The American Journal of Human Genetics, Volume *93*

Supplemental Data

Detecting Identity by Descent

and Estimating Genotype Error Rates in Sequence Data

Brian L. Browning and Sharon R. Browning

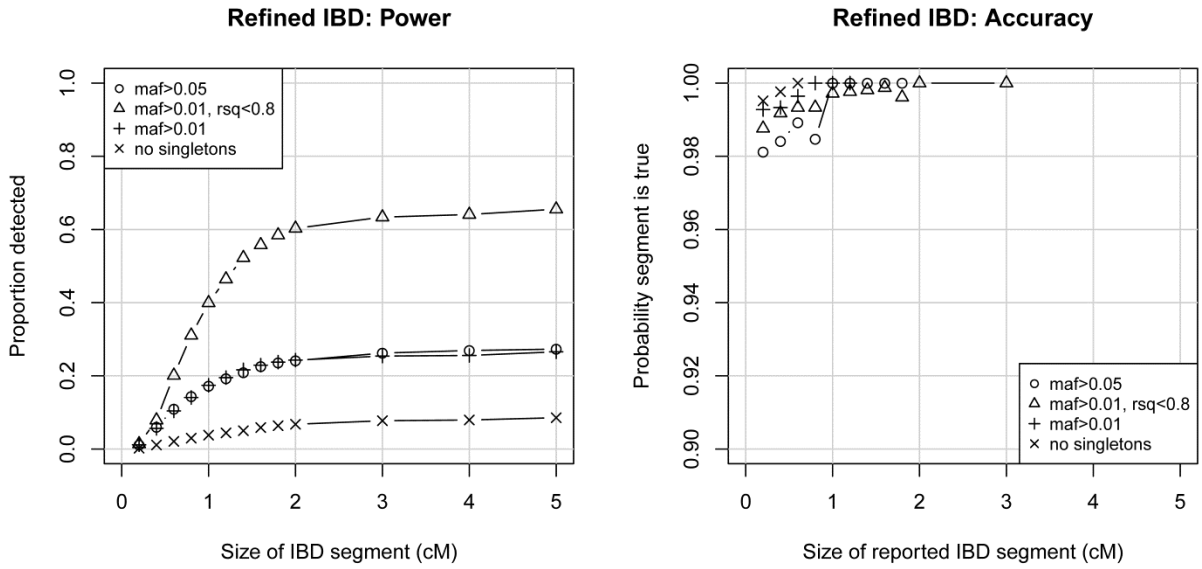

Figure S1: Power and accuracy with Refined IBD

As in Figure 1, power (proportion detected) is the average proportion of a true IBD segment of given length that overlaps with reported IBD segments. Accuracy (probability a segment is true) is the proportion of reported segments of given length for which there is a true segment that overlaps at least half of the reported segment. Results are binned by segment size: bins extend 0.05 cM on either side of the x-axis value for x-axis values ≤1 cM; 0.1 cM either side for x-axis values ≤2 cM; and 0.5 cM either side for x-axis values > 2 cM. We applied Refined IBD with several variant filtering strategies. We filtered on MAF and on r^2 correlation between variants ("rsq"). All analyses used a LOD threshold of 2.0, and a minimum IBD length of 0.2 cM. Note that the y-axis scale of the Refined IBD accuracy plot differs from that for the other methods (Figures S2-S4 below) due to the much greater accuracy with Refined IBD.

