Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data using TRUFFLE

Apostolos Dimitromanolakis[1,2], Andrew D. Paterson[3,4,5*], Lei Sun[1,5*]

[1] Department of Statistical Sciences, University of Toronto, Ontario, M5S 3G3, Canada

[2] Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 60 Murray Street, Toronto, Ontario MT5 3L9, Canada

[3] Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, M5G 0A4, Canada

[4] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Ontario, M5T 3M7, Canada

[5] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Ontario, M5T 3M7, Canada

ADP: https://orcid.org/0000-0002-9169-118X

LS: https://orcid.org/0000-0002-5640-937X

[*] Correspondence to ADP or LS: Andrew D. Paterson, Research Institute, Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, 686 Bay Street, Toronto, Ontario, M5G

0A4, Canada. E-mail: andrew.paterson@sickkids.ca; Lei Sun, Department of Statistical Sciences, 100 St George Street, University of Toronto, Toronto, Ontario, M5S 3G3, Canada. E-mail: sun@utstat.toronto.edu.

Abstract

Relationship estimation and segment detection between individuals is an important aspect of disease gene mapping. Existing methods are either tailored for computational efficiency, or require phasing to improve accuracy. We developed TRUFFLE, a method that integrates computational techniques and statistical principles for the identification and visualization of identity-by-descent (IBD) segments using un-phased data. By skipping the haplotype phasing step and, instead, relying on a simpler region-based approach, our method is computationally efficient while maintaining inferential accuracy. In addition, an error model corrects for segment break-ups that occur as a consequence of genotyping errors. TRUFFLE can estimate relatedness for 3.1 million pairs from the 1000 Genomes Project data in a few minutes on a typical laptop computer. Consistent with expectation, we identified three second cousin or closer pairs across different populations, while commonly used methods identified over 15,000 such pairs. Similarly, within populations, we identified much fewer related pairs. Benchmarking to methods relying on phased data, TRUFFLE has a favorable accuracy profile but is drastically faster. We also identified specific local genomic regions that are commonly shared within populations, suggesting selection. When applied to pedigree data, we observed 99.7% accuracy in detecting 1st to 5th degree relationships. As genomic datasets become much larger, TRUFFLE can enable disease gene mapping through implicit shared haplotypes by accurate IBD segment detection.

Keywords

IBD, segment sharing, relationship, un-phased genetic data, implicit haplotype inference, 1000 Genomes Project, software, high-throughput

Introduction

Estimating relatedness and co-ancestry among pairs of individuals is a commonly encountered task in genetic studies. Traditionally, likelihood-based methods (e.g. PREST[1; 2]) or method-of-moments estimators (e.g. PLINK[3]) were used in linkage or association studies, respectively. KING[4] introduced a computationally efficient kinship estimation approach (KING-kinship) that does not explicitly require allele frequency estimation and presumably could be more robustly applied to relationship inference in non-homogeneous population samples. The method is widely used for the inference of *close* relationships in large-scale genetic data, although it can have a higher error rate for distantly related pairs[5]; also see Results below.

The availability of increasing marker densities in studies using genotyping arrays or sequencing technologies makes a different class of methods that perform identical-by-decent (IBD) segment detection more attractive. These methods estimate recent shared ancestry between pairs of individuals by identifying shared chromosomal segments, and they have been implemented in software such as GERMLINE[6], ERSA[7], and BEAGLE Refined and fastIBD[8; 9]. However, although these methods can provide more refined estimates of shared ancestry, identify long-distance relationships and assist disease mapping, they typically require orders of magnitude more computational time and need accurate phasing of the input data. The resulting application complexities prevent their broader use in large-scale genetic studies.

Methods for IBD segment detection in un-phased data have been proposed, including IBDSeq[10], Parente2[11], and the recent shared segment method implemented in the KING software (KING-segment)[12]. Among those, IBDSeq and Parente2 are not fast enough for application to large studies and do not provide genome-wide IBD estimation in a single step. The accuracy profile of the KING-segment method has yet to be reported.

We developed TRUFFLE, a practical method that enables faster and yet accurate identification of IBD1 and IBD2 segments shared between individuals, calculation of averaged

IBD sharing probabilities across the genome (or kinship coefficients), and visualization of shared segments using un-phased genetic data. By skipping the haplotype phasing step and, instead, relying on a simpler region-based approach, the proposed method is less computationally intensive and much easier to apply. In addition, a built-in error model corrects for segment break-ups that can occur as a consequence of genotyping errors. Finally, an integrated variant filtering allows direct application of TRUFFLE to raw variant calls from VCF files that are typically provided in sequencing studies, without the need for external linkage disequilibrium (LD) pruning of markers.

Materials and Methods

Without loss of generality, let us consider a single chromosome on which we identify IBD1 and IBD2 segments for a pair of individuals (a, b) based on available un-phased bi-allelic single nucleotide polymorphisms (SNPs) data. As in convention, the markers are arranged by physical position, and the genotypes of individual a at marker j are $G^a(j) = \{0,1,2\}$, where a ranges from 1 to n, and j ranges from 1 to m. For every pair of individuals (a, b), we define

$$IBS_2^{a,b}(j) := 1 \text{ if } G^a(j) = G^b(j), \text{ and } = 0 \text{ otherwise,}$$

$$IBS_{12}^{a,b}(j) := 1 \text{ if } \mathrm{abs}(G^a(j) - G^b(j)) < 2, \text{ and } = 0 \text{ otherwise.}$$

$IBS_2^{a,b}(j)$ tracks if the genotypes at marker j are identical between the two individuals, i.e. two alleles shared identity-by-state (IBS), IBS2, while $IBS_{12}^{a,b}(j)$ tracks IBS *at least* one. If either of the genotypes at SNP j is missing, then both values are defined to be 1 to prevent segment break-up and keep the same set of markers for all pairs analyzed. It is assumed that long continuous stretches of missing data do not exist in datasets that underwent standard quality control.

