The American Journal of Human Genetics, Volume *107*

Supplemental Data

Probabilistic Estimation of Identity by Descent Segment Endpoints and Detection of Recent Selection

Sharon R. Browning and Brian L. Browning

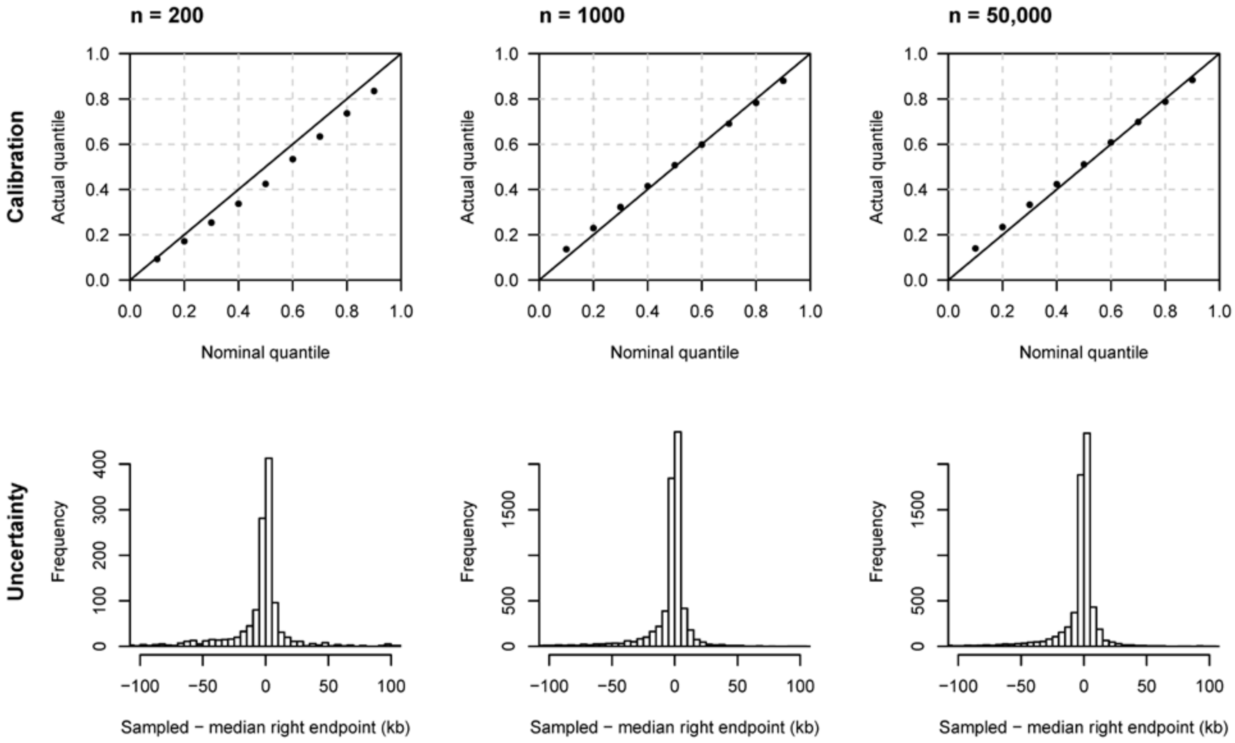


Figure S1: Performance on UK-like simulated data with varying sample sizes. The input data comprise sequence data on individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%, and true haplotype phase. The top row shows quantile-quantile plots which assess 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 posterior median right endpoint. 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 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.

