

Improving IBD detection algorithms for genetic genealogy databases

Gregory R. Magoon
Aerodyne Research, Inc.
Billerica, Massachusetts, United States

September 2018

Copyright ©2018 All Rights Reserved.

Abstract

In recent years, with the proliferation of personal ancestry testing through microarray-based genotyping, detection of identity-by-descent (IBD) has become an important tool in genetic genealogy, enabling, for example, the identification and analysis of previously unknown biological relationships. The large amounts of data involved (hundreds of thousands to millions of individuals, and hundreds of thousands or more genotyped loci per individual) have prompted commercial databases and a prominent open, public database (GEDmatch) to adopt computationally streamlined approaches that are able to efficiently accommodate steadily incoming data, as new individuals are incorporated into the comparison databases. The tradeoff for computational efficiency with these approaches is a reduced ability to discriminate between IBD and non-IBD compared to more advanced approaches, especially when a) the two datasets being compared have limited density or limited overlap in genotyped loci, b) IBD segments of interest are small ($< \sim 7$ cM) or c) genotypes have high error rates. Here, we consider several variations of these algorithms, and use Bayesian analysis to gain insight into the underlying statistics and develop improved algorithms. Of particular note, the analysis demonstrates the heretofore unappreciated importance of heterozygous-to-heterozygous situations that arise in the pairwise genotype comparisons. The results suggest several possible refinements to current approaches, and two specific possibilities are discussed and demonstrated with both simulated and real data. One approach involves a simple modification to the type of genotype comparisons that contribute to meeting locus count thresholds for candidate IBD segments. The second approach takes advantage of allele frequency knowledge to increase detection power, and incorporates a genotype error model to enable error tolerance in a tunable, statistically justified manner; importantly, this is done without adding significant computational expense or algorithmic complexity.

1 Introduction

Detection and analysis of identity-by-descent (IBD) between pairs of individuals or among cohorts of individuals has been the subject of extensive research, with wide-ranging applications across population genetics and medical genetics, including genome-wide association studies, or GWAS (e.g. [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11]). A relatively recent, but powerful application of IBD detection is in the growing field of genetic genealogy.[12] Very recently, these new IBD-based genetic genealogy tools have also entered the field of forensic genetics.[13]

The main IBD comparison databases include those of commercial ancestry testing services, such as 23andMe and AncestryDNA, as well as GEDmatch, a free and open database accepting raw genotyping data from the various ancestry testing services, enabling cross-company comparisons.[14]

Due to the extreme number of pairwise comparisons involved, and the need to run new pairwise comparisons as new samples continue to be added to a steadily growing comparison pool, these databases have adopted variations on a low-computational-cost IBD detection approach first described by Henn et al.[12] The algorithm involves looking for a contiguous set of genotyped markers lacking "opposite homozygotes" (i.e. AA vs. BB for a biallelic locus with possible symbolic alleles A and B). In addition to requiring that the set of markers span a minimal genetic distance (e.g. 7 cM), identifying a segment as IBD requires that the region lacking opposing homozygotes also span a minimum number of markers (e.g. 400 SNPs). This criterion helps to guard against detecting regions where the genotyped (but unphased) markers lack opposite homozygotes due to coincidental identity-by-state (IBS). The algorithm is elegant in its simplicity, compared to more advanced approaches, and it has apparently been independently implemented (with some variation in details, but sharing the same core framework) in at least four commercial databases (23andMe, AncestryDNA, My Heritage, and Family Tree DNA), as well as several third-party databases, including GEDmatch.

Although the algorithm's low computational cost and ease of implementation are very favorable features, it is not without limitations. Commenting on the algorithm described by Henn and colleagues, Huang et al. have noted that the algorithm lacks sensitivity compared to more advanced approaches (only reliably identifying segments larger than about 7 cM), and adopts an heuristic approach to account for sporadic genotype errors.[5] So the tradeoff for the simplicity of the algorithm is that it has limited ability to discriminate between IBD and non-IBD in more challenging use cases, such as when making comparisons between individuals genotyped on different microarray platforms, where there may be limited overlap between the two sets of genotyped markers. A similar sort of challenging use case arises when trying to use the algorithm with genotype sets derived from confident calls in low-coverage whole-genome sequencing (WGS) data. In such situations, a user must either accept a high rate of false negatives (due to the inability of genuine IBD segments to meet the marker count threshold), or reduce the marker count threshold (if given the option) in order to reduce false negatives, but with the side effect of increasing false

positives related to IBS.