For each marker j in the chromosome we also define $p.IBS2(j)$ and $p.IBS12(j)$ to be the expected probability of having, respectively, two alleles and at least one allele shared IBS between a pair of unrelated individuals. These probabilities depend on the minor allele frequency (MAF) of the bi-allelic marker, maf_j, such that $p.IBS2(j) = (maf_j)^4 + (1-maf_j)^4 + (2maf_j(1-maf_j))^2$ and $p.IBS12(j) = 1 - 2(maf_j)^2(1-maf_j)^2$. We can then define p_2 and p_{12} as the averaged, respectively, $p.IBS2(j)$ and $p.IBS12(j)$ across all markers. The values of p_2 and p_{12} would in turn depend on the distribution of the MAFs for the panel of bi-allelic markers used.

Base model with no genotype error consideration

As a basic model we consider scanning a pair of individuals for long stretches of IBS2 or IBS12. A stretch of multiple consecutive IBS markers are likely to be IBD if it (1) is unlikely to occur by chance, and (2) extends beyond the LD of the region that is being considered.

For criterion (1), in a typical whole-genome dataset (either by genotyping or sequencing, e.g. the 1000 Genomes Project[13]) and considering only common variants (MAF > 5%, defined globally), the average $p.IBS2(j)$ over approximately 1,000 markers from a randomly selected region ranges from 0.46 to 0.51, with a mean of $p_2 = 0.48$ over the whole genome (Supplementary Data). A consecutive stretch of k independent markers that are all IBS2 has a probability of occurring by chance of approximately 0.48^k. Thus, when ignoring LD between markers we would set a length threshold $l2$ for declaring IBD2 segments, such that $p_2^{l2} < 10^{-8}$. The detection of IBD1 segments through long stretches of IBS sharing is harder, as the probability of at least one allele shared IBS (i.e. IBS12) by chance is substantially higher than IBS2 alone. For example, the average $p.IBS12(j)$ in a randomly selected region ranges from 0.92 to 0.94, with a mean of $p_{12}=0.93$ (Supplementary Data). To establish a low probability of a false positive for IBD1, as before, we would set the minimum length $l1$ (typically substantially greater than $l2$) such that $p_{12}^{l1} < 10^{-8}$ for independent markers.

For criterion (2), ideally, using a model that takes into account local LD structure can guide a selection of minimum required segment length for a particular region. However, LD-based hidden Markov models (HMMs) pose a serious computational burden and are typically thousands of times slower than non-LD models[5]. The need for an accurate and high-resolution genetic map also limits their applicability to individuals of mostly European descent. To reduce the effects of LD without incurring a significant computational time penalty, we consider a basic pruning approach such that markers closer than a specific number of base pairs are removed. We performed sensitivity analysis of the minimum length parameters, *l1* and *l2*, to identify the cutoff values for robust estimation of overall IBD1 and IBD2 sharing (Supplementary Data – Section 1). As a result, segments shorter than 5Mb for IBD1 or 2Mb for IBD2 are not considered. Filtering of segments by genetic distance in centiMorgan can also be done, with a set of post-processing scripts. Although these default parameter values can be revised by the user, we have found that in practice (e.g. Results below) they work well for datasets with a variety of ancestral compositions, variant densities, and sequencing or array-based platforms.

Model with genotyping error

A common problem in segment detection is the presence of genotyping error, which breaks apart segments and can easily cause false negatives in segment detection. In addition, *de novo* mutation events can generate spurious errors that will further increase the rate of segment break-ups. For example, with an error rate of 0.5%, two individuals on average will have at least 25 markers with genotyping error in a segment of 5000 markers. Previous analysis[10] showed that methods accounting for genotyping error like IBDseq[10] have better power than methods that do not, like Refined IBD[14] (Figure 2 of [10]).

The error model implemented in TRUFFLE was developed to cope with error rates up to 1%, which might be the case in low-depth sequencing data or imputed genotypes depending on

the reference sample size and MAF, among others. In its essence, the proposed approach is a finite deterministic state space model with an unbounded number of states (figure 1).

For the case of identifying IBD1 segments for a pair of individuals (a, b), the genome is scanned sequentially from marker 1 to m. A set of four states is kept at each marker position: $S_j = \{s_1, s_2, s_3, e\}$, $j = 1,\ldots,m$, with the initial values being $S_0 = \{0,0,0,0\}$. Intuitively, these four states correspond to the lengths of the last three IBS12 segments (s_1, s_2, s_3) that were found and an error load (e) of the currently considered segment. A description of the algorithm for identifying IBD1 segments is shown in figure 1. The algorithm for IBD2 segment detection is identical, but with different values for the five tuning parameters: A, B, r_0, r_1 and r_2.

The parameter setting of $A = 1$, $B = 0$ and $r_0, r_1, r_2 = \infty$ corresponds to a no genotyping error model. In contrast, given a non-zero B, approximately B/A errors are allowed for a shared segment before it is considered broken, while short segments are joined together if at least one of them is long enough. In practice, this model is computationally efficient as no complex mathematical operations are required.

The default values for parameters A, B, r_0 , r_1, and r_2 are different for detecting IBD1 and IBD2, and they were optimized by simulations. Briefly, for IBD1 we simulated regions containing 20,000 independent variants, with an average probability of IBS12 of 0.93 to 0.95 (representative of typical genotyping datasets, for example see supplementary Figure S5). Artificial IBD shared segments of 500 markers were added by copying one or two haplotypes of the region from one individual to another; this segment size corresponds to a proportion of IBD1 of 0.125% in a 400k-marker panel. An exhaustive search for the best parameter values was then performed, selecting the ones that maximized the detection power at a false positive error rate of 0.001. For the case of IBD2, similar datasets were simulated, with the markers having an average probability of IBS2 of 0.5, and artificial IBD2 segments of sizes 1 to 5Mb were added.

Segment visualization

A significant benefit of IBD segment detections algorithms is that they can provide the exact locations of IBD1 and IBD2 segments, across the genome. TRUFFLE aids the visualization of such segments by including a set of scripts to create interactive images showing the chromosomes and shared segments; see Results below.

