Genetics: Early Online, published on July 24, 2017 as 10.1534/genetics.117.1122


# Benchmarking relatedness inference methods with genome-wide data from thousands of relatives

**Monica D. Ramstetter**[*,1], **Thomas D. Dyer**[†], **Donna M. Lehman**[†], **Joanne E. Curran**[†], **Ravindranath Duggirala**[†], **John Blangero**[†], **Jason G. Mezey**[*,‡] **and Amy L. Williams**[*,1]

[*]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA, [†]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA and Edinburg, TX 78539, USA, [‡]Department of Genetic Medicine, Weill Cornell Medicine, New York, NY 10065, USA

**ABSTRACT** Inferring relatedness from genomic data is an essential component of genetic association studies, population genetics, forensics, and genealogy. While numerous methods exist for inferring relatedness, thorough evaluation of these approaches in real data has been lacking. Here, we report an assessment of 12 state-of-the-art pairwise relatedness inference methods using a dataset with 2,485 individuals contained in several large pedigrees that span up to six generations. We find that all methods have high accuracy ($\sim 92\% - 99\%$) when detecting first and second degree relationships, but their accuracy dwindles to less than 43% for seventh degree relationships. However, most IBD segment-based methods inferred seventh degree relatives correct to within one relatedness degree for more than 76% of relative pairs. Overall, the most accurate methods are ERSA and approaches that compute total IBD sharing using the output from GERMLINE and Refined IBD to infer relatedness. Combining information from the most accurate methods provides little accuracy improvement, indicating that novel approaches—such as new methods that leverage relatedness signals from multiple samples—are needed to achieve a sizeable jump in performance.

**KEYWORDS** relatedness estimation; identical by descent; admixture

The recent explosive growth in sample sizes of genetic studies has led to an increasing proportion of individuals with at least one close relative in a dataset, necessitating relatedness detection. As the number of pairs in a sample grows quadratically in its size, for a constant rate of relatedness among pairs, proportionately more individuals will have close relatives in large datasets. This pervasiveness has relevance to nearly every genetic analysis performed in moderate to large scale data, including trait mapping and population genetics. In particular, inferring relatedness between samples (Weir *et al.* 2006; Thompson 2013; Speed and Balding 2015) is essential to avoid spurious signals in genetic association studies (Marchini *et al.* 2004; Hirschhorn and Daly 2005; Voight and Pritchard 2005); empowers linkage analysis by enabling the correct specification of pedigree structures (O'Connell and Weeks 1998; Ott 1999; Epstein *et al.* 2000); facilitates identification of relatives in the context of forensic genetics (Jobling and Gill 2004; Weir *et al.* 2006; Kayser and de Knijff 2011); and is needed to account for or remove relatives in population genetic analyses (Queller and Goodnight 1989; Hurst 2009; Schraiber and Akey 2015). Relatedness estimation has also drawn the interest of the general public via companies that offer genetic testing services and advertise their ability to find customers' relatives, thus allowing individuals to explore their ancestry and genealogy. The broad utility of relatedness detection has motivated the development of numerous methods for such inference. These methods work by estimating the proportion of the genome shared identical by descent (IBD) between individuals (Weir *et al.* 2006; Speed and Balding 2015) or a closely-related quantity, where an allele in two or more individuals' genomes is said to be IBD if those individuals inherit it from a recent common ancestor (Thompson 2013). Characterizing the true relatedness of two or more samples is challenging for several reasons, including chance sharing of alleles between individuals who are only distantly related, and the fact that the distributions of IBD proportions for different relatedness classes overlap (Hill and Weir 2011; Thompson 2013) (e.g., first cousins and half-first cousins).


doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Monday 17th July, 2017%
[1]Correspondence: mdr232@cornell.edu (M.D.R); alw289@cornell.edu (A.L.W.)

| Degree | Number of Pairs |
|---|---|
| 1 | 4,969 |
| 2 | 6,625 |
| 3 | 8,241 |
| 4 | 7,636 |
| 5 | 3,794 |
| 6 | 816 |
| 7 | 73 |
| Unrelated | 3,051,598 |
| Total | 3,083,752 |

**Table 1** Number of pairs of individuals in the SAMAFS dataset that passed sample filters (Supplemental Note) and are reported to have relatedness between first and seventh degree or as unrelated. We combined reported monozygotic (MZ) twins with the set of first degree relatives.