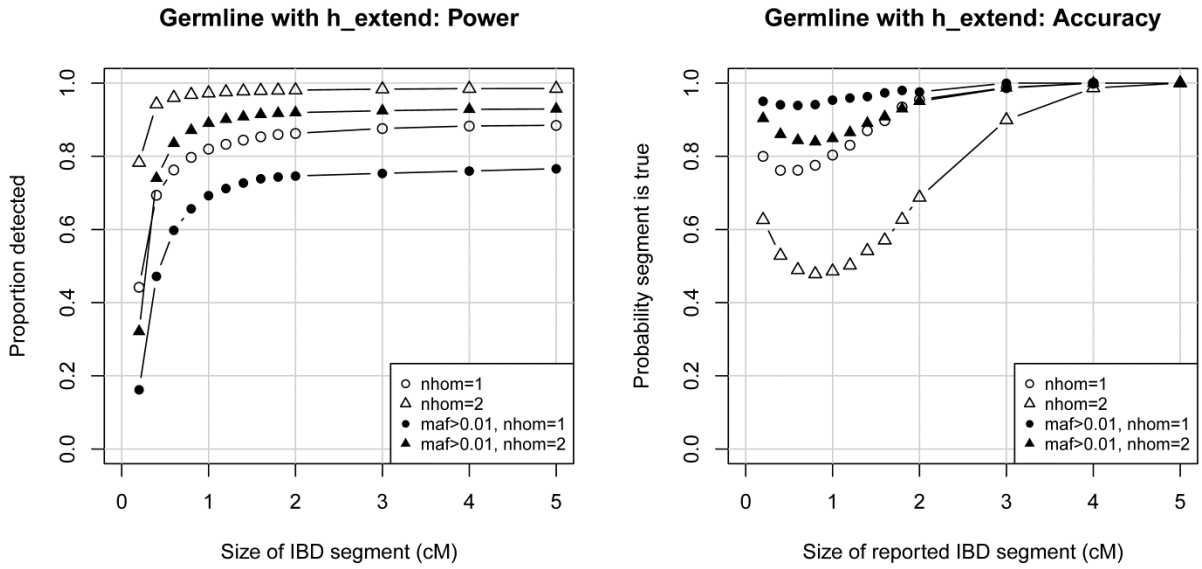

Figure S2: Power and accuracy of GERMLINE with h_extend

See Figure S1 for a description of axis labels. We used GERMLINE with the "bits" parameter set to 128. We analyzed the data with the "h_extend" option and with the error parameter "nhom" set to 1 or 2. We used a "min_m" parameter (minimum segment length) value of 0.2. The value of the nhet parameter made no difference in power or accuracy when using the h_extend parameter (data not shown). We compare a MAF threshold of 0.01 with removal of singletons only. Data were phased using Beagle version 4 with the settings used in our runs of Refined IBD. Marker exclusions (removal of singletons or MAF thresholding) were applied prior to phasing.