Another issue is that expanding to broader marker sets (for example, arising from genotyping known variants based on WGS, tends to introduce a significant portion of loci with rare variants (allele frequencies close to zero or unity). As noted by Huang et al.,[5] these markers are usually not particularly informative in distinguishing IBD from non-IBD. The strong allele frequency bias associated with broader marker sets would be expected to lead to unreliable behavior in the algorithm as originally described by Henn and colleagues, at least without adjustments to the marker count threshold and the heuristics for handling genotyping error. One would want to incorporate allele frequency information into the algorithm, either indirectly (by excluding markers not meeting some minimum minor allele frequency threshold, for example) or directly (through, for example, a weighting approach as described later).

A key point worth noting in the context of this paper is that the different implementations of this algorithm have adopted a variety of approaches related to which genotype match situations they consider as counting towards meeting the marker count threshold (with the threshold correspondingly also varying in accordance with the inclusion or exclusion of certain genotype match situations). For a biallelic marker, in addition to the previously-discussed opposite homozygotes situation (inconsistent with half-IBD), three genotype match situations, all consistent with half-IBD, are possible: 1) matching homozygotes (e.g. AA vs. AA), 2) matching heterozygotes (AB vs. AB), and 3) one heterozygote vs. one homozygote (e.g. AB vs. AA). The approach described in the original paper of Henn and colleagues requires at least one of the two samples have a homozygous genotype in order for the locus to count towards the 400 SNP threshold, essentially excluding the second genotype match situation (het-to-het) from consideration. The original GEDmatch implementation considered all three types of genotype match situations in counting toward the SNP threshold (default of 500 SNPs for "one-to-one" pairwise comparisons). An implementation on GEDmatch's new "Genesis" platform only considers matching homozygotes, excluding the second and third genotype match situations (het-to-het and het-to-hom), while reducing the SNP threshold.

Here, we will use Bayesian statistical methods to argue that the most effective allele-frequency-agnostic approach involves considering the second genotype match situation (het-to-het) but excluding the third genotype match situation (het-to-hom), constituting a variation on the algorithm not used in any previous implementation at the time of this writing, to our knowledge. We will also demonstrate how consideration of allele-frequencies during match evaluation, wherein appropriate weighting factors are applied in the consideration of all three types of genotype match situations (in addition to the opposite homozygote mismatch situation), can lead to significant improvements in IBD vs. non-IBD discrimination power as well as improvements in accuracy of estimated IBD segment boundaries.

2 Approach

To analyze, with Bayesian statistical methods, the genotype match situations that can be considered within the framework of the IBD-detection algorithm of Henn and colleagues, we use the logarithm of likelihood ratios for half-IBD relative to non-IBD scenarios. These log likelihood ratios allow the contribution of each locus comparison to be considered independently and in an additive manner. The summation of these log likelihood ratios might be considered to provide an indication of the likelihood of a given region being half-IBD rather than non-IBD. Thus, the summed log likelihoods might be considered as serving the same purpose as the marker threshold in the algorithm of Henn and colleagues, representing an aggregate body of evidence to rule out coincidental IBS. A similar log likelihood approach was adopted in the development of the Parente and Parente2 algorithms.[11] [6] The ratios are derived as a function of each locus’ (possibly unknown) allele frequency and the specific genotype match situation arising in a particular pairwise match at that locus.

For the purposes of this paper, we restrict the scope to considering detection of half-IBD only, without considering the possibility of full-IBD. Full-IBD will typically only be seen in close (e.g. sibling) relationships or in cases where the parents of the individuals being compared are from the same endogamous population. And, in any case, full-IBD situations will generally be identified by half-IBD detection algorithms.