Motivated by the substantial need to identify relatives in modern samples, we present an evaluation of 12 state-of-the-art pairwise relatedness methods, each capable of scaling to analyze thousands of individuals, including seven that directly infer genome-wide relatedness measures (Chang *et al.* 2015; Manichaikul *et al.* 2010; Thornton *et al.* 2012; Li *et al.* 2014; Moltke and Albrechtsen 2014; Sun and Dimitromanolakis 2014; Conomos *et al.* 2016) and five IBD segment detection methods (Gusev *et al.* 2009; Browning and Browning 2011a, 2013a,b; Durand *et al.* 2014) that we utilized to infer these quantities. To assess these methods, we used SNP array genotypes from Mexican American individuals contained in large pedigrees from the San Antonio Mexican American Family Studies (SAMAFS) (Mitchell *et al.* 1996; Duggirala *et al.* 1999; Hunt *et al.* 2005). Our analysis sample included 2,485 individuals genotyped at 521,184 SNPs (Supplemental Note) within pedigrees that span up to six generations, and with genotype data from as many as five generations of individuals. Given this large sample, including 13 pedigrees with >50 individuals (Supplemental Figure 1), numerous relatives exist, and we used these to evaluate the inference methods. In particular, we analyzed >3,700 pairs of individuals within each of the first through fifth degree relatedness classes, 816 and 73 sixth and seventh degree relatives, respectively, and more than three million pairs of individuals that are reported as unrelated (Table 1). Prior evaluations of relatedness inference methods included only a subset of the methods we evaluate, and either considered simulated data (Manichaikul *et al.* 2010; Thornton *et al.* 2012; Moltke and Albrechtsen 2014; Sun and Dimitromanolakis 2014; Conomos *et al.* 2016) (which may not fully capture the complexities of real data), used small sample sizes (Manichaikul *et al.* 2010; Huff *et al.* 2011; Thornton *et al.* 2012; Conomos *et al.* 2016), or did not consider sixth and seventh degree relatives (Manichaikul *et al.* 2010; Thornton *et al.* 2012; Moltke and Albrechtsen 2014; Conomos *et al.* 2016). This analysis of real data from large numbers of up to sixth degree relatives, as well as dozens of seventh degree relative pairs, provides a comprehensive evaluation of existing pairwise relatedness inference methods.

The performance metric for this study is the rate at which each method infers the pairs of samples to have the same degree of relatedness as that reported in the SAMAFS pedigrees. These reported relationships are generally reliable, and we filtered out relative pairs whose degree of relatedness is potentially inflated due to cryptic relatedness between their ancestors (Supplemental Note). Some programs directly infer the degree of relatedness (Li *et al.* 2014), while others infer a kinship coefficient (Manichaikul *et al.* 2010; Thornton *et al.* 2012; Moltke and Albrechtsen 2014) or a coefficient of relatedness (Chang *et al.* 2015; Conomos *et al.* 2016) (which is two times the kinship coefficient (Wright 1922)), and the remainder instead detect IBD segments (Gusev *et al.* 2009; Browning and Browning 2011a, 2013b,a; Durand *et al.* 2014) (Table 2). To infer the degree of relatedness from an estimated kinship coefficient, we use the mapping recommended in the KING paper (Manichaikul *et al.* 2010) (Supplemental Table 1), which are ranges that use differences in powers of two for the relatedness degree intervals and are generally consistent with simulations (Manichaikul *et al.* 2010).

For IBD detection methods that report the number of IBD segments shared at a locus (Gusev *et al.* 2009; Browning and Browning 2013b)—denoted IBD0, IBD1, and IBD2 for the corresponding number of copies that are IBD—it is straightforward to calculate a kinship coefficient (Thompson 2013). This coefficient, $\phi_{ij}$, between a pair of samples $i, j$ denotes the probability that a randomly selected allele in individual $i$ is IBD with a randomly selected allele from the same genomic position in $j$. Let $k_{ij}^{(0)}$, $k_{ij}^{(1)}$, and $k_{ij}^{(2)}$ denote the proportion of their genomes that individuals $i, j$ share IBD0, IBD1, and IBD2 respectively; then the kinship coefficient is $\phi_{ij} = \frac{k_{ij}^{(1)}}{4} + \frac{k_{ij}^{(2)}}{2}$. The proportions $k_{ij}^{(1)}$ and $k_{ij}^{(2)}$ are simply the sum of the genetic lengths of the IBD1 and IBD2 segments, respectively, between samples $i, j$ divided by the total genetic length of the genome analyzed. For the IBD detection methods (Browning and Browning 2011a, 2013a; Durand *et al.* 2014) that do not distinguish between regions that are IBD1 from IBD2, the proportion of the genome that is inferred to be IBD0 provides an alternate means of estimating the degree of relatedness (Supplemental Table 1), with the ranges of values here again from the KING paper (Manichaikul *et al.* 2010). We classified pairs of individuals with lower kinship coefficients or higher IBD0 rates than indicated for the eighth degree range as unrelated.

