Appendix S3: comparative study of isoRelate and hmmIBD

Methods

Summary of files and variables used

The following summarizes results generated from a comparative study of hmmIBD and isoRelate (Henden L, et al. BioRxiv. 2016). Analyses were based on data generated by artificial recombination (details below). The steps, data and scripts required to reproduce this study are as follows.

1. Download the hmmIBD_benchmark repository from https://github.com/artaylor85/hmmIBD_benchmark and unzip the pf3k_data directory.
2. Install hmmIBD following instructions at https://github.com/glipsnort/hmmIBD/releases (v2.0.0).
3. Install isoRelate following instructions at https://github.com/bahlolab/isoRelate/releases (results here based on v0.1.0 installed Aug 9th 2017).
4. Set working directory to this source file location.
5. Run `Simulate_chimeric_genotypes.R`.
6. Run `Run_isolate_hmmIBD.R`.
7. Run `Post_process_results.R`.
8. Run/knit this file.

Simulation of artificially recombined data

We used artificially recombined data to compare results generated under hmmIBD and isoRelate to a known truth that was not generated under either model. Artificially recombined data were based on the MalariaGen Pf3k samples, pilot release 5.0 (https://www.malariagen.net/projects/pf3k). These data were filtered prior to their use in this comparative study, leaving only single nucleotide polymorphisms (SNPs) in the accessible genome (as defined by Manske M, et al. Nature 2012), and those with a high probability of being monogenomic (ad defined by DEploid from Zhu SJ, Almagro-garcia J, Mcvean G. BioRxiv. 2017). The filtered data can be found in `pf3k_data`. Using `Simulate_chimeric_genotypes.R` we:

1. Extracted samples from sites with 100 or more samples (Thies, Kassena, Pursat).
2. For each site, removed multiallelic SNPs (unsupported by isoRelate) and those with minor allele frequency ≤ 0.01, leaving 57307, 41992, 69438 SNPs per sample from Kassena, Pursat, Thies, respectively.
3. Calculated and saved allele frequencies and data sets based on the non recombined data to ensure frequencies were not based on chimeric samples.
4. For each pairwise comparison within a site, calculated the average identity-by-state, IBS (one minus the genome-wide average SNP difference), and plotted.
5. Extracted unrelated pairs (those with IBS $<$ 1 percentile of the empirical IBS distribution).
6. Artificially recombined each unrelated sample pair to create a “chimeric child”. Recombination was simulated by sampling crossover positions (in base pairs) from an exponential distribution with mean equal to the recombination rate (in M/bp) (see `functions.R`).
7. Recorded the parent of each DNA segment in each chimeric child, and plotted the number of crossovers per chromosome averaged over all the chimeric children per site.

Parameter values used to run the HMMs

For each site, IBD segments between 50 “chimeric children” and each of their two parents were inferred under isoRelate and hmmIBD using the parameter values listed in the table below, some of which differ to the defaults provided in order to more closely match the two methods. Timing experiments were done separately on the first 50 samples per site (including non-recombined parents and chimeric children), and repeated 3 times on a MacBook Air laptop with 1.7 GHz Intel Core i7 processor.

Table 1: Specified parameter values. NA denotes not applicable. †In isoRelate, the "recombination rate" is a function of distance in Morgans (M). The equivalent fixed rate in M per base pair (bp) for hmmIBD was thus based on the empirical relationship between positions in bp and centimorgans provided in the png_pedmap data set of the isoRelate package.

