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 (H-S) from niece/nephew–avuncular (N/N-A) pairs—both sharing approximately 25% of the genome—remains a significant challenge. Current SNP-based methods rely on aggregate 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 H-S and N/N-A pairs using haplotype-level sharing features derived from across-chromosome phasing. By modeling these features with a multivariate Gaussian mixture model (GMM), we demonstrate exceptional classification performance in biobank-scale data. Ground-truth validation labels were established through a multi-step inference process within a family graph constructed from high-confidence first-degree relationships. To overcome the scarcity of genotyped half-sibling parents, we structurally validated H-S pairs by identifying specific pedigree configurations involving first cousins that are logically incompatible with an avuncular relationship. Our method achieves a sensitivity of 96.9% and a specificity of 99.7% on these structurally-validated pairs. Furthermore, we show that these identified relationships serve as superior phase anchors, that measurably 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 [4]. 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 [5]. Robust relationship subtyping is essential for controlling cryptic relatedness in genome-wide association studies (GWAS), refining demographic models, and enhancing forensic and genealogical investigations [6, 8].

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 (H-S) and niece/nephew–avuncular (N/N-A) pairs [21]. Distinguishing between these configurations is computationally non-trivial because the genome-wide proportion of IBD sharing is stochastically similar across these classes [4].

Accurately distinguishing H-S from N/N-A relationships is critical for several domains of genomic research. First, in heritability and phenotypic analysis, misclassification can lead to biased estimates; half-siblings typically share a more similar childhood environment and possible maternal effects compared to pairs of avuncular and niece/nephew, making it essential to partition these components correctly when modeling shared variance [19, 20].

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

Second, in forensic genetics and missing persons cases, resolving this ambiguity is a primary challenge, as the two relationships imply different pedigree structures for lineage reconstruction [21, 22]. Third, for computational phasing, H-S pairs serve as superior "phase anchors." Because H-S pairs share a parent directly, they provide long-range structural constraints that link haplotypes across chromosomes more effectively than N/N-A sharing, which is fragmented by an additional generation of recombination [23]. Finally, in clinical diagnostics, identifying the exact relationship is necessary for accurate risk assessment and the identification of carriers in recessive disease pedigrees [24].

Current methodologies primarily rely on the distribution of IBD segments; half-siblings theoretically inherit sharing from a single shared parent, whereas niece/nephew–avuncular pairs inherit through an additional meiosis, typically leading to a higher count of shorter segments [5]. Despite these theoretical differences, empirical distributions of segment length and count exhibit substantial overlap, exacerbated by the stochastic nature of recombination [18]. 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 move beyond these aggregate IBD metrics to propose a novel computational framework that utilizes across-chromosome phasing to resolve second-degree relationships. By deriving a 4-dimensional feature vector, we capture the specific inheritance logic of parental homologues. We demonstrate that modeling these features with a multivariate Gaussian mixture model (GMM), fitted via the Expectation-Maximization (EM) algorithm [25], provides near-perfect separation between H-S and N/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 [11], 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 [11]. Genetic principal components (PCs) were computed on a linkage-disequilibrium–pruned set of common autosomal SNPs [5]. 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 327,744 SNPs for relatedness estimation and downstream analyses [11].

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 [1]. 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 [21]. 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 [4]. The exclusion of compound pedigrees [11, 16, 17] relied on the genome-wide IBD2 sharing estimated by PLINK. Since IBD2 is theoretically zero for both N/N-A and H-S 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 H-S and N/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 (H-S)**: pair of individuals sharing one parent.
- **N/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 395 N/N-A pairs but only 5 H-S pairs. To increase the number of H-S examples, additional H-S 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 H-S pairs:

- **Shared first cousin**: pairs in which two individuals ($H-S1$ and $H-S2$) share a first cousin CFC as illustrated in Figure 3. Such pairs cannot be N/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 H-S.
- **Unrelated first cousins**: pairs in which each individual ($H-S1$ and $H-S2$) 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/N-A pair and are therefore classified as H-S, 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 H-S 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.

