use of a genetic map is recommended, hap-IBD can be used with Mb units simply by supplying a genetic map with a recombination rate of 1 cM = 1 Mb.

Two output files are produced: one containing within-individual segments of homozygosity by descent (HBD) and one containing between-individual IBD segments. Each output line contains the pair of samples, the specific haplotypes, the starting base position, the ending base position, and the genetic length of the HBD or IBD segment.

Hap-IBD Parameters

The minimum seed length parameter has a large influence on computation time. Increasing the minimum seed length reduces computation time because fewer seed IBS segments will be considered. Decreasing the minimum seed length can increase power to detect short IBD segments that have discordant alleles on the pair of shared haplotypes.

The maximum gap length and minimum extension length allow reported IBD segments to contain discordant alleles due to genotype error, mutation, or gene conversion. The hap-IBD software also has an option for excluding input markers that have low minor allele counts.

The minimum markers parameter controls the minimum number of markers in IBS seed and extension segments. The number of reported IBD segments should be approximately constant throughout the genome; however, regions with low marker density can produce local spikes in the number of reported IBD segments (see Results). These spikes contain many IBS segments that satisfy the genetic length requirements, but contain relatively few markers. The spikes can be reduced or eliminated by post-processing[23,56] or by requiring seed and extension IBS segments to contain a minimum number of markers.

UK Biobank Genotype Data

We downloaded the UK Biobank genotype data from the European Genome-phenome Archive[57] (dataset accession: EGAD00010001497). The UK Biobank data contain 488,377 individuals and 784,256 autosomal markers.[49] We excluded markers with more than 5% missing genotypes ($n = 70{,}246$), markers that had only one individual carrying a minor allele ($n = 5{,}123$), and markers that failed one or more of the UK Biobank's batch quality control tests ($n = 1{,}527$).[49] After we excluded 72,601 markers that failed one or more of these filters, there were 711,655 autosomal markers.

We then excluded 968 individuals that were identified by the UK Biobank as being outliers for their proportion of missing genotypes or proportion of heterozygous genotypes, and we excluded nine individuals that were identified by the UK Biobank as showing third-degree or closer relationships with more than 200 individuals (indicating sample contamination).[49] After these exclusions, there were 487,400 individuals.

We identified parent-offspring trios using the kinship coefficients and the proportion of markers that share no alleles (IBS0) that are reported by the UK Biobank.[49,58] First-degree relatives were considered to be pairs of individuals with a kinship coefficient between $2^{-2.5}$ and $2^{-1.5}$. Among first-degree relatives, parent-offspring relationships were assumed to be the first-degree relative pairs with IBS0 < 0.0012. These are the same kinship coefficient and IBS0 thresholds used by the UK Biobank to identify parent-offspring relationships.[49] We considered an individual to be the offspring in a parent-offspring trio if the individual had a parent-offspring relationship with exactly one male and one female individual, and if the male and female first-degree relatives were not in the set of related pairs of individuals reported by the UK Biobank, which is the set of pairs of individuals with estimated kinship coefficient greater than $2^{-4.5}$. In this case, we considered the male and female first-degree relatives to be the offspring's parents. Using this procedure, we identified 1,064 parent-offspring trios.

The 1,064 trio offspring have 2,054 distinct parents. We excluded these parents from the data before phasing and IBD segment detection so that phasing accuracy in the trio offspring would more closely match phasing accuracy in unrelated individuals. After we excluded the trio parents, there were 485,346 remaining individuals. We listed the 1,064 trio offspring followed by the remaining samples in random order. By taking the corresponding number of samples from the top of this list, we created five telescoping genotype datasets that included 5,000, 15,000, 50,000, 150,000, and all 485,346 individuals. We then used Beagle 5.1 to phase each dataset.[59]

Because we use trio genotypes to determine the genotype phase in the offspring, we selected the 850 trios that had the lowest genotype error rate, as measured by the number of autosomal sites with Mendelian inconsistent genotypes.[60] The number of inconsistent sites in the 1,064 trios ranged from 57 to 5,102 sites per trio, and the number of inconsistent sites in the 850 trios with the lowest genotype error rate ranged from 57 to 456 sites per trio. We phased the 850 trio offspring at all heterozygous genotypes for which phase could be determined from parental genotypes and Mendelian inheritance constraints (82.4% of heterozygous genotypes), and we masked genotypes at Mendelian inconsistent sites in this phased data. We used these estimated haplotypes to evaluate false-positive and false-negative rates for IBD segment detection as described below.

After we excluded trio parents, there were 43 remaining parent-offspring pairs who were not part of a trio in the 50,000 individual subset of the UK Biobank data. We used these 43 remaining pairs to compute the mean proportion of chromosome 20 covered by detected IBD segments in parent-offspring pairs.

Simulated Data

In order to test the performance of hap-IBD and other methods on sequence data, we generated 60 Mb of data for 50,000 individuals from a demographic model that simulates the present UK European population.[47] This 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.

We used forward simulation with SLiM v3.3[61,62] to simulate the ancestral recombination graph for the most recent 5,000 generations. Gene conversion tracts were initiated at a rate of 2×10^{-8} per base pairs (bp) per generation, and had geometrically distributed lengths with mean 300 bp, giving an overall gene conversion rate of 6×10^{-6}.[27,63] A constant recombination rate of 1×10^{-8} was used. We then used coalescent simulation in msprime (v0.7.1) to add mutations (at rate 1.38×10^{-8}) and simulate the more distant past.[64] This hybrid strategy of using SLiM and msprime enables utilization of msprime's computational efficiency for large datasets, while incorporating biologically realistic settings, such as gene conversion, that are implemented in SLiM but not currently implemented in msprime.[65] Our simulation only includes gene conversion events in the most recent 5,000 generations, but it is the more recent gene conversions that have the greatest potential impact on haplotype phase accuracy and that can create discordances between identical-by-descent haplotypes.

We determined the true IBD segments for 1,000 simulated individuals from the simulated ancestral recombination graphs. IBD segments are required to have the same ancestral node along their length, except for short breaks due to gene conversion.

We added genotype error at a rate of 0.02%, which is the error rate that produces the observed 0.04% rate of discordance at SNVs that pass quality control in the TOPMed Freeze 5 whole-genome-sequence data.[50] We then removed variants with frequencies less than 10% and used Beagle 5.1 to phase the remaining genotypes.[59] We also separately phased a subset of 5,000 individuals with the same minor allele frequency threshold of 10%. Low-frequency variants are not very informative for IBD because most individuals are homozygous for the major allele, and because allele discordance at low-frequency variants in IBD segments could be due to genotype error, recent mutation, or phasing error, rather than indicating that they are non-IBD variants. Other methods for IBD detection in sequence data have used a minor allele frequency filter. The application of GERMLINE to the Genomes of the Netherlands whole-genome sequence data used a minor allele frequency filter of 1%.[56] The TRUFFLE analysis of 1000 Genomes sequence data used minor allele frequency filters of 5% and 10%.[42]

Parameter Settings

Each method has an option for setting the minimum length of reported IBD segments. All methods, except TRUFFLE, measure distance in cM units. For TRUFFLE, we substituted Mb units for cM units. All genetic distances are interpolated from the HapMap genetic map.[66]

We required all UK Biobank chromosome 20 analyses to complete within two days of wall-clock time on the compute nodes used for these analyses. Parameter settings for analysis of UK Biobank and simulated sequence data are based on previously published analyses of SNP array[5,42–44] and sequence data.[42,43,56] Parameter settings for each method are reported in Tables S1 and S2.

The hap-IBD parameter settings in Tables S1 and S2 are recommended parameter settings for analysis of SNP array data (Table S1) and sequence data (Table S2) that have marker densities, rates of genotype error per Mb, and effective population sizes similar to those of the data considered in this paper. We used a lower minimum seed length and extension length for analysis of the sequence data because IBS segments are shorter in the sequence data due to the much higher number of genotype errors per Mb. We use the same 2.0 cM minimum output IBD length for both SNP array data and sequence data because both datasets are derived from outbred, human populations, and previous work has shown that allele identity at this scale is an accurate proxy for identifying IBD segments in outbred populations.[45] If one is analyzing data from an inbred population, then use of a longer minimum output IBD length could be appropriate due to the increased probability in an inbred population that a 2.0 cM IBS segment is actually a conflation of multiple shorter IBD segments.