We also restrict to consideration of loci recognized as biallelic, for the sake of simplicity. These loci will typically be single-nucleotide polymorphisms (SNPs) in applications of the algorithms discussed, but this need not necessarily be the case, and small insertions or deletions (indels) might also be considered, for example.

2.1 Simple refined approach

To gain insight into the statistics underlying the genotype comparisons used in the algorithm of Henn and colleagues, we initially consider likelihood ratios based on a simple allele frequency model, without accounting for the possibility of genotyping error or *de novo* mutation. The log likelihood ratios, $\ln K_{mn,i}$, for marker i between individuals m and n , consider the ratio of the probability of observing a given genotype combination, $G_{mn,i}$, given that the marker is within an IBD segment, to the probability of observing the same genotype combination given that the marker is not within an IBD segment, as follows:

$$\ln K_i = \ln K_{mn,i} = \frac{P(G_{mn,i}|\text{half-IBD})}{P(G_{mn,i}|\text{non-IBD})} \quad (1)$$

The log likelihood ratios, $\ln K_{mn,i}$, will be denoted $\ln K_i$ for simplicity in subsequent discussion. The allele-frequency-dependent probabilities considered here are equivalent to those presented in Table 1 of Epstein et al.,[15] except that we consider order-agnostic genotype combinations rather than ordered genotype pairs, and also restrict to a biallelic model. Although we model allele

frequencies here for purposes of analysis, we consider the allele frequencies to be unknown, in this section, for purposes of developing a refined algorithm. This corresponds to the situation with the algorithm of Henn and colleagues, where allele frequencies are not taken into consideration.

In the following figure, we plot the log likelihood ratios for the three matching genotype situations (plus two additional situations with the A and B alleles swapped) as a function of the allele frequency of the B allele. In the figure, positive values correspond to a likelihood ratio favoring IBD, while negative values favor non-IBD. The curves for the situations with the A and B alleles swapped are simply the reflection of the curve for the original situation about $x_{B,i} = \frac{1}{2}$, where $x_{B,i}$ is the B allele frequency.

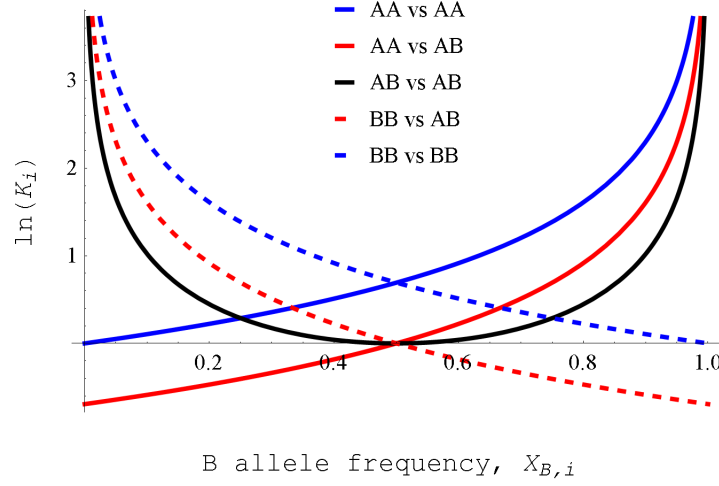


Figure 1: Log likelihood ratios for half-IBD relative to non-IBD in various matching genotype situations.

From the figure, it can be seen that the observed genotype situations favor IBD in most cases, with a few exceptions. The log likelihood ratios for the het-to-het and hom-to-het situations are zero with allele frequencies of $\frac{1}{2}$, indicating that these observations, at a marker with equal allele frequencies, do not provide any IBD-relevant information in the context of the algorithm. Similarly, the hom-to-hom situation also provides no new information when the observed allele has frequency of unity. Next, and perhaps contrary to what one might intuitively expect, the hom-to-het situation has negative log likelihood ratio when the frequency of the shared allele (i.e. B in BB vs. AB) is above $\frac{1}{2}$, indicating that this observation is actually suggestive of non-IBD. In contrast, it can be seen that the het-to-het situation is strongly suggestive of IBD when one of the alleles is rare.