Implementation

TRUFFLE was implemented in C++. It is readily applicable to genome-wide datasets with high marker densities, even on typical laptop computers. Support for parallel execution in multi-core computers and variant filtering (e.g. MAF, missing rate, and minimal distance between markers in base pairs) is integrated. The input file is a multi-sample VCF file [15], which can be phased or un-phased and contains all or some autosomal chromosomes. If necessary, users can define parameter values for the minimum segment detection length and the reporting threshold for related pairs. TRUFFLE is available free for non-commercial use (see Web Resources).

Results

1000 Genomes Project data

We applied TRUFFLE to the 20130502 release of the 1000 Genomes phase III data [13]. The dataset consists of variant calls for 2,504 individuals from 26 populations (five major ethnic groups: 661 Africans, 347 Admixed Americans, 504 East Asians, 489 South Asians, and 503 Europeans). The total number of variants is approximately 88 million before any filtering is applied. These variants were derived from a combination of low and high coverage whole

genome sequencing data, high coverage exome sequence, and genome-wide association study (GWAS) array data from two platforms[13].

For our analyses and method evaluations, two subsets of the bi-allelic markers were generated. The first dataset (A) mirrors what is typically used for relatedness estimation by selecting bi-allelic markers with global MAF > 10% across all populations, and performing LD pruning using PLINK v1.90b3.44 [3]; the indep-pairwise procedure with parameter values 2000, 200, and 0.1 for the number of markers in window, shift, and r^2 criteria. A total of 63,126 markers remained in dataset (A). The second dataset (B) was derived by selecting markers with MAF > 5% and with a spacing of at least 5kb between two consecutive markers, resulting in 469,470 markers remaining. Dataset (B) was generated to evaluate the performance of TRUFFLE, when the computationally expensive step of LD-pruning is avoided. Unlike dataset (A), dataset (B) can be internally generated by TRUFFLE in a single step from a multi-sample VCF file, streamlining the cryptic relatedness analysis for whole-genome sequencing studies. In addition, due to the higher marker density dataset (B) allows for detection of shorter shared segments.

As a comparison, the two datasets were also analyzed using the two different approaches implemented in KING version 2.1.2, i.e. KING-kinship [4], and the more recent KING-segment.

Despite over 3 million pairs of individuals to be examined, the running time of TRUFFLE was 1.6 minutes, using 8 cores of a 2Ghz Xeon CPU processor for dataset (A). The running time of the KING-kinship procedure was only 9.7 seconds but came at the cost of robustness; see below. TRUFFLE running time for dataset (B) was 9.7 minutes because of the increased number of markers, however it does not require the LD-pruning step that involved 115 minutes of computing time using the indep-pairwise procedure in PLINK. KING time for dataset B was 40 seconds.

For relationship estimation, KING identified a significant number of distant relationship pairs and appeared to be quite sensitive to the density of markers, in contrast with TRUFFLE

(Table 1, Supplementary Figures S3-S4). With a kinship coefficient cutoff of > 0.035 for declaring second cousin or closer relatedness, the KING-kinship method reported 573,326 pairs while the KING-segment method reported 217 pairs using the low-density LD-pruned marker dataset (A). When using the higher-density bp-pruned dataset (B), KING-segment also reported an unusually high number of 46,759 second cousin or closer related pairs, among which 26,961 pairs are across populations (Table 1). In contrast, TRUFFLE estimated 172 pairs with kinship coefficient equal or greater to second cousin using (A) and 221 using (B), among which only two and three pairs are across populations and 165 pairs are overlapping.

For first-degree relative pairs, results of all three methods (KING-kinship or segment, and TRUFFLE) agree: there were four full-sib pairs and eight parent-offspring pairs reported (Table 1); KING-segment using dataset (A) reported only 11 first-degree pairs missing one parent-offspring pair. Looking closer, the estimated IBD2 sharing by the KING-segment method are 0.2087, 0.1497, 0.177 and 0.2109, respectively, for the identified four full-sib pairs using dataset (A), with a mean of 0.187. This is noticeably different from the mean value of 0.255 using dataset (B) also based on KING-segment. In contrast, the mean values based on TRUFFLE are, 0.255 and 0.252, using respectively datasets (A) and (B).

Visualization of the exact locations of detected IBD segments can be generated using the provided TRUFFLE post-processing scripts. Figure 1 illustrates segment locations for two selected related pairs from the 1000 genomes dataset, obtained in dataset (B).

Short shared segment analysis

Genomic sharing among unrelated individuals is common and has been described previously. For example, analyses of HapMap II data revealed patterns of segment sharing among seemingly unrelated individuals[16]. It was estimated that, on average, any two individuals from the same

population share approximately 0.5% of their genome through recent IBD, and 10% to 30% of the pairs share at least one region of their genome IBD.

We performed a scan of all the 2,504 individuals from the 1000 Genomes Project for IBD1 and IBD2 segments using TRUFFLE and dataset (B). To this end, a minimum length of 2,000 markers was used as a cutoff to detect IBD1 segments and 1,000 markers for IBD2 segments. To understand the characteristics of locally shared IBD segments between apparently unrelated individuals, we removed segments between 574 pairs that are closely related (estimated average IBD1+IBD2 > 0.02). Among the remaining pairs, the minimum segment length detected was 5.4Mb for IBD1 and 1.3Mb for IBD2, and the maximum lengths are, respectively, 68.0Mb (for a within PUR population pair: HG00641-HG01162) and 11.9Mb (pair HG02348-HG01967, population PEL). In total, 956,577 IBD1 and 1,062,316 IBD2 segments were identified.

For within-population pairs, 8.5% of them share at least one IBD1 segment, in contrast to 0.06% for between-population pairs. For IBD2 segments, the proportions were 5.2% and 2.3%, respectively, for within- and between-populations. The proportion of individuals sharing at least one IBD1 segment varies widely across the populations.