Comparison of Methods

For the simulated data, coalescent trees for 1,000 simulated samples were used to determine true IBD segments exceeding 1.5 cM in length for those samples. For the UK Biobank data, we considered true IBD segments to be IBS segments exceeding 1.5 cM in length among the 850 trio offspring which were phased using parental genotypes and Mendelian inheritance rules.

False-Positive Rate Estimation

We divided detected IBD segments into bins according to the detected segment length (2–3, 3–4, 4–6, 6–10, 10–18, and >18 cM). For each detected IBD segment, we identified the cM length of the portion of the detected segment that is not covered by any true IBD segment with length >1.5 cM, and we calculated the sum of these false-positive segment lengths. The false-positive rate for a bin is the sum of the false-positive segment lengths divided by the sum of the detected segments' lengths.

False-Negative Rate Estimation

We divided true IBD segments into bins according to the true segment length. The length bins and number of true UK Biobank IBD segments in each length bin are: 2.5–3 cM (2,492), 3–4 cM (1,360), 4–6 cM (551), 6–10 cM (160), 10–18 cM (55), and >18 cM (64). For each true IBD segment, we identified the cM length of the portion of the true segment that is not covered by any detected IBD segment with length ≥2.0 cM, and we calculated the sum of these false-negative segment lengths. The false-negative rate for the bin is the sum of the false-negative segment lengths divided by the sum of the true segments' lengths.

ROC Analysis

In order to account for inter-method differences in determining IBD end-points, differences which affect the reported lengths of IBD segments, we calculated false-positive and false-negative rates for each method over a range of detected segment length thresholds (1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, and 2.4 cM). We calculated the false-positive rate for each threshold as described above using all true segments having length >1.5 cM, and we calculated the false-negative rate as described above using all true segments having length >2.5 cM. For each method, we then generated a receiver operating characteristic (ROC) curve that shows the true-positive rate (which is one minus the false-negative rate) and false-positive rate for each detected segment length threshold. For length thresholds <2.0 cM, the hap-IBD minimum seed length was set to 1.6 cM.

We also generated ROC curves for each method for 5 cM segments. For this analysis, we used detected segment length thresholds of 4.6, 4.8, 5.0, 5.2, and 5.4 cM. We calculated false-positive rates using all true segments having length >1.5 cM, and we calculated false-negative rates using all true segments having length >5.5 cM.

Computation Time

All analyses were run on a 12-core 2.6 GHz computer with Intel Xeon ES-2630 processors and 128 GB of memory. Computation time was measured using the Unix time command, which returns a "real," a "system," and a "user" time. The wall-clock time is the "real" time, which is the length of time the program was running. The central processing unit (CPU) time is the sum of the "system" and "user" times. For multi-threaded compute jobs, the CPU time includes the sum of the CPU time for each computational thread, so that it represents the total CPU resources consumed by the program. A maximum of 2 days of wall-clock time was allowed for each analysis of UK Biobank chromosome 20 data or of the 60 Mb of simulated sequence data, with no results reported if the analysis did not complete within this time frame.

Results

Computational Feasibility

Figure 2 shows CPU times for subsets of the UK Biobank chromosome 20 data (5,000 to 485,346 individuals) for 2 cM and 5 cM output length thresholds. For the full UK Biobank data with 485,346 individuals, hap-IBD detected 3.43 billion IBD segments on chromosome 20 at the 2 cM threshold, and 106 million segments at the 5 cM threshold. GERMLINE could not analyze subsets of 50,000 or more individuals because it required more than 128 GB of memory. TRUFFLE could not analyze subsets of 150,000 or more individuals on our compute nodes. iLASH could not analyze subsets of 150,000 or more individuals at the 2 cM output threshold because it needed more than 128 GB of memory, but it could analyze the full dataset at the 5 cM output threshold. RaPID could not analyze the full chromosome 20 dataset at the 2 cM

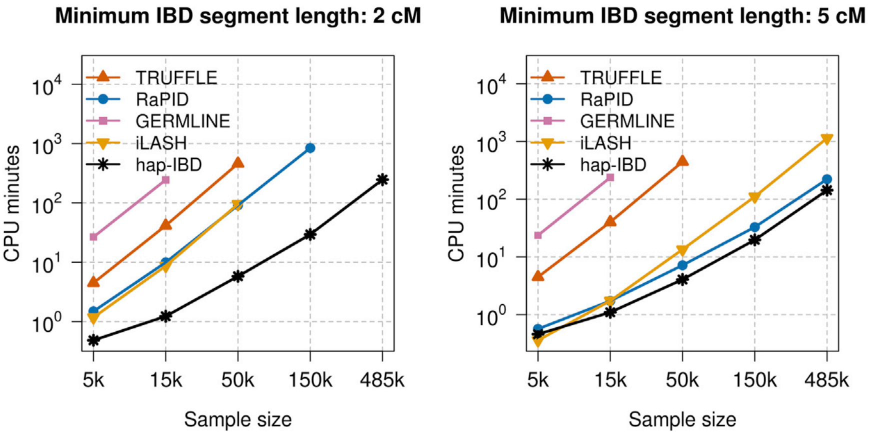

Figure 2. CPU Time
Central processing unit (CPU) time for detecting IBD segments with length ≥2 centiMorgan (cM) (left panel) and ≥5 cM (right panel) on chromosome 20 in samples of 5,000, 15,000, 50,000, 150,000, and 485,346 individuals from the UK Biobank. CPU time is the sum of the computation time for each CPU core. All programs used 12 computational threads, except RaPID and GERMLINE, which are limited to one computational thread each. The CPU times for using Beagle 5.1 to phase the 5,000, 15,000, 50,000, 150,000, and 485,346 individuals were 168, 525, 2,008, 6,595, and 20,567 min.

threshold within the permitted two days of wall-clock time, but it could analyze the full dataset at the 5 cM output threshold.

Three of the five methods can use multiple computational threads, and running these methods on a 12-core computer leads to an approximate 10-fold reduction in computing time compared to single-threaded analysis (Figure S1). This degree of speedup is important for analysis of large datasets. For example, the single-threaded RaPID program required 223.6 min of wall-clock time to output IBD segments ≥5 cM for all samples on chromosome 20, but hap-IBD required only 13.4 min when using 12 computational threads.

Overall, we see that hap-IBD is the fastest program except when analyzing the smallest sample size (5,000 individuals) using the largest output threshold (5 cM output threshold); for this combination, iLASH is faster. In our experiments, hap-IBD was the only method that could analyze the full UK Biobank chromosome 20 data on our compute servers in less than 2 days when using a 2 cM output length threshold.

We also performed a genome-wide analysis of the 22 autosomes for the UK Biobank data. Genome-wide analysis of 485,346 UK Biobank samples using hap-IBD with 12 computational threads detected 231.5 billion autosomal IBD segments in 24.4 h.

Accuracy

Some methods have a low false-positive rate (Figure 3 and Figure S2) but a high false-negative rate (Figure 4 and Figure S3), or vice versa. The methods apply different algorithms for determining the end points of IBD segments, and this results in different methods reporting different lengths for a true IBD segment. Because false-positive rates and false-negative rates can be traded off by changing the output length threshold, we constructed ROC curves by varying the output IBD segment length threshold for each method in order to assess the true-positive versus false-positive trade-off. The true-positive rate is one minus the false-negative rate. An ideal method would have a true-positive rate of 1 and a false-positive rate of 0. For 2 cM IBD segments (Figure 5) and for 5 cM IBD segments (Figure S4) hap-IBD shows the best performance on these ROC curves. In particular, hap-IBD has much lower false-positive rates than RaPID and much higher true-positive rates than iLASH. The IBD segment detection method for unphased genotype data (TRUFFLE) has high error rates for these short IBD segments.

We also investigated the proportion of chromosome 20 in parent-offspring pairs that was covered by detected IBD segments with length ≥2 cM in 43 parent-offspring pairs in the set of 50,000 UK Biobank samples. The proportions were 0.978 for iLASH, 0.987 for hap-IBD, 0.994 for RaPID, and 1.0 for TRUFFLE. GERMLINE was not evaluated because it could not analyze 50,000 individuals on our compute server. All methods detected IBD across all or nearly all of the chromosome in the parent-offspring pairs. For haplotype-based methods, the methods with higher false-positive rates (Figure 3) detected slightly higher numbers of IBD segments in the parent-offspring pairs. Genotype-based methods are not affected by haplotype phase errors, and the genotype-based method (TRUFFLE) had the highest detection rate for these chromosome-length shared haplotypes.

