bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A performance assessment of relatedness inference methods using genome-wide data from thousands of relatives

Monica D. Ramstetter[1,*], Thomas D. Dyer[2], Donna M. Lehman[2], Joanne E. Curran[2], Ravindranath Duggirala[2], John Blangero[2], Jason G. Mezey[1,3], and Amy L. Williams[1,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
[2]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA and Edinburg, TX 78539, USA
[3]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 methods in real data has been lacking. Here, we report an assessment of 11 state-of-the-art 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$93% − 99%) when reporting first and second degree relationships, but their accuracy dwindles to less than 60% for fifth degree relationships. However, the inferred relationships were correct to within one relatedness degree at a rate of 83% − 99% across all methods and considered relationship degrees. Furthermore, most methods infer unrelated individuals correctly at a rate of $\sim$99%, suggesting a low rate of false positives. Overall, the most accurate methods were ERSA 2.0 and approaches that classify relationships using the IBD segments inferred by Refined IBD and IBDseq. Combining results from the most accurate methods provides little accuracy improvement, indicating that novel approaches for relatedness inference may be needed to achieve a sizeable jump in performance.

The recent explosive growth in sample sizes of genetic datasets has led to an increasing proportion of close relatives hidden within these large studies, necessitating relatedness detection. Inferring relatedness between samples[1–3] is an essential step in performing genetic association studies[4–6] and linkage analysis[7–9], is a powerful tool for forensic genetics[1,10,11], and is needed to account for or remove relatives in population genetic analyses[12–14]. Relatedness estimation has also drawn the interest of the general public via companies such as 23andMe and AncestryDNA which advertise their ability to find and report relatives, allowing individuals to explore their ancestry and genealogy. The broad utility of relatedness estimation 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[1,3] 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[2]. As previously shown, the distributions of IBD proportions for different relatedness classes (such as first cousins and half-first cousins) are expected to overlap[2,15], posing a challenge for these inference procedures.

Here, we present a rigorous evaluation of 11 state-of-the-art methods that can scale to large study sizes, including seven that directly infer genome-wide relatedness measures[16–22] and four IBD segment detection methods[23–26] that we utilized to infer these quantities. To assess each of these methods, we used SNP array genotypes from Mexican American individuals contained in large pedigrees from the San Antonio Mexican American Family Studies (SAMAFS)[27–29]. Our analysis sample included 2,485 individuals genotyped at 521,184 SNPs (Supplemental Note) within pedigrees that span up to six generations with genotype data

*Correspondence: mdr232@cornell.edu (M.D.R.), alw289@cornell.edu (A.L.W)

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Degree	Number of Pairs
1	4,969
2	6,590
3	8,244
4	7,950
5	4,501
Unrelated	3,025,035
Total	3,057,289

Table 1: Numbers of pairs of individuals from the SAMAFS dataset reported to have relatedness between first and fifth degree and counts of unrelated pairs used for the evaluation. Only individuals from distinct pedigrees are considered unrelated.

from as many as five generations of individuals. Given this large sample, including 13 pedigrees with >50 individuals (Supplemental Figure 1), numerous close relatives exist, and we used these to evaluate each of the inference methods. In particular, there are >4,500 pairs of individuals within each of the first through fifth degree relatedness classes that we evaluated, and we further considered more than three million pairs of individuals that are in distinct pedigrees and hence assumed unrelated (Table 1). Prior analyses of relatedness inference methods considered either simulated data[17,18,20–22]—which may not fully capture the complexities of real data—or used small sample sizes[17,18,22,30]. Our analysis using real data for large numbers of up to fifth degree relatives provides a comprehensive evaluation of these relatedness inference methods.