Greater than 30% of the pairs in the Puerto Ricans from Puerto Rico population (PUR) share at least 0.5% of their genomes IBD 1 (Figure 3, panel (a)) and 62% of pairs share at least one segment of length >10cM (Supplementary figure S9). The sharing is even more extensive for segments 5cM or greater, where more than 82% of pairs share at least one segment. These findings align with previous analyses using Refined IBD of BEAGLE[14], for example[13] showed that Puerto Ricans have one of the lowest effective population sizes.

The Finnish in Finland sample (FIN) also showed extensive segment sharing, to an extent higher than the other three European populations (CEU, TSI and GBR). More than 18% of the pairs in the population share a segment of length >10cM, in contrast to 0.7% for the CEU population.

Distribution of segments in panel (b) of figure 3 also suggests sharing hotspots across the genome, likely due to reduced recombination rates in those locations. Similar to our analysis, other approaches have found and reported such hotspots[17; 18]. A high proportion of the identified segments fall in specific genomic regions across multiple different populations (e.g. CEU and CHS in figure 3); see Figure S8 for all 26 populations. Some of the hotspot regions match centromeres of specific chromosomes, indicative of reduced recombination rates at those regions but perhaps also low SNP density. These patterns are less pronounced in African populations (e.g. GWD and YRI), possibly reflecting their higher genetic diversity[19].

Comparison with genotyping array data

To assess the applicability of TRUFFLE to genotyping array data, we used individuals genotyped on the Illumina Omni2.5 array as part of the 1000 Genomes Project [13]. Quality control has been previously performed[20], and the post-quality control data were downloaded from the TCAG website, consisting of 2,318 individuals and 1,989,184 SNPs.

To mirror the dataset (B) generated from the 1000 Genomes combined sequencing and array data, we applied TRUFFLE to 322,849 bi-allelic markers with MAF > 5% and having minimum distance of 5kb between markers. Among the 2,318 individuals in the array dataset, 1,693 were common with those in dataset (B). Thus, we compared the kinship estimates for all the pairs involving those 1,693 individuals.

The correlation of TRUFFLE kinship estimates, using array or combined sequencing and array data, was very high with a sample correlation of 0.998 for pairs estimated as having kinship coefficient > 0.01 in either of the two datasets. Essentially, the inference of relatives closer than 3nd cousin is identical between array and sequencing data. Among all pairs, the sample correlation was 0.932 (Supplementary data – Figure S8), with a mean difference in kinship estimates of 2.9×10^{-4} (standard error of 8.2×10^{-4}).

Comparison with the BEAGLE Refined IBD method

The Refined IBD procedure in BEAGLE [14] is a hidden HMM approach for detecting IBD segments that accounts for LD structure in phased genotype data. Previously, shared segment analysis using BEAGLE version 4.1 was conducted for the 1000 Genomes phase III data and reported by the 1000 Genomes Project [13]. In their analysis, bi-allelic SNPs with more than 10 copies of the minor allele were used and results were post-processed to delete small gaps between segments (as Refined IBD does not account for genotyping error). We compared the reported results with those of TRUFFLE, as well as with the estimates from the KING-segment method applied to the dataset (A) or (B) as previously. The BEAGLE results were obtained from the 1000 Genomes project ftp site (Supplementary Data). Because the reported segments in BEAGLE did not distinguish between IBD1 and IBD2, we compared the total length of all IBD segments in each individual pair, which is proportional to the estimated *p.IBD1+p.IBD2*.

The agreement between the Refined BEAGLE IBD segment estimation and TRUFFLE is very high with a sample correlation of 0.95 for dataset (A) and 0.973 for dataset (B) (Figure 4 and Figure S7). In contrast, the correlation with KING-segment was 0.948 for (A) and 0.312 for (B). Essentially, TRUFFLE is a compromise between statistical and computational efficiency. KING-kinship is faster, but TRUFFLE provides a better approximation of the shared segment lengths inferred from the accurate BEAGLE Refined IBD procedure, but without the necessities of haplotype phasing that is computationally costly, or a detailed genetic map that may not be available for a population of interest

Estimation accuracy in pedigree data

We analyzed Affymetrix 6.0 array data from 822 genotyped individuals from 173 pedigrees. The data were part of the Genetic Analysis Workshop 20 (GAW20) project, and provided by the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study [21; 22]. The GOLDN study recruited European American pedigrees with at least two siblings from the communities of Minneapolis, MN, and Salt Lake City. The average pedigree size was 17.8 individuals, with an average of 4.75 genotyped individuals per pedigree. The numbers of reported relationship pairs within the pedigrees are shown in supplementary table S10. Individuals from different pedigrees are presumed to be unrelated.

As part of the GAW20 data release, 718,542 autosomal bi-allelic SNPs were available for analysis. We applied TRUFFLE to a reduced variant set of 418,672 markers, selected as having MAF > 6% and minimum distance of 300bp between two consecutive markers. The TRUFFLE analysis of 337,431 pairs required 1 minute on a Core-i7 desktop computer (including both across and within pedigree pairs).

Overall, the TRUFFLE kinship estimates closely matched the reported relationships (figure 5), even using this non-LD pruned variant subset. Overall, 99.6% of the relationships were estimated correctly to within one degree, where the estimated degree of relationship is computed from the estimated kinship $\hat{k}$ as the closest integer to $-\log_2 \hat{k} - 1$ (Supplementary Table S11).

Even though the estimated relationship was in line with the specified one overall, 11 pairs of 3rd degree or closer related individuals appeared to have mis-specified relationship, with a ratio of reported vs. estimated kinship greater than 2 (or ½). In addition, 12 pairs showed strong evidence of inbreeding, having estimated *p.IBD2* > 2%. Among the 684 presumed unrelated within-pedigree pairs, the average estimated kinship was 0.00017. However, among the 334,431 between-pedigree presumably unrelated pairs, 27 showed estimated relationship of 5th degree or

closer. These individuals share 4.4% to 12.0% of their genome as IBD1, with shared segments occurring in multiple locations across their genomes and an average of 7.5 shared segments per pair; they are likely to be true relatives.

