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

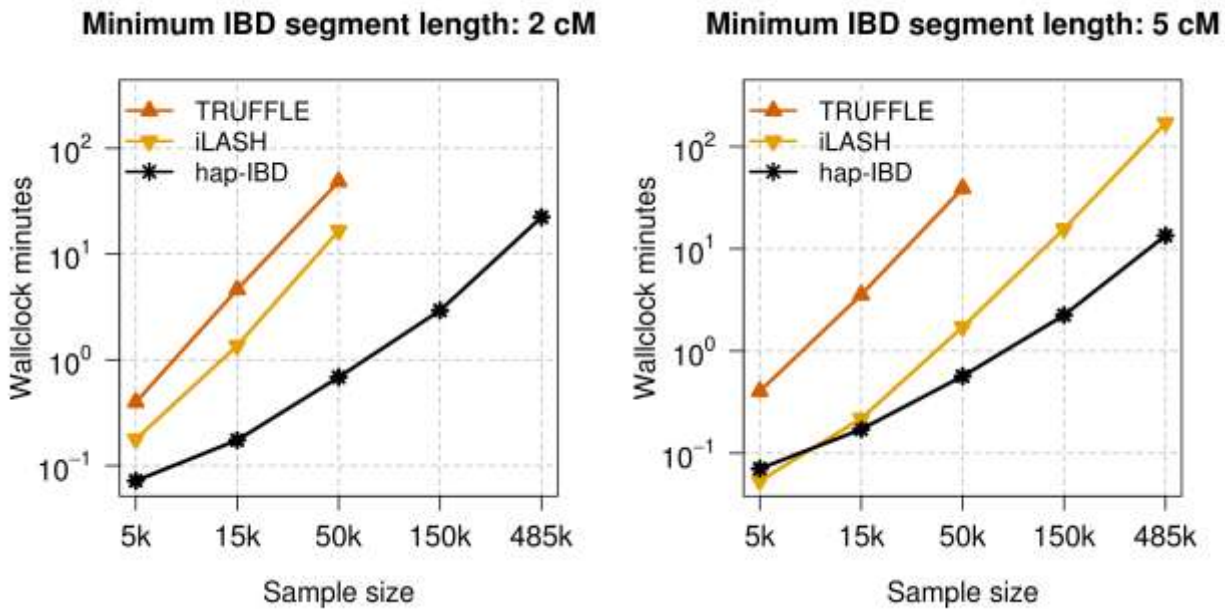

Figure S1. Wall-clock compute time. Wall-clock time for multi-threaded programs when using 12 CPU cores for detecting IBD segments with length ≥ 2 cM (left panel) and ≥ 5 cM (right panel) on chromosome 20 in samples of 5000, 15,000, 50,000, 150,000, and 485,346 individuals from the UK Biobank. All programs 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.