The results from the analysis are shown in Figure 1, which depicts the proportion of sample pairs inferred to be within each of the degree classes that we considered (first through eighth degree and unrelated), separated according to their reported relatedness degree. All methods perform well when inferring first and second degree relatives, with accuracies ranging from 98.8% to 99.5% for first degree relatives, and from 92.8% to 98.6% for second degree relatives. However, the methods' accuracies diverge for more distant relatedness, with the IBD segment-based methods generally having higher accuracy than those that rely on allele frequencies of independent markers. For example, for sixth and seventh degree relatives, the top performing IBD segment-based method has 58.1% and 42.5% accuracy, respectively, while the highest performing allele frequency-based method has an accuracy of only 44.6% and 27.4%, respectively. This general pattern applies to fourth and fifth degree relatives as well, although with less discrepancy between these two inference approaches for these closer relatives. The decreased inference accuracy of all methods for higher relatedness de-

| Method | Version | Citation | Type | Output | Parallelized? | Runtime (×cores used if >1) | Requires independent markers | Input required from outside program | Accounts for population structure |
|---|---|---|---|---|---|---|---|---|---|
| **ERSA** | 2.0 | Li *et al.* (2014) | IBD segment-based | Degree of relatedness | N | 14.3h + 96.3h (×16)* | N | IBD segments | NA |
| **fastIBD** | Beagle 3.3.2 | Browning and Browning (2011a) | IBD segment-finding | IBD segments | N | 55.2h | N | NA | NA |
| **GERMLINE** (-haploid) | 1.5.1 | Gusev *et al.* (2009) | IBD segment-finding (Distinguishes IBD1 and IBD2) | IBD segments | N | 19.2m + 96.0h (×16)† | N | Phased genotypes | NA |
| **HaploScore** | NA | Durand *et al.* (2014) | IBD segment-based | IBD segments | N | 2.4h + 96.3h (×16)* | N | IBD segments; phased genotypes | NA |
| **IBDseq** | r1206 | Browning and Browning (2013a) | IBD segment-finding | IBD segments | Y | 33.1h (×16) | N | NA | NA |
| **KING** (KING-robust) | 1.4 | Manichaikul *et al.* (2010) | Allele frequency-based IBD estimate | IBD 0,1,2 proportions | N | 4.6m | Y | NA | Y |
| **PC-Relate** | 2.0.1 | Conomos *et al.* (2016) | Allele frequency-based IBD estimate | IBD 0,1,2 proportions | N | 8.9h + 4.6m‡ | Y | Pairwise kinship coefficients | Y |
| **PLINK 1.9** | 1.90b2k | Chang *et al.* (2015) | Allele frequency-based IBD estimate | IBD 0,1,2 proportions | N | 18.1s | Y | NA | N |
| **PREST-plus** | 4.1 | Sun (2012) | Allele frequency-based; uses linkage model | IBD 0,1,2 proportions | N | 178.9h | N | NA | N |
| **REAP** | 1.2 | Thornton *et al.* (2012) | Allele frequency-based IBD estimate | IBD 0,1,2 proportions | N | 3.8h + 2.8h§ | Y | Ancestral population allele frequencies; sample ancestry proportions | Y |
| **Refined IBD** | Beagle 4.1 | Browning and Browning (2013b) | IBD segment-finding (Distinguishes IBD1 and IBD2) | IBD segments | Y | 96.0h (×16) | N | NA | NA |
| **RelateAdmix** | 0.1 | Moltke and Albrechtsen (2014) | Allele frequency-based IBD estimate | IBD 0,1,2 proportions | Y | 15.8h (×16) + 2.8h§ | Y | Ancestral population allele frequencies; sample ancestry proportions | Y |

**Table 2** Properties of the 12 relationship inference methods we analyzed. Type indicates the inference methodology the program uses. Runtime is wall clock time to run the program with any additional time to run programs needed for input as indicated. We ran parallelized programs using the numbers of cores indicated in parentheses: total compute time for the parallelized programs is the runtime multiplied by the number of cores used. Input required from outside program indicates extraneous information needed to run the program. Programs that use either principal components, sample ancestral population proportions, or that use a model designed for multiple populations are indicated as accounting for population structure. "Y" indicates yes, "N" indicates no, and "NA" indicates not applicable. Runtimes are from a machine with four AMD Opteron 6176 2.30 GHz processors (64 cores total) and 256 GB memory. *Additional time to phase the data using Beagle 4.1 and run GERMLINE. †Additional time to phase the data using Beagle 4.1. ‡Additional time to obtain KING relatedness estimates; base PC-Relate time is the sum of time to run this method and PC-AiR (Conomos *et al.* 2015). §Additional time to obtain ancestral population proportions using ADMIXTURE (Alexander *et al.* 2009).

grees is likely due to the exponential drop in mean pairwise IBD shared and an increased coefficient of variation for more distant relationships (Hill 1993; Visscher 2009; Hill and Weir 2011).