We examined the effect of varying the hap-IBD min-seed and min-extend parameters when detecting 5 cM segments in the UK Biobank chromosome 20 data for 5,000 individuals. Accuracy was unchanged when using min-seed values between 1 cM and 3 cM (Figure S5). The false-positive rate doubled when the min-extend parameter decreased from 1.0 cM to 0.1 cM (Figure S6), although the absolute increase in the false-positive rate is small (approximately 0.5%).

In genome-wide analysis of the UK Biobank data, we found regions in which IBD detection methods reported inflated levels of IBD segments. These are generally regions with large gaps in marker coverage, or very low marker density, and often occur around centromeres. Figure 6 shows results for chromosomes 1 and 20 for the methods with the highest accuracy for short IBD segments (the four haplotype-based methods) for the 5,000 individuals' UK Biobank data. Around the chromosome 1 centromere,

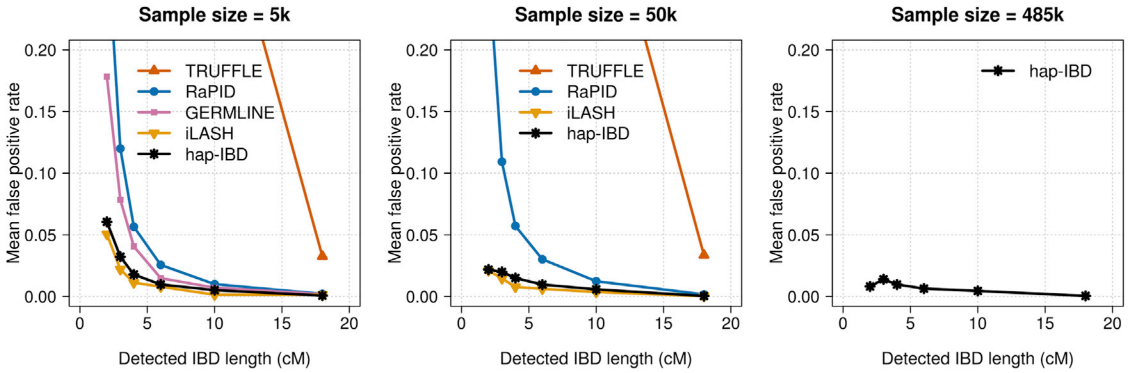

Figure 3. False-Positive IBD Segment Detection in UK Biobank Chromosome 20 Data
False-positive rates for identity by descent (IBD) segment detection for 5,000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥2 centiMorgan (cM) were detected with each method. Detected IBD segments were assigned into bins of 2–3, 3–4, 4–6, 6–10, 10–18, and >18 cM according to their segment length. The false-positive rate is the proportion of detected IBD segments in a bin that are not covered by any true IBD segment >1.5 cM in length. Hap-IBD is the only method shown for the full UK Biobank analysis (485,346 individuals) because other methods were unable to complete the analysis with a 2 cM output threshold within the memory and time constraints (see the Computational Feasibility subsection in Results). The x-coordinate of each data point is the left bin end point (e.g., 2 cM for the 2–3 cM bin). For the full range of y-coordinate values, see Figure S2.

the methods found IBD segments at a rate 40 to 3,000 times greater than the baseline level. The inflation was worse for RaPID and iLASH than for GERMLINE and hap-IBD. Figure S7 shows that the inflated detection can be reduced by increasing hap-IBD's min-markers parameter. However, the use of overly high values of this parameter will reduce power to detect short IBD segments. Alternatively, regions with high rates of IBD segment discovery can be identified after IBD segment detection and excluded.[67]

We also assessed accuracy by using simulated sequence data. There are several important differences between the UK Biobank analysis and the simulated sequence data analysis. First, the approach to assessing accuracy differs. In the UK Biobank data, we determined the true phase of trio offspring, and we used that to determine identity by state at the haplotype level, which we used as a proxy for true IBD. The genotype error rate was extremely low in these data (with a duplicate discordance rate of 6.7×10^{-5}),[49] but genotype errors can disrupt both the true IBD and the estimated IBD in the UK Biobank analysis. In contrast, in the simulated data, the true IBD status was obtained directly from the simulation (defined as no change in common ancestor across a segment except in tracts of gene conversion), and mis-called alleles may have disrupted the estimated IBD but did not affect the ascertainment of true IBD. Second, the marker density was much higher for the simulated sequence data. Although we removed markers with minor allele frequency <10% (see Material and Methods), the marker density was still five times greater than that of the UK Biobank data (97,890 markers with minor allele frequency ≥10% in the simulated 60 Mb region, compared with 18,424 total UK Biobank markers on chromosome 20). Third, the genotype error rate in the simulated sequence data was much higher than for the UK Biobank data. With current technology, error rates tend to be higher for sequence data than for SNP array data, even with high sequence coverage and careful processing. We added genotype error to the simulated sequence data at a rate that generates the level of duplicate discordance observed in the TOPMed data, which is 4×10^{-4} for SNVs passing quality control.[50] This level of duplicate discordance is six times higher than for the UK Biobank SNP data. There are also important similarities between the two analyses, which include the length of the region (approximately 60 Mb for the simulated analysis and for the UK Biobank chromosome 20 analysis), large sample size (up to 50,000 for the simulated data and up to 485,346 for the UK Biobank data), and demographic history (UK-like simulation versus actual UK population).

In the simulated sequence data, we used settings which the authors of the GERMLINE, RaPID, and TRUFFLE methods have used in published analyses of sequence data, while for iLASH, we used the same settings as for SNP array data (see Material and Methods for details). To compare accuracy, we produced ROC curves for detection of 2 cM segments. We considered sample sizes of 5,000 (Figure 7) and 50,000 individuals (Figure S8). We found that hap-IBD and GERMLINE have a very similar accuracy profile (for the 5,000 individuals only, because GERMLINE could not analyze the 50,000 individuals with the available computer memory). iLASH had reduced power to detect 2 cM IBD segments, while TRUFFLE had very low power to detect these segments, and RaPID had a high false-positive rate. Overall, these results are similar to those seen in the UK Biobank analysis, except that the relative accuracy of GERMLINE is improved in these simulated sequence data. The parameters that we used for GERMLINE in the simulated sequence analysis may be a better match for these data than were the parameters that we used for the UK Biobank data, although we used published parameter settings in both instances.

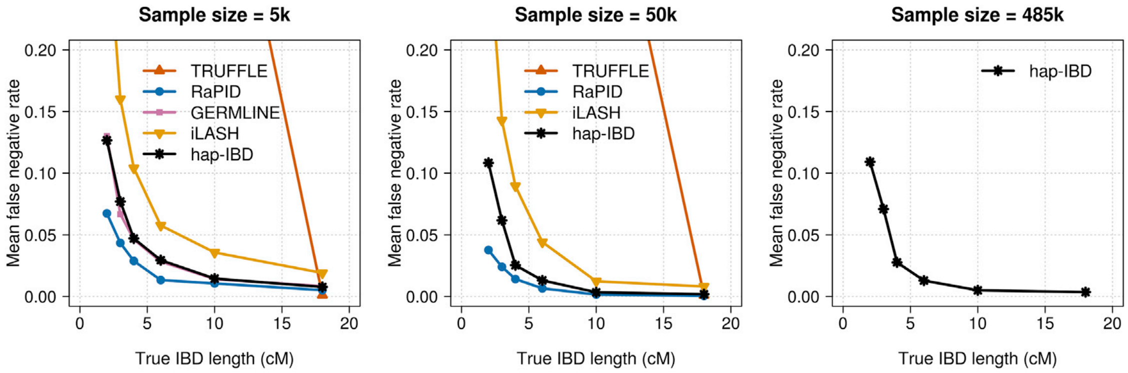

Figure 4. False-Negative IBD Segment Detection in UK Biobank Chromosome 20 Data
False-negative rates for identity by descent (IBD) segment detection for 5,000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥2 centiMorgan (cM) were detected with each method. True IBD segments with length >2.5 cM were assigned into bins of 2.5–3, 3–4, 4–6, 6–10, 10–18, and >18 cM according to their segment length. The false-negative rate is the proportion of true IBD segments in a bin that are not covered by any detected IBD segment ≥2 cM in length. Hap-IBD is the only method shown for the full UK Biobank analysis (485,346 individuals) because other methods were unable to complete the analysis with a 2 cM output threshold within the memory and time constraints (see the Computational Feasibility subsection in Results). The x-coordinate of each data point is the left bin end point (e.g., 2.5 cM for the 2.5–3 cM bin). For the full range of y-coordinate values, see Figure S3.

