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: Large-scale genomic biobanks contain thousands of second-degree relatives with missing pedigree metadata. Accurately distinguishing half-sibling (HS) from niece/nephew–avuncular (N/A) pairs—both sharing approximately 25% of the genome—remains a significant challenge. Current SNP-based methods rely on Identical-By-Descent (IBD) segment counts and age differences, but substantial distributional overlap leads to high misclassification rates. There is a critical need for a scalable, genotype-only method that can resolve these "half-degree" ambiguities without requiring observed pedigrees or extensive relative information.

Results: We present a novel computational framework that achieves near-complete separation of HS and N/A pairs using only genotype data. Our approach utilizes across-chromosome phasing to derive haplotype-level sharing features that summarize how IBD is distributed across parental homologues. By modeling these features with a Gaussian mixture model (GMM), we demonstrate near-perfect classification accuracy ($> 98\%$) in biobank-scale data. Furthermore, we show that these high-confidence relationship labels can serve as long-range phasing anchors, providing structural constraints that improve the accuracy of across-chromosome homologue assignment. This method provides a robust, scalable solution for pedigree reconstruction and the control of cryptic relatedness in large-scale genomic studies.
Contact: emmanuel.sapin@colorado.edu, kristen.kelly@colorado.edu, and matthew.c.keller@colorado.edu

1 Introduction

Genome-wide datasets routinely contain pairs of individuals more closely related than expected under random population sampling, often in the absence of explicit pedigree records [Ramstetter et al., 2017]. The accurate inference of these relationships has become a cornerstone of modern human genetics, particularly as sample sizes in biobank-scale resources have scaled to hundreds of thousands of individuals [Browning & Browning, 2012]. Robust relationship subtyping is essential for controlling cryptic relatedness in genome-wide association studies (GWAS), refining demographic models, and enhancing forensic and genealogical investigations [Yang et al., 2014, Conomos et al., 2017].

A persistent challenge in relatedness inference is the ambiguity between "half-degree" relationships. While first-degree relatives (parent–child and full siblings) share approximately 50% of their genome identical-by-descent (IBD), second-degree relatives share approximately 25%. However, several distinct pedigree configurations result in this same 25% sharing, most notably half-siblings (HS) and niece/nephew–avuncular (N/A) pairs [Thompson, 2013]. Distinguishing between these configurations is computationally non-trivial because the genome-wide proportion of IBD sharing is stochastically similar across these classes [Ramstetter et al., 2017].

Current methodologies primarily rely on the distribution of IBD segments; half-siblings theoretically inherit sharing from a single shared parent, whereas avuncular pairs inherit through an additional meiosis, typically leading to a higher count of shorter segments [Browning & Browning, 2012]. Despite these theoretical differences, empirical distributions of segment length and count exhibit substantial overlap, exacerbated by the stochastic nature of recombination [Williams et al., 2025]. Furthermore, while demographic metadata such as age differences can provide heuristic support, they are often insufficient for high-confidence classification in cohorts with restricted recruitment windows or non-traditional family structures.

In this work, we propose a three-stage computational strategy to resolve second-degree relationships using across-chromosome phasing. By moving beyond diploid SNP similarity to a haplotype-aware framework, we derive features that capture the inheritance logic of parental homologues.

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

We demonstrate that these features, when modeled via a Gaussian mixture model (GMM), provide near-perfect separation between HS and N/A pairs. Finally, we show that these identified relationships can be utilized as phase anchors, establishing a feedback loop that improves the structural integrity of across-chromosome phasing. This approach offers a scalable, genotype-only solution that eliminates the dependency on explicit pedigree metadata.

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).

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 within the European subset, we estimated the proportions of the genome sharing 0, 1, or 2 haplotypes identical-by-descent (IBD_0, IBD_1, IBD_2) using PLINK [Purcell et al., 2007]. Additionally, we calculated the standard measure of diploid SNP similarity, $\hat{\pi}^{(f,i)}$, as 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.