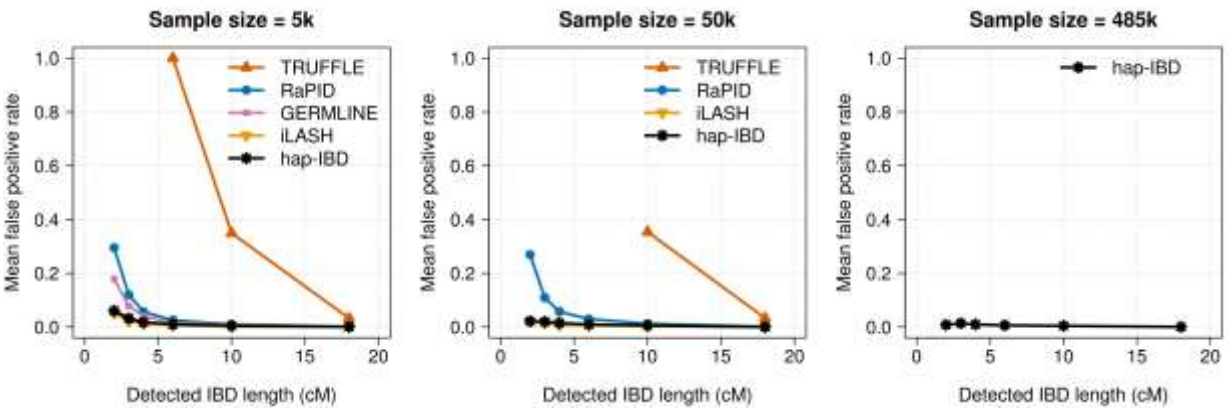

Figure S2. False-positive IBD segment detection in UK Biobank chromosome 20 data. As for Figure 2, but zoomed out to show the full range of false-positive rates. False-positive rates for IBD segment detection for 5000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥ 2 cM were detected 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).

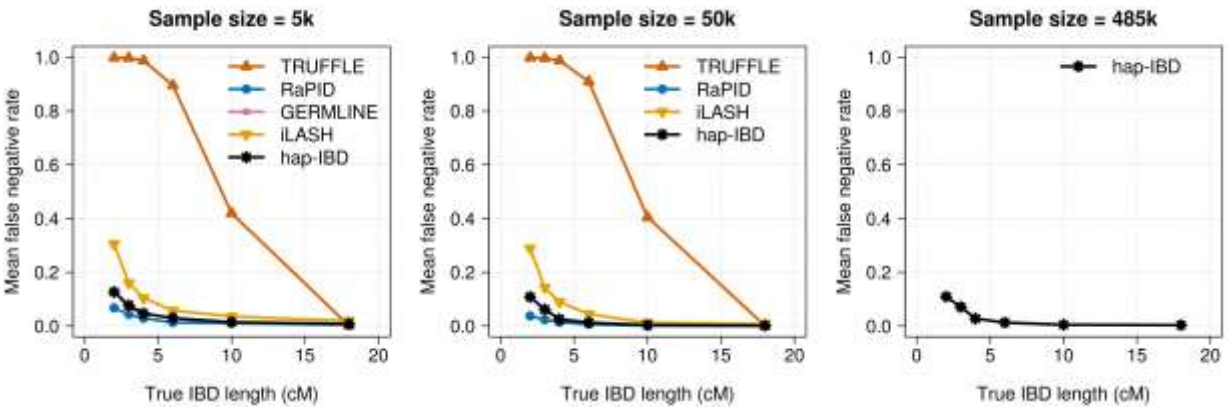

Figure S3. False-negative IBD segment detection in UK Biobank chromosome 20 data. As for Figure 3, but zoomed out to show the full range of false-negative rates. False-negative rates for IBD segment detection for 5000, 50,000, and 485,346 UK Biobank samples. IBD segments with length ≥ 2 cM were detected 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).