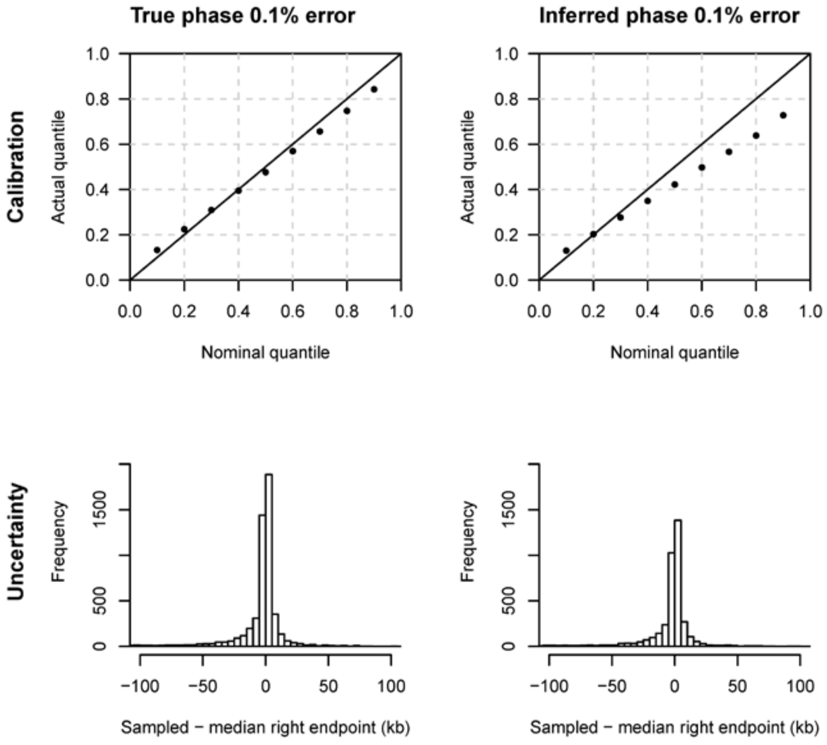


Figure S2: Performance on UK-like simulated data with a 0.1% rate of genotype error. The data comprise sequence data on 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.1%. Results were calculated using true IBD segment endpoints for all pairs 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 posterior median right endpoint. 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.

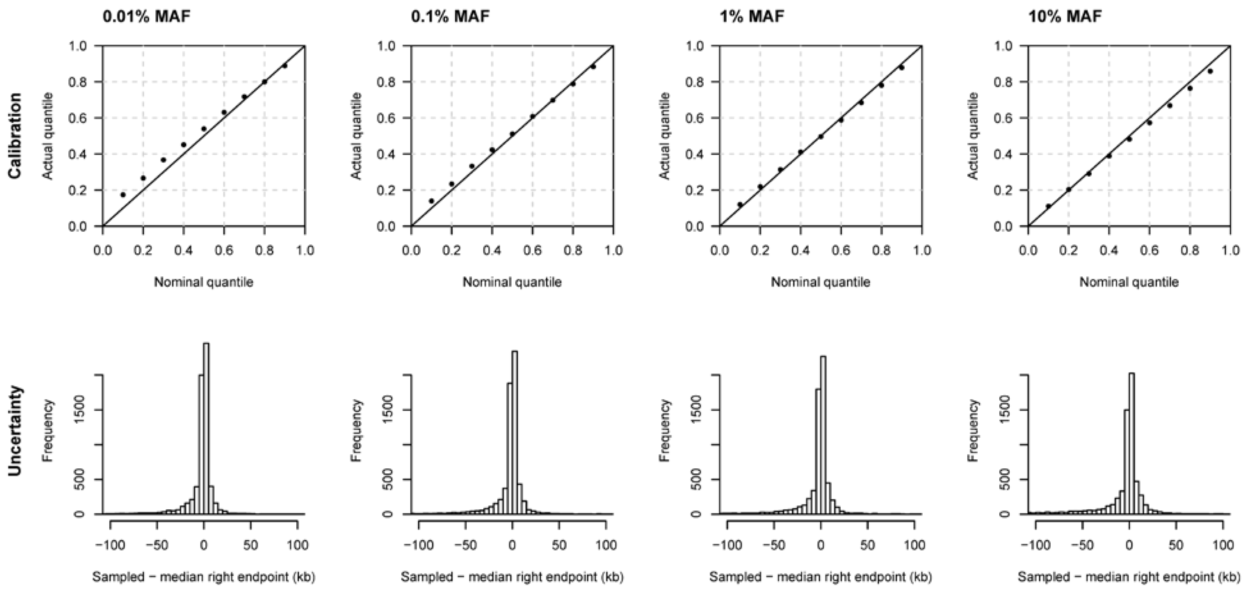


Figure S3: Performance on UK-like simulated data with varying minor allele frequency filters. The data comprise sequence data on 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. The top row shows quantile-quantile plots which assess the 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 posterior median right endpoint. 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.