While the accuracies for exact inference of distant relatives are fairly low among all methods, the IBD segment-based methods (excluding fastIBD) are correct to within one degree of the reported relationship at a rate of ≥95.3% for sixth degree relatives and ≥76.7% for seventh degree relatives. At the same time, ERSA, GERMLINE, and Refined IBD classify ≥80.4% pairs of unrelated individuals correctly, and several other methods also correctly infer ∼80% pairs of unrelated individuals, although many of these methods perform poorly when classifying reported relatives. The inference of ∼20% of the more than three million unrelated samples as eighth degree or closer relatives suggests the presence of a non-trivial fraction of unreported relationships in these data. Alternatively, and perhaps more likely, many of these may be false positive relationships, as distinguishing pairs of unrelated individuals from fairly distant relatives is difficult. With the lower bound for eighth degree relatives being a total of 19.5 cM of IBD segments shared between individuals, spurious inferences at this level are possible, with IBD segments detected in regions subject to historical selection (Albrechtsen *et al.* 2010) or with low SNP density potentially leading to inflated IBD proportions. In that regard, we note that some analyses of IBD reweight segments that overlap regions with excess IBD sharing in order to improve the reliability of overall sharing rates (Browning and Browning 2013c; Ball *et al.* 2016). Additionally, analyses that consider relatedness among the parents and/or children of inferred distant relatives have the potential to avoid some of these issues, and indeed, the recently developed relatedness classification method PADRE does analyze familial relatedness signals and shows improved accuracy (Staples *et al.* 2016).

Overall, the most accurate programs for first through seventh degree and unrelated classification are ERSA, GERMLINE, and Refined IBD—all IBD segment-based methods. The improved accuracy of these methods may be due to their focus on identifying long stretches of identical haplotype segments that more readily discriminate recent shared relatedness from chance sharing of alleles. The IBDseq method, while performing well for inferring first through seventh degree relatives, infers a much larger fraction of pairs of individuals as related that are reported as unrelated, suggesting it may be biased towards detecting higher levels of IBD sharing than the other methods.

Noting that the SAMAFS consist of admixed Mexican American individuals, we examined the accuracy results among the allele frequency-based methods, several of which account for population structure. While IBD segment-based methods generally have the best performance and do not directly account for population structure, inferring IBD segments is computationally demanding, and considering the performance of more efficient allele frequency-based methods is of interest. Among all these methods, PC-Relate has the highest accuracy across all levels of relatedness, and it accounts for population structure using principal components (PCs) inferred from a set of samples with low relatedness (Conomos *et al.* 2016). However, PREST-plus has only slightly lower performance than PC-Relate even though it does not account for population structure. PREST-plus implements a hidden Markov model (HMM) that enables it to leverage linkage signals to identify regions that are likely to be IBD between samples (Sun and Dimitromanolakis 2014). Therefore, although PREST-plus does not explicitly detect IBD segments, it leverages similar signals to the IBD segment-based approaches, which might enable it to be less susceptible to biases caused by ignoring the effects of population structure. Relatedness estimation that ignores population structure in admixed samples can produce either a positive or negative bias (Conomos *et al.* 2016). Consistent with this, PLINK infers many sample pairs to be more related than they are reported to be, and, at the same time, infers substantial fractions of fourth through seventh degree pairs as unrelated. KING also dramatically underestimates relatedness, presumably because it assumes that all samples derive from one of several homogeneous populations—a model that is inappropriate for recently admixed samples (Manichaikul *et al.* 2010). We also examined results from the version of KING that assumes a single homogeneous population and its accuracy profile more closely resembles that of PLINK (not shown).

Because the relatedness within SAMAFS has the potential to confound methods that characterize population structure (Conomos *et al.* 2015), we further analyzed the performance of several methods using a dataset consisting of the SAMAFS samples together with a diverse set of HapMap individuals (International HapMap 3 Consortium and others 2010) (Supplemental Note; Supplemental Figure 4). This combined dataset yields inferences of sample ancestry proportions that are strongly correlated with those inferred in a reduced dataset that has only low level relatedness (Supplemental Note). Using this sample, the accuracies of both REAP and RelateAdmix improve significantly, suggesting that either high levels of relatedness or limited ability to discriminate the ancestral populations in the admixed-only SAMAFS data adversely affected the initial inference. Based on this augmented analysis, REAP and RelateAdmix have closer accuracies to that of PC-Relate yet remain somewhat less accurate (Supplemental Note; Supplemental Figure 4). The accuracy of PC-Relate and of KING are quite similar between the two analyses, with the exception that PC-Relate has improved accuracy for seventh degree relatives in the larger sample. Given this improvement and the fact that PC-Relate is the highest performing allele frequency-based method overall, we tested it further by varying its input parameters and the kinship values it uses to detect the set of individuals it uses to infer PCs. All these analyses resulted in similar accuracies except for different rates of inferred seventh degree relatives (Supplemental Note; Supplemental Figure 5); the variation in seventh degree relatedness inference may be due to stochastic factors and the relatively small numbers of these relatives in the dataset.