Our analysis considered each method's ability to correctly infer the degree of relatedness between the pairs of samples based on their reported relationships. These reported relationships are extremely reliable and in most cases we can validate them via first degree connections among samples in the densely-genotyped SAMAFS pedigrees. Some methods directly infer the degree of relatedness[19] while others infer a kinship coefficient[17,18,20], a coefficient of relatedness[16,22] (which is two times the kinship coefficient[31]), or instead detect IBD segments[23–26] (Table 2). To infer the degree of relatedness from an estimated kinship coefficient for a pair of samples, we use the ranges of estimated kinship values from the KING method[17] (Table 3). These ranges use differences in powers of two for the relatedness degree intervals, which is generally consistent with simulations[3]. For IBD detection methods that report the number of IBD segments shared at a locus[23,26]—denoted IBD0, IBD1, and IBD2 for the corresponding number of copies that are IBD—it is straightforward to calculate a kinship coefficient[2]. This coefficient, ϕ_{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 $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ denote the proportion of their genomes that individuals i, j share IBD1 and IBD2 respectively; then the kinship coefficient is $\phi_{ij} = \frac{p_{ij}^{(1)}}{4} + \frac{p_{ij}^{(2)}}{2}$. The proportions $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ 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. (Note if $i = j$, then $\phi_{ii} = \frac{1}{2}(1 + f_i)$ where f_i is the kinship coefficient between the parents of i which is equivalent to the inbreeding coefficient of individual i.) For the IBD detection methods that do not distinguish between regions that are IBD1 from IBD2[24,25], the proportion of the genome that is inferred to be IBD0 provides an alternate means of estimating the degree of relatedness (Table 3), with the ranges of values here again from the KING paper[17]. We classified individuals with lower kinship coefficients or higher IBD0 rates than indicated for the fifth degree range as unrelated.

Using the SAMAFS sample, we assessed the performance of each program by using them to classify all pairs of individuals. Figure 1 shows the proportion of sample pairs inferred to be within each of the degree classes that we considered (first through fifth degree and unrelated), with results separated according to the reported and inferred relatedness degrees of the pairs. All methods perform well when inferring first and second degree relatives, with the accuracy ranging from 98.4% to 99.5% for first degree relatives, and from 93% to 98.6% for second degree relatives. For more distant relatedness, the IBD-based methods have

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Method	Version	Citation Number	Type	Output	Parallelized?	Runtime (×cores used if >1)	Requires independent markers	Input required from outside program	Accounts for population structure
ERSA	2.0	19	IBD segment-based	Degree of relatedness	N	14.5h	N	IBD segments	NA
fastIBD	Beagle 3.3.2	24	IBD segment-finding	IBD segments	N	55.5h	N	NA	NA
GERMLINE (-haploid)	1.5.1	23	IBD segment-finding (Distinguishes IBD1 and IBD2)	IBD segments	N	20m	N	Phased genotypes	NA
IBDseq	r1206	25	IBD segment-finding	IBD segments	Y	33.5h (×16)	N	NA	NA
KING (KING-robust)	1.4	17	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	5m	Y	NA	Y
PC-Relate	2.0.1	22	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	9h	Y	Pairwise kinship coefficients	Y
PLINK 1.9	1.90b2k	16	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	20s	Y	NA	N
PREST-plus	4.1	21	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	179h	N	NA	N
REAP	1.2	18	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	4h	Y	Ancestral population proportions	Y
Refined IBD	Beagle 4.1	26	IBD segment-finding (Distinguishes IBD1 and IBD2)	IBD segments	Y	91h (×16)	N	NA	NA
RelateAdmix	0.1	20	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	Y	16h (×16)	Y	Ancestral population proportions	Y

Table 2: Properties of the 11 relationship inference methods we analyzed. Type indicates the inference methodology the program uses. Runtime is wall clock time to run the program; 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 or ancestral population proportions 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.

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

			Expected				Accepted range for:	
Relationship	Degree	# Meiosis	IBD0	IBD1	IBD2	ϕ	ϕ	P(IBD=0)
Parent-child	1	1	0	1	0	$\frac{1}{2^2}$	$(\frac{1}{2^{3/2}}, \frac{1}{2^{1/2}}]$	< 0.1
Full siblings (not MZ twin)	1	2 (×2)	1/4	1/2	1/4	$\frac{1}{2^2}$	$(\frac{1}{2^{3/2}}, \frac{1}{2^{1/2}}]$	$[0.1, 0.365)$
Grandparent	2	2	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Avuncular	2	3	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Double-cousins	2	4 (×2)	9/16	3/8	1/16	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Half sibling	2	2	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
First Cousin	3	4	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Double half-cousins	3	5 (×2)	23/32	7/32	1/64	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Great-grandparent	3	3	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Half-/grand-avuncular	3	4	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
First cousin once removed	4	5	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
Great-great-grandparent	4	4	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
Half-grand-/great-grand-avuncular	4	5	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
First cousin twice removed	5	6	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$
Second cousin	5	6	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$
GGG-grandparent	5	5	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$