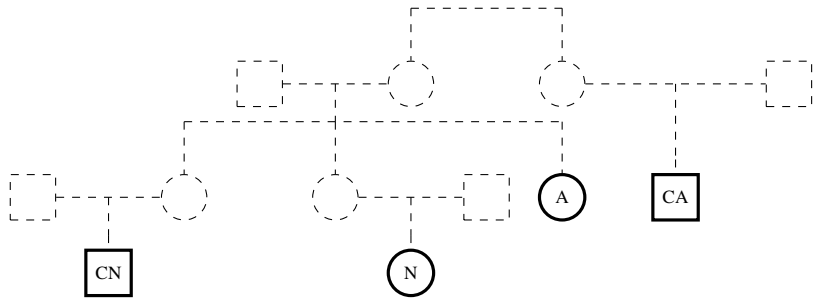

Figure 1: Example pedigree showing a niece/nephew N and their avuncular relative A, with first cousins (CN and CA) used to verify the relationship labels.

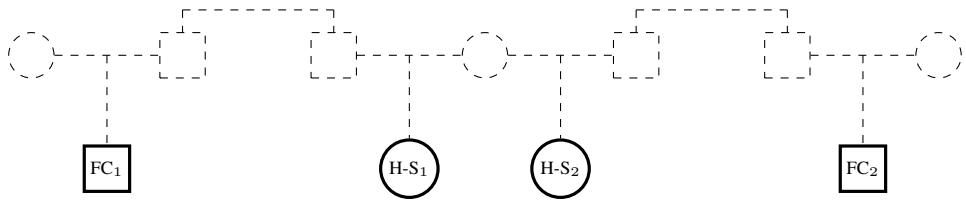

Figure 2: Example pedigree with two half-siblings (H-S_1, H-S_2) sharing a mother and each having a paternal first cousin (FC_1, FC_2), illustrating a configuration that is incompatible with an N/N-A relationship between H-S_1 and H-S_2.

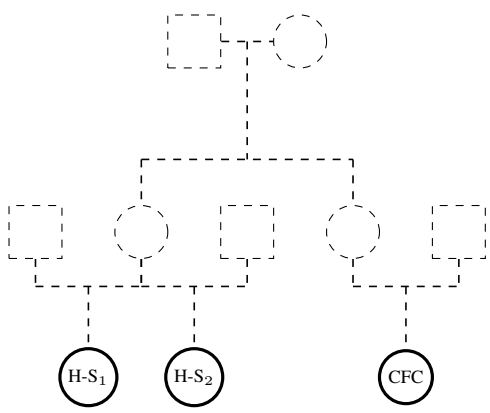

Figure 3: Refined pedigree showing two half-siblings ($H-S_1$, $H-S_2$) sharing a mother and a maternal first cousin (CFC).

2.4 Across-chromosome phasing

All UK Biobank individuals were initially phased within chromosomes using Shapeit2 [11]. To resolve the phase across different chromosomes, we utilized a novel across-chromosome phasing algorithm that employs a window-based SNP-similarity metric as described in **?**]. This method identifies which haplotypes on different chromosomes originate from the same parent by analyzing correlations of SNP-similarity values across genomic windows defined by recombination hotspots. For each window, the algorithm compares each of the focal individual's haplotypes to both haplotypes of every non-focal individual in a sample of the UK Biobank defined in Section 2.1, selecting the maximum similarity to mitigate within-chromosome switch errors in the reference panel. By iteratively merging windows based on a consolidated measure of across-window correlation, the method establishes a consistent parental homologue (p_1 or p_2) across the entire autosomal genome without requiring explicit relative data or IBD segment calling.