Besides considerations related to detecting population structure, the presence of many relatives in SAMAFS may lead to biased allele frequency estimates. Furthermore, haplotype phasing and therefore IBD inference accuracy might be greater than would be achieved in a sample composed mostly of unrelated individuals. To ensure the performance results presented here also apply to analyses of non-pedigree datasets, we identified a set of only distantly related individuals using FastIndep (Abraham and Diaz 2014) and merged these samples with pairs of related individuals to form 1,000 datasets (Supplemental Note). Each reduced dataset contains at most one related pair of samples from any distinct SAMAFS pedigree, limiting the potential for bias. When classifying sample pairs included in at least one reduced dataset, PLINK's inference accuracy differs by less than 3% for the first through fifth relatedness degrees compared to the full dataset (Supplemental Figure 2), suggesting that allele frequency biases are small and only minimally impact inference accuracy. In order to test the IBD detection methods, we increased the

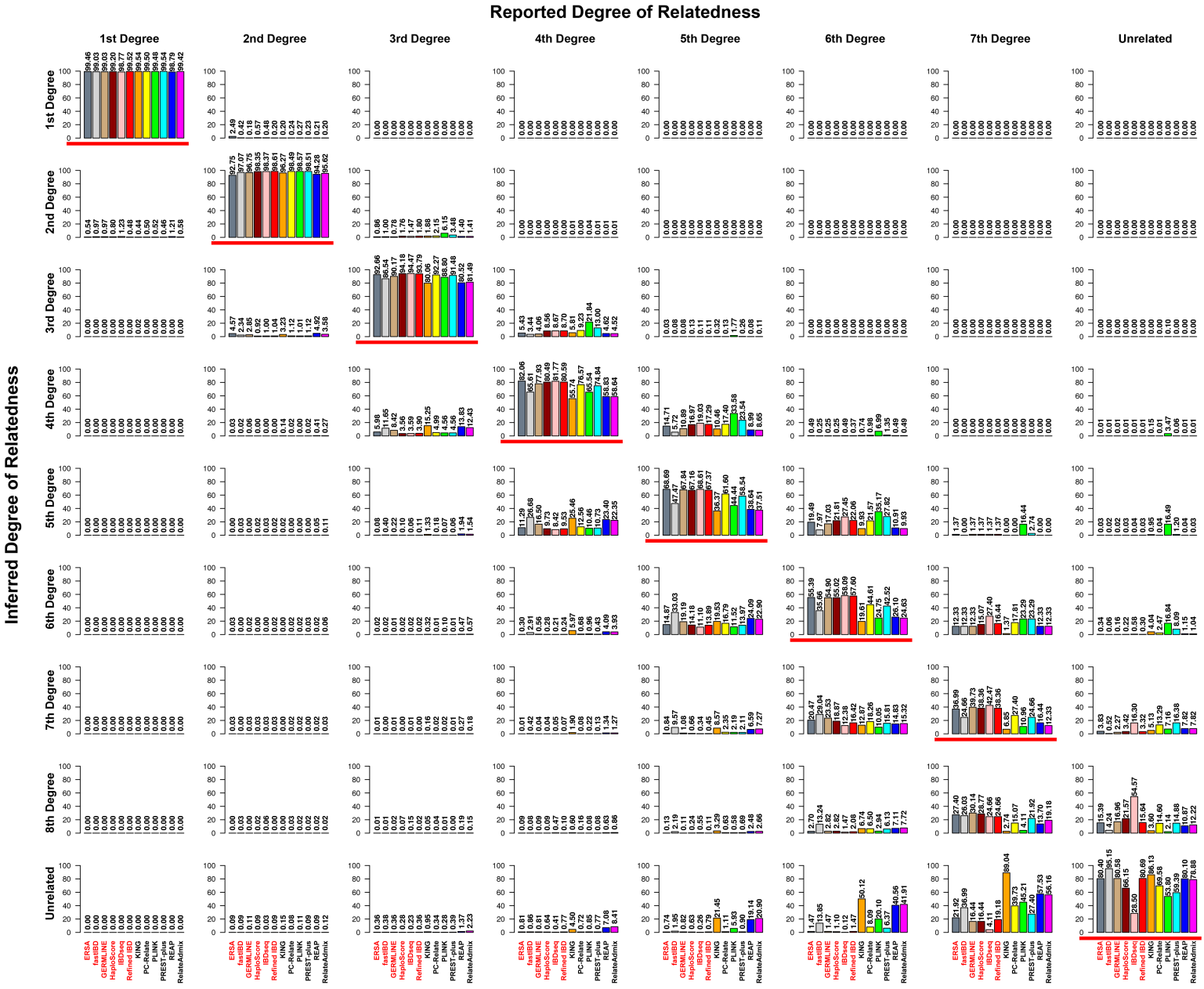


**Figure 1** Performance comparison of the evaluated methods using the SAMAFS dataset. Bar plots indicate the percentage of sample pairs that are reported to have a given degree of relatedness and that are inferred to be related as the indicated degree. The bar plots are separated on the horizontal axis by the reported relatedness degree and on the vertical axis by inferred relatedness degree. For clarity, the plots list above each bar the inferred percentage that the corresponding bar depicts. Program names listed in red are IBD segment-based methods while those in black utilize allele frequencies for inference. Red horizontal bars under a bar plot indicate that the corresponding inferences agree with the reported relationships.