Power study

Although method comparison using a real dataset such as the 1000 Genomes Project is useful, the true numbers of related pairs and shared segments are not known. To better understand the statistical properties of TRUFFLE (and KING) to identify shared segments across distant relatives, we pursued simulations using simuPOP v1.1.3 [23], following the simulation design of [24]; the exact simulation scripts are provided in the Supplementary Data.

A single chromosome was simulated, with 38,174 bi-allelic markers having MAF > 5%. The simulation used the HapMap phase III populations TSI and LWK as the initial population composition[3], simulating a heterogeneous dataset of 3,000 individuals with equal numbers from the two populations. Artificial IBD1 segments of varying sizes were then injected into pairs of individuals within each population.

Under the null condition (before adding the artificial IBD segments), the false positive rate of the KING-segment method was found to be 4.6×10^{-4}. Although this appears to be small, there were 2,066 false positives for the 3,000 individuals analyzed because there were about 4.5×10^{6} pairs. Since KING does not allow adjustable sensitivity for segment detection, we set the false positive rate of TRUFFLE at the same level by selecting a detection threshold for IBD1 segments that would give the same number of false positives; we then use this threshold for power study.

For each dataset simulated under the alternative, 100 individuals were randomly selected and 100 artificial IBD1 segments, of lengths ranging from 2Mb to 14Mb, were created by copying these 100 segments into another 100 randomly selected individuals. In addition,

genotype errors based on an error rate of 0.9% were added to the shared segments. In total, 15,000 datasets were simulated.

While both methods accurately detect large segments (power > 80% for > 10Mb), TRUFFLE is much more powerful than KING for smaller segments (Figure 6). For example, TRUFFLE had > 90% power to detect segments of size 5.3Mb compared to < 1% for KING (Figure 6). For segments < 4Mb, power is < 5% for both methods.

Discussion

In applications to population[13] and family-based pedigree data[21], TRUFFLE provides accurate IBD1 and IBD2 estimation within a few minutes of computer time for a complete scan of all pairs in a sample using un-phased genome-wide data. Although it is likely that HMM based models, such as Refined IBD [14], will ultimately have more power in detecting short (1-3Mb) segments, their computational burden and requirement for phased data prohibits their widespread use.

Our power and pedigree studies showed that TRUFFLE has high accuracy in providing pedigree relationship estimation and distinguishing distant cousin pairs sharing > 5 Mb segments (corresponding to a putative 10th degree relative pair). Our applications also demonstrated its applicability to both sequencing and array-based studies. The visualization of the exact locations of detected IBD segments is another useful feature of TRUFFLE.

Although it is easy to apply TRUFFLE to studies with up to 20,000 individuals, further enhancements and speed improvements would be needed to make application to large-scale, population-based genetic studies routine. Hashing and dictionary-based approaches are useful future directions by means of avoiding the all-pairs quadratic number of comparisons. Although such methods have been previously applied to segment detection in phased data[6], application of

such methods to un-phased data is not trivial and would require new algorithmic techniques and inferential methods.

The relatedness from the X-chromosome can be wildly different from the autosomes, as it follows a different inheritance pattern. Because of the lower recombination rate[25], the X-chromosome will require different models for the analysis and discovery of shared segments. The pseudo-autosomal regions of the X-chromosome (PAR1-3) would also require specific handling, which is of future research interest.

Overall, TRUFFLE provides a significant improvement in the applicability of IBD segment detection methods to many types of genetic studies. The combination of ease of use, accurate IBD estimation for both distant and close relationships, and segment location visualization, greatly extend the goal of traditional relationship inference methods. TRUFFLE can enable future disease mapping and population genetics through implicit shared haplotypes by accurate IBD segment detection.

Supplemental Data

Supplemental data include 10 figures, 1 table and accompanying text.

Acknowledgements

The authors have no conflict of interest to declare.

We sincerely thank the investigators of the GOLDN project and the Genetic Analysis Workshop advisory committee for permission to use GAW20 data. This research was funded by the Canadian Institutes of Health Research (CIHR, MOP-310732-G-CEAA-117978) and the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-250053) to LS.

Web Resources

TRUFFLE v1.34: https://adimitromanolakis.github.io/truffle-website/

PLINK v1.90b3.44: https://www.cog-genomics.org/plink2

KING v2.1.2: http://people.virginia.edu/~wc9c/KING/

1000 Genomes VCF: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

1000 Genome array data: http://www.tcag.ca/tools/1000genomes.html

BEAGLE IBD Results:

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/ibd_by_pair/

References

1. Sun, L., and Dimitromanolakis, A. (2014). PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. BMC Proc 8, S23.
2. McPeek, M.S., and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66, 1076-1094.
3. 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., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.
4. 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.
5. 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.
6. 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.

7. 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.
8. Browning, B.L., and Browning, S.R. (2011). A fast, powerful method for detecting identity by descent. Am J Hum Genet 88, 173-182.
9. 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.
10. 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.
11. 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.
12. Chen, W.M., Manichaikul, A., Nguyen, J., Onengut-Gumuscu, S., and Rich, S.S. (2017). Integrated inference that accurately identifies close relatives in > 1 million samples. In Annual Meeting of the American Society of Human Genetics 2017. (Orange County Convention Center, Orlando, Florida.
13. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature 526, 68-74.
14. 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.
15. 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. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
16. 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. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.
17. 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.
18. 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.
19. Campbell, M.C., and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9, 403-433.
20. Roslin, N., Li, W., Paterson, A.D., and Strug, L. (2016). Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes. biorxiv 078600.
21. Aslibekyan, S., Goodarzi, M.O., Frazier-Wood, A.C., Yan, X., Irvin, M.R., Kim, E., Tiwari, H.K., Guo, X., Straka, R.J., Taylor, K.D., et al. (2012). Variants identified in a GWAS meta-analysis for blood lipids are associated with the lipid response to fenofibrate. PLoS One 7, e48663.