2.5 Haplotype–haplotype $\hat{\pi}$ and feature extraction

For each candidate 2nd-degree pair, we define four haplotype–haplotype $\hat{\pi}$ ($\hat{\pi}^{\text{hh}}$) values, corresponding to the sharing between each phased parental homologue of the two individuals. While the standard diploid $\hat{\pi}$ (Eq. 1) measures aggregate similarity, the haploid version evaluates the probability of identity-by-descent between specific haplotypes. For haplotype a of individual i and haplotype b of individual j, we define the haploid SNP similarity as:

$$\hat{\pi}^{\text{hh}}(a,b) = \frac{1}{m}\sum_{k}\frac{(h_{ka}-p_k)(h_{kb}-p_k)}{p_k(1-p_k)} \tag{2}$$

where h_{ka} and h_{kb} are the haploid genotypes (coded as 0 or 1) at SNP k of haplotype a of individual i and haplotype b of individual j, p_k is the reference allele frequency, and m is the total number of SNPs.

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/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 H-S and N/N-A pairs are summarized in Table 2.

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

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

H-S

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

N/N-A

For each candidate 2nd-degree pair, we derive a 4-dimensional feature vector $\mathbf{x} = (x_1, x_2, x_3, x_4)$ from the 2×2 matrix of haplotype–haplotype sharing values ($\hat{\pi}^{\text{hh}}$) calculated only on SNPs where both individuals are heterozygous to maximize informativeness of phased sharing and reduce noise from uninformative loci, because these doubly heterozygous loci carry direct information about which phased allele is shared on which haplotype.

2.6 Multivariate GMM and Posterior Probability $P(\text{H-S})$

To classify pairs, we fit a two-component multivariate normal mixture model to the 4-D feature vectors of the labeled ground-truth set using the Expectation-Maximization (EM) algorithm (Implemented in C and R, available at [26]). The joint density for a pair with feature vector $\mathbf{x}$ is modeled as:

$$f(\mathbf{x}|\boldsymbol{\Theta}) = \pi\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \quad (3)$$

where π represents the mixing proportion, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the 4-dimensional multivariate normal distribution for component $k \in \{\text{H-S, N/N-A}\}$. For every pair with a complete 4-D feature vector, we compute the posterior probability $P(\text{H-S})$ as the probability of the H-S component given the observed vector $\mathbf{x}$:

$$P(\text{H-S}|\mathbf{x}) = \frac{\pi\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{H-S}}, \boldsymbol{\Sigma}_{\text{H-S}})}{\pi\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{H-S}}, \boldsymbol{\Sigma}_{\text{H-S}}) + (1-\pi)\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{N/N-A}}, \boldsymbol{\Sigma}_{\text{N/N-A}})} \quad (4)$$

This value is clamped to the range $[0, 1]$ and serves as the primary metric for relationship subtyping and the identification of high-confidence phase anchors.

3 Results

3.1 Unsupervised cluster identification and validation

The classification of second-degree relatives was performed using an unsupervised framework to minimize bias and avoid circularity in validation. The mixture model EM algorithm identifies two distinct clusters within the 459 ground-truth (GT) pairs based solely on the 4-D feature vector; critically, the relationship labels (H-S or N/N-A) are not utilized by the algorithm during the cluster discovery phase [25].

Following the unsupervised estimation of cluster parameters, the GT labels are utilized to assign the biological identity of each component (e.g., identifying which cluster corresponds to the H-S class) and to estimate the sensitivity and specificity of the approach on this labeled set. This methodology ensures that the performance metrics—showing a sensitivity of 96.9% and a specificity of 99.7%—reflect the model's ability to recover known biological structures without over-fitting to pre-defined labels.

The 4-D multivariate GMM demonstrates near-perfect classification performance when applied to the full UK Biobank candidate set. As shown in Figure 4, the distribution of the posterior probability $P(H-S)$ exhibits sharp bimodal peaks at 0 and 1, reflecting high model certainty in distinguishing between the two pedigree configurations.

Using a stringent classification threshold of $P(H-S) > 0.999995$ to ensure the highest confidence for long-range phasing anchors, the model achieves exceptional separation (Figure 4). Of the 395 ground-truth N/N-A pairs, 394 were correctly classified ($TN = 394$), with only a single pair exceeding the threshold ($FP = 1$). For the ground-truth H-S pairs, 62 out of 64 were correctly identified ($TP = 62$), while two pairs fell below the threshold ($FN = 2$). These results correspond to a sensitivity of 96.9% and a specificity of 99.7% on ground-truth data.