sample size of these reduced datasets by further merging 580 HapMap samples (Supplemental Note). Results from running the IBD segment-based methods on these datasets show a reduction in accuracy that ranges between $0\% - 9.6\%$ for first through fifth degree relatives, indicating that relatedness in SAMAFS may impact the inference accuracy (Supplemental Figure 3). Yet the results are still consistent with those of the larger analysis as the IBD segment-based methods generally have higher performance than allele frequency-based methods. This is true even in the reduced datasets that have no more than 1,204 samples and therefore are subject to a non-trivial rate of phasing error (Browning and Browning 2011b).

In comparison to previous method evaluations, our results show some notable differences. For example, using real data from 30 pedigrees, ERSA reported lower accuracies for first through sixth degree relatives than we observe (Li *et al.* 2014), with differences ranging from 8.9% to nearly 21%. We believe this is attributable to differences in sample size, as the ERSA analysis considered only 304 individuals compared to 2,485 here. This—in addition to the accuracy reductions of IBD segment-based methods in the reduced datasets described above—indicates that sample size can have a dramatic impact on the quality of IBD segment-based methods. Thus smaller studies may wish to use allele frequency-based methods such as PC-Relate or, for non-admixed individuals, KING-robust, which in fact considers data from each sample pair separately rather than estimating allele frequencies from the full data (Manichaikul *et al.* 2010). The authors of PC-Relate (Conomos *et al.* 2016) find that KING and PLINK each tend to both overestimate and underestimate relatedness when analyzing admixed individuals, which is consistent with our results. They also report that PC-Relate generally outperforms REAP and RelateAdmix, matching our findings even after we incorporate additional HapMap individuals to aid detection of population structure (Supplemental Note). To our knowledge, other evaluations of relatedness inference have not included methods that directly detect IBD segments, and our results indicate that these are promising methods to apply in this setting.

As current methods provide only moderate accuracy when classifying third through seventh degree relatives, we evaluated the potential for increasing performance by combining inference results from the top three programs: ERSA, GERMLINE, and Refined IBD. We first used an approach that calls the degree of relatedness for a pair only when all three programs unanimously agree on the relatedness degree, providing no classification for other pairs (3,012 relative pairs and 632,615 reported unrelated pairs are unclassified). In comparison to the most accurate method's performance in each degree class, the inference accuracy using this strategy increases only slightly for related pairs (+0.01%, +0.13%, +2.6%, +1.5%, +3.4%, +2.2%, and +1.1%, respectively, for first through seventh degree), but increases by 9.0% for unrelated pairs of individuals. This indicates a high level of discordance among the inferred relatedness status for a large fraction of pairs that are reported as unrelated. Many of these unrelated pairs must therefore have borderline inferences, and indeed most methods infer a sizeable fraction as only eighth degree relatives (Figure 1). We also considered a majority vote between the three programs, discarding cases in which all three programs inferred a different degree (only five relative pairs had such variable inferences while 110,848 pairs reported as unrelated are so discrepant). With this approach, there is a slight decrease in performance overall (-0.04%, -0.6%, -1.3%, -0.7%, -0.2%, -2.3%, and 0% for first through seventh degree relatives and +1.6% for unrelated samples). These results suggest that while there is room for improvement in the specificity of relatedness inference methods, dramatic improvement is likely to be achieved only with novel approaches and not composites of current methods. Of interest in this regard are recently developed methods that combine information across related individuals in order to infer a pedigree structure and/or improve relatedness accuracy (Staples *et al.* 2014, 2016; Ko and Nielsen 2017). Importantly, each of these methods relies on a pairwise relatedness approach, highlighting the continued relevance of pairwise inference methodologies even as new methods arise for addressing multi-way relatedness inference.

As an application of these findings, we leveraged the high accuracy of IBD segment-based methods to explore pairs of samples inferred to be closely related but reported as unrelated in the SAMAFS dataset. We used the top performing methods, ERSA, GERMLINE, and Refined IBD, to characterize unreported relatives. These three methods all infer a small number of first through third degree relationships that connect individuals from different pedigrees within SAMAFS (Supplemental Figure 7; Supplemental Note). Overall, we found six pairs of pedigrees with at least five sample pairs between them that the methods unanimously infer to have first through third degree relatedness. Additionally, these three methods agree on the inference of 235 and 744 pairs of fourth and fifth degree relatives between the pedigrees (not shown), and suggest instances of reported first and second degree relatives likely to have the reverse relatedness class or to have much lower relatedness (Supplemental Table 3; Supplemental Note). These results highlight the necessity of checking reported or for unreported relatedness among samples in all cohorts and indicate that there can be sizeable numbers of unknown relatives across a range of relatedness degrees even in well-studied samples.