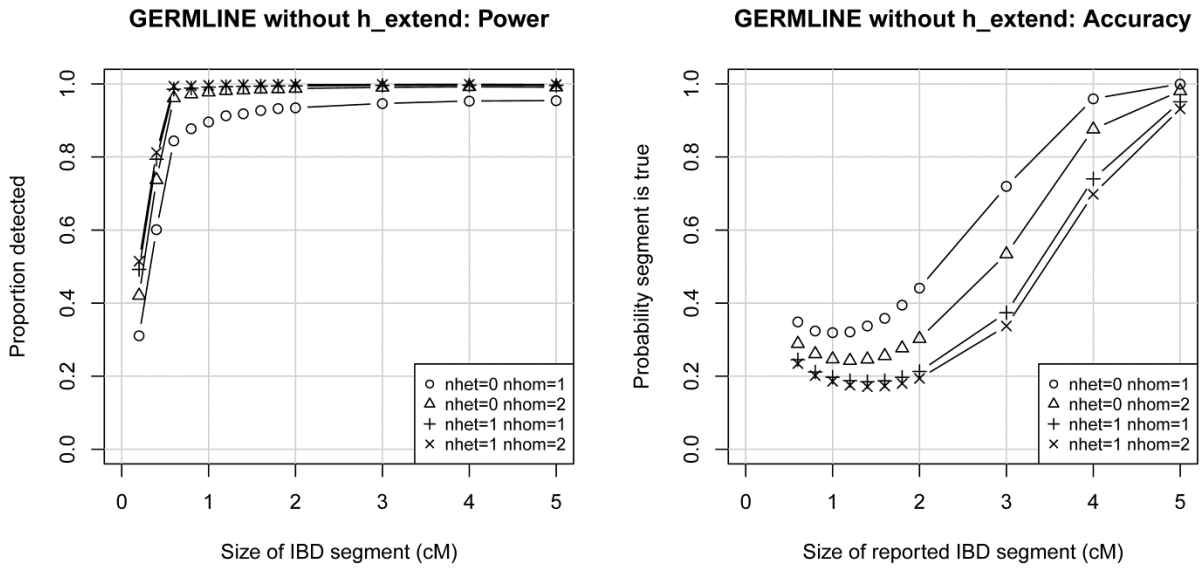

Figure S3: Power and accuracy of GERMLINE without h_extend.

See Figure S1 for a description of axis labels. We used GERMLINE with the "bits" parameter set to 128. We removed singletons only. We analyzed the data without the "h_extend" option, with the error parameter "nhom" set to 1 or 2, and with the "nhet" parameter set to 0 or 1. We used a "min_m" parameter (minimum segment length) value of 0.5. Although we used a minimum reported segment length of 0.5 cM, some true segments of length < 0.5 cM were reported with estimated length > 0.5 cM, hence there is some power to detect these shorter segments as shown in the power (left) plot for x-axis values (size of IBD segment) < 0.5 cM.

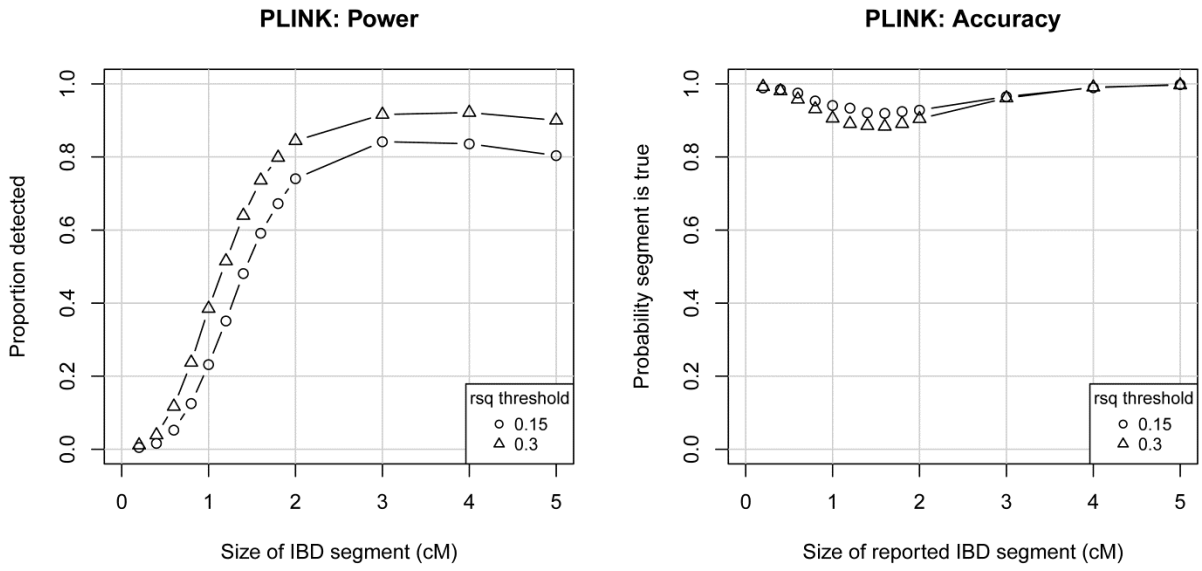

Figure S4: Power and accuracy with PLINK

See Figure S1 for a description of axis labels. We investigated two levels of r^2-based variant filtering ("rsq threshold").