Beyond the ground-truth set, the model identified 800 previously unlabeled pairs in the highest probability bin ($P(H-S) > 0.999995$), while 5,657 unlabeled pairs were predicted as N/N-A (Figure 4). This demonstrates the power of the multivariate approach to recover a substantial number of high-confidence phase anchors from an unlabeled biobank-scale population.

3.2 Impact of across-chromosome phasing on misclassification

The high classification accuracy of the multivariate framework suggests that haplotype-level features are robust discriminators; however, the remaining edge cases highlight the method's dependence on the quality of across-chromosome phasing. Our analysis indicates that misclassifications are primarily driven by noise in the across-chromosome phasing process rather than a failure of the multivariate model or within-chromosome switch errors.

For the single false positive (FP), a validated N/N-A pair yielded an extremely high posterior probability of $P(H-S) = 0.9999966$. This occurs when across-chromosome phasing noise incorrectly assigns shared IBD segments from different chromosomes to the same parental homologue. For an N/N-A pair, independent assortment typically distributes sharing across both parental homologues of the focal individual. However, if the across-chromosome phasing erroneously groups these disjoint shared segments into a single parental strand, the resulting 4-D feature vector will stochastically mimic the single-parent inheritance pattern of a half-sibling, leading to an artificially inflated $P(H-S)$.

Similarly, noise in the across-chromosome phasing is the likely cause for the two false negatives (FN). These validated H-S pairs (with $P(H-S)$ values of 0.999918 and 0.977821) fell just below our stringent 0.999995 threshold. In these instances, the across-chromosome phasing failed to maintain the continuity of the truly shared parental haplotype, effectively distributing the sharing across both homologues in the focal individual. This fragmentation reduces the observed x_1 ($\hat{\pi}_{\text{Max}}$) while increasing the values of the remaining features, thereby deflating the posterior probability of the H-S component.

These results underscore that while the multivariate GMM is highly effective at resolving the biological logic of inheritance, its precision is ultimately capped by the performance of the across-chromosome phasing algorithm. Improving the stability of across-chromosome homologue assignment remains a key area for further refinement in biobank-scale pedigree reconstruction.

3.3 Resolution limits and third-degree relationships

To evaluate the generalizability of the haplotype-aware framework, we applied the $\hat{\pi}^{\text{hh}}_{\text{Max}}$ metric to pairs sharing approximately 12.5% of their genome (third-degree relatives). This degree of relatedness encompasses first-degree cousins (FDC), who share a pair of grandparents, and half niece/nephew–avuncular (Half N/N-A) pairs, who share a single grandparent (i.e., a parent of the half-avuncular is a grandparent of the half-niece/nephew). Similar to the second-degree case, these configurations are often indistinguishable using diploid IBD measures alone.

The difficulty in resolving third-degree relationships can be illustrated by the expected haplotype-sharing matrices, analogous to those presented