Table 3: For a range of relationship types, the corresponding degree of relatedness of the individuals; the number of meioses that separate them, with (×2) indicating samples that are related along two lines of descent (such as full siblings) that have the listed meiotic distance on both lines; proportions of the genome that are expected to be IBD0, IBD1, and IBD2 between the samples; and expected kinship coefficient ϕ. For inferring a degree of relatedness from a kinship coefficient, the range of values that map to the given degree are listed. Likewise for inference using IBD0, the proportions of IBD0 values that map to each degree are shown. The list does not include all possible relationship types for the degrees of relatedness listed.

higher accuracy than those that rely on allele frequencies of independent markers—for example, for fifth degree relatives, the top performing IBD-based method has 59.4% accuracy while the highest performing allele frequency-based method has only 53.8% accuracy. Overall, the most accurate programs are ERSA 2.0, Refined IBD, and IBDseq. The improved accuracy of IBD-based methods may be due to their focus on identifying long stretches of identical segments that more readily discriminate recent shared relatedness from chance sharing of alleles.

Noting that the SAMAFS consist of admixed Mexican American individuals, we examined the accuracy results among the allele frequency-based methods, of which several account for population structure. Of all these methods, PC-Relate has the highest accuracy across all levels of relatedness, and it does account for population structure using principal components. Overall, the results are mixed with regards to accounting for population structure and accuracy, with PC-Relate, REAP, RelateAdmix, and KING all incorporating population structure into their models, and PREST-plus and PLINK ignoring this structure. Because relatedness structure can confound methods that detect population structure, we employed a procedure designed to locate true ancestral population proportions for the input supplied to REAP and RelateAdmix (Supplemental Note). PC-Relate, by contrast, addresses these concerns by performing population structure analysis internally using a set of samples with low levels of relatedness. However, IBD detection methods do not directly account for population structure and generally have the best performance.

The inference accuracy of all methods decreases for higher relatedness degrees, likely due to the exponential drop in mean pairwise IBD shared and an increased coefficient of variation as relatedness decreases[15,32,33]. In particular, for fifth degree relatives, the accuracy rates for all methods are very low at less than 60%. However, in nearly all cases ($\geq$ 83.8%), the programs correctly inferred the degree of relatedness to within

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

one degree of that reported in the SAMAFS pedigrees. IBDseq has the highest within-one-degree accuracy for reported fourth degree pairs (the relationship class with the lowest accuracies for off-by-one inference) at 98.7%. At the same time, the methods classify an average of 97.9% of pairs of unrelated individuals correctly, averaged across all programs (99.7% when PLINK is excluded), with few instances of fifth or greater degree of relatedness inferred for these pairs. These results suggest that, when methods do detect relatedness—even as far distant as fifth degree—the individuals are likely to be truly related.

Because the SAMAFS data consist of many closely related individuals, the allele frequencies derived from it have the potential to be biased. Furthermore, haplotype phasing and therefore IBD inference accuracy might be greater than would be achieved in a more outbred sample. To ensure the performance results presented here also apply to analyses of non-pedigree datasets, we identified a set of unrelated individuals using FastIndep[34] and merged these samples with pairs of related individuals to form 1,000 datasets that include different pairs of relatives (Supplemental Note). Each reduced dataset contains at most one pair of samples from any distinct SAMAFS pedigree, limiting the potential for bias. When classifying the related individuals included in at least one of these reduced datasets, PLINK's inference accuracy differs by less than 3% 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 further merged 580 HapMap samples[35] with each of the reduced datasets (Supplemental Note). Results from running IBD detection methods on these datasets show a reduction in accuracy that ranges between $0\% - 8\%$, yet the results are still consistent with those of the larger analysis (Supplemental Figure 3). Specifically, the IBD segment-finding methods tend to have higher performance than allele frequency-based methods, supporting the conclusion that IBD segment-based methods provide the highest accuracy. This is true even in the reduced datasets that have no more than 1,204 samples and therefore are subject to a relatively high level of phasing errors.

We examined the pairs of samples that were inferred to be related but were reported as unrelated (in distinct pedigrees) in the SAMAFS dataset. ERSA 2.0, Refined IBD, and IBDseq all inferred a small number of first through third degree relationships that connect individuals from different pedigrees within SAMAFS (Figure 2). Overall, we found 48 pairs of pedigrees with at least five pairs of relatives between them which all three methods unanimously infer to have the same degree of relatedness. Additionally, these three methods agreed on the inference of 374 and 1,632 pairs of fourth and fifth degree relatives between the pedigrees (not shown). These results highlight the importance of checking for relatedness among samples in all cohorts, and indicate that there can be sizable numbers of relatives across a range of degrees even in well-studied samples.