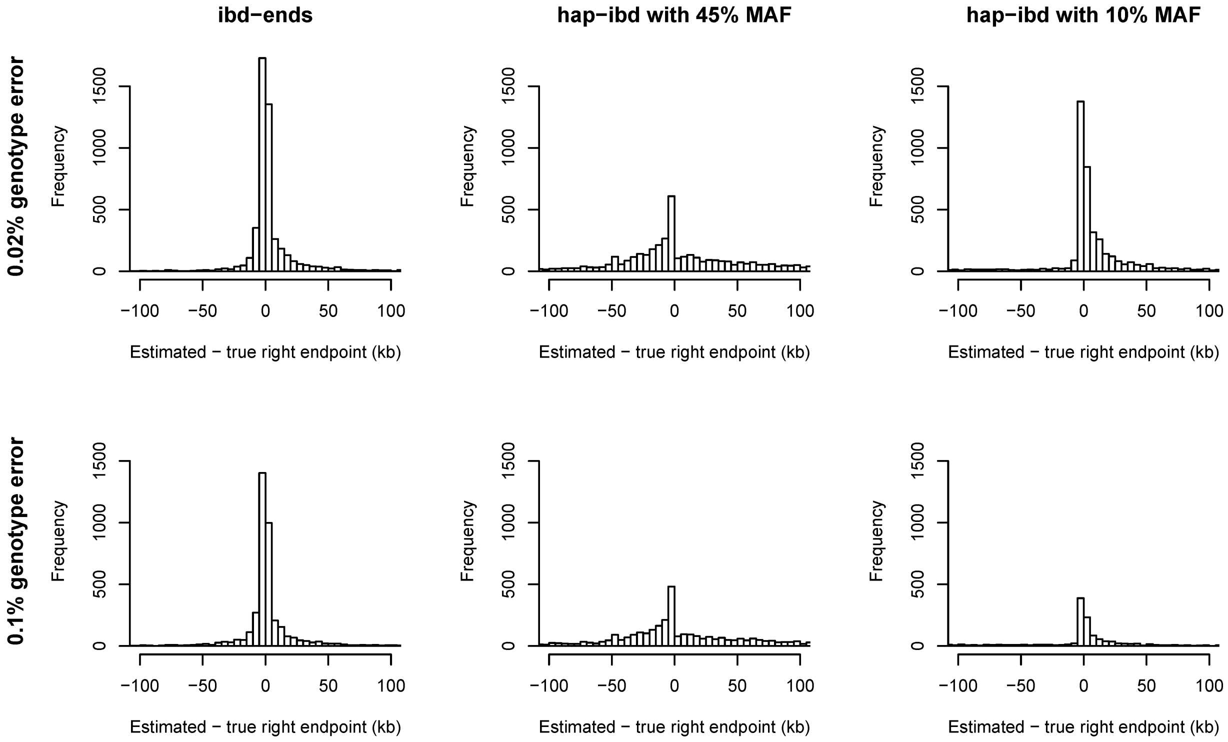


Figure S4: Comparing precision of estimated endpoints between ibd-ends and hap-ibd. The data comprise 50,000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. True IBD segment endpoints were determined for 1000 individuals, and these individuals were used to generate the results in this figure. Analyses are restricted to segments with true length > 2 cM. The histograms show estimated right endpoints of IBD segments minus the true right endpoints of the segments. In the left column, the estimated endpoints are the posterior medians from ibd-ends. In the middle column, the estimated endpoints are from hap-ibd with a minor allele frequency filter of 45% (the hap-ibd segments used as input to ibd-ends). In the right column, the estimated endpoints are from hap-ibd with parameters recommended for sequence data (min-seed=1.0, min-extend=0.2, and a minor allele frequency filter of 10%). Histogram bin widths are 5 kb. All analyses use true haplotype phase. The top row has a low genotype error rate (0.02%), while the bottom row has a high genotype error rate (0.1%).