Discussion

We have presented an IBD segment detection method for large-scale genotype data that is substantially faster and more accurate than four state-of-the-art competing methods (GERMLINE, iLASH, RaPID, and TRUFFLE). We applied hap-IBD to 485,346 samples from the UK Biobank[49] and detected 231.5 billion autosomal IBD segments having length ≥2 cM in less than 24.4 h of wall-clock time on a compute server with 12 CPU cores.

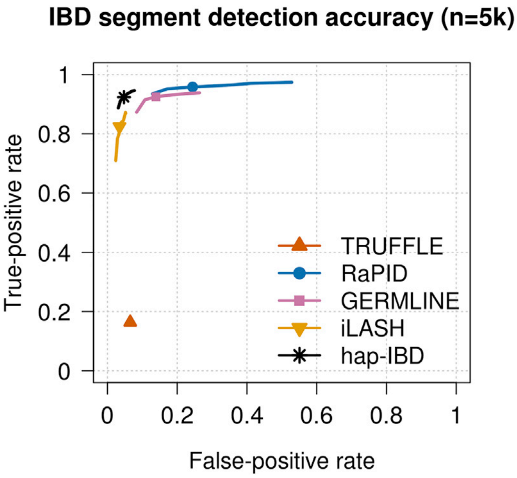

Figure 5. Receiver Operating Characteristic Curves for 2 cM IBD Segment Detection in UK Biobank Chromosome 20 Data
False-positive and false-negative rates for detection of identity by descent (IBD) segments over a range of output length thresholds around 2 centiMorgan (cM) for 5,000 UK Biobank samples. In order to allow for some discrepancy between reported and true lengths, false positives are assessed using true segments having length >1.5 cM, and false negatives are assessed using true segments having length >2.5 cM. IBD segments were detected with each method using length thresholds of 2 cM (plotted symbol) and with other thresholds between 1.6 and 2.4 cM (plotted lines; see Material and Methods). Figure S4 shows a similar plot for 5 cM.

An attractive feature of hap-IBD is its simplicity. All seed IBS segments that exceed a specified length are identified and then extended if possible. The extension process allows for sporadic non-IBS alleles due to mutation, genotype error, or gene conversion. The hap-IBD parameters define the minimum length of IBS seed and extension segments and the maximum length of non-IBS gaps. These parameters have a simple and direct relationship to the IBD segments that are reported, and this enables the correctness of the output results to be confirmed. In contrast, some methods utilize a large number of tuning parameters which have only an indirect relationship to output IBD segments, such as iLASH's seven parameters for controlling locality-sensitive hashing: perm_count, shingle_size, shingle_overlap, bucket_count, match_threshold, interest_threshold, and minhash_threshold.[44]

The hap-IBD method shares some similarities with the GERMLINE method: both methods search for long-IBS segments via a seed and extend algorithm, and both methods allow for the presence of some discordant alleles in a reported IBD segment.[40] However, hap-IBD achieves much greater computational efficiency and greater accuracy than GERMLINE does, because hap-IBD employs the PBWT instead of a hash table, and hap-IBD identifies seeds that exceed a specific genetic length rather than a specified number of markers.

In our tests, hap-IBD consistently required less CPU time than competing methods did. One factor contributing to hap-IBD's fast computation time is that it only considers shared haplotype segments that exceed a relatively long minimum seed length, and these segments are efficiently detected using the PBWT. The hap-IBD method also includes internal parallelization that can yield wall-clock compute times that are a fraction of the total CPU time on multi-core processors.

The hap-IBD method requires phased genotype data. In practice, nearly all large genotype datasets are phased

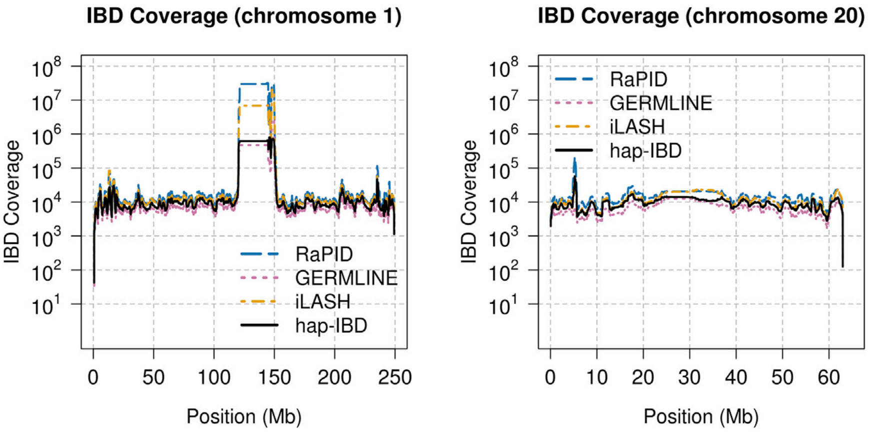

Figure 6. Chromosome-wide IBD Segment Detection in the UK Biobank
The methods were run on 5,000 UK Biobank samples on chromosome 1 and chromosome 20 in order to detect identity by descent (IBD) segments having length ≥2 centiMorgan. Each chromosome is divided into non-overlapping 10 kb intervals. For each interval, all IBD segments (from all pairs of individuals) that intersect the interval are each weighted by the proportion of the 10 kb interval that is covered by the IBD segment, and the sum of the weights is plotted as the IBD coverage.

because phased data are required to obtain the highest accuracy for many downstream analyses, including IBD segment detection, relationship inference,[13] local ancestry inference,[68] population demography inference,[21–23] and detection of selection.[10] Phased data are also required for computationally efficient and accurate genotype imputation.[53,69] With state-of-the-art methods, the effort and computational cost required to phase large datasets is modest when using a small compute cluster. We phased the UK Biobank genome-wide data with Beagle 5.1 in less than two days using 16 compute servers, each with 20 CPU cores.

Our results confirm that IBD segment detection methods for phased genetic data can detect much shorter IBD segments than can methods for unphased genetic data. In our tests, the method for unphased data (TRUFFLE) could not accurately detect segments with length <10 cM, but most methods for phased data could accurately detect IBD segments with length ≥2 cM (Figure 5). Furthermore, haplotype-based methods identify the shared allele sequence, whereas genotype methods cannot identify the shared allele when the individuals both carry an identical heterozygote genotype. However, genotype-based IBD detection methods, such as TRUFFLE, have some advantages. Genotype-based methods can detect first- and second-degree relationships with high accuracy before haplotypes are estimated,[13,42] and this can be useful during initial data quality control. In addition, genotype-based methods can be used when genotype data cannot be accurately phased due to non-uniform marker coverage or high rates of genotype error, as can be the case for exome and low-coverage sequence data. If the sample size is less than 50,000, if one is interested only in accurately detecting long (≥15 cM) IBD segments, and if haplotypes have not been estimated, then TRUFFLE can detect long IBD segments in less CPU time than is required for phasing and haplotype-based IBD segment detection. The CPU time for the TRUFFLE analysis of the chromosome 20 UK Biobank data for 50,000 individuals was 4.4 times faster than the combined CPU time for Beagle 5.1 phasing and hap-IBD analysis (461 min versus 2,014 min). However, pairwise IBD detection scales quadratically with sample size, while phasing with Beagle 5.1 scales approximately linearly. Thus, an extrapolated TRUFFLE CPU time for all 485,346 UK Biobank individuals is more than twice as long as the total CPU time for Beagle 5.1 phasing and hap-IBD analysis (43,447 min versus 20,815 min).