22. Higgins, M., Province, M., Heiss, G., Eckfeldt, J., Ellison, R.C., Folsom, A.R., Rao, D.C., Sprafka, J.M., and Williams, R. (1996). NHLBI Family Heart Study: objectives and design. Am J Epidemiol 143, 1219-1228.
23. Peng, B., and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21, 3686-3687.
24. Peng, B., and Amos, C.I. (2010). Forward-time simulation of realistic samples for genome-wide association studies. BMC Bioinformatics 11, 442.
25. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928-933.

Figures and Tables

```
function computeSharedLengthIBD1:
    Lshared ← 0
    {s_1, s_2, s_3, e} ← {0, 0, 0, 0}
    for j in 1 to m
      s_3 ← s_3 + 1
      if IBS12^{a,b}(j) = 1 then
        e ← max{e − 1, 0}
      else
        e ← e + A
      end if
      if e + A > B then
        length ← s_3
        if s_1 > r_0 or s_2 > r_0 then length ← s_3 + s_2 + s_1; end if
        if s_1 > r_1 or s_2 > r_2 then s_1 ← s_1 + s_2 else s_1 ← s_2 ; end if
        if length > l_1 then Lshared ← Lshared + length ; end if
        s_2 ← s_3
        e ← 0
      end if
    end for
    return Lshared
```

Figure 1: The TRUFFLE algorithm for IBD1 detection with error model; for IBD2 replace $IBS_{12}^{a,b}(j)$ with $IBS_{2}^{a,b}(j)$.

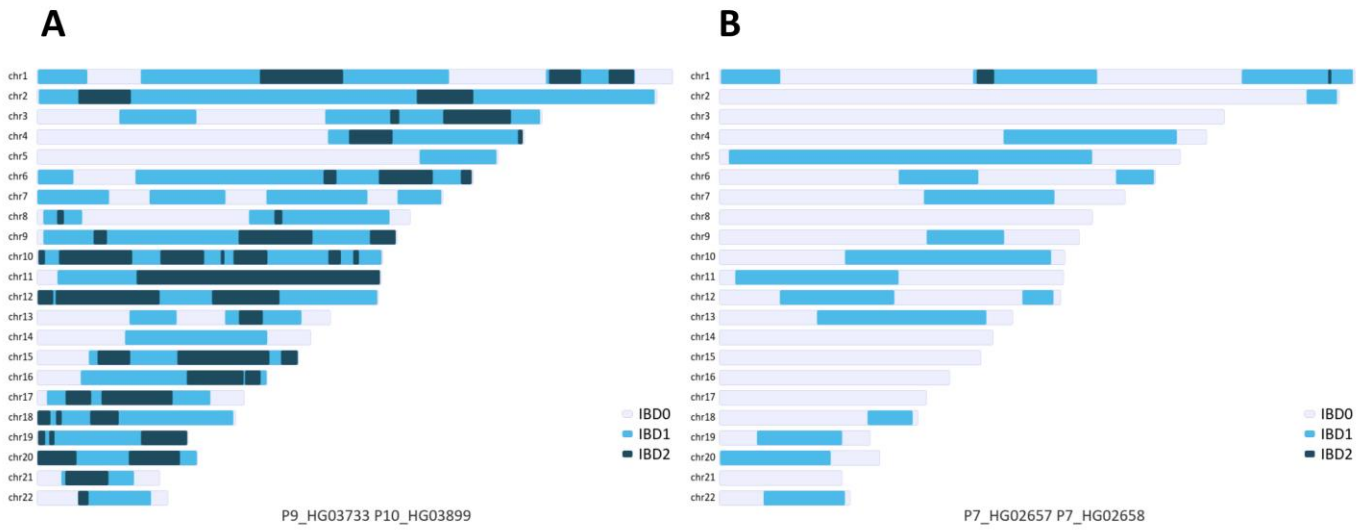

Figure 2: Locations of shared segments identified by TRUFFLE in two pairs from the 1000 Genomes data. (a) A putatively full-sib pair from the STU population showing numerous *IBD1* and *IBD2* shared segments. (b) A putatively more distant related pair from the PJL population with estimated *IBD1* of 32% across the genome and *IBD2* of 0.48%.