As current methods provide only moderate accuracy when classifying third through fifth degree relatives, we evaluated the potential for increasing performance by combining inference results from the top three programs. We 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. The resulting inference accuracy increased only negligibly (0.15%, 0.22%, 1.6%, 3.1%, 1.8%, and 0.01%, respectively for first through fifth degree and unrelated pairs) in comparison to the most accurate method's performance in each degree class. We also considered a majority vote between the three programs, discarding the cases in which all three programs inferred a different degree (only two cases were of this class). With this approach, there is a slight decrease in performance overall (-0.46%, -0.26%, -1.4%, -1.5%, +0.28%, +0.01%). 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.

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 fifth degree relatives as well as numerous individuals that are reported to be unrelated. All the methods we assessed reliably identify first and second degree relatives as well as unrelated pairs (accuracy $\sim 93\% - 99\%$), but their accuracy falls precipitously when classifying third to fifth 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. Despite these challenges, the inferred relationship was within one degree of the reported relationship at a rate of $83\% - 99\%$ for all programs and relationship degrees (Figure 1). Misreported or unknown relationships in

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (Supplemental Table 4) or as unrelated (Figure 2). We find that IBD-based methods outperform other approaches for more distantly-related pairs, though notably these packages require substantially more compute time to run which may limit their utility in some applications (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 having greater accuracy than other approaches. Therefore, the results presented here should be generalizable and indicate overall properties of relationship inference methodologies: approaches that use IBD segments outperform other methods for third degree and more distant relatives; and the specificity of relatedness inference, even in a dataset where phase accuracy may be relatively high, is inhibited for all but the closest relatives.