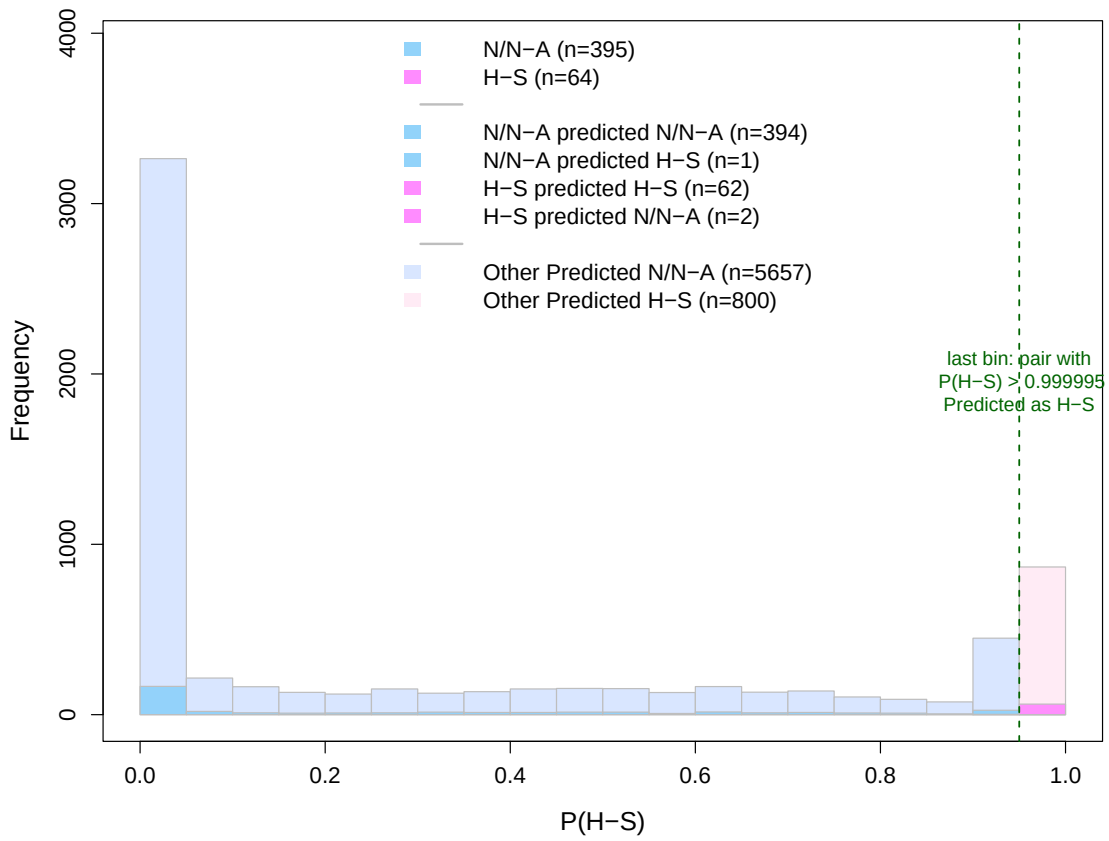

Figure 4: Distribution of the posterior probability $P(H-S)$ for ground-truth and unlabeled second-degree relative candidates. The dashed line indicates the stringent threshold for half-sibling identification; the last bin represents pairs with $P(H-S) > 0.999995$.

in Table 2 for second-degree relatives. We compare the expected $\hat{\pi}^{\text{hh}}$ values for Half N/N-A and FDC pairs in Table 3.

Table 3. Expected haplotype–haplotype $\hat{\pi}^{\text{hh}}$ values for a Half Niece/Nephew–Avuncular (left) and a First-Degree Cousin (right) pair.

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

Half N/N-A

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

FDC

In both cases, the pairs are connected via a relationship between two ancestors sharing 50% of the genome. For Half N/N-A pairs, parent p_1 of individual i and parent p_1 of individual j are in a parent–offspring relationship, while for FDC, parent p_1 of individual i and parent p_1 of individual j are full siblings. Because both parental configurations involve 50% genomic sharing and IBD2 cannot be assessed here as only one haplotype of each parent's individual is known, the resulting $\hat{\pi}^{\text{hh}}_{\text{Max}}$ in the offspring generation is theoretically identical. Without across-chromosome phasing of the grandparental generation to anchor the shared segments to a specific homologue, the expected haplotype sharing remains undiscriminating. This is empirically observed in Figure 6, where a sample of 12,755 pairs (randomly selected from 54,765 total pairs sharing between 9.375% and 18.75% in our sample) shows no bimodal separation in the distribution of $\hat{\pi}^{\text{hh}}_{\text{Max}}$ calculated only on SNPs where both individuals are heterozygous.