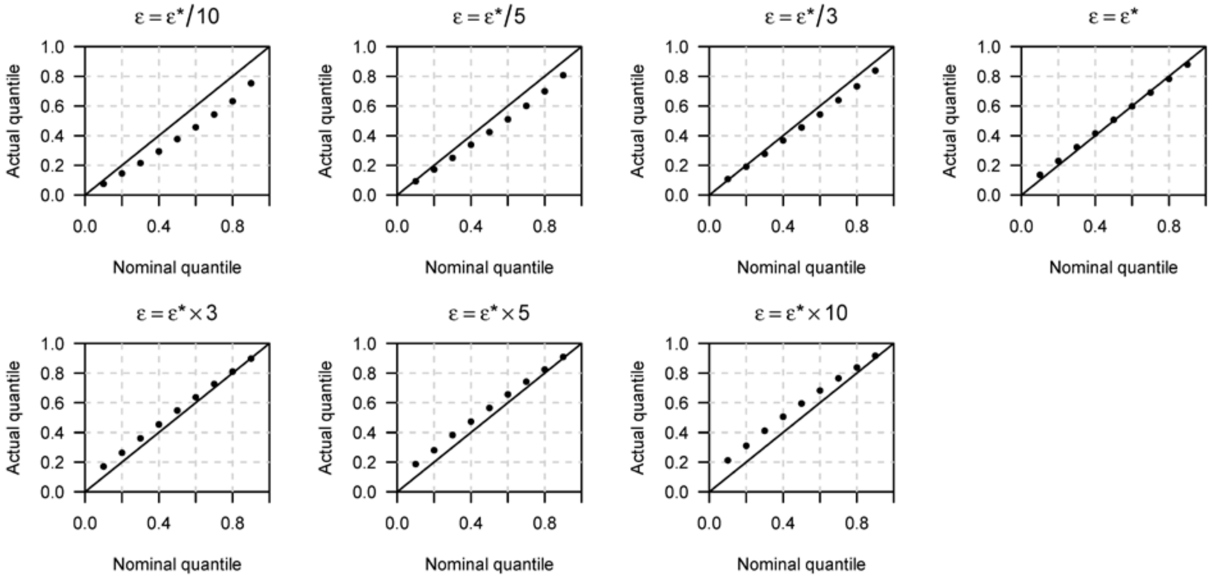


Figure S5: Performance on UK-like simulated data with varying analysis error rates. The data comprise sequence data on 1000 individuals simulated from a UK-like demographic history (see Methods), with a genotype error rate of 0.02%. Over each plot is the analysis error rate ε, as a function of the near-optimal rate for these data of $\varepsilon^* = 0.0005$. The plots are 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 true haplotype phase was used in all analyses.