Parameter	isoRelate	hmmIBD
genotyping error	0.001	0.001
recombination rate	5.83e-07 M/bp†	5.83e-07 M/bp
minimum no. SNPs per segment	0	NA
minimum length (bp) per segment	0	NA
Minimum marker spacing (bp)	NA	0
Minimum informative sites per genome	NA	0

Results

Table 2: Clocktime (sec) per 50 samples

	isoRelate	hmmIBD
Kassena 1	1710.868	70.789
Pursat 1	1287.921	50.731
Thies 1	2108.324	77.843
Kassena 2	1719.406	71.102
Pursat 2	1285.698	51.404
Thies 2	2108.073	77.129
Kassena 3	1715.745	70.887
Pursat 3	1289.467	52.115
Thies 3	2174.474	78.257

Table 3: CPU time (sec) per 50 samples

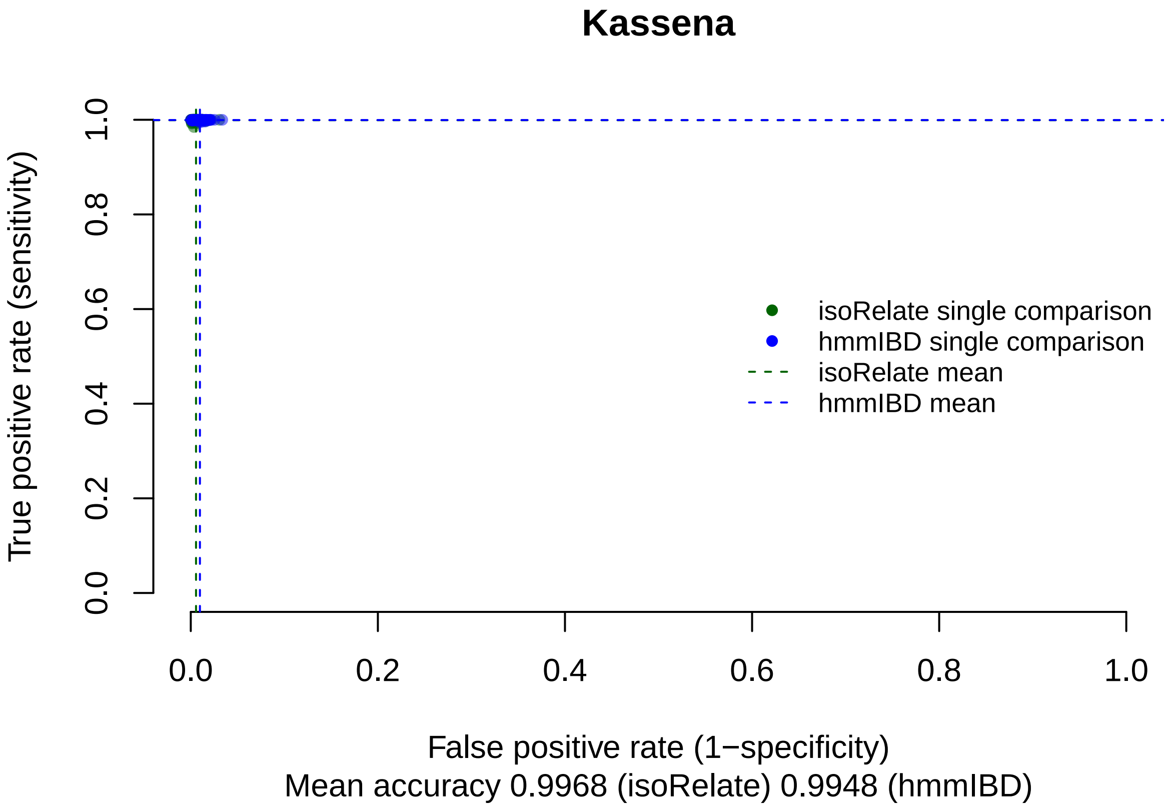
	isoRelate	hmmIBD
Kassena 1	1658.248	70.304
Pursat 1	1243.584	49.188
Thies 1	2034.809	77.163
Kassena 2	1664.926	70.682
Pursat 2	1241.516	50.788
Thies 2	2034.305	76.684
Kassena 3	1662.618	70.462
Pursat 3	1245.173	50.790
Thies 3	2086.543	77.399