4 Application to Phasing

Beyond subtyping second-degree relationships, the inferred H-S and N/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 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

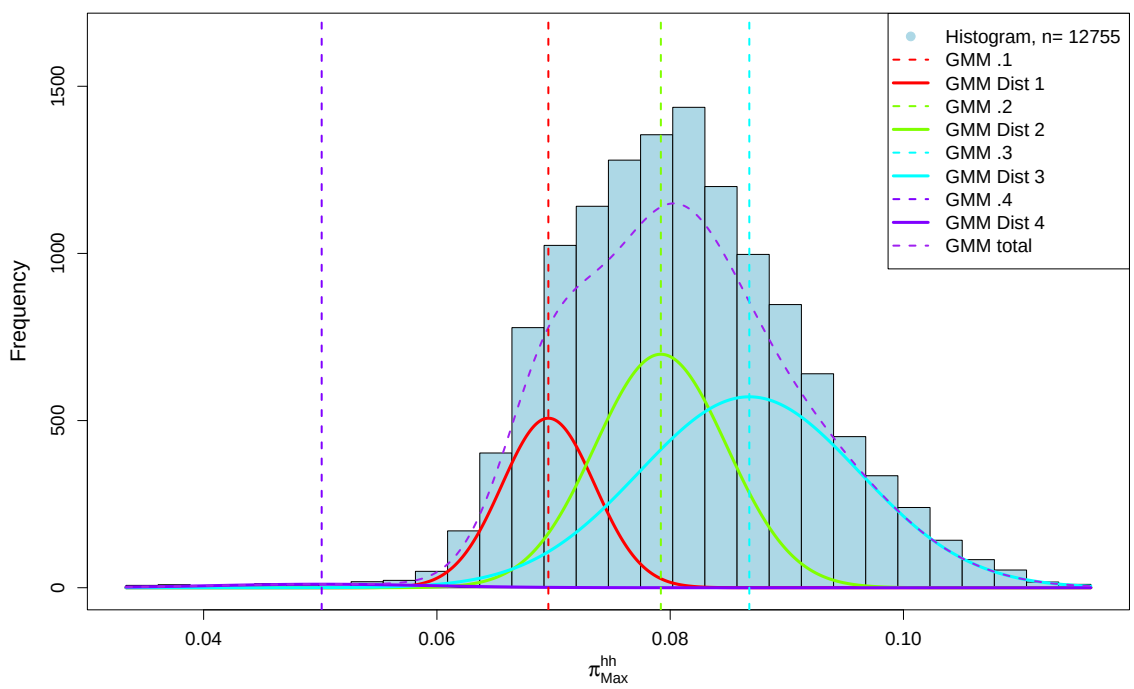

Figure 5: Distribution of $\hat{\pi}^{hh}_{Max}$ for 12,755 pairs sharing between 9.375% and 18.75% of their genome (third-degree relatives).

detected H-S 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. Specifically, this approach can be applied to the niece/nephew of all pairs detected as N/N-A. To assess the resulting improvement, we present the results for nieces/nephews in the ground-truth set where a parent is also genotyped, allowing for direct verification of the phase. As shown in the right panel of Table 2, where the avuncular relative is identified as individual i, our results demonstrate that ground-truth avuncular relatives are detected with 100% accuracy for these 392 N/N-A pairs.

For the niece/nephew within each N/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.

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 demonstrates that a multivariate Gaussian mixture model (GMM) applied to haplotype-level sharing features provides a robust solution for distinguishing half-sibling (H-S) from niece/nephew–avuncular (N/N-A) pairs. By utilizing a 4-dimensional sharing vector derived from across-chromosome phasing, the method effectively leverages the structural inheritance logic of parental homologues to resolve ambiguities that diploid SNP-based methods cannot.

Our results in the UK Biobank show that the model achieves exceptional specificity (99.7%), identifying 800 high-confidence half-sibling pairs and 5,657 high-confidence niece/nephew–avuncular pairs that were previously unlabeled. These results not only clarify the pedigree structure of large-scale biobanks but also provide superior phase anchors that measurably improve the accuracy of across-chromosome genomic reconstruction.

While the current framework provides a scalable, genotype-only solution for large cohortslike the UK biobank. To confirm the generalizability and robustness of the $P(H - S)$ decision boundary, future work should include validation on independent datasets with diverse ancestry and different recombination landscapes. Furthermore, extending this haplotype-aware logic to resolve third-degree relatives remains a critical direction for refining pedigree reconstruction in biobank-scale research.