The hap-IBD method performs well across a range of haplotype switch error rates. In the UK Biobank data, the switch error rate for 5,000 samples is more than an order of magnitude higher than the switch error rate for 485,346 samples.[60] However, even for the 5,000-sample subset of the UK Biobank data, the IBD-detection accuracy is very high and is sufficient to identify close relatives in the data. Furthermore, one can increase the accuracy of phase estimates in small samples by phasing the samples together with a reference panel of sequenced individuals.[70]

Segments of IBD are genomic regions over which the pair of haplotypes share a recent ancestor. The expected length of the segment is related to the number of generations to the common ancestor, so the genetic length of the segment acts as a proxy for how recent the ancestry is. When using a genetic length threshold to determine which segments to report, as is done in this paper, the choice of threshold determines the degree of recentness. If the length threshold is increased, only more recent ancestry will be included, while if the threshold is decreased, slightly less recent ancestry will be included. In this paper, we used a 2 cM threshold. Segments of IBD resulting from shared ancestry 25 generations ago have an expected length of 2 cM, while the typical 2 cM segment has a common ancestor around 50–100 generations ago,[16] because there are many more segments with less recent ancestry, and some of these are by chance relatively long for their age and pass the length threshold. Thus “recent” in the context of this paper essentially means “within the past 100 generations.” One problem with the length-based approach is that informativeness of the genotype data varies considerably across the genome, primarily due to gaps in marker coverage around repetitive regions, leading to increased false positive detection in some regions. We found that all of the methods investigated in this paper are affected by this issue.

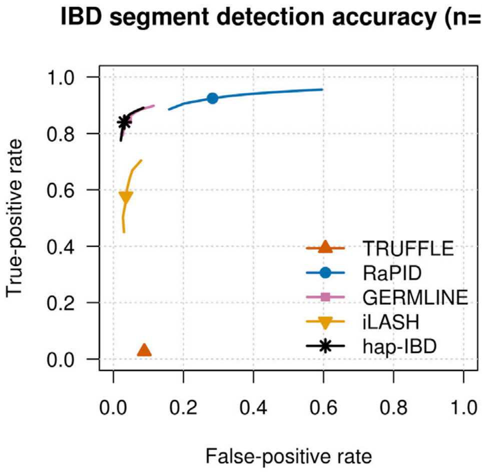

Figure 7. Receiver Operating Characteristic Curves for IBD Segment Detection in Simulated Sequence Data
False-positive and false-negative rates for detection of identity by descent (IBD) segments over a range of output length thresholds around 2 centiMorgan (cM) for 5,000 simulated samples. In order to allow for some discrepancy between reported and true lengths, false positives are assessed using true segments having length >1.5 cM, and false negatives are assessed using true segments having length >2.5 cM. IBD segments were detected with each method using length thresholds of 2 cM (plotted symbol) and with other thresholds between 1.6 and 2.4 cM (plotted lines; see Material and Methods). Figure S8 shows a similar plot for 50,000 samples.

Reported IBD segments in these regions typically do not result from recent shared ancestry. Previous attempts to address this issue of variable informativeness across the genome include methods that are based on haplotype frequency modeling, such as fastIBD[11] and RefinedIBD,[46] however, these methods do not scale to analysis of large datasets.

A general limitation of IBD segment detection methods that rely on IBS is that there is some degree of error in determination of segment endpoints. The IBS interval can extend beyond the endpoints of a contained IBD segment. Consequently, IBD detection methods that report the full IBS interval will often overextend the IBD segment ends. Such methods can also miss some regions at the end of IBD segments when genotype error, mutation, or gene conversion near the end of the IBD segment causes the IBS segment to end before the actual end of the IBD segment. If the genetic distance between the truncated end of the IBS segment and the true end of the IBD region is short, it is not possible to determine with confidence whether or not the IBD segment extends past the end of the IBS segment. Development of IBD segment detection methods that are robust to genotype error, recent mutation, and gene conversion that occur near the ends of IBD segments is an area for future research.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.02.010.

Acknowledgments

Research reported in this publication was supported by the National Human Genome Research Institute and National Institute of General Medical Sciences of the National Institutes of Health under award numbers HG008359, HG005701, and GM075091. 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.

Declaration of Interests

The authors declare no competing interests.

Received: December 11, 2019
Accepted: February 12, 2020
Published: March 12, 2020

Web Resources

hap-IBD, https://github.com/browning-lab/hap-ibd

The European Genome-Phenome Archive, https://www.ebi.ac.uk/ega/home

References

1. Houwen, R.H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A., and Freimer, N.B. (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat. Genet. *8*, 380–386.
2. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
3. Kenny, E.E., Gusev, A., Riegel, K., Lütjohann, D., Lowe, J.K., Salit, J., Maller, J.B., Stoffel, M., Daly, M.J., Altshuler, D.M., et al. (2009). Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. Proc. Natl. Acad. Sci. USA *106*, 13886–13891.
4. Moltke, I., Albrechtsen, A., Hansen, T.V., Nielsen, F.C., and Nielsen, R. (2011). A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. Genome Res. *21*, 1168–1180.
5. Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D.M., Friedman, J.M., Breslow, J.L., and Pe'er, I. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. *88*, 706–717.
6. Browning, S.R., and Thompson, E.A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics *190*, 1521–1531.
7. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.
8. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.

9. Browning, S.R., and Browning, B.L. (2013). Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. Hum. Genet. *132*, 129–138.

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

12. 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 Res. *21*, 768–774.

13. 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.

14. Qiao, Y., Sannerud, J., Basu-Roy, S., Hayward, C., and Williams, A.L. (2019). Distinguishing pedigree relationships using multi-way identical by descent sharing and sex-specific genetic maps. bioRxiv. https://doi.org/10.1101/753343.

15. Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P., and Pe'er, I. (2012). The architecture of long-range haplotypes shared within and across populations. Mol. Biol. Evol. *29*, 473–486.

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

17. Fu, W., Browning, S.R., Browning, B.L., and Akey, J.M. (2016). Robust inference of identity by descent from exome-sequencing data. Am. J. Hum. Genet. *99*, 1106–1116.

18. Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany, A.R., Myres, N.M., Barber, M.J., et al. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat. Commun. *8*, 14238.

19. Taylor, A.R., Schaffner, S.F., Cerqueira, G.C., Nkhoma, S.C., Anderson, T.J.C., Sriprawat, K., Pyae Phyo, A., Nosten, F., Neafsey, D.E., and Buckee, C.O. (2017). Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. PLoS Genet. *13*, e1007065.

20. Henden, L., Lee, S., Mueller, I., Barry, A., and Bahlo, M. (2018). Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. *14*, e1007279.

21. 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.

22. Palamara, P.F., and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. Bioinformatics *29*, i180–i188.

23. 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.

24. 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.

25. 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.

26. 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.

27. Palamara, P.F., Francioli, L.C., Wilton, P.R., Genovese, G., Gusev, A., Finucane, H.K., Sankararaman, S., Sunyaev, S.R., de Bakker, P.I., Wakeley, J., et al.; Genome of the Netherlands Consortium (2015). Leveraging distant relatedness to quantify human mutation and gene-conversion rates. Am. J. Hum. Genet. *97*, 775–789.

28. 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.

29. Zhou, Y., Browning, B.L., and Browning, S. (2019). Population-specific recombination maps from segments of identity by descent. bioRxiv. https://doi.org/10.1101/868091.

30. 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.

31. Cai, Z., Camp, N.J., Cannon-Albright, L., and Thomas, A. (2011). Identification of regions of positive selection using Shared Genomic Segment analysis. Eur. J. Hum. Genet. *19*, 667–671.

32. Han, L., and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. Eur. J. Hum. Genet. *21*, 205–211.

33. 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.

34. Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E.A. (2003). Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. *73*, 516–523.

35. 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.

36. 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. Genet. Epidemiol. *33*, 266–274.

37. Han, L., and Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. *35*, 557–567.

38. Brown, M.D., Glazner, C.G., Zheng, C., and Thompson, E.A. (2012). Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics *190*, 1447–1460.

39. Thompson, E.A. (2008). The IBD process along four chromosomes. Theor. Popul. Biol. *73*, 369–373.

40. 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.