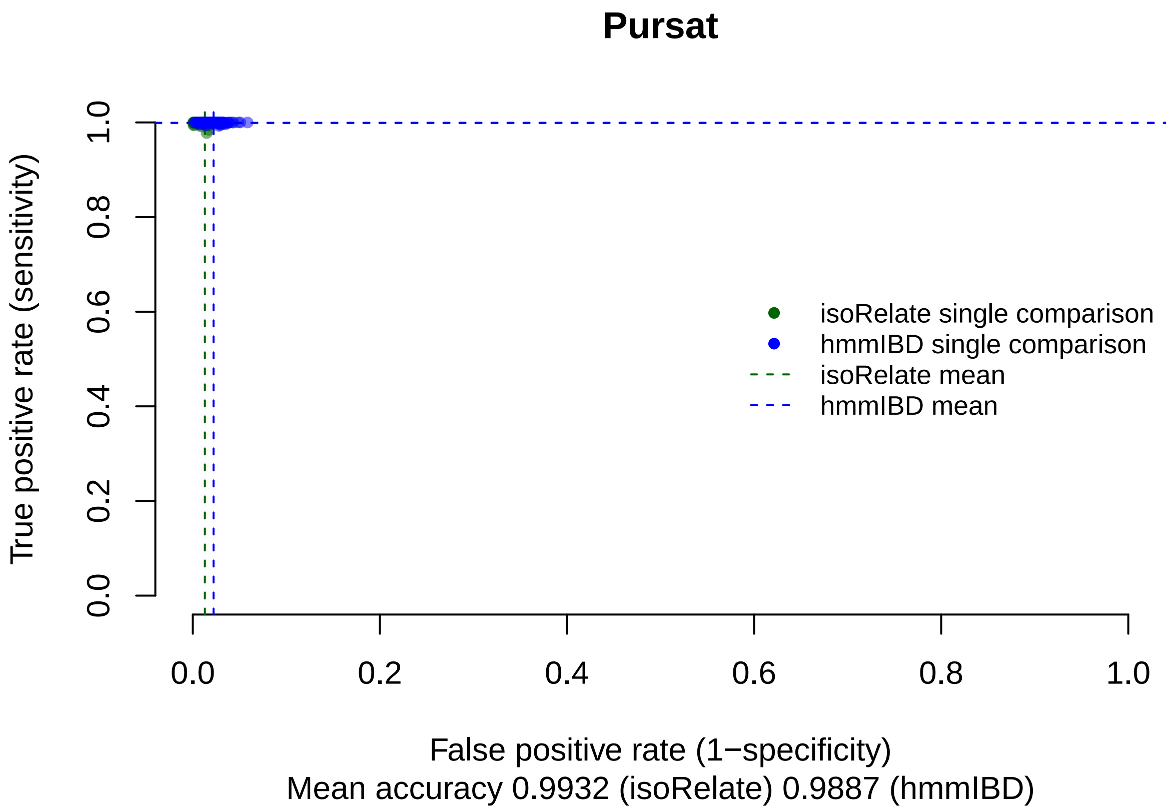
Accuracy, sensitivity and specificity were calculated as follows, where for a given pairwise comparison and SNP, a true positive is an IBD observation given an IBD state, and a true negative is a not IBD (nIBD) observation given a nIBD state,