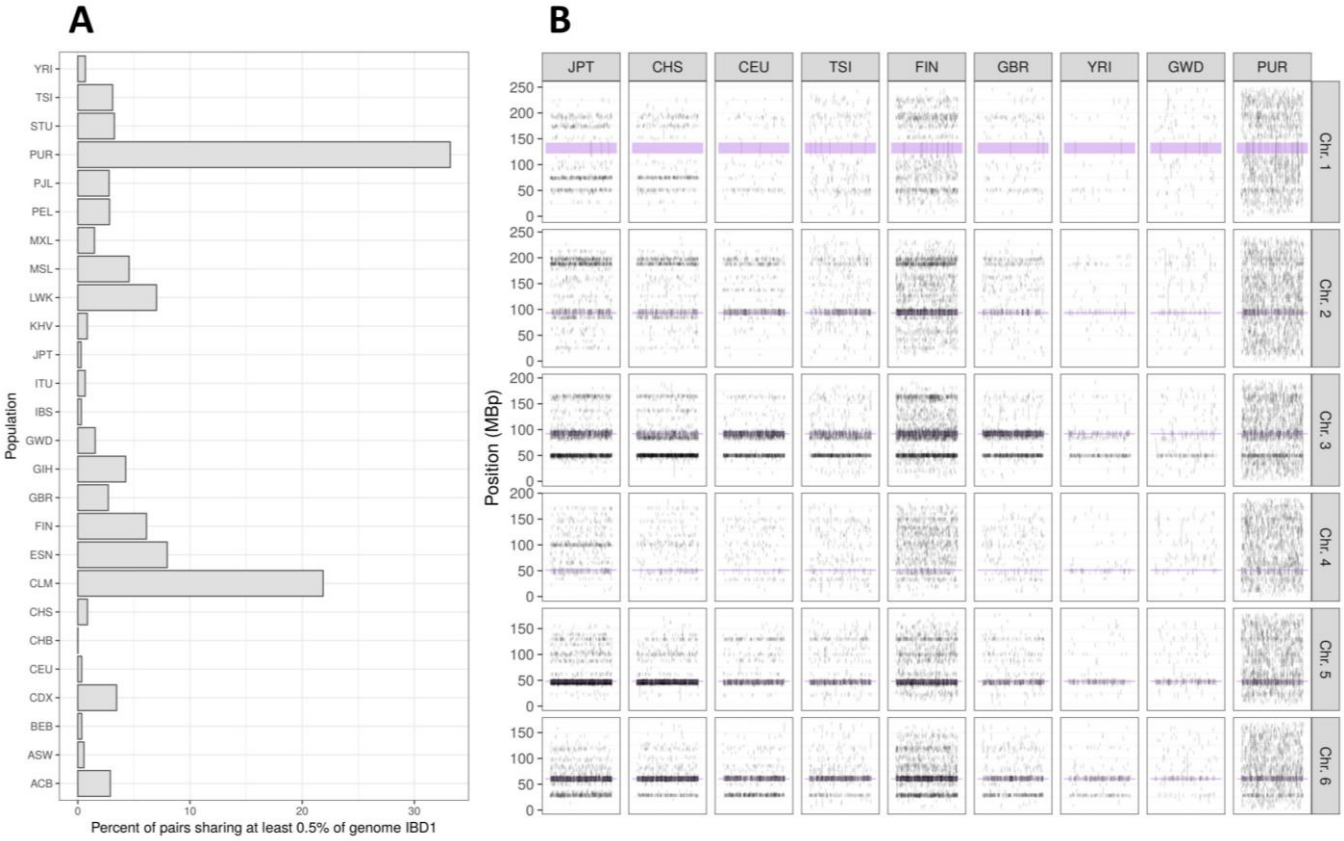

Figure 3: Recent co-ancestry among the 1000 Genomes populations. (a) Proportion of pairs within a population that share at least 0.5% of their genome as IBD1. (b) Distribution of segment locations identified within pairs of the same population for the first six chromosomes and nine selected populations. The segments are positioned randomly on the x-axis to aid visualization and reduce over-plotting. The centromere location is denoted with a purple segment.

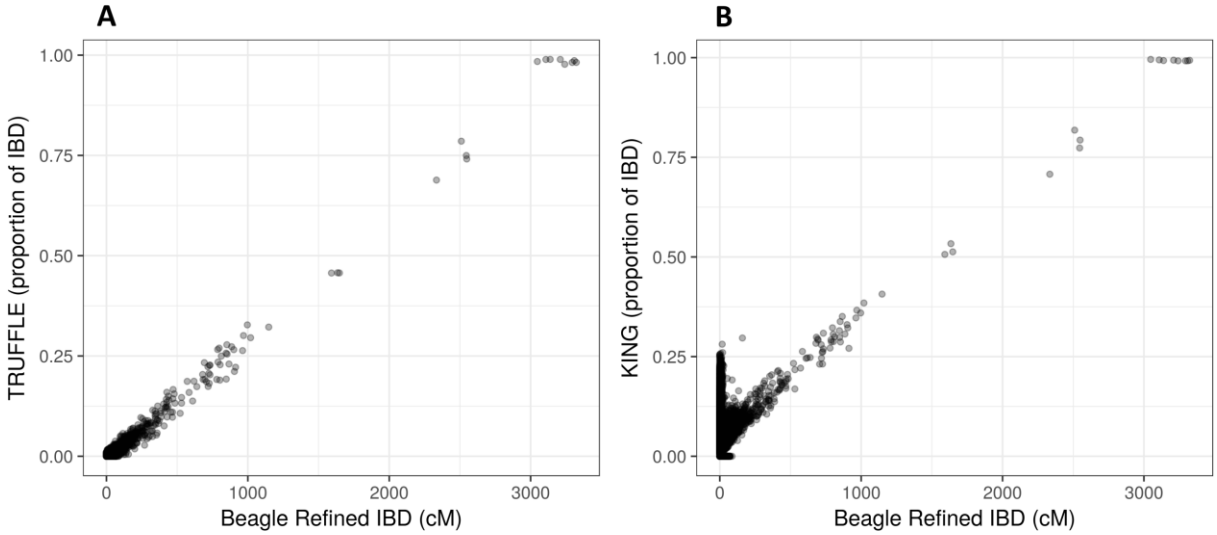

Figure 4: Comparison of shared segment length identified by (a) TRUFFLE vs. BEAGLE, and (b) KING-segment vs. BEAGLE using the 1000 Genomes variant set (B). BEAGLE results were downloaded from the 1000 Genomes project ftp site. Because BEAGLE did not distinguish between IBD1 and IBD2, the Y-axis shows the estimated *p.IBD1+p.IBD2* by TRUFFLE for comparison. We did not convert the cM segment sizes to Mb for BEAGLE in the X-axis, as it would require population specific genetic maps.