41. 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. Nat. Genet. *40*, 1068–1075.
42. Dimitromanolakis, A., Paterson, A.D., and Sun, L. (2019). Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE. Am. J. Hum. Genet. *105*, 78–88.
43. Naseri, A., Liu, X., Tang, K., Zhang, S., and Zhi, D. (2019). RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. Genome Biol. *20*, 143.
44. Shemirani, R., Belbin, G.M., Avery, C.L., Kenny, E.E., Gignoux, C.R., and Ambite, J.L. (2019). Rapid detection of identity-by-descent tracts for mega-scale datasets. bioRxiv. https://doi.org/10.1101/749507.
45. Chiang, C.W., Ralph, P., and Novembre, J. (2016). Conflation of short identity-by-descent segments bias their inferred length distribution. G3: Genes, Genomes. G3 (Bethesda) *6*, 1287–1296.
46. Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics *194*, 459–471.
47. Browning, B.L., and Browning, S.R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. *93*, 840–851.
48. Rodriguez, J.M., Bercovici, S., Huang, L., Frostig, R., and Batzoglou, S. (2015). Parente2: a fast and accurate method for detecting identity by descent. Genome Res. *25*, 280–289.
49. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
50. 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. https://doi.org/10.1101/563866.
51. Weedon, M.N., Jackson, L., Harrison, J.W., Ruth, K.S., Tyrrell, J., Hattersley, A.T., and Wright, C.F. (2019). Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing. bioRxiv. https://doi.org/10.1101/696799.
52. Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics *30*, 1266–1272.
53. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. *103*, 338–348.
54. Jeffreys, A.J., and May, C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. *36*, 151–156.
55. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
56. Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.
57. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. *47*, 692–695.
58. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
59. 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.
60. 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*, 5436.
61. Haller, B.C., and Messer, P.W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. Mol. Biol. Evol. *34*, 230–240.
62. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol. Biol. Evol. *36*, 632–637.
63. 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.; T2D-GENES Consortium (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. eLife *4*, e04637.
64. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. *12*, e1004842.
65. 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.
66. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
67. Li, H., Glusman, G., Hu, H., Shankaracharya, Caballero, J., Hubley, R., Witherspoon, D., Guthery, S.L., Mauldin, D.E., Jorde, L.B., et al. (2014). Relationship estimation from whole-genome sequence data. PLoS Genet. *10*, e1004144.
68. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
69. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.
70. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

The American Journal of Human Genetics, Volume *106*

Supplemental Data

A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data

Ying Zhou, Sharon R. Browning, and Brian L. Browning

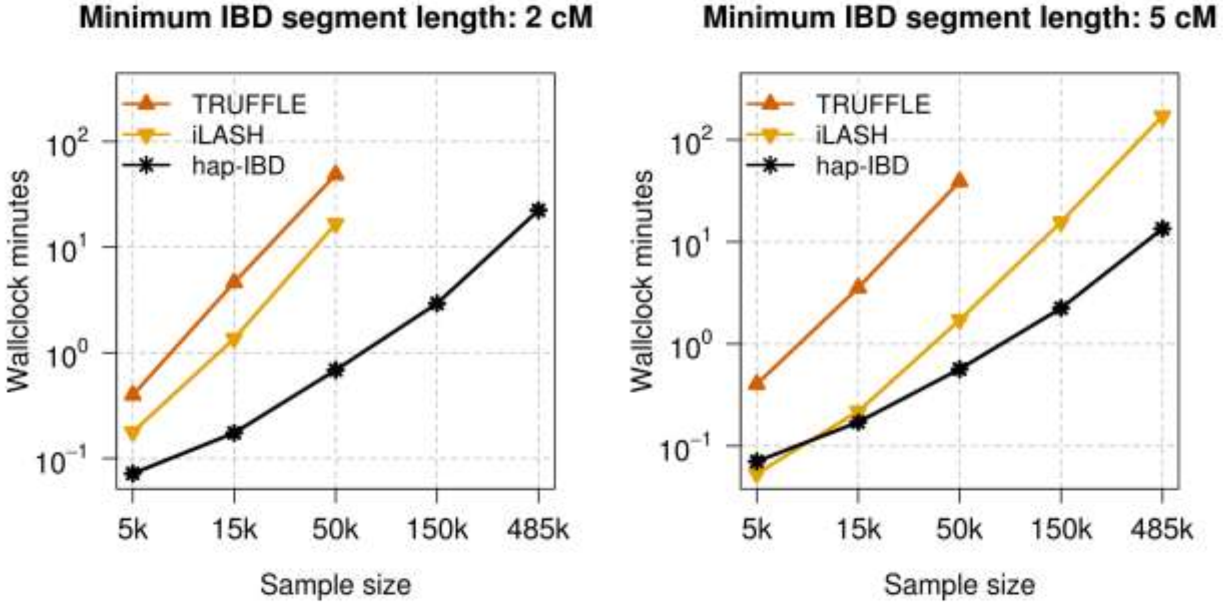

Figure S1. Wall-clock compute time. Wall-clock time for multi-threaded programs when using 12 CPU cores for detecting IBD segments with length ≥ 2 cM (left panel) and ≥ 5 cM (right panel) on chromosome 20 in samples of 5000, 15,000, 50,000, 150,000, and 485,346 individuals from the UK Biobank. All programs used 12 computational threads. The wall-clock times for phasing the 5000, 15,000, 50,000, 150,000, and 485,346 individuals using Beagle 5.1 were 14, 44, 169, 554, and 1729 minutes.

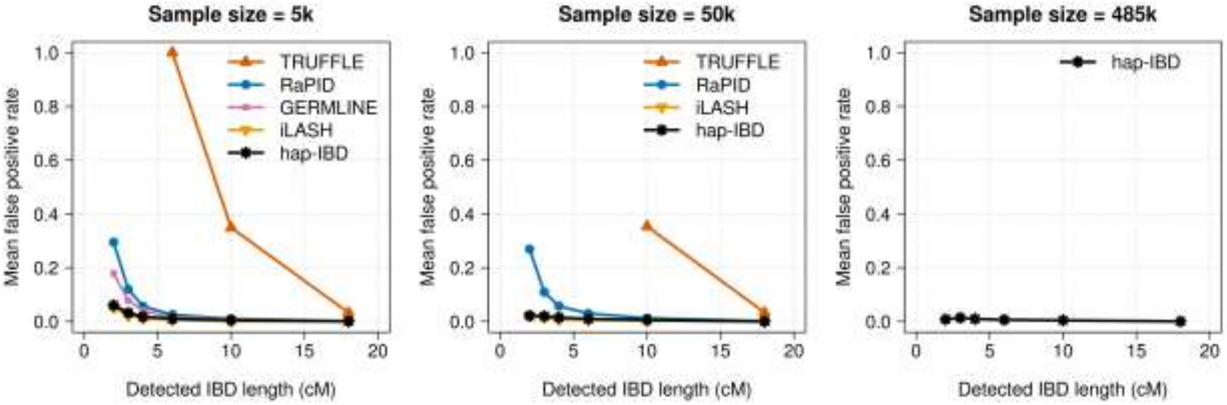

Figure S2. False-positive IBD segment detection in UK Biobank chromosome 20 data. As for Figure 2, but zoomed out to show the full range of false-positive rates. False-positive rates for IBD segment detection for 5000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥ 2 cM were detected with each method. Detected IBD segments were assigned into bins of 2-3, 3-4, 4-6, 6-10, 10-18, and >18 cM according to their segment length. The false-positive rate is the proportion of detected IBD segments in a bin that are not covered by any true IBD segment > 1.5 cM in length. Hap-IBD is the only method shown for the full UK Biobank analysis (485,346 individuals) because other methods were unable to complete the analysis with a 2 cM output threshold within the memory and time constraints (see Computational Feasibility Results). The x-coordinate of each data point is the left bin end point (e.g. 2 cM for the 2-3 cM bin).

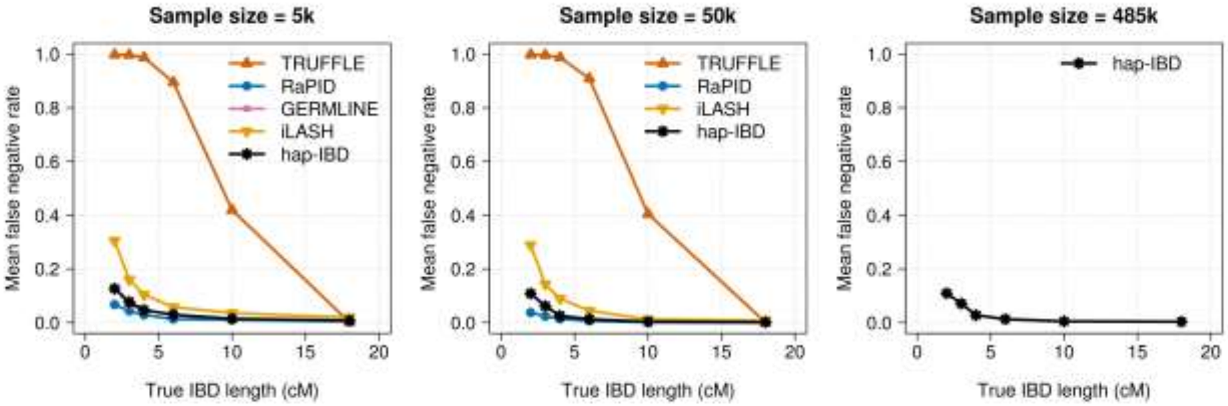

Figure S3. False-negative IBD segment detection in UK Biobank chromosome 20 data. As for Figure 3, but zoomed out to show the full range of false-negative rates. False-negative rates for IBD segment detection for 5000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥ 2 cM were detected with each method. True IBD segments with length > 2.5 cM were assigned into bins of 2.5-3, 3-4, 4-6, 6-10, 10-18, and >18 cM according to their segment length. The false-negative rate is the proportion of true IBD segments in a bin that are not covered by any detected IBD segment ≥ 2 cM in length. Hap-IBD is the only method shown for the full UK Biobank analysis (485,346 individuals) because other methods were unable to complete the analysis with a 2 cM output threshold within the memory and time constraints (see Computational Feasibility Results). The x-coordinate of each data point is the left bin end point (e.g. 2.5 cM for the 2.5-3 cM bin).