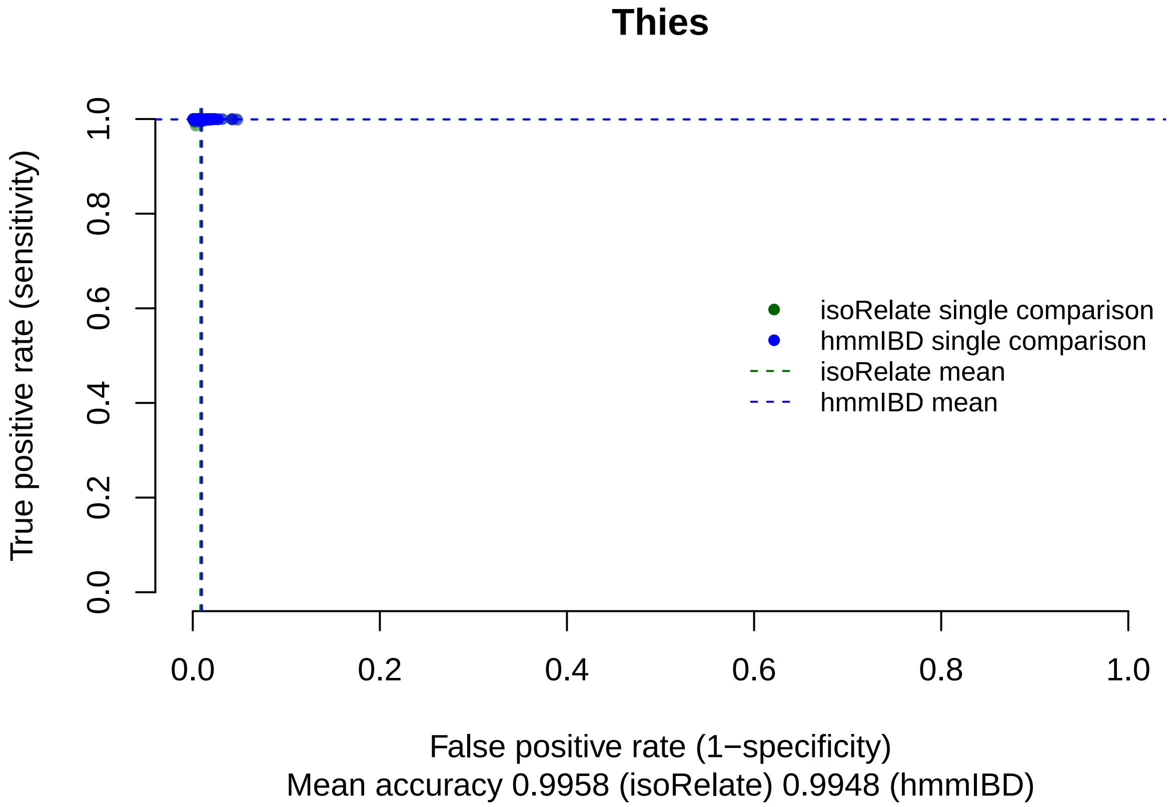
$$\text{Accuracy} = \frac{\sum \text{ True positive} + \sum \text{ True negative}}{\text{Number of SNPs}}, \tag{1}$$

$$\text{Sensitivity} = \frac{\sum \text{ True positive}}{\sum \text{ IBD states}}, \tag{2}$$