The log likelihood curve for the hom-to-het situation deserves further discussion. As noted previously, when the shared allele has frequency above $\frac{1}{2}$,

the log likelihood ratio is slightly negative, suggestive of non-IBD rather than IBD. The log likelihood ratio passes through zero for allele frequencies of $\frac{1}{2}$, and becomes positive when the shared allele has frequency below $\frac{1}{2}$. Now, we expect that most of the times a hom-to-het situation is observed will involve sharing of the more common allele, suggesting that the negative portion of the curve may dominate in practice.

Indeed, with certain assumptions on the anticipated allele frequencies, we can calculate expectation values for the log likelihood ratios for each genotype match situation. These expectation values provide an indication of the IBD-relevant information each type of genotype match situation can provide in an allele-frequency-agnostic detection algorithm within the framework of the approach described by Henn and colleagues.

Here, we derive allele frequency probability distributions for each genotype match situation by first assuming a uniform, Beta(1,1) prior on the B allele frequency, $X_{B,i}$. Then, we incorporate information from the observed allele counts in the two samples (assuming either half-IBD or non-IBD situations) to derive a posterior probability distribution for the B allele frequency. Since the conjugate prior to the binomial genotype combination probability model is a beta distribution, this posterior distribution is also a beta distribution.

The resulting allele frequency probability distributions then allow us to compute the integrals needed to determine the expectation value for the log likelihood ratio for each genotype match situation. The results are shown in the following table, assuming either half-IBD or non-IBD scenarios. For each scenario, the table indicates the modeled probability distribution for the B allele frequency, as well as the corresponding expectation value for the log likelihood ratio.

Table 1: Expectation values of log likelihood ratios, with uniform prior on B allele frequency, $X_{B,i}$

Genotype match		assuming non-IBD		assuming half-IBD	
Type	Alleles	$X_{B,i}$	$E[\ln K_i]$	$X_{B,i}$	$E[\ln K_i]$
hom-to-hom	AA vs. AA	Beta(1,5)	0.20	Beta(1,4)	0.25
	BB vs. BB	Beta(5,5)		Beta(4,4)	
het-to-het	AB vs. AB	Beta(3,5)	0.18	$\frac{1}{2}$ (Beta(2,4) + Beta(3,4))	0.28
hom-to-het	AA vs. AB	Beta(2,5)	-0.24	Beta(2,4)	-0.11
	BB vs. AB	Beta(4,5)		Beta(3,4)	

We see from the table that the expected log likelihood ratio, under the given assumptions, for the het-to-het and hom-to-hom situations are positive and similar in magnitude. Furthermore, most marker sets would be expected to follow an allele frequency spectrum more strongly biased towards zero and unity than the uniform prior considered here. Adopting a more rare-variant-biased prior would further increase the expectation value for the het-to-het log likelihood ratio.

The expected log likelihood ratio for the hom-to-het situation is negative and comparable in magnitude to the positive het-to-het and hom-to-hom values.

Taken together, this all suggests a variation on the algorithm of Henn and colleagues, wherein hom-to-hom and het-to-het situations are considered to count towards meeting the marker count threshold (appropriately adjusted to achieve favorable balance between false negatives and false positives), while hom-to-het situations are ignored. Indeed, the negative expectation value for the hom-to-het log likelihood ratio might argue for actually subtracting from the marker count when these genotype match situations are encountered. On the other hand, the log likelihood ratio for hom-to-het genotype match situations is not universally negative, and can actually be strongly suggestive of IBD when the shared allele is rare. This naturally leads to a second, more advanced refinement to the algorithm, discussed in the next section, wherein knowledge of allele frequencies is taken into consideration.

2.2 Allele-frequency-aware, error-rate-aware refined approach

The preceding discussion suggests an approach that directly considers the allele frequencies and the associated log likelihood ratios in evaluating the aggregate evidence for IBD in a particular genomic region. Such an approach offers the potential for more substantial improvements in IBD detection accuracy and endpoint estimation. In fact, an approach using an underlying mathematical model along these lines has been developed and implemented by Rodriguez et al. in the development of Parente.[11] Despite the favorable features of their approach, the adoption of Parente’s methods among genetic genealogy databases has apparently been impeded by the complexity and limited accessibility of the underlying algorithms, and the non-negligible increase in computational expense relative to the approach of Henn and colleagues.