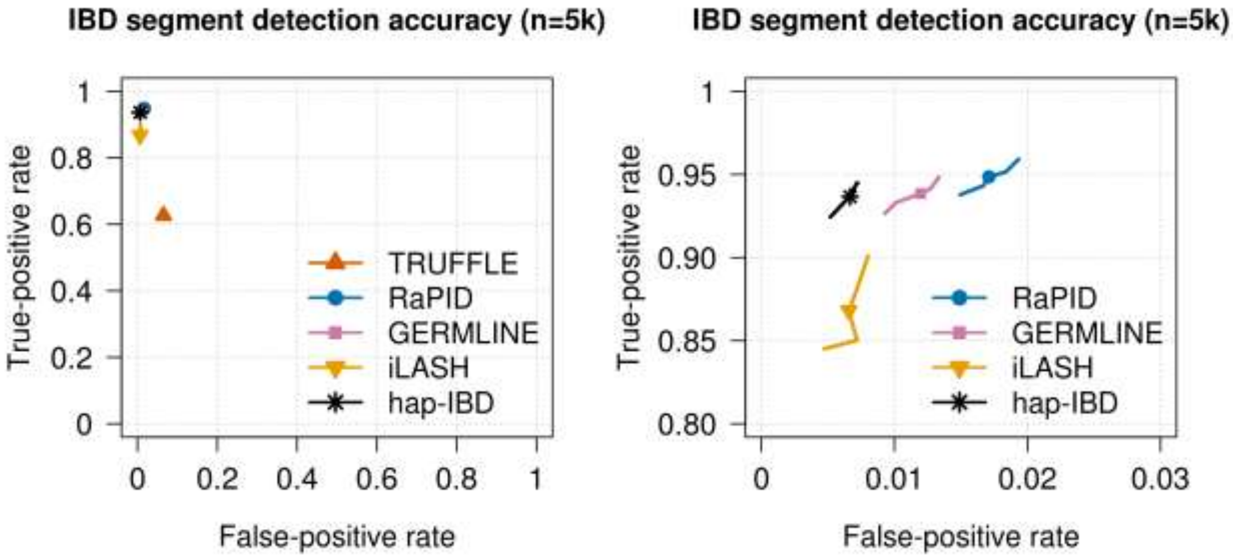

Figure S4. ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. The right panel is a zoomed-in version of the left panel. False positives are assessed using true segments having length > 1.5 cM, and false negatives are assessed using true segments having length > 5.5 cM in order to allow for some discrepancy between reported and true lengths. IBD segments were detected with each method using length thresholds of 5 cM (plotted symbol) and with other thresholds between 4.6 and 5.4 cM (plotted lines; see Methods).

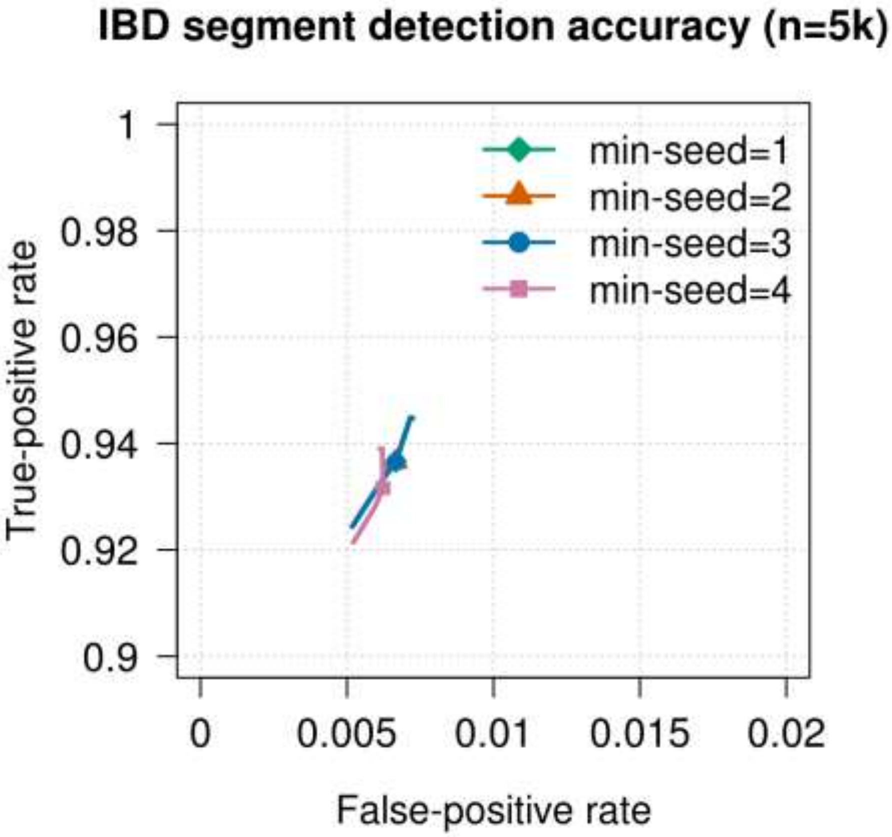

Figure S5. Effect of varying hap-IBD min-seed parameter on ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. False positives are assessed using true segments having length > 1.5 cM, and false negatives are assessed using true segments having length > 5.5 cM in order to allow for some discrepancy between reported and true lengths. IBD segments were detected using length thresholds of 5 cM (plotted symbol) and with other thresholds between 4.6 and 5.4 cM (plotted lines; see Methods). The ROC curves for minimum seed lengths of 1, 2, and 3 cM are indistinguishable.

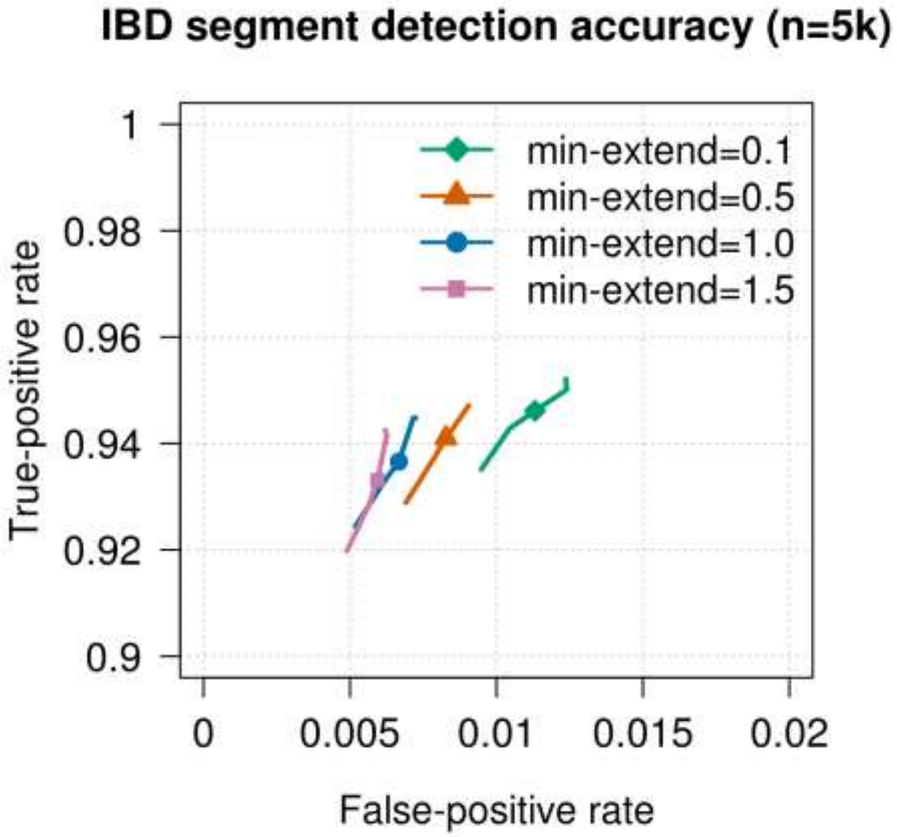

Figure S6. Effect of varying hap-IBD min-extend parameter on ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. False positives are assessed using true segments having length > 1.5 cM, and false negatives are assessed using true segments having length > 5.5 cM in order to allow for some discrepancy between reported and true lengths. IBD segments were detected using length thresholds of 5 cM (plotted symbol) and with other thresholds between 4.6 and 5.4 cM (plotted lines; see Methods).