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 0.03 was used.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. Since IBD2 is theoretically zero for both N/A and HS pairs, we employed an empirical filter where pairs were retained only if their estimated IBD2 sharing fell within the range of -0.005 to 0.03.

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)**: pair of individuals sharing one parent.
- **N/A pairs**: pair 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.
- **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.

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 $0.115 < \hat{\pi} < 0.14$, 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.

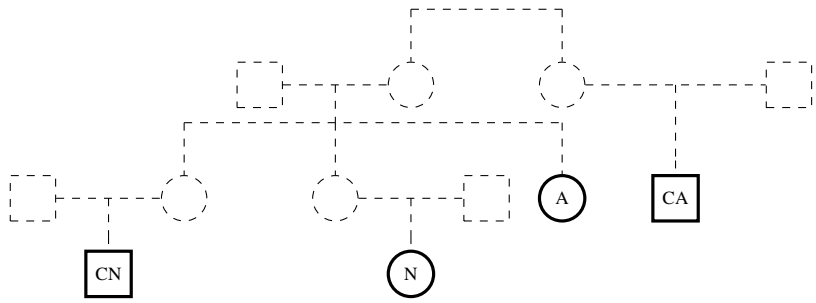

Figure 1: Example pedigree showing a niece/nephew N and their avuncular relative A, with first cousins (CN and CA) used to verify the relationship labels.

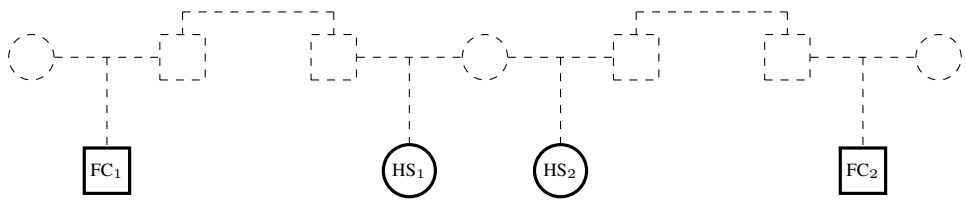

Figure 2: Example pedigree with two half-siblings (HS_1, HS_2) sharing a mother and each having a paternal first cousin (FC_1, FC_2), illustrating a configuration that is incompatible with an N/A relationship between HS_1 and HS_2.

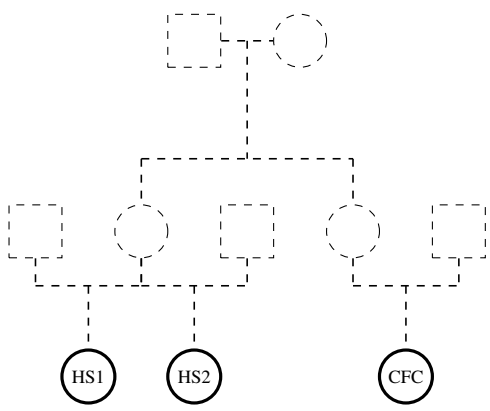

Figure 3: Refined pedigree showing two half-siblings ($HS1$, $HS2$) sharing a mother and a maternal first cousin (CFC).

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

The GMM identified two distinct components. The higher-mean component matches the expected distribution for HS pairs, where sharing is concentrated on a single parental haplotype, while the lower-mean component aligns with the pattern expected for N/A pairs. The fitted mixture curves closely followed the empirical histogram of $\hat{\pi}^{\text{hh}}_{\text{Max}}$, indicating that a two-component Gaussian model provides a robust description of the observed haplotype-level sharing.