2.2.1 Mathematical model

Here, we aim to develop an approach with similar theoretical underpinnings to that of Parente, but without significant increases in algorithmic complexity or computational expense relative to the approach of Henn and colleagues. As has been noted by Rodriguez and colleagues,[11][6] this model neglects linkage disequilibrium (i.e. marker genotype correlation) considerations. Since straightforward approaches for adopting a more advanced model considering linkage disequilibrium can be expected to add significant computational expense (e.g. due to the need to consider multiple markers jointly), we consider this an acceptable tradeoff for use with large genetic genealogy databases. In any case, even the modeling of allele-frequencies and genotyping error rates would represent a significant improvement in modeling fidelity relative to the status quo with the Henn et al. algorithm.

As alluded to previously, for this allele-frequency-aware approach, we also incorporate a genotype error model, again, similar to the underlying model used

in Parente and speedB.[11] [5] However, here we generalize somewhat to distinguish two types of error: a) het-vs-hom error and b) hom-vs-opposite-hom error. The reason for this separation is that many assays will be more prone to the former error type than the latter. Here we neglect second-order error terms (where both samples have erroneous genotype) for simplicity. For reasonably small genotyping error rates characteristic of microarray assays, this omission should have negligible impact on the computed log likelihood ratios. It is also worth noting that a further generalization, to consider sample-dependent error rates, is possible. Modeling such differences in error rates might help, for example, when raw genotypes called from low-coverage whole-genome sequencing (relatively high error rate) are compared to microarray-based genotypes (relatively low error rate). Another conceivable generalization, not investigated here, is the possibility of modeling different error rates for het-to-hom genotyping errors relative to hom-to-het errors. Such asymmetric transitions could arise in raw calls from low-coverage whole-genome sequencing, for example, where true heterozygotes are susceptible to be erroneously called as homozygotes, whereas true homozygotes are generally less likely to be erroneously called as heterozygotes.

2.2.2 IBD detection algorithm

Now, we aim to leverage the log likelihood ratios at each site, based on this underlying mathematical model, to efficiently and accurately identify IBD segments for a given pair of individuals.

The computation of the log likelihood ratios themselves constitutes a non-negligible computational expense, but we note that these can be pre-computed to significantly improve efficiency. Specifically, half-IBD - to - non-IBD log likelihood ratios can be evaluated for each of the six pairwise genotype combinations possible for each biallelic marker. Subsequently, an arbitrarily high number of pairwise comparisons can access these pre-computed quantities, stored in memory, with negligible computational expense.

Now, given these per-site log likelihood ratios (LLRs), the IBD detection algorithm presented here aims to identify regions defined by contiguous marker sets where the aggregate (summed) LLR exceeds some (positive) user-specified confidence threshold, without any contiguous subsets where the aggregate LLR drops below some negative value (another user-specified threshold). The magnitude of the latter threshold is assumed to be less than or equal to that of the former threshold. The presented algorithm accomplishes this by taking a cumulative sum of LLRs from the chromosome "start" to "end", while tracking the running draw-up and draw-down relative to candidate segment boundaries associated with local minima and maxima in the cumulative LLR. Candidate segment boundary information is stored in a handful of tracking variables that are updated as the algorithm runs down the length of the chromosome, and memory usage (excluding allele frequency or pre-computed LLR data) is $O(1)$ and negligible (independent of marker density). CPU requirements, as for many IBD detection algorithms, are $O(N)$. However, the computations required per

site here are particularly sparse, involving at most two floating point operations (additions/subtractions) per locus, assuming pre-computed LLRs. Furthermore, the algorithmic simplicity is apparent from the fact that the core functionality (genotype comparison + LLR computation + aggregate LLR analysis) has been implemented in about 110 lines of Python code. The algorithm has been implemented in a manner to scale variables (the cumulative log likelihood ratio in particular) such that precision is not critical for the additions / subtractions. Relative to the use of double precision floating point arithmetic, use of single or even half-precision should not generally have a significant effect on the results.