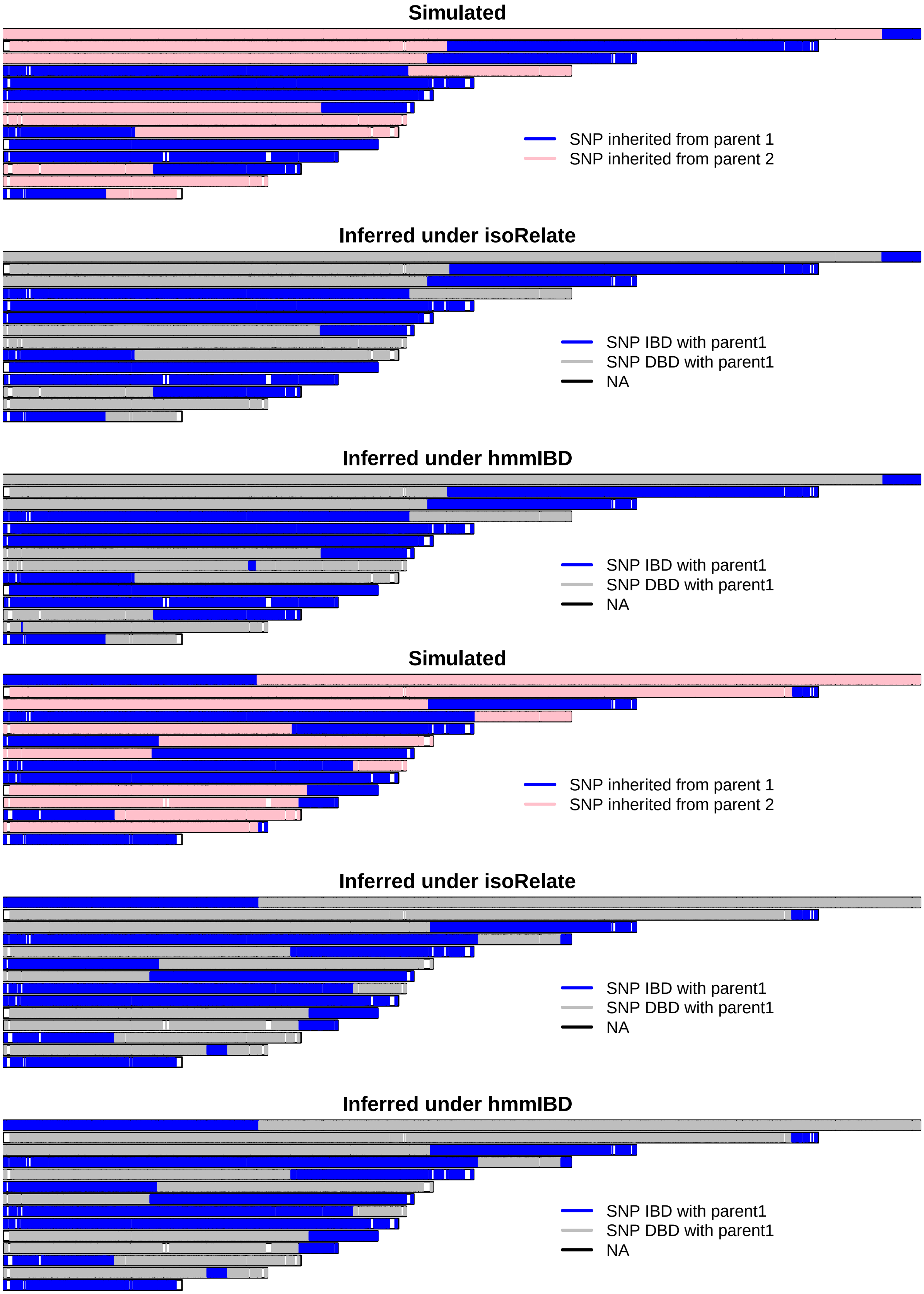
$$\text{Specificity} = \frac{\sum \text{ True negative}}{\sum \text{ nIBD states}}. \tag{3}$$




Pursat
True positive rate (sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
isoRelate single comparison
hmmIBD single comparison
isoRelate mean
hmmIBD mean
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate (1−specificity)
Mean accuracy 0.9932 (isoRelate) 0.9887 (hmmIBD)