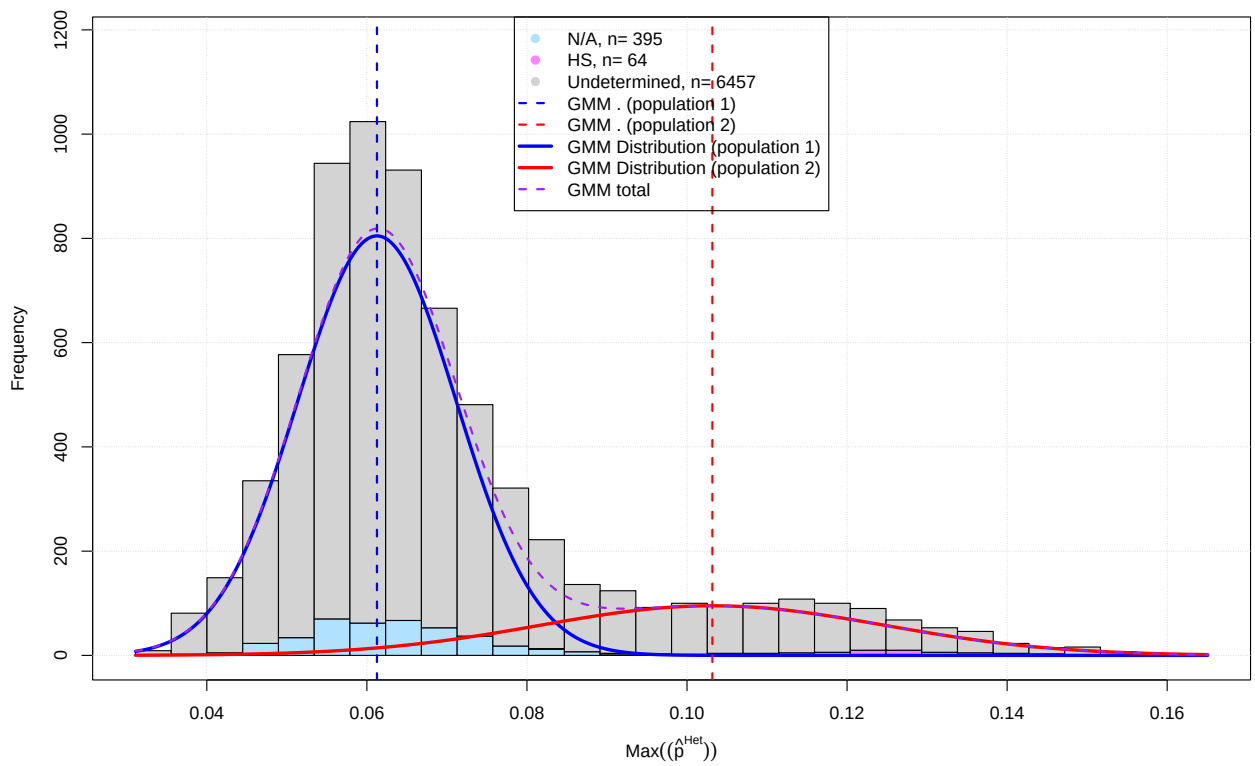

Figure 4: Distribution of $\hat{\pi}^{\text{hh}}_{\text{Max}}$ for all candidate 2nd-degree pairs, with the fitted Gaussian mixture model. The two components correspond to N/A-enriched (lower mean) and HS-enriched (higher mean) populations, separated by a cut-off at $\hat{\pi}^{\text{hh}}_{\text{Max}} = 0.0927$.

3.2 Performance on ground truth pairs

To evaluate the classification power of the GMM, we applied the derived cutoff of $\hat{\pi}^{\text{hh}}_{\text{Max}} = 0.0927$ to a validation set of ground-truth relationships. When applied to the ground-truth set, the model achieved near-perfect separation (Figure 5), correctly identifying 62 out of 64 half-sibling pairs (TP: 62, FN: 2) and 392 out of 395 avuncular pairs (TN: 392, FP: 3).