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

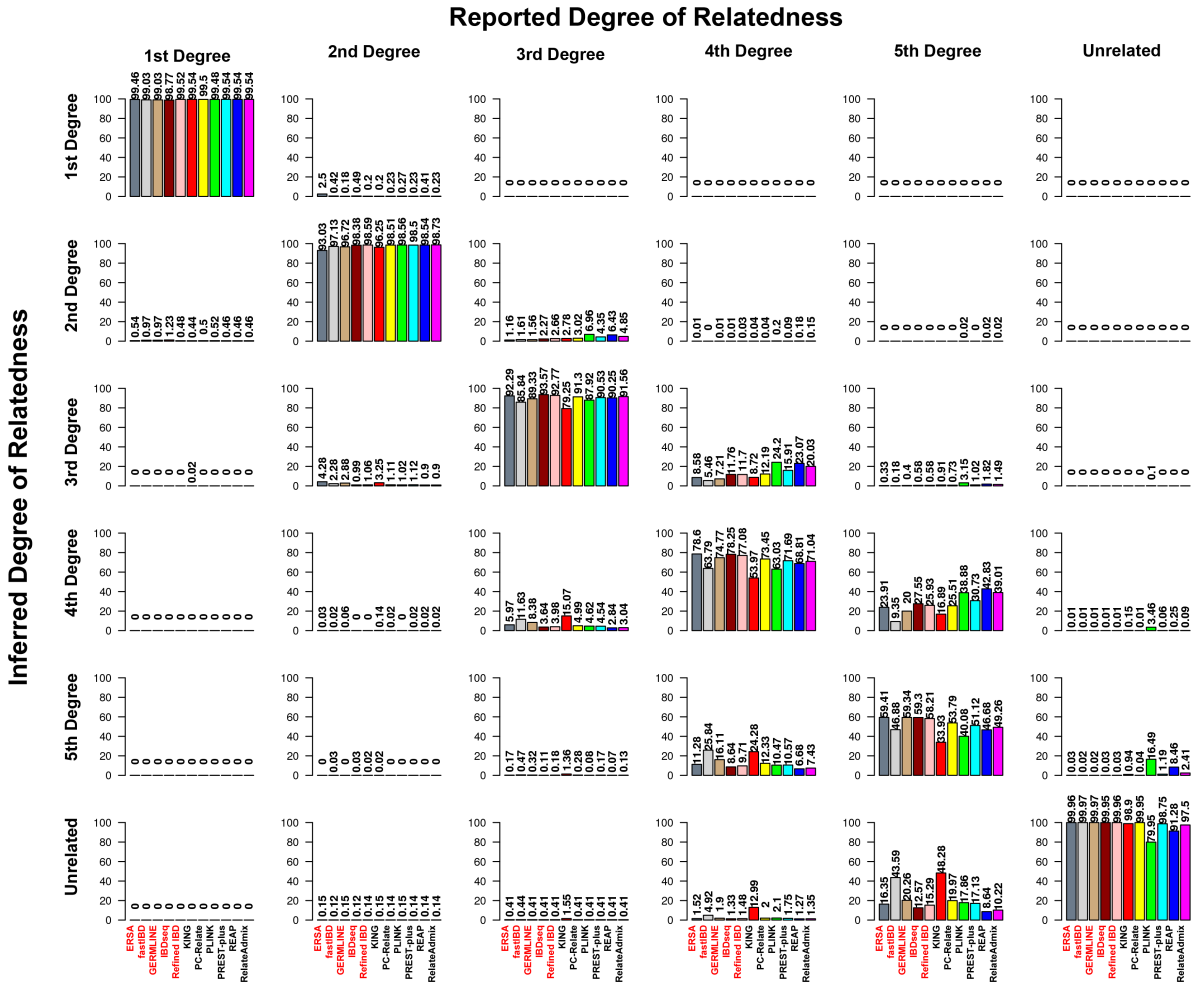

Figure 1: Performance comparison of the evaluated methods using the SAMAFS dataset. Bar plots indicate the percentage of pairs of samples that are reported to have a given degree of relatedness and who are inferred to be in each degree class. The bar plots are separated on the horizontal axis by the reported relatedness degree and on the vertical axis by inferred relatedness degree. For clarity, the plots list above each bar the percentage number that the corresponding bar depicts. Program names listed in red are IBD-based methods while those in black utilize allele frequencies for inference.