Thus, the simplicity, flexibility, and relatively low computational expense of the algorithm might be expected to facilitate its adoption among genetic genealogy databases. Although there are no special hardware requirements, the relative simplicity, low overhead, and low precision requirements of the algorithm offer the opportunity for relatively straightforward performance tuning, potentially leading to highly-efficient implementations customized to various computing architectures, including GPUs. The negligible memory use associated with each pairwise comparison with this approach allows ample room for in-memory storage of allele frequencies or pre-computed log likelihood ratios in highly-parallelized implementations, for example.

Our implementation does not consider genetic distance during IBD segment identification, but filtering by cM is straightforward to apply either in post-processing of the identified segments, or in the algorithm directly (with the additional overhead associated with an in-memory recombination map).

This approach also provides some favorable features related to IBD segment endpoint estimation, relative to the current approaches based on the algorithm of Henn and colleagues. The approach of Henn et al. suffers from some uncertainty in reported endpoints due to chance matching near the boundaries of a segment. Although this approach can have similar issues, we expect that they should be mitigated through consideration of the negative log likelihood ratios associated with many hom-to-het situations. Furthermore, the probability-based framework provides a straightforward means to determine confidence intervals on the endpoints. Specifically, these can be determined from the positions where the draw-up or draw-down in the cumulative LLR, relative to the local minimum and maximum pair respectively, exceeds some threshold.

[***]

3 Results

3.1 Refined algorithm performance tests

[to be added]

4 Conclusion

We have demonstrated here the importance of heterozygous-to-heterozygous situations in variations on the IBD detection algorithm of Henn et al. [to be added]

5 Disclosure

Through his employer, G.R.M. works with Full Genomes Corporation (Harrisonburg, VA), a company utilizing massively parallel / "next-generation" sequencing in the fields of consumer genomics, genetic genealogy, and forensic genetics.
...

References

- [1] 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.
- [2] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.
- [3] Anders Albrechtsen, Thorfinn Sand Korneliussen, Ida Moltke, Thomas van Overseem Hansen, Finn Cilius Nielsen, and Rasmus Nielsen. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, 33(3):266–274, 2009.
- [4] Sharon R Browning and Brian L Browning. High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*, 86(4):526–539, 2010.
- [5] Lin Huang, Sivan Bercovici, Jesse M. Rodriguez, and Serafim Batzoglou. An effective filter for IBD detection in large data sets. *PLOS ONE*, 9(3):1–10, 03 2014.
- [6] Jesse M. Rodriguez, Sivan Bercovici, Lin Huang, Roy Frostig, and Serafim Batzoglou. Parente2: a fast and accurate method for detecting identity by descent. *Genome Research*, 25(2):280–289, 2015.
- [7] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

- [8] 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, Scott R. Woodward, and Lynn B. Jorde. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, 21(5):768–774, 2011.
- [9] Hong Li, Gustavo Glusman, Hao Hu, Shankaracharya, Juan Caballero, Robert Hubley, David Witherspoon, Stephen L. Guthery, Denise E. Mauldin, Lynn B. Jorde, Leroy Hood, Jared C. Roach, and Chad D. Huff. Relationship estimation from whole-genome sequence data. *PLOS Genetics*, 10(1):1–12, 01 2014.
- [10] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068, 2008.
- [11] Jesse M. Rodriguez, Serafim Batzoglou, and Sivan Bercovici. An accurate method for inferring relatedness in large datasets of unphased genotypes via an embedded likelihood-ratio test. In Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, editors, *Research in Computational Molecular Biology*, pages 212–229, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [12] Brenna M. Henn, Lawrence Hon, J. Michael Macpherson, Nick Eriksson, Serge Saxonov, Itsik Pe’er, and Joanna L. Mountain. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE*, 7(4):1–13, 04 2012.
- [13] Chris Phillips. The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics*, 36:186 – 188, 2018.
- [14] GEDmatch. Available at <https://www.gedmatch.com>.
- [15] Michael P Epstein, William L Duren, and Michael Boehnke. Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, 67(5):1219–1231, 2000.