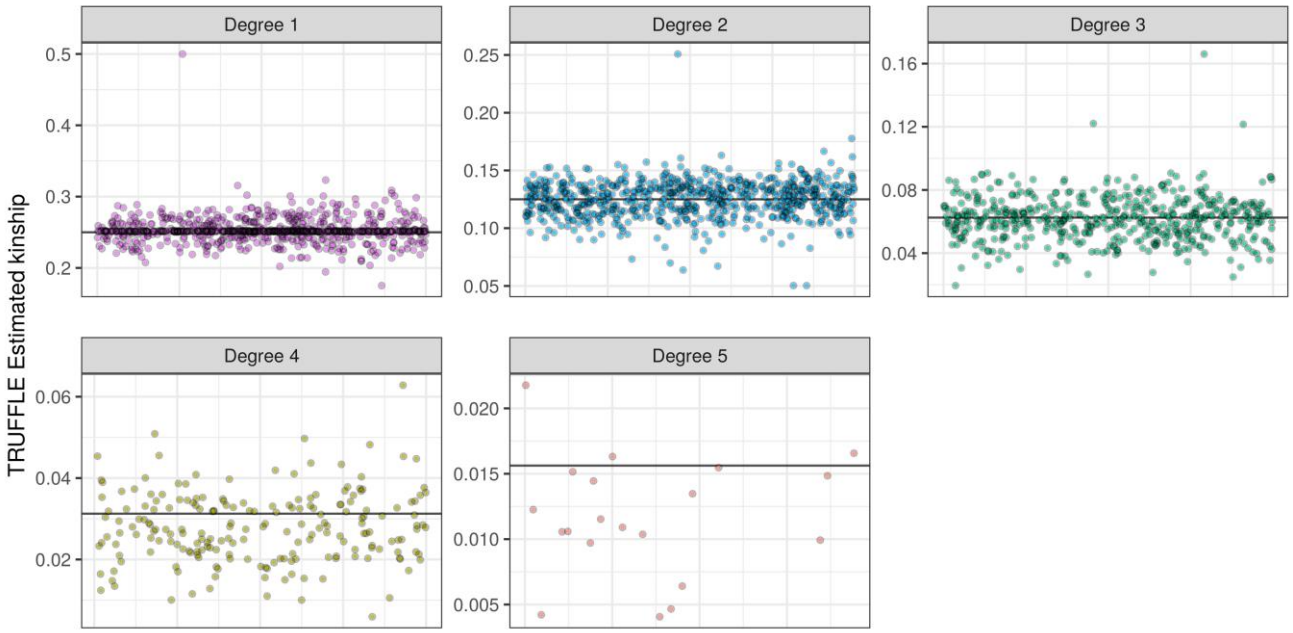

Figure 5: Kinship estimation in the GAW20 data. Only relative pairs, as specified in the study pedigree, are shown. Results are grouped by the degree of relationship based on the given pedigree structure, and the horizontal line shows the expected kinship coefficient for each panel. With each group, pairs are randomly ordered on the X-axis. Horizontal Line shows kinship as specified in the study pedigrees.

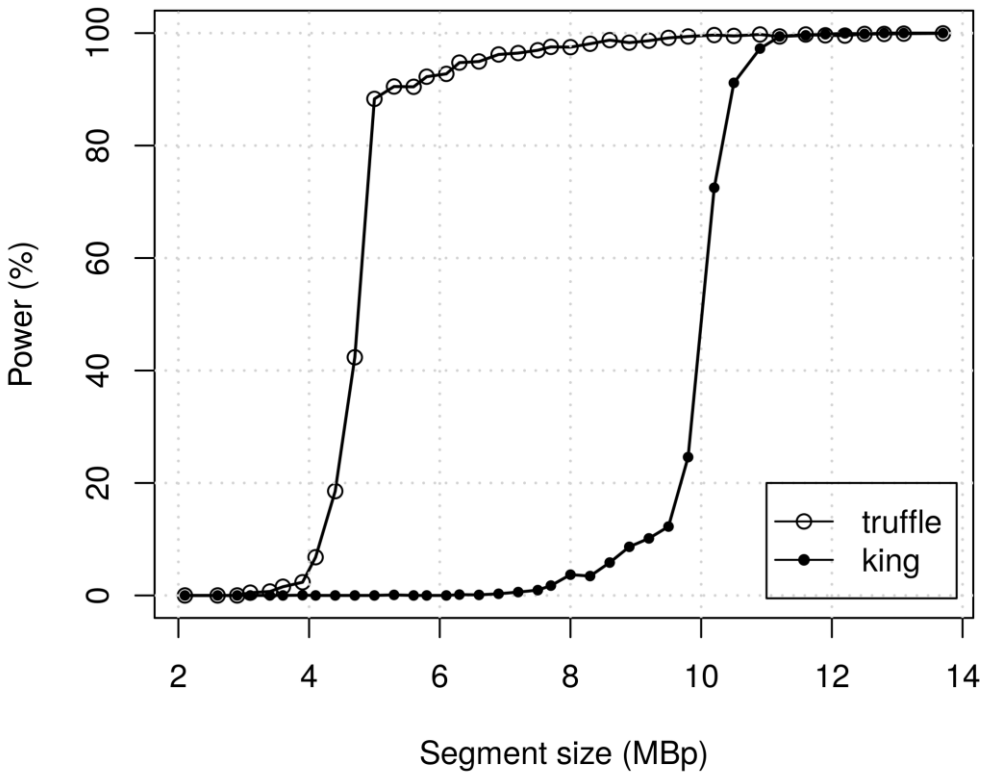

Figure 6: Power of IBD 1 segment detection ($\alpha = 4.6 \cdot 10^{-4}$) of TRUFFLE and KING stratified by segment size. True shared segments (IBD1) of varying sizes were inserted in simulated variant data (38,174 markers) using simuPOP and the simuGWAS scripts.

		Dataset (A)			Dataset (B)		
Relationship	**Cutoff [1]**	**TRUFFLE**	**KING kinship**	**KING segment**	**TRUFFLE**	**KING kinship**	**KING segment**
All pairs							
Full sibling or parent-offspring	0.1875	12	12	11	12	12	12
First cousin or closer	0.035	54	14201	61	61	90	1701
Second cousin or closer	0.0097	172	573326	217	221	34927	46759
Third cousin or closer	0.0024	874	684657	1904	2534	172800	64265
Pairs observed across populations							
Full sibling or parent-offspring	0.1875	1	1	1	1	1	1
First cousin or closer	0.035	2	6524	2	2	8	181
Second cousin or closer	0.0097	2	467149	2	3	16746	26961
Third cousin or closer	0.0024	7	573081	20	20	110274	39614

Table 1: Top: Kinship estimation using TRUFFLE and KING for the two datasets (A) and (B) generated from the 1000 Genomes phase III data[13]. The numbers of pairs that fall under four different kinship cutoffs are shown. **Bottom:** The corresponding numbers of pairs where the two individuals belong to two different populations are also shown. Large numbers of such pairs are more likely to be false positives. [1] Cutoff chosen as the midpoint between the kinship of the relationship in consideration and the kinship of the next more distant relationship considered in this table.