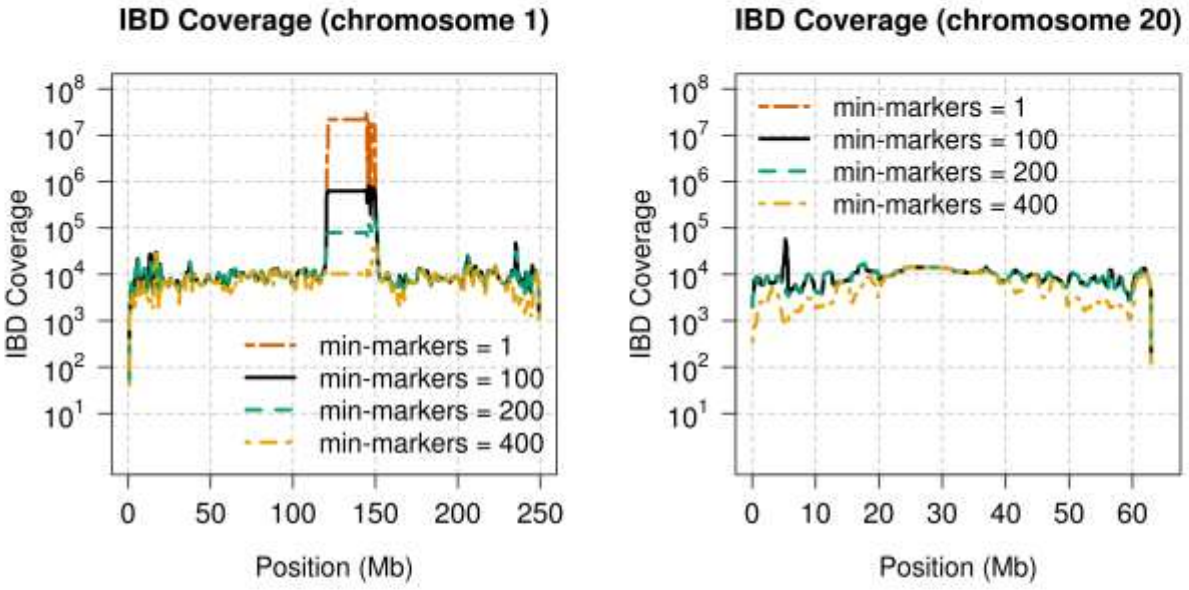

Figure S7: Effect of marker thresholds on IBD segment detection in UK Biobank. The hap-IBD program was run on 5000 UK Biobank samples on chromosome 1 (left panel) and chromosome 20 (right panel) with the min-markers parameter set to 1, 100, 200, and 400 markers. The min-markers parameter controls the minimum number of markers that must be present in a reported seed IBD segment. All other hap-IBD parameters were set at their default values. Each chromosome is divided into non-overlapping 10 kb intervals. For each interval, the IBD segments intersecting the interval are each weighted by the proportion of the 10 kb interval that is covered by the IBD segment, and the sum of weights is plotted as the IBD coverage.

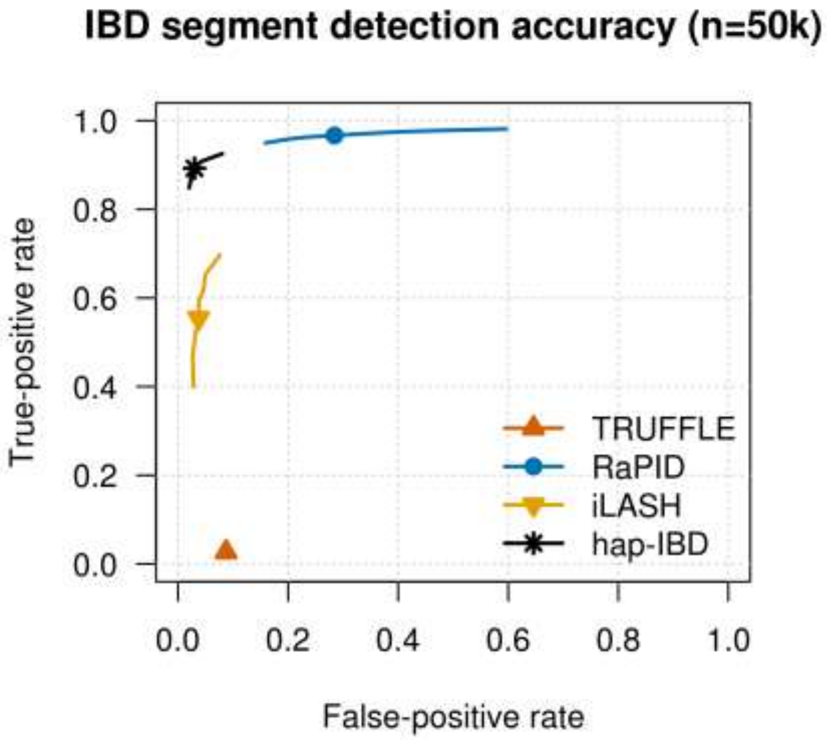

Figure S8. ROC curves for IBD segment detection in simulated sequence data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 2 cM for 50,000 simulated samples. False positives are assessed using true segments having length > 1.5 cM, and false negatives are assessed using true segments having length > 2.5 cM in order to allow for some discrepancy between reported and true lengths. IBD segments were detected with each method using length thresholds of 2 cM (plotted symbol) and with other thresholds between 1.6 and 2.4 cM (plotted lines; see Methods).

Method	Parameters
hap-IBD v1.0	Default options (`min-seed=2.0 max-gap=1000 min-extend=1.0 min-output=2.0 min-mac 2 min-markers=100`)
TRUFFLE v1.38	`-mindist 5000 -maf 0.05 -segments -L 1` (output was filtered to exclude segments <2.0 cM in length)
GERMLINE 1.5.3	`-haploid -bits 32 -w_extend -min_m 2.0`
RaPID v1.7	`-r 10 -s 2 -w 3 -d 2.0`
iLASH (commit de697321*)	`perm_count 12 shingle_size 20 shingle_overlap 0 bucket_count 4 max_thread 12 match_threshold 0.99 interest_threshold 0.7 max_error 0 min_length 2.0 auto_slice 1 cm_overlap 1.4`

Table S1: Parameters used for analysis of UK Biobank data with 2 cM minimum IBD segment length.

Parameters that control the minimum output IBD segment length are in red.

Method	**Parameters**
hap-IBD v1.0	`min-seed=1.0 max-gap=1000 min-extend=0.2 min-output=2.0 min-markers=100`
TRUFFLE v1.38	`-mindist 5000 -maf 0.1 -segments -L 1` (output was filtered to exclude segments <2.0 cM in length)
GERMLINE 1.5.3	`-haploid -bits 75 -w_extend -min_m 2.0 -err_hom 2`
RaPID v1.7	`-r 10 -s 2 -w 80 -d 2.0`
iLASH (commit de697321*)	`perm_count 12 shingle_size 20 shingle_overlap 0 bucket_count 4 max_thread 12 match_threshold 0.99 interest_threshold 0.7 max_error 0 min_length 2.0 auto_slice 1 cm_overlap 1.4`

Table S2: Parameters used for analysis of simulated sequence data with minimum 2 cM IBD segment length. Parameters that control the minimum output IBD segment length are in red.