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 without pedigree data. Accurately distinguishing half-siblings from avuncular 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 overlap in these distributions leads to high misclassification rates. There is a need for a scalable, genotype-only method that can resolve these second-degree ambiguities without requiring observed pedigrees or phenotypic information.

Results: We present a novel computational framework that achieves near-complete separation of half-siblings and avuncular pairs using haplotype-level sharing features derived from across-chromosome phasing. These features also allow differentiation of nieces/nephews from aunts/uncles within avuncular pairs. By modeling these features with a multivariate Gaussian mixture model (GMM), we demonstrate exceptional classification performance in biobank-scale data. Ground-truth labels for validation were established through a multi-step inference process within a family graph constructed from high-confidence first-degree relationships. All 395 ground-truth avuncular pairs were correctly classified, and 60 of 61 half-sibling pairs were correctly identified. Treating (e.g.) half-sibling status as the positive class, this corresponds to 98.4% sensitivity and 100% specificity. Our results suggest that the framework could perform comparably on other datasets of similar size. 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 closely related individuals for whom the precise pedigree relationship is unknown and must be inferred from genotype data alone [5]. As sample sizes in biobank-scale cohorts have grown to hundreds of thousands of individuals, accurately inferring these relationships has become increasingly important for reconstructing pedigree structure, improving haplotype phasing, and supporting downstream analyses in human genetics, including demographic inference and genealogical applications [6, 7, 9].

A persistent challenge in relatedness inference is distinguishing among relationships beyond the first degree. Parent–offspring and full sibling pairs can be readily distinguished because, although both share approximately 50% of their genome identical-by-descent (IBD), parent–offspring pairs share no genomic segments where both alleles are inherited from the same ancestor (IBD2), whereas full siblings share 25% of their genomes IBD2 as a result of inheriting the same haplotypes from both parents. This clear difference in IBD state patterns