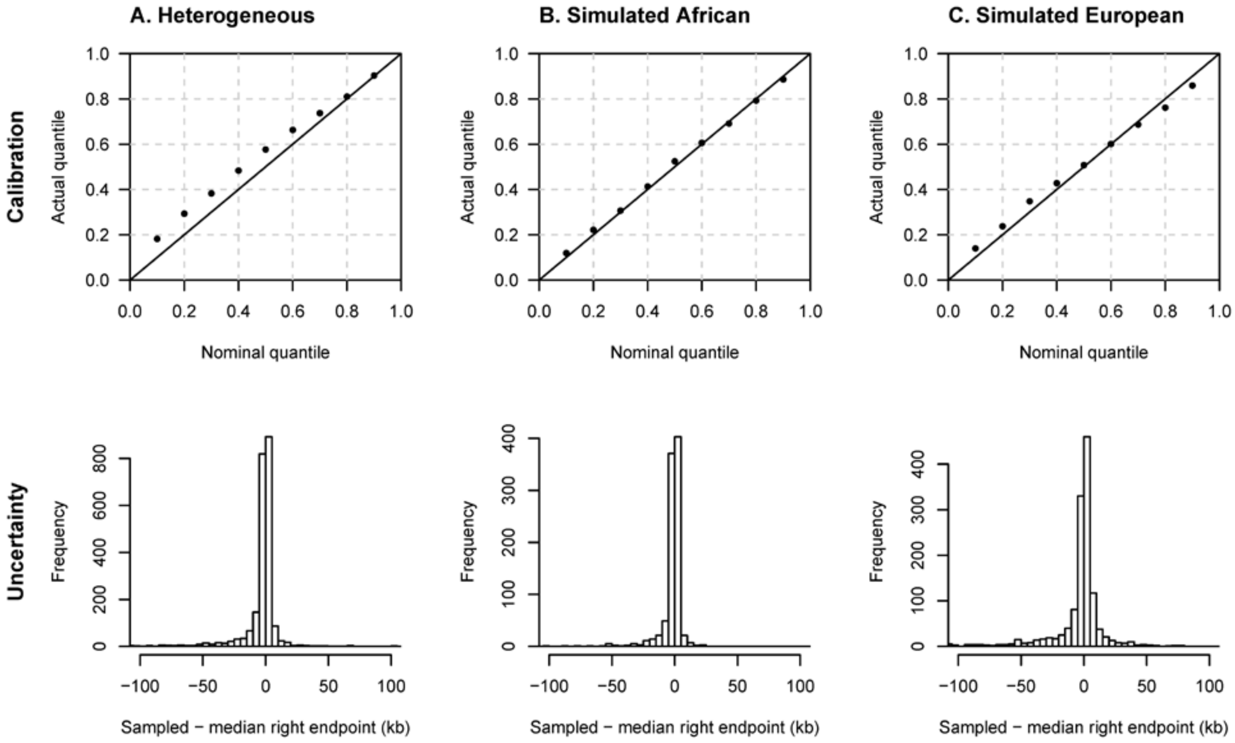
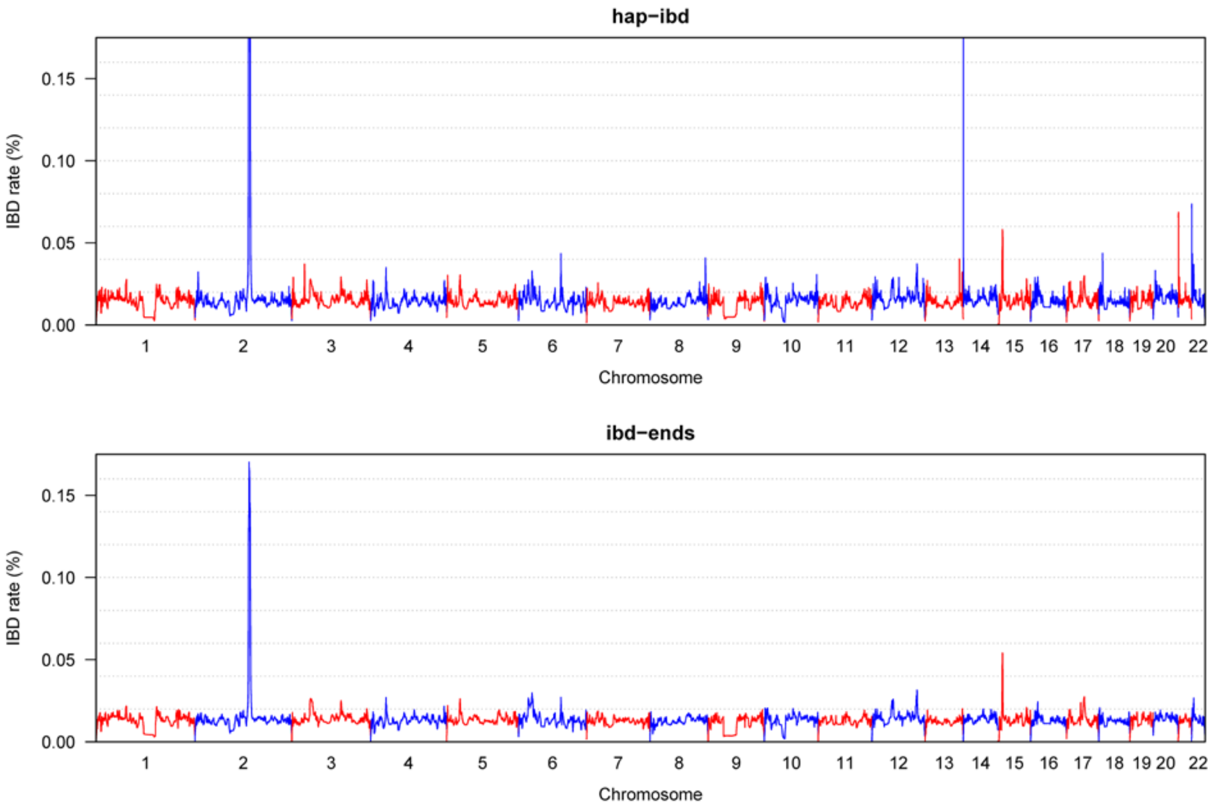


Figure S6: Performance in a simulated two-population model. The data comprise sequence data on 500 individuals each from simulated African and European demographic histories. The top row shows quantile-quantile plots which assess the calibration of the estimated endpoint uncertainty. The $y = x$ line 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 posterior median right endpoint. 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.



hap−ibd

ibd−ends

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 an IBD segment with length > 2 cM for the haplotype pair. IBD segment endpoints are posterior medians. 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.