Important factors for determining which analysis method to use in a study are its accuracy and its computational demands, and the runtimes of the methods evaluated here vary over several orders of magnitude (Table 2). PLINK is the fastest program with a runtime of only 18.1 seconds, while the IBD segment-based methods require up to 64 compute days in total (parallelized across 16 cores in our analyses). In general, we observe a trade-off between runtime and accuracy, with the top-performing methods being those that require the largest compute time, and with PLINK being one of the least accurate methods. Given the uniformly high accuracy of all methods for inferring first and second degree relatives, applications that are focused only on identifying close relatives have the option of using an efficient allele frequency-based method such as PLINK or PC-Relate to perform inference, the latter being an accurate program that is more computational intensive than PLINK but much faster than IBD segment-based methods. A further consideration is the ethnic group of the analysis cohort. PLINK and KING have biased results for distant relatives in the admixed SAMAFS data we focus on, but are expected to perform well in homogeneous populations or, for KING, collections of unadmixed samples from multiple homogeneous populations. On the other hand, for applications in which the aims include locating more distant relatives, the use of IBD segment-based methods should produce improved results. Although beyond the scope of this paper, recently developed methods for phasing extremely large samples (Loh *et al.* 2016) should improve upon the computational requirements of several methods (GERMLINE, ERSA,

and HaploScore) and extend their utility to much larger datasets than the one we consider here.

We have presented a detailed comparison of state-of-the-art relatedness inference methods using thousands of pairs of individuals that range from first to seventh degree relatives as well as numerous sample pairs that are reported to be unrelated. All the methods we assessed reliably identify first and second degree relatives (accuracy $\sim$92% $-$ 99%), but their accuracy falls precipitously when classifying third to seventh degree relatives. This is unsurprising given the increased coefficient of variation as well as greater skewness in the proportion of genome shared as the meiotic distance between two relatives increases (Hill and Weir 2011). Despite these challenges, several IBD segment-based methods infer relatedness correct to within one degree of the reported relationship at a rate of $\geq$76.7% for all relationship degrees (Figure 1). Misreported or unknown relationships in the SAMAFS dataset likely explain some of the inference errors, particularly since even some confidently inferred first degree relationships were likely misreported as a more distant relationship or as unrelated (Supplemental Table 3; Supplemental Figure 7). We find that IBD segment-based methods outperform other approaches for more distantly related pairs, though notably these packages require substantially more compute time to run (Table 2). While the precise performance results presented here are specific to the SAMAFS sample, we find that reducing the sample size still produces similar results, with methods that leverage IBD segments generally having greater accuracy than other approaches. Therefore, the results presented here should be generalizable to moderate and large scale studies and indicate overall properties of pairwise relationship inference methodologies: approaches that use IBD segments outperform other methods for third degree and more distant relatives; and the specificity of the inferences, even in a dataset where phase accuracy may be relatively high, are limited for all but the closest relatives.

## Data availability

The SAMAFS sample data are available on dbGaP under accession numbers phs000847 and phs001215. A script to extract pairwise IBD1 and IBD2 proportions from the output of Refined IBD can be found at http://github.com/MonicaRamstetter/bakeoff.

## Acknowledgments

We thank the San Antonio Mexican American Family Study participants that made this analysis possible. We also thank Shai Carmi for helpful comments. This work was supported by a National Science Foundation Graduate Research Fellowship grant number DGE-1144153 to M.D.R.; Qatar National Research Fund grant NPRP 7-1425-3-370 to J.G.M.; an Alfred P. Sloan Research Fellowship, and a seed grant from Nancy and Peter Meinig to A.L.W.

## Literature Cited

Abraham, K. J. and C. Diaz, 2014 Identifying large sets of unrelated individuals and unrelated markers. Source code for biology and medicine **9**: 1.

Albrechtsen, A., I. Moltke, and R. Nielsen, 2010 Natural selection and the distribution of identity-by-descent in the human genome. Genetics **186**: 295–308.

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**: 1655–1664.

Ball, C. A., M. J. Barber, J. Byrnes, P. Carbonetto, K. G. Chahine, *et al.*, 2016 Ancestry DNA matching white paper .

Browning, B. L. and S. R. Browning, 2011a A fast, powerful method for detecting identity by descent. American Journal of Human Genetics **88**: 173–182.

Browning, B. L. and S. R. Browning, 2013a Detecting identity by descent and estimating genotype error rates in sequence data. American Journal of Human Genetics **93**: 840–851.

Browning, B. L. and S. R. Browning, 2013b Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194**: 459–471.

Browning, S. R. and B. L. Browning, 2011b Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**: 703–714.

