Advance Access Publication Date: 2 April 2015
Applications Note

Genome analysis

Near-perfect identification of half-sibling versus niece/nephew–avuncular pairs without pedigree information or genotyped relatives

Emmanuel Sapin*, Kristen M. Kelly, and Matthew C. Keller

Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Genome-wide biobank datasets contain many second-degree relatives whose precise pedigree relationships are not recorded, yet accurate distinction between half-sibling (HS) and niece/nephew–avuncular (N/A) pairs is critical for downstream genetic and epidemiological analyses. Standard SNP-based relatedness measures identify "25% relatives" but provide only modest power to distinguish HS from N/A pairs because segment number, length, and age differences show substantial overlap between these two categories. There is therefore a need for methods that can subtype 2nd-degree relationships into HS versus N/A using only genotype data, without requiring fully observed pedigrees or extensive relative information in the sample.

Results: We introduce a framework to classify 2nd-degree pairs as HS or N/A in UK Biobank. We first identify ground-truth HS and N/A pairs using information from close relatives, yielding labeled examples of both categories. We then apply across-chromosome phasing to derive a haplotype-level relatedness statistic which, when modeled with a Gaussian mixture model, yields near-complete separation of ground-truth HS and N/A pairs with high accuracy, sensitivity, and specificity, enabling robust, scalable subtyping of 2nd-degree relatives in large cohorts without explicit pedigree data.
Contact: emmanuel.sapin@colorado.edu, kristen.kelly@colorado.edu, and matthew.c.keller@colorado.edu

1 Introduction

Genome-wide genetic datasets routinely contain pairs of individuals who are more closely related than expected under simple population sampling, even in the absence of explicit pedigree information [Ramstetter et al., 2017]. Methods for inferring degrees of relatedness from genotypes have therefore become a core component of modern human genetics, particularly as sample sizes have increased from thousands to hundreds of thousands of individuals in biobank-scale resources [Ramstetter et al., 2017, Browning & Browning, 2012]. Accurate degree-of-relatedness inference enables downstream applications including control of relatedness in genome-wide association studies, estimation of SNP-based heritability and inbreeding, and refinement of demographic and population-structure analyses, as well as forensic and genealogical investigations in large genotype databases [Yang et al., 2014, Broman, 2015, Conomos et al., 2017, Emery et al., 2022, Edge & Coop, 2020].

Large population-based cohorts with genome-wide genotype data, such as the UK Biobank, contain substantial numbers of closely related individuals who are not explicitly recorded as relatives in the study data [Bycroft et al., 2018]. In this context, a key challenge is that "half-degree" relationships have the same expected proportion of genome shared as more direct (numerically lower) degrees, because they involve only half the number of shared ancestors [Thompson, 2013]. For example, half-siblings (HS) can be viewed as "half degree 1" but share on average 25% of their genome and thus fall in the same expected genome-wide relatedness range as 2nd-degree relatives such as niece/nephew–avuncular (N/A) pairs, and half-avuncular relationships can produce identity-by-descent (IBD) patterns similar to those of full first cousins [Thompson, 2013, Ramstetter et al., 2017]. Among recorded and unrecorded relatives in UK Biobank and comparable resources, 2nd-degree pairs are common and arise from several distinct pedigree relationships, including half-siblings, N/A pairs, and compound pedigrees such as double first cousins and more complex multi-path configurations [Bycroft et al., 2018, Pinto et al., 2010, Brennan, 2019].

In principle, IBD segment number and length should differ between HS and N/A pairs, because half-siblings inherit all shared IBD from a single parent, whereas avuncular pairs inherit through an additional generational step [Browning & Browning, 2012]. Empirical analyses and simulations of 25% relationships indicate that half-siblings tend to share fewer but longer IBD segments, whereas N/A pairs tend to share a larger number of shorter

© The Author . Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

segments, even though both categories have similar total IBD sharing of about 25% of the autosomal genome [Browning & Browning, 2012, Ramstetter et al., 2017]. However, the corresponding segment-count and length distributions overlap substantially across relationship types [Williams et al., 2025, **?**]. Age differences can also be informative, since avuncular pairs typically span a wider age gap than half-siblings drawn from the same cohort, but age alone does not cleanly separate the two categories in biobank settings with restricted recruitment ages. As a result, segment number, segment length, and age differences considered individually provide only modest discriminative power for separating HS from N/A pairs among 2nd-degree relatives [Williams et al., 2025, **?**].

Here, we develop a three-stage strategy for classifying 2nd-degree pairs in UK Biobank into half-sibling and N/A relationships. First, candidate 2nd-degree pairs are detected using a standard genome-wide relatedness estimator based on SNP sharing. Second, close relatives present in the dataset (primarily parent–child, full siblings, and 3rd-degree relationships such as first cousins) are used to reconstruct partial family graphs, from which a subset of 2nd-degree pairs can be assigned high-confidence labels as HS or N/A and compound pedigrees can be identified and excluded from training. Third, genome-wide phasing is used to derive across-chromosome haplotypes for each individual, and haplotype-level sharing metrics are computed that summarize how IBD is distributed across the four haplotypes of a pair. These haplotype-based features, together with diploid relatedness estimates and age, are then used in a probabilistic classifier that assigns HS vs N/A labels to the remaining unlabeled 2nd-degree pairs, leveraging the fact that HS pairs concentrate sharing on a single parental haplotype, whereas N/A pairs exhibit more diffuse sharing across haplotypes despite similar total IBD [Browning & Browning, 2011, Delaneau et al., 2012, Staples et al., 2016].

2 Methods

2.1 UK Biobank sample and ancestry definition

The method was developed and evaluated using data from the UK Biobank [Bycroft et al., 2018], focusing on individuals of European descent. This choice was motivated by the altered performance of relatedness estimation in mixed-ancestry samples and by the fact that individuals of European descent constitute the overwhelming majority of the cohort [Bycroft et al., 2018]. Genetic principal components (PCs) were computed on a linkage-disequilibrium–pruned set of common autosomal SNPs [Browning & Browning, 2012]. The smallest 4-dimensional hypercube in PC space containing all individuals with self-described white British ancestry was then identified, and all individuals lying within this hypersphere, regardless of self-described ancestry, were retained as the eligible sample (435,187 individuals) [Bycroft et al., 2018].

Next, we selected biallelic autosomal SNPs that were present on both the UK BiLEVE and UK Biobank Axiom arrays, had a minor allele frequency (MAF) of at least 5%, and had a Hardy–Weinberg equilibrium chi-square test p-value greater than 0.0005. This resulted in a final dataset of 330,005 SNPs for relatedness estimation and downstream analyses [Bycroft et al., 2018].

2.2 Detection of 2nd-degree relatives

For all pairs of individuals in the European subset, the proportion of the genome for which the pair shares 0, 1 or 2 haplotypes identical by descent is denoted IBD0, IBD1 and IBD2 Zhou et al. [2020], which were estimated with the standard software PLINK [**?**] as well as the widely used measure of diploid SNP similarity $\hat{\pi}^{(f,i)}$ defined in equation 1.

$$\hat{\pi}(f,i) = \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{kf}-p_k)(x_{ki}-p_k)}{2p_k(1-p_k)}, \tag{1}$$

where f denotes the focal individual, i indexes the non-focal individuals, p_k is the minor allele frequency at SNP k, x_k^f and x_k^i represent the diploid genotypes (coded as 0, 1, or 2 copies of the reference allele) for individuals f and i at SNP k, and m is the total number of SNPs included in the calculation [Browning & Browning, 2012].

Pairs with $\hat{\pi}$ close to 0.25 are typically classified as 2nd-degree relatives, consistent with the theoretical expectation that 2nd-degree relatives share one quarter of their genome IBD on average [Thompson, 2013]. Here, a conservative window of $\hat{\pi}$ between 0.20 and 0.325, with thresholds chosen empirically, was used to define the candidate 2nd-degree set, excluding 1st-degree pairs (with $\hat{\pi}$ near 0.5) and more distant pairs [Ramstetter et al., 2017]. The exclusion of compound pedigrees [Bycroft et al., 2018, Pinto et al., 2010, Brennan, 2019] relied on the genome-wide IBD2 sharing estimated by PLINK. Because IBD2 should be close to 0 for N/A and HS pairs, an empirical threshold of 0.025 was used.

Because of the age structure of UK Biobank, true grandparent–grandchild pairs are expected to be rare, so this candidate 2nd-degree set is assumed to be dominated by half-sibling and N/A pairs.

2.3 Ground truth labeling

To obtain labeled examples of specific 2nd-degree types, a family graph was constructed from high-confidence 1st-degree relationships. Parent–child and full-sibling relations were identified using $\hat{\pi}$ and genome-wide IBD2 sharing. From these relationships, the following labels could be assigned:

- **Half-siblings (HS)**: pairs sharing exactly one inferred parent in the graph.
- **N/A pairs**: pairs in which one individual is inferred to be a child of a sibling of the other individual.

This procedure led to the identification of 484 N/A pairs but only 5 HS pairs. To increase the number of HS examples, additional HS pairs were sought using information from other degrees of relatedness, despite the additional uncertainty. Two configurations involving first-cousin relationships (with $\hat{\pi}$ around 0.125) can be used to identify HS pairs:

- **Shared first cousin**: pairs in which two individuals ($HS1$ and $HS2$) share a first cousin CFC as illustrated in Figure 3. Such pairs cannot be N/A, because,as illustrated in Figure 1, a first cousin CB of an avuncular individual B would be a parent's first cousin of the niece/nephew A (with expected $\hat{\pi}$ around 0.0625) . Therefore, these pairs are classified as HS, Figure 3.
- **Unrelated first cousins**: pairs in which each individual ($HS1$ and $HS2$) has a first cousin ($FC1$ and $FC2$) who is unrelated to the other individual in the pair as illustrated Figure 2. As shown in Figure 1, a first cousin of an avuncular individual B is a parent's first cousin of the niece/nephew A, and therefore cannot be unrelated to the niece/nephew. Consequently, pairs in which each individual has a first cousin who is unrelated to the other individual in the pair cannot be an N/A pair and are therefore classified as HS, as in Figure 2.

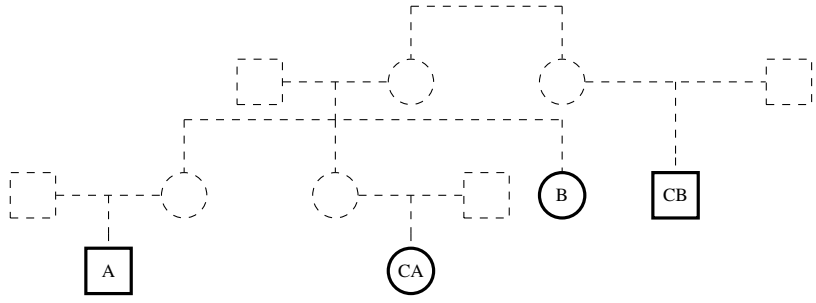

Figure 1: Example pedigree showing a niece A and her aunt/uncle B, each with a first cousin (CA and CB) in the sample.

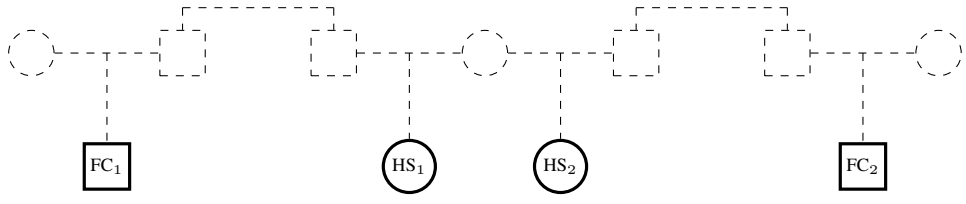

Figure 2: Example pedigree with two half-siblings (HS_1, HS_2) sharing a mother and each having a paternal first cousin (FC_1, FC_2), illustrating a configuration that is incompatible with an N/A relationship between HS_1 and HS_2.

The remaining candidate 2nd-degree pairs that could not be labeled with high confidence, typically because they were isolated from informative relative in the sample, were left unlabeled at this stage.

When detecting HS pairs, relatives were considered first cousins following the conservative threshold of $0.11 < \hat{\pi} < 0.14$, and a cousin of one individual was considered unrelated to the other individual in the pair only if their pairwise $\hat{\pi}$ was lower than 0.025.

2.4 Across-chromosome phasing

All UK Biobank individuals were initially phased within chromosomes using Shapeit2 [Bycroft et al., 2018]. We then performed across-chromosome phasing following previously described procedures for constructing genome-wide haplotypes from chromosome-wise phase estimates [**?**], so that each individual's two haplotypes could be treated as approximate parental homologues across the entire autosomal genome without explicit assignment to maternal or paternal origin.

2.5 Haplotype–haplotype $\hat{\pi}$

For each candidate 2nd-degree pair, four haplotype–haplotype $\hat{\pi}$ ($\hat{\pi}^{\text{hh}}$) values were defined, corresponding to each phased haplotype of one individual against each phased haplotype of the other individual. The $\hat{\pi}^{\text{hh}}$ metric has the same functional form as the diploid $\hat{\pi}$, but with p_k interpreted as the allele frequency on a single haplotype (half the diploid frequency at SNP k) and x_{kf} and x_{ki} representing haploid genotypes (coded as 0 or 1 copies of the reference allele) for individuals f and i at SNP k. This definition ensures that the sum of the four $\hat{\pi}^{\text{hh}}$ values equals the diploid–diploid $\hat{\pi}$ for the pair.

To maximize informativeness of phased sharing and reduce noise from uninformative loci, haplotype–haplotype $\hat{\pi}$ was computed using only SNP sites at which both individuals were heterozygous, because these doubly heterozygous loci carry direct information about which phased allele is shared on which haplotype. This procedure yielded a 2×2 matrix of $\hat{\pi}^{\text{hh}}$ values for each pair, as shown in Table 1. The ordering of haplotypes p_1 and p_2 and of individuals i and j is chosen so that $\hat{\pi}^{\text{hh}}_{\text{Max}}$ is the largest of the four entries and $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ is larger than $\hat{\pi}^{\text{hh}}_{\text{R_Min}}$. The remaining cell, $\hat{\pi}^{\text{hh}}_{\text{Opp}}$, is defined as the entry opposite $\hat{\pi}^{\text{hh}}_{\text{Max}}$ in the 2×2 matrix.

Table 1. Haplotype–haplotype $\hat{\pi}$ values for a given pair of individuals, with haplotypes ordered so that $\hat{\pi}^{\text{hh}}_{\text{Max}}$ is the maximum value.

		Individual i	
		p_1	p_2
Individual j	p_1	$\hat{\pi}^{\text{hh}}_{\text{Max}}$	$\hat{\pi}^{\text{hh}}_{\text{R_Max}}$
	p_2	$\hat{\pi}^{\text{hh}}_{\text{R_Min}}$	$\hat{\pi}^{\text{hh}}_{\text{Opp}}$

When the two individuals share a parent (as in the case of half-sibling), the haplotypes inherited from that parent are expected to yield $\hat{\pi}^{\text{hh}}_{\text{Max}} \approx 0.25$, while the other three haplotype pairs have mean values close to 0. When two individuals form an N/A pair, $\hat{\pi}^{\text{hh}}_{\text{Max}}$ and $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ are both expected to be around 0.125 (due to independent assortment), while the remaining two entries have expected value 0. The four expected $\hat{\pi}^{\text{hh}}$ values for HS and N/A pairs are summarized in Table 2.

Table 2. Expected haplotype–haplotype $\hat{\pi}^{\text{hh}}$ values for a half-sibling (HS, left) and a niece/nephew–avuncular (N/A, right) pair.

		Individual i	
		p_1	p_2
Individual j	p_1	0.25	0
	p_2	0	0

HS

		Individual i	
		p_1	p_2
Individual j	p_1	0.125	0.125
	p_2	0	0

N/A

Among the four $\hat{\pi}^{\text{hh}}$ values, $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ provided the greatest power to distinguish HS from N/A pairs and was therefore chosen as the primary input to the Gaussian mixture model.

3 Results

3.1 Gaussian mixture model on haplotype sharing

The distribution of $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ is expected to exhibit two populations corresponding to half-sibling (HS) pairs and niece/nephew–avuncular (N/A) pairs. To test this hypothesis, a Gaussian mixture model (GMM) was fitted to the distribution of $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ across all candidate 2nd-degree pairs.

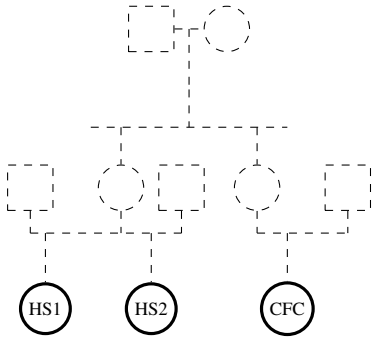

Figure 3: Exemple pedigree with two half-sibling $HS1$ and $HS2$ sharing a mother and having in a sample a maternal first cousin CFC.

The GMM identified two components with a cut-off at $\hat{\pi}^{\text{hh}}_{\text{R_Max}} = 0.0209$, comprising 1,423 pairs in the lower-mean component and 6,633 pairs in the higher-mean component (Figure 4). The higher-mean component is lower than the expected $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ values for N/A pairs because of errors in across chromosome phasing, while the lower-mean component matches the expected distribution for HS pairs, in line with the theoretical patterns summarized in Table 2. The fitted mixture curves closely followed the empirical histogram of $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$, indicating that a two-component Gaussian model provides a good description of the observed distribution.

3.2 Performance on ground truth pairs

To evaluate classification performance, the GMM-based component assignment was compared to the graph-derived ground truth labels (treating HS as the positive class). Among labeled pairs, the model achieved 87 true positives (TP), 428 true negatives (TN), 3 false positives (FP), and 6 false negatives (FN), corresponding to an overall accuracy of 0.9828, a sensitivity of 0.9355, and a specificity of 0.993 (Figure 5). Most misclassified pairs, including 2 of the 3 FP and 4 of the 6 FN, lay in the region between the two mixture components near the cut-off at $\hat{\pi}^{\text{hh}}_{\text{R_Max}} = 0.0209$, where the posterior probabilities of HS and N/A are most similar. The number of 1,423 HS pairs is consistent with what demographic studies would predict for a sample of this type [Andersson, 2020].

These results show that $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ provides near-complete separation between HS and N/A pairs among 2nd-degree relatives that pass the filtering. In particular, the low number of FP and FN relative to the total number of labeled pairs indicates that, for most pairs, the haplotype-level sharing signal is strong enough to yield a high-confidence classification into HS or N/A.

4 Conclusion

This study shows that a simple haplotype-level sharing statistic from across chromosome phasing, $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$, combined with a Gaussian mixture model, can robustly distinguish half-sibling from niece/nephew–avuncular pairs among 2nd-degree relatives in UK Biobank. The method leverages existing family-graph structure to define high-confidence HS and N/A examples, and then generalizes to unlabeled pairs without requiring explicit full-pedigree reconstruction. In this setting, the model yields clean separation of the two classes with only a small number of misclassified pairs concentrated near the decision boundary.

This study is a direct application of across chromosome phasing that have direct implications for downstream analyses that require accurate relationship subtyping, including the construction of refined pedigrees, the control of cryptic relatedness in association studies, and the interpretation of shared genetic and environmental variance in large cohorts.

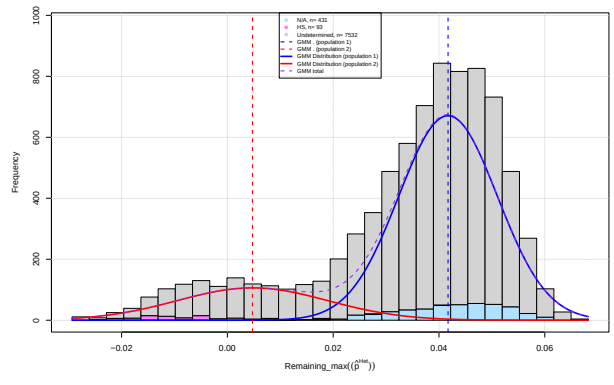

Figure 4: Distribution of $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ for all candidate 2nd-degree pairs, with the fitted Gaussian mixture model. The two components correspond to HS-enriched (lower mean) and N/A-enriched (higher mean) populations, separated by a cut-off at $\hat{\pi}^{\text{hh}}_{\text{R_Max}} = 0.0209$.

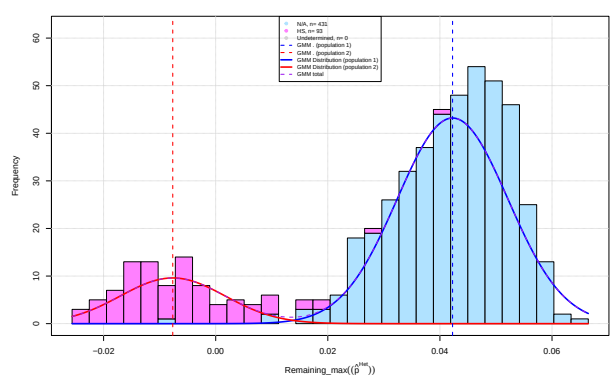

Figure 5: Distribution of $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$ restricted to ground truth HS and N/A pairs, with Gaussian mixture components overlaid.

Future work could be to generalize this approach to distinguish pairs of first degree cousins and pairs of half niece/nephew–avuncular. Within and across chromosome phasing could be improve as, the integration of explicit information from close relatives could further improve phasing accuracy. For example, all genomic segments shared between an avuncular

relative and a nice or nephew must originate from the same parental lineage providing a strong constraint on phase orientation.

5 Funding

This publication and the work reported in it are supported in part by the National Institute of Mental Health Grant 2R01 MH100141 (PI: M.K.) and R01 MH130448 (PI: M.K.).

References

Zhou Y, Browning SR, Browning BL. IBDkin: fast estimation of kinship coefficients from identity by descent segments. Bioinformatics. 2020 Aug 15;36(16):4519-4520. doi: 10.1093/bioinformatics/btaa569. PMID: 32844204; PMCID: PMC7750976.

https://www.demographic-research.org/volumes/vol43/16/43-16.pdf

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., & Göring, H. H. H. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(1), 75–89.

Browning, S. R., & Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, 46, 617–633.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2), 100–106.

Broman, K. W. (2015). Identification of sample mix-ups and mixtures in high-throughput genotyping data. *American Journal of Human Genetics*, 97(4), 555–564.

Conomos, M. P., Thornton, T. A., Gogarten, S. M., *et al.* (2017). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 41(4), 303–315.

Emery, L. S., *et al.* (2022). Forensic applications of identity-by-descent and relatedness inference. *Annual Review of Genomics and Human Genetics*, 23, 129–153.

Edge, M. D., & Coop, G. (2020). How lucky was the genetic investigation in the Golden State Killer case? *Genetics*, 215(4), 939–944.

Bycroft, C., Freeman, C., Petkova, D., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209.

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

Staples, J., Qiao, D., Croteau-Chonka, D. C., *et al.* (2016). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, 99(1), 210–221.

Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2), 173–182.

Delaneau, O., Marchini, J., & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2), 179–181.

Pinto, N., Gusmão, L., Amorim, A., *et al.* (2010). General derivation of the sets of pedigrees with the same kinship coefficients. *Journal of Theoretical Biology*, 263(1), 61–69.

Brennan, N. (2019). Genetic and Quantitative Aspects of Genealogy: Types of Collateral Relationships. Available at https://www.genetic-genealogy.co.uk/.

Williams, X., *et al.* (2025). Title of rapid accurate pedigree inference in endogamous populations. *Genetics*, 230(4), 2131–2148.