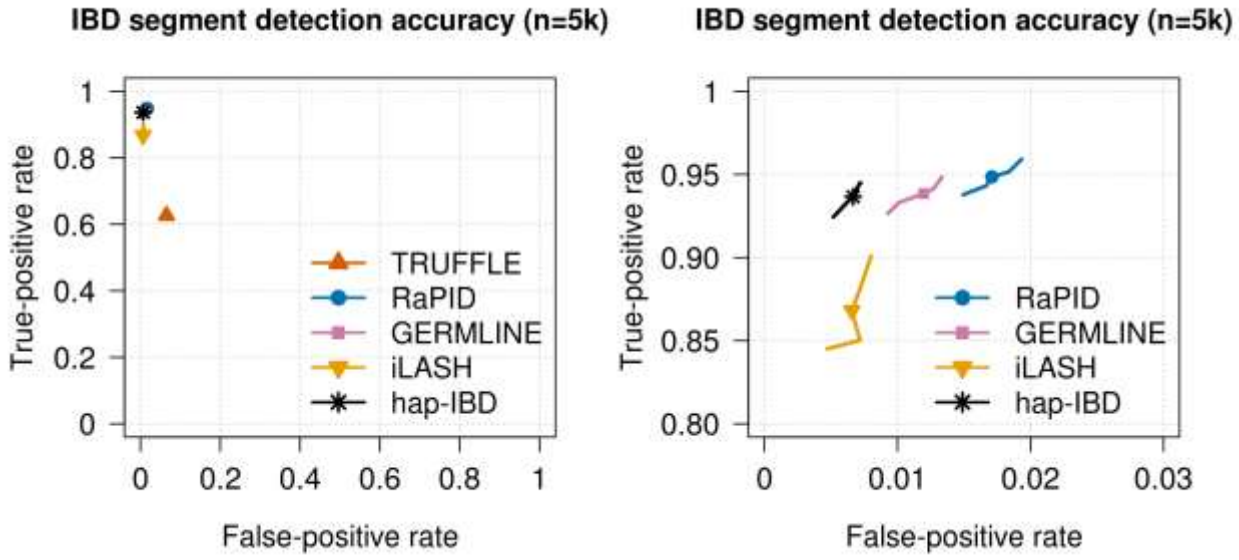

Figure S4. ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. The right panel is a zoomed-in version of the left panel. False 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).

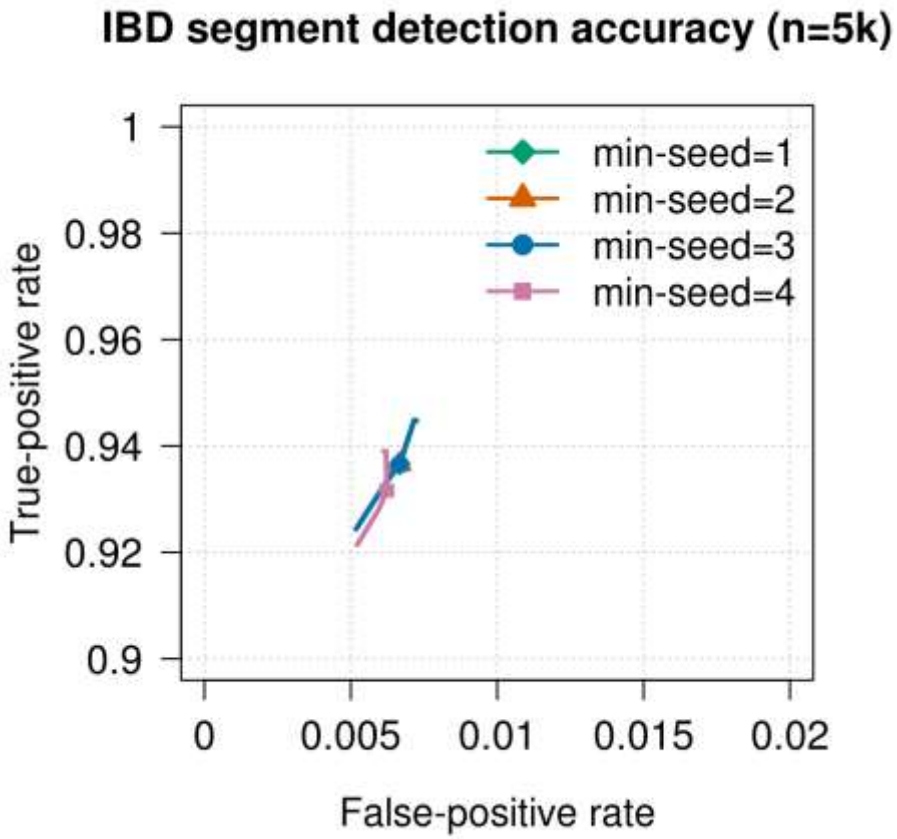

Figure S5. Effect of varying hap-IBD min-seed parameter on ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. False positives 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.