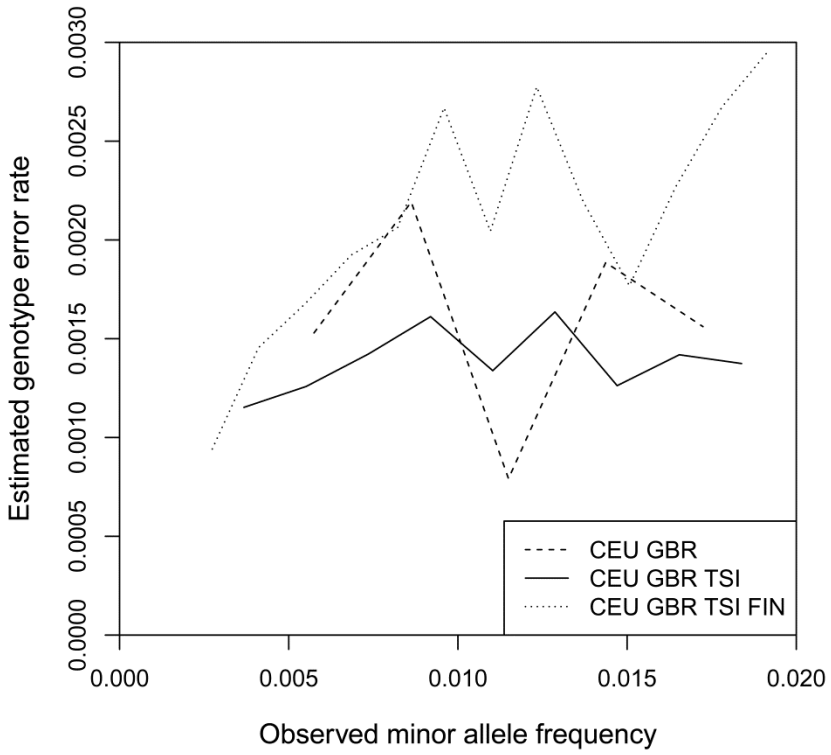

Figure S5: Estimated genotype error rates in the 1000 Genomes European data

We analyzed groups of European populations from the 1000 Genomes Phase 1 data. For each group, the observed minor allele frequency (x-axis value) is the minor allele frequency in that group. Thus, for example, for the "CEU GBR" group the minor allele frequency is the minor allele frequency in the combined CEU and GBR samples, which may differ from the allele frequency in other European populations and from the allele frequency in the 1000 Genomes data as a whole. All autosomal diallelic SNVs with more than one copy of the minor allele in the specified group of populations were used for IBD detection with IBDseq. An analysis error rate of 0.001 was used with IBDseq. IBD segments of size 2 cM or greater were used with SEQERR to estimate the genotype error rates.

Software	Command	Comments
Macs 0.4f	macs 4000 10000000 -seeds <rep> -T -t 1.167 -r 0.8453 -h 1000 -G 21133448 -eG 1.183e-07 5072027 -eG 7.098e-07 1183473 -eN 3.549e-06 0.0001420 -eN 5.915e-05 0.001136	<rep> takes values 1, 2, … 10.
Refined IBD r1106	java -jar b4.r1106.jar gt=simdata.vcf map=simdata.map out=refinedibd_out excludemarkers=simdata.refinedibd_excl window=200000 ibd=true ibdlength=0.2 ibdlod=2.0	simdata.map is a map file converting bp to cM using 1 Mb = 1 cM. simdata.refinedibd_excl is a list of markers with MAF < threshold.
Germline 1.5.1	germline -min_m 0.8 -err_hom 2 -bits 128 -h_extend	
PLINK 1.07	plink --ped simdata.ped --map simdata.map --maf 0.0004 --exclude simdata.plink_excl --cm --genome --out plink_out plink --ped simdata.ped --map simdata.map --maf 0.0004 --exclude simdata.plink_excl --cm --segment --all-pairs --read-genome plink_out.genome --out plink_out --segment-length 50 --segment-snp 50	The first command assesses average relatedness. The second command performs IBD segment detection using average relatedness as a prior.
IBDseq r1117	java -jar ibdseq.r1117.jar gt=simdata.vcf out=ibdseq_out nthreads=12 errormax=0.0025	

Table S1: Software command lines

For coalescent simulation and IBD detection methods, we list the command lines used to generate results in the main text.