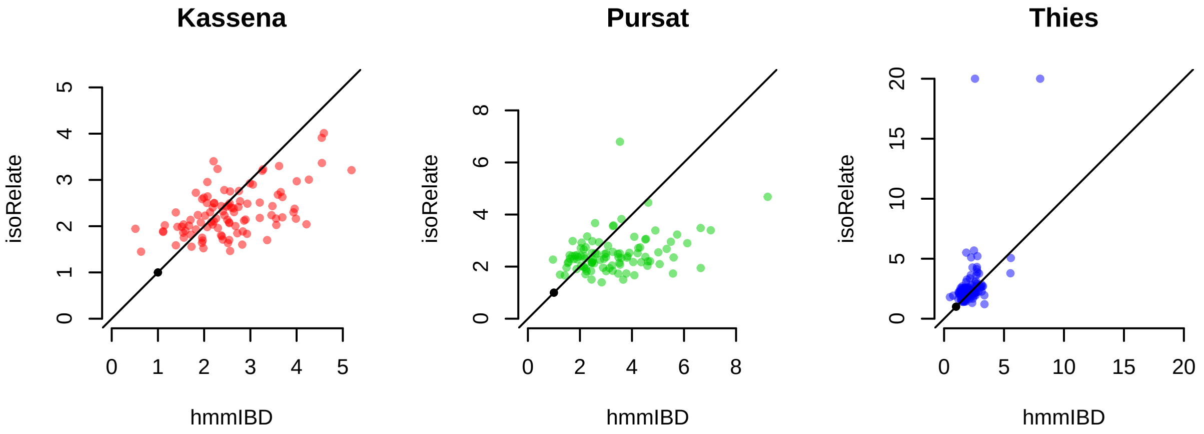

Thies
True positive rate (sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
isoRelate single comparison
hmmIBD single comparison
isoRelate mean
hmmIBD mean
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate (1−specificity)
Mean accuracy 0.9958 (isoRelate) 0.9948 (hmmIBD)

Illustrative assignment plots for two random pairwise comparison from Kassena

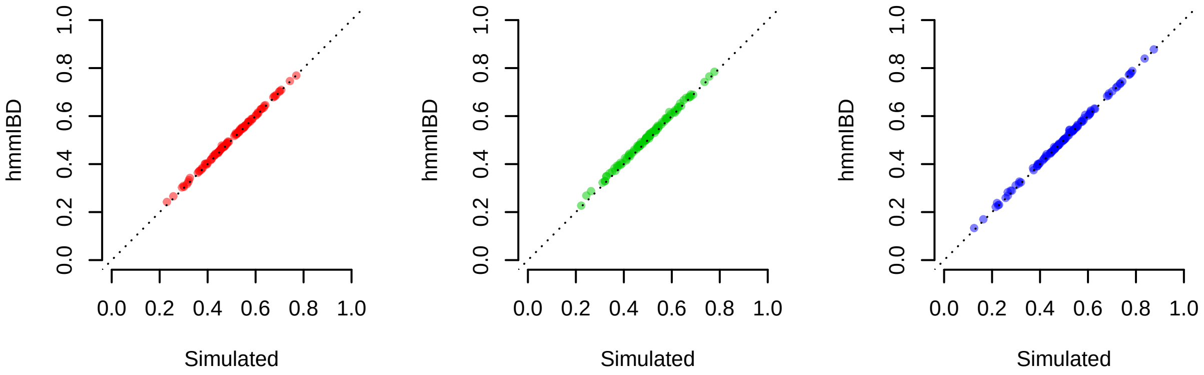


Estimates of numbers of generations

Both methods over estimate the number of generations:



hmmIBD posterior probability of IBD versus proportion simulated IBD



Summary

Both isoRelate and hmmIBD are highly accurate, sensitive and specific. In addition to IBD segments, hmmIBD returns the posterior IBD proportion (a measure of relatedness that integrates over all possible IBD segment assignments). Under this present version of isoRelate, posterior proportion are not readily accessible, but many auxiliary functions for visualizing model output and assessing significance are provided. On average, hmmIBD was 25 times faster in user CPU time than isoRelate.

Table 4: Summary of average scores (standard deviations). Times correspond to 50 samples on a MacBook Air with 1.7 GHz Intel Core i7 processor.

	Clock time (sec)	CPU time (sec)	Accuracy	Sensitivity	Specificity
isoRelate	1711.108 (365.377)	1652.414 (350.474)	0.995 (0.004)	0.999 (0.002)	0.991 (0.008)
hmmIBD	66.695 (11.842)	65.94 (12.116)	0.993 (0.005)	0.999 (0.001)	0.986 (0.011)