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.

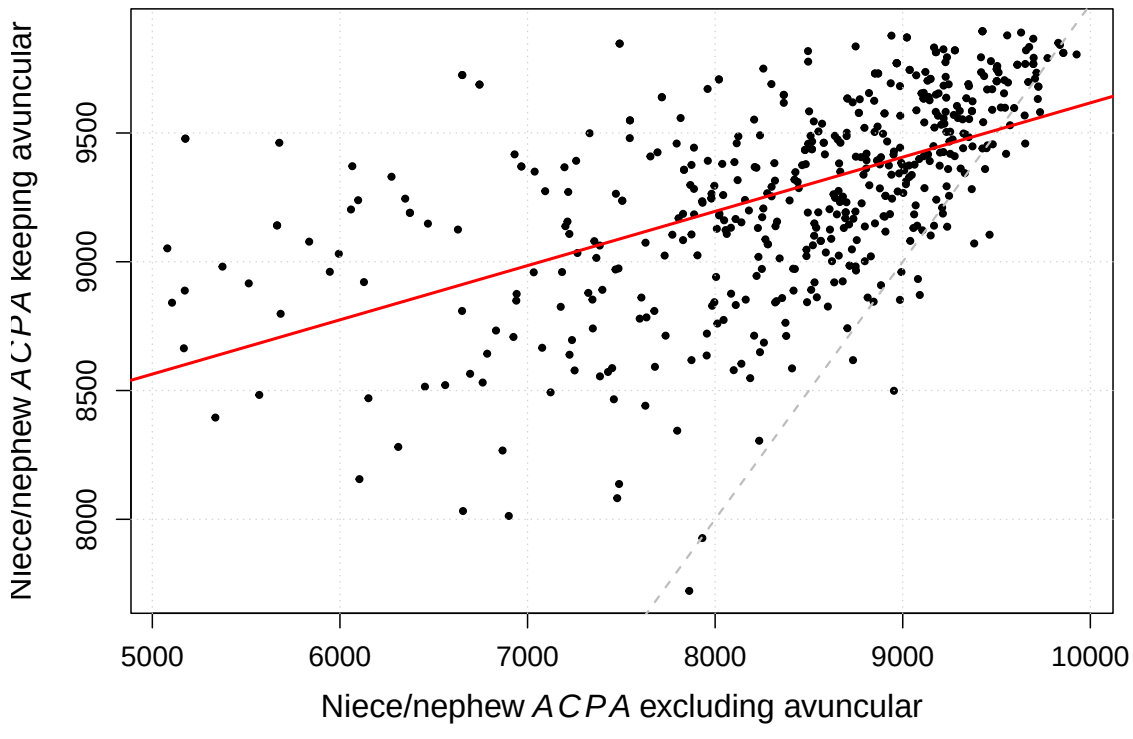

Figure 6: Phasing accuracy improvement for nieces/nephews when using avuncular relatives as phase anchors, compared against ground-truth parental haplotypes.

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

[1]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.

[2]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.

[3]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.

[4]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.

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

[6]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.

[7]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.

[8]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.

[9]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.

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

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

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

[13]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.

[14]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.

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

[16]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.

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

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

[19]Visscher, P. M., *et al.* (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3), e41.

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

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

[22]Kling, D., *et al.* (2012). Identifying individuals by means of their relatives. *Forensic Science International: Genetics*, 6(4), 528–534.

[23]Kong, A., *et al.* (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9), 1068–1075.

[24]Sun, L., *et al.* (2002). Distinguishing distant relationships of unknown etiology using gene identity by descent. *The American Journal of Human Genetics*, 70(6), 1562–1571.

[25]Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

[26]E. Sapin, "N-N-A against H-S: 4-D Multivariate GMM for Pedigree Classification," GitHub repository, 2026. [Online]. Available: https://github.com/emmanuelsapin/N-N-AagainstH-S