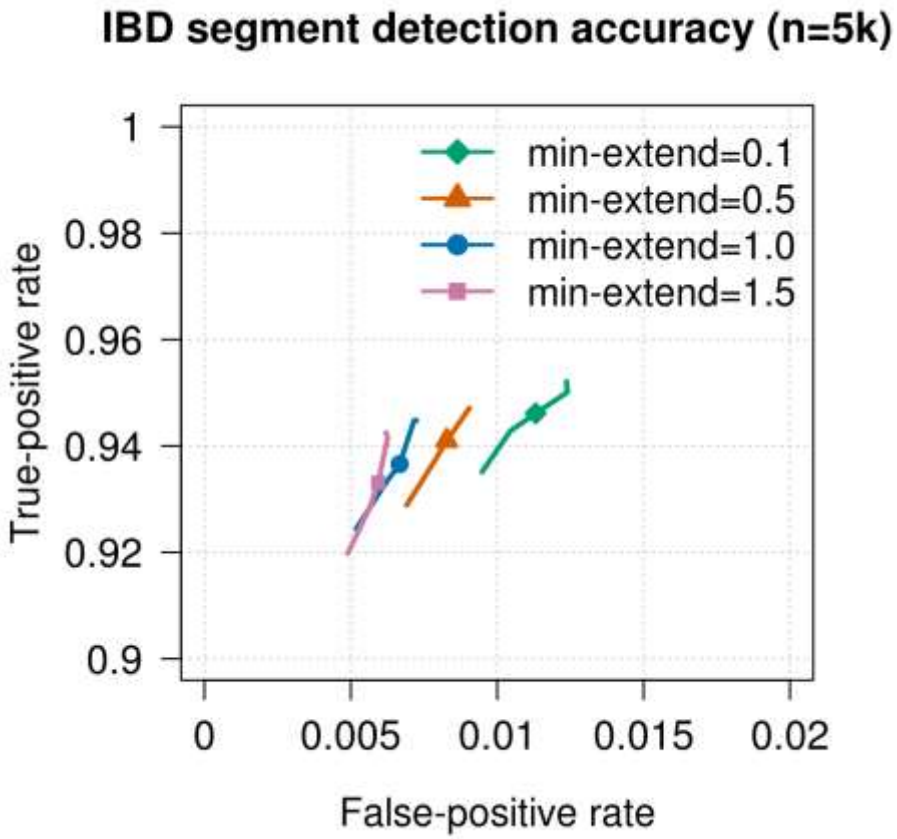

Figure S6. Effect of varying hap-IBD min-extend parameter on ROC curves for 5 cM IBD segment detection in UK Biobank chromosome 20 data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 5 cM for 5000 UK Biobank samples. False positives 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).