Browning, S. R. and B. L. Browning, 2013c Identity-by-descent-based heritability analysis in the northern finland birth cohort. Human genetics **132**: 129–138.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4**: 1.

Conomos, M. P., M. B. Miller, and T. A. Thornton, 2015 Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic Epidemiology **39**: 276–293.

Conomos, M. P., A. P. Reiner, B. S. Weir, and T. A. Thornton, 2016 Model-free estimation of recent genetic relatedness. American Journal of Human Genetics **98**: 127–148.

Duggirala, R., J. Blangero, L. Almasy, T. D. Dyer, K. L. Williams, *et al.*, 1999 Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. American Journal of Human Genetics **64**: 1127–1140.

Durand, E. Y., N. Eriksson, and C. Y. McLean, 2014 Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. Molecular biology and evolution p. msu151.

Epstein, M. P., W. L. Duren, and M. Boehnke, 2000 Improved inference of relationship for pairs of individuals. American Journal of Human Genetics **67**: 1219–1231.

Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, *et al.*, 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Research **19**: 318–326.

Hill, W. and B. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research **93**: 47–64.

Hill, W. G., 1993 Variation in genetic identity within kinships. Heredity **71**: 652–653.

Hirschhorn, J. N. and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6**: 95–108.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins, *et al.*, 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Research **21**: 768–774.

Hunt, K. J., D. M. Lehman, R. Arya, S. Fowler, R. J. Leach, *et al.*, 2005 Genome-wide linkage analyses of type 2 diabetes in Mexican Americans. Diabetes **54**: 2655–2662.

Hurst, L. D., 2009 Genetics and the understanding of selection. Nature Reviews Genetics **10**: 83–93.

International HapMap 3 Consortium and others, 2010 Integrat-

ing common and rare genetic variation in diverse human populations. Nature **467**: 52–58.
Jobling, M. A. and P. Gill, 2004 Encoded evidence: DNA in forensic analysis. Nature Reviews Genetics **5**: 739–751.
Kayser, M. and P. de Knijff, 2011 Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics **12**: 179–192.
Ko, A. and R. Nielsen, 2017 Composite likelihood method for inferring local pedigrees. bioRxiv p. 106492.
Li, H., G. Glusman, H. Hu, *et al.*, 2014 Relationship estimation from whole-genome sequence data. PLOS Genetics **10**.
Loh, P.-R., P. F. Palamara, and A. L. Price, 2016 Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics .
Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics **26**: 2867–2873.
Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly, 2004 The effects of human population structure on large genetic association studies. Nature Genetics **36**: 512–517.
Mitchell, B. D., C. M. Kammerer, J. Blangero, M. C. Mahaney, D. L. Rainwater, *et al.*, 1996 Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. Circulation **94**: 2159–2170.
Moltke, I. and A. Albrechtsen, 2014 RelateAdmix: a software tool for estimating relatedness between admixed individuals. Bioinformatics **30**: 1027–1028.
O'Connell, J. R. and D. E. Weeks, 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. American Journal of Human Genetics **63**: 259–266.
Ott, J., 1999 *Analysis of human genetic linkage*. JHU Press.
Queller, D. C. and K. F. Goodnight, 1989 Estimating relatedness using genetic markers. Evolution pp. 258–275.
Schraiber, J. G. and J. M. Akey, 2015 Methods and models for unravelling human evolutionary history. Nature Reviews Genetics **16**: 727–740.
Speed, D. and D. J. Balding, 2015 Relatedness in the post-genomic era: is it still useful? Nature Reviews Genetics **16**: 33–44.
Staples, J., D. Qiao, M. H. Cho, E. K. Silverman, D. A. Nickerson, *et al.*, 2014 PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. American Journal of Human Genetics **95**: 553–564.
Staples, J., D. J. Witherspoon, L. B. Jorde, D. A. Nickerson, J. E. Below, *et al.*, 2016 PADRE: Pedigree-aware distant-relationship estimation. The American Journal of Human Genetics **99**: 154–162.
Sun, L., 2012 Detecting pedigree relationship errors. Statistical Human Genetics: Methods and Protocols pp. 25–46.
Sun, L. and A. Dimitromanolakis, 2014 PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. BMC Proceedings **8**: S23.
Thompson, E. A., 2013 Identity by descent: variation in meiosis, across genomes, and in populations. Genetics **194**: 301–326.
Thornton, T., H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, *et al.*, 2012 Estimating kinship in admixed populations. American Journal of Human Genetics **91**: 122–138.
Visscher, P. M., 2009 Whole genome approaches to quantitative genetics. Genetica **136**: 351–358.
Voight, B. F. and J. K. Pritchard, 2005 Confounding from cryptic relatedness in case-control association studies. PLOS Genetics **1**: e32.
Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics **7**: 771–780.
Wright, S., 1922 Coefficients of inbreeding and relationship. The American Naturalist **56**: 330–338.