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

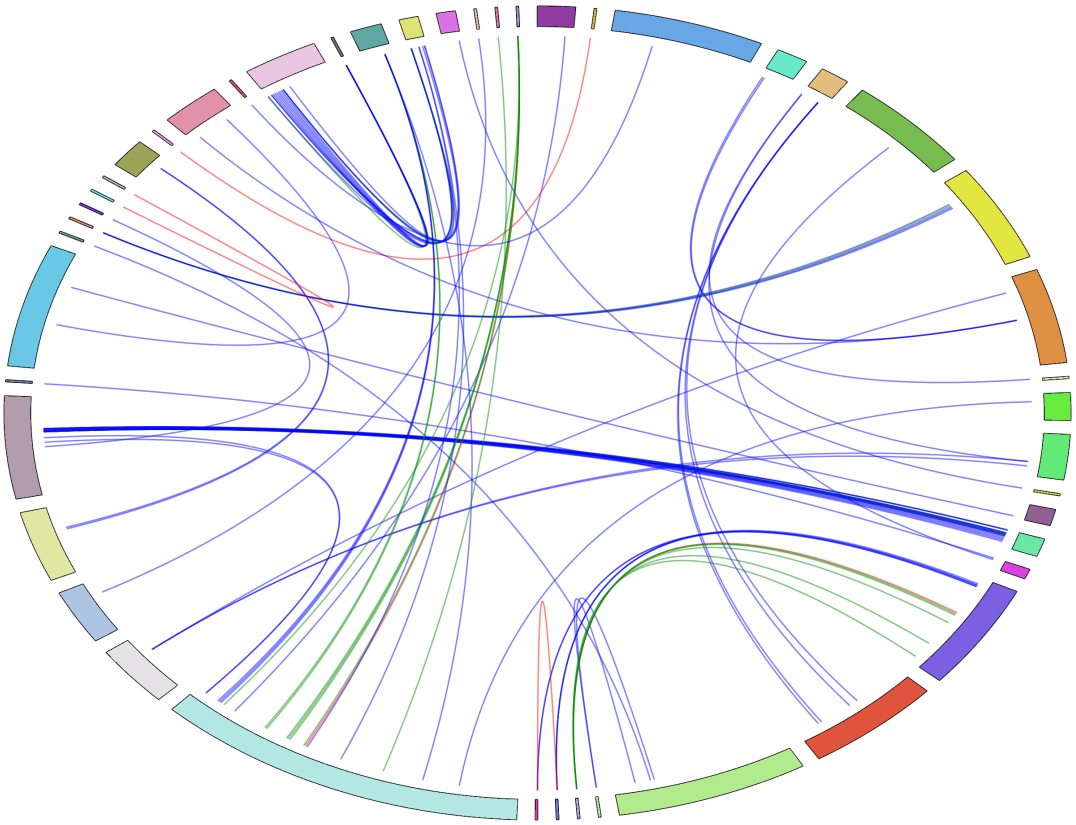

Figure 2: Relationships discovered between individuals from different SAMAFS pedigrees. Bands on the perimeter of the elliptical plot indicate distinct pedigrees within SAMAFS with band size proportional to the number of individuals in the pedigree. Curves between two bands correspond to discovered relative pairs with color indicating the degree of relatedness: red for first degree, green for second degree, and blue for third degree. Points where the curves end correspond to specific individuals, and a single point may have multiple curves running to it, indicating several relationships between that individual and others in the dataset.

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[2] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[3] Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

[4] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.

[5] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[6] Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLOS Genetics*, 1(3):e32, 2005.

[7] Jeffrey R O'Connell and Daniel E Weeks. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, 63(1):259–266, 1998.

[8] Jurg Ott. *Analysis of human genetic linkage*. JHU Press, 1999.

[9] Michael P Epstein, William L Duren, and Michael Boehnke. Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics*, 67(5):1219–1231, 2000.

[10] Mark A Jobling and Peter Gill. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10):739–751, 2004.

[11] Manfred Kayser and Peter de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[12] David C Queller and Keith F Goodnight. Estimating relatedness using genetic markers. *Evolution*, pages 258–275, 1989.

[13] Laurence D Hurst. Genetics and the understanding of selection. *Nature Reviews Genetics*, 10(2):83–93, 2009.

[14] Joshua G Schraiber and Joshua M Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015.

[15] WG Hill and BS Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(01):47–64, 2011.

[16] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):1, 2015.

[17] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[18] Timothy Thornton, Hua Tang, Thomas J Hoffmann, Heather M Ochs-Balcom, Bette J Caan, and Neil Risch. Estimating kinship in admixed populations. *American Journal of Human Genetics*, 91(1):122–138, 2012.

[19] Hong Li, Gustavo Glusman, Hao Hu, et al. Relationship estimation from whole-genome sequence data. *PLOS Genetics*, 10(1), 2014.

bioRxiv preprint doi: https://doi.org/10.1101/106013; this version posted February 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[20] Ida Moltke and Anders Albrechtsen. RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, 30(7):1027–1028, 2014.

[21] Lei Sun and Apostolos Dimitromanolakis. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. *BMC Proceedings*, 8(Suppl 1):S23, 2014.

[22] Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, 98(1):127–148, 2016.

[23] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, 2009.

[24] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2):173–182, 2011.

[25] Brian L Browning and Sharon R Browning. Detecting identity by descent and estimating genotype error rates in sequence data. *American Journal of Human Genetics*, 93(5):840–851, 2013.

[26] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[27] Braxton D Mitchell, Candace M Kammerer, John Blangero, Michael C Mahaney, David L Rainwater, Bennett Dyke, James E Hixson, Richard D Henkel, R Mark Sharp, Anthony G Comuzzie, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation*, 94(9):2159–2170, 1996.

[28] Ravindranath Duggirala, John Blangero, Laura Almasy, Thomas D Dyer, Kenneth L Williams, Robin J Leach, Peter O'Connell, and Michael P Stern. 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(4):1127–1140, 1999.

[29] Kelly J Hunt, Donna M Lehman, Rector Arya, Sharon Fowler, Robin J Leach, Harald HH Göring, Laura Almasy, John Blangero, Tom D Dyer, Ravindranath Duggirala, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans. *Diabetes*, 54(9):2655–2662, 2005.

[30] Chad D Huff, David J Witherspoon, Tatum S Simonson, Jinchuan Xing, W Scott Watkins, Yuhua Zhang, Therese M Tuohy, Deborah W Neklason, Randall W Burt, Stephen L Guthery, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, 21(5):768–774, 2011.

[31] Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

[32] William G Hill. Variation in genetic identity within kinships. *Heredity*, 71:652–653, 1993.

[33] Peter M Visscher. Whole genome approaches to quantitative genetics. *Genetica*, 136(2):351–358, 2009.

[34] Kuruvilla Joseph Abraham and Clara Diaz. Identifying large sets of unrelated individuals and unrelated markers. *Source code for biology and medicine*, 9(1):1, 2014.

[35] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.