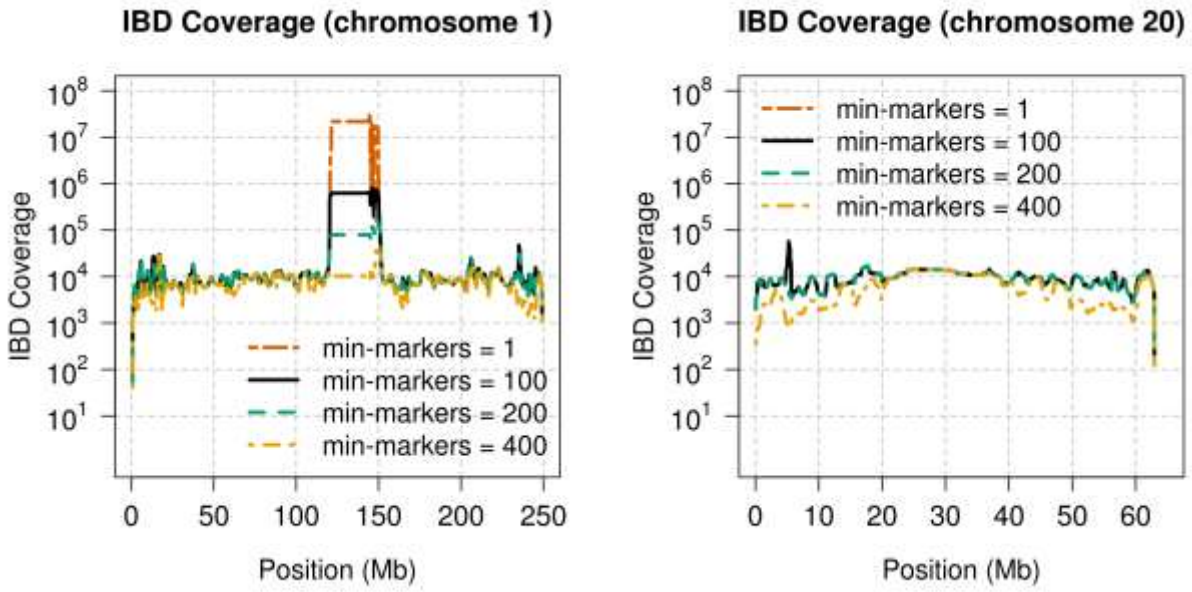

Figure S7: Effect of marker thresholds on IBD segment detection in UK Biobank. The hap-IBD program was run on 5000 UK Biobank samples on chromosome 1 (left panel) and chromosome 20 (right panel) with the min-markers parameter set to 1, 100, 200, and 400 markers. The min-markers parameter controls the 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.

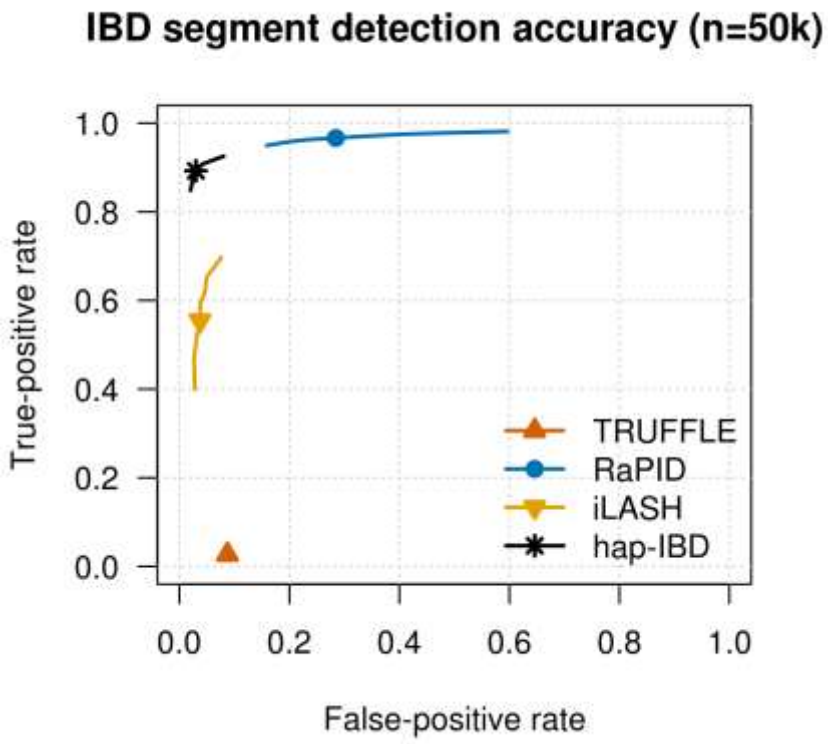

Figure S8. ROC curves for IBD segment detection in simulated sequence data. False-positive and false-negative rates for detection of IBD segments over a range of output length thresholds around 2 cM for 50,000 simulated samples. False positives are assessed using true segments having length > 1.5 cM, 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.