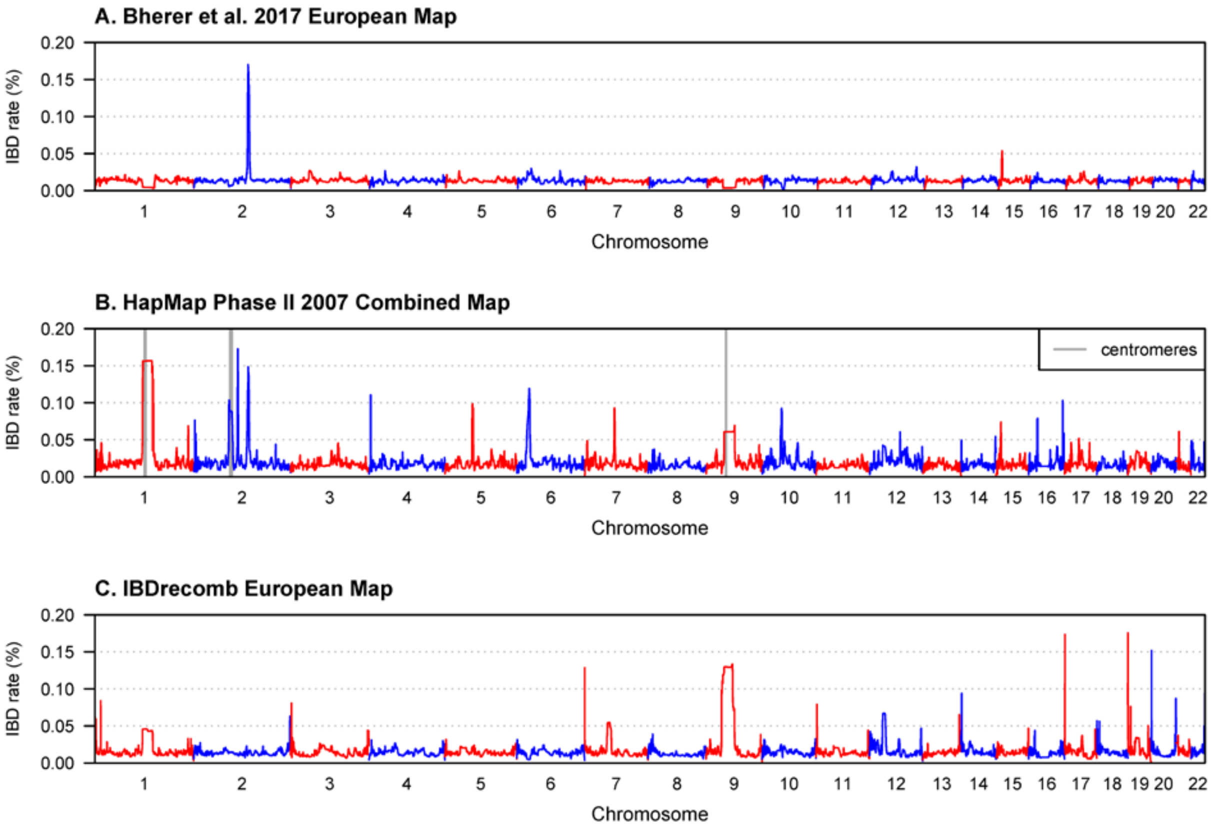


Figure S8: Effect of genetic map on inferred IBD rate. We estimated endpoint uncertainty for 50,000 White British individuals from the UK Biobank using three different maps. A) Bherer et al.'s refined European map.[1] B) The HapMap Phase II combined ancestry map,[2] with the centromeres for chromosomes 1, 2, and 9 notated. C) The IBDrecomb European map[3] 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 an IBD segment with length > 2 cM for the haplotype pair. IBD segment endpoints are posterior medians. The IBD rate is calculated at 10 kb intervals.

Supplemental References

1. 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.

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

3. 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.