These results correspond to a sensitivity of 96.9% and a specificity of 99.2%, yielding an overall accuracy of 98.9%. The high specificity is particularly critical, as it ensures that the relationships identified as half-siblings—which are subsequently used as high-confidence long-range phasing anchors—are virtually free of misclassified avuncular noise. The minimal number of False Positives ($N = 3$) indicates that even in the presence of stochastic recombination, the $\hat{\pi}^{\text{hh}}_{\text{Max}}$ metric remains a robust discriminator of second-degree pedigree structures.

These results correspond to a sensitivity of 96.9% and a specificity of 99.2%, yielding an overall accuracy of 98.9%. The high specificity is particularly significant, as it demonstrates that the model reliably distinguishes half-siblings from avuncular pairs with minimal cross-classification. The low number of False Positives ($N = 3$) indicates that the $\hat{\pi}^{\text{hh}}_{\text{Max}}$ metric remains a robust discriminator of second-degree pedigree structures.

The high classification accuracy (98.9%) achieved by our haplotype-aware GMM framework underscores the power of using across-chromosome phasing to resolve pedigree ambiguities. However, the presence of three false positives (N/A misclassified as HS) and two false negatives (HS misclassified as N/A) warrants further investigation into the edge cases of $\hat{\pi}^{\text{hh}}_{\text{Max}}$-based classification.

The three false positives ($FP = 3$) represent avuncular pairs whose haplotype-sharing signal exceeded the 0.0927 threshold. In these cases, the shared IBD segments likely failed to undergo sufficient recombination across the maternal and paternal homologues, resulting in a sharing pattern that stochastically resembles the single-parent inheritance of a half-sibling. Conversely, the two false negatives ($FN = 2$) may be attributed to two primary factors. First, extreme recombination events can occasionally fragment the IBD sharing of a half-sibling across both parental homologues in the focal individual. Second, switch errors in the initial across-chromosome phasing process can artificially deflate the $\hat{\pi}^{\text{hh}}_{\text{Max}}$ value by incorrectly assigning shared segments to the opposite homologue, as illustrated in the error analysis in Figure **??**.

4 Application to Phasing

Beyond subtyping second-degree relationships, the inferred HS and N/A pairs can also be used to improve both within- and across-chromosome phasing. In particular, different second-degree configurations contribute unequal information about parental origin: some effectively mix maternal and paternal lineages, whereas others constrain sharing to a single parent and therefore act as strong phase anchors. Using relationship labels derived from haplotype-level sharing, phasing algorithms can incorporate additional long-range constraints that link homologous segments across individuals and chromosomes.

Niece/nephew–avuncular pairs offer little direct information for refining the avuncular individual's phase, because the niece or nephew is

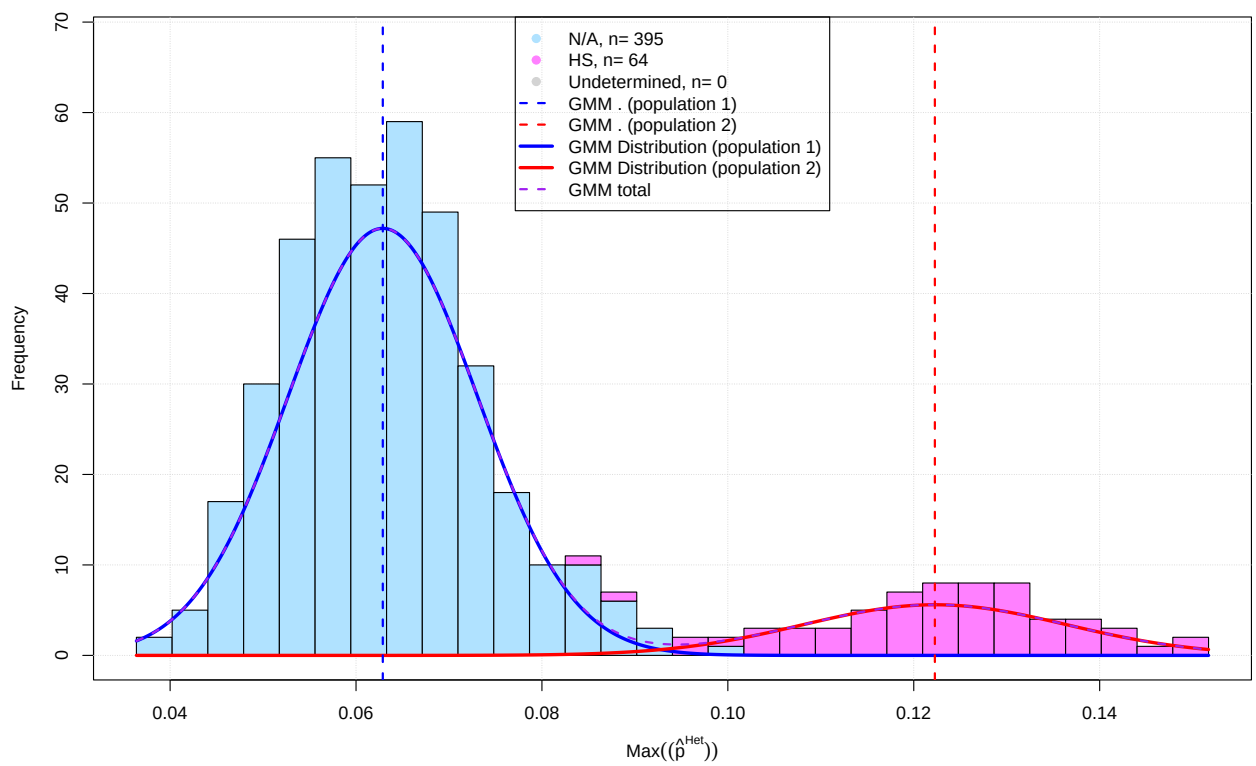

Figure 5: Distribution of $\hat{\pi}^{\text{hh}}_{\text{Max}}$ restricted to ground-truth HS and N/A pairs, with Gaussian mixture components overlaid.

related to both parents of the avuncular individual and the shared genome can originate from either parental lineage. As a result, shared regions between a niece or nephew and an avuncular relative cannot, on their own, be cleanly attributed to one of the avuncular individual's parents. By contrast, half-siblings to each other and avuncular relatives of a given niece or nephew share genetic material through a single parental line, so confidently detected HS pairs and avuncular individuals provide phase-consistent segments that are expected to trace back to the same parental haplotype within each family.

These phase-informative relationships can be incorporated into phasing pipelines in several ways. In this work, they are integrated into the across-chromosome phasing algorithm described in **?** by including detected relatives in the reference population against which the focal individual is compared. In the right panel of Table 2, the avuncular individual is identified as individual i; our results show that ground-truth avuncular relatives are detected with 100% accuracy for the 392 true negative pairs.

For the niece/nephew within each detected N/A pair, the algorithm of **?** is executed in two configurations: first, using a $\hat{\pi}^{(f,i)}$ threshold of 0.1875 to exclude all first- and second-degree relatives from the reference panel, and second, using a threshold of 0.33 to exclude only first-degree relatives, thereby explicitly incorporating the second-degree avuncular relative as a phase anchor. To provide an objective measure of phasing accuracy, we evaluate the results for nieces/nephews in the ground-truth set who have at least one genotyped parent. Crucially, this parent is utilized only as a reference to determine the true within- and across-chromosome phase and is excluded from the phasing process itself, following the validation framework of **?**.

The resulting improvement in phasing metrics when avuncular information is utilized demonstrates that relationship-informed constraints can measurably enhance chromosome-scale phasing in biobank-scale datasets. The across-chromosome phasing accuracy ($ACPA$), as defined in **?**, is calculated for the two configurations (with and without avuncular constraints) for every niece/nephew in the ground truth set, as illustrated in Figure 6.

5 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.

Future work could be to generalize this approach to distinguish pairs of first degree cousins and pairs of half niece/nephew–avuncular or other degree against half degree is an important direction for future work.

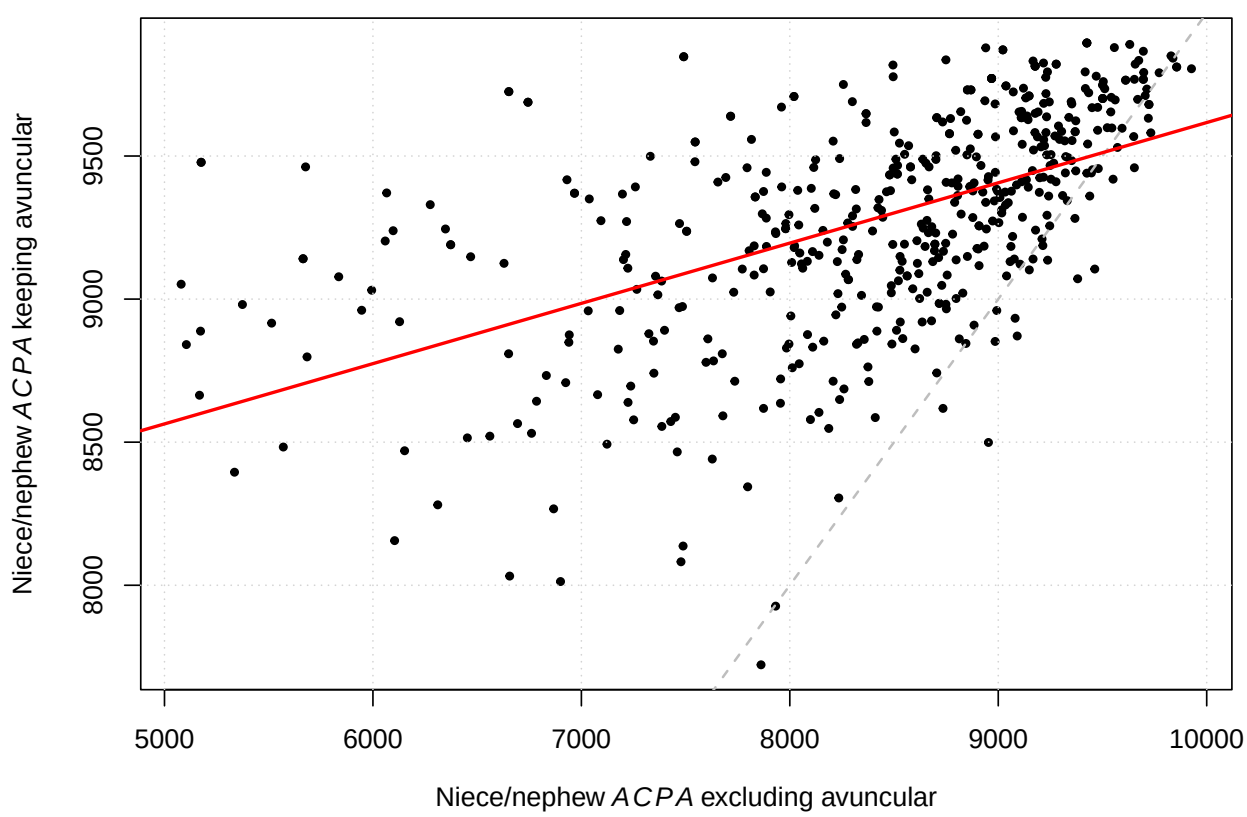

Figure 6: Phasing accuracy improvement for nieces/nephews when using avuncular relatives as phase anchors, compared against ground-truth parental haplotypes.

6 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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007 Sep 1;81(3):559-75.

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.

Andersson G. The role of sibling configurations in the formation of family networks. *Demographic Research*. 2020;43(16):437–464. doi: 10.4054/DemRes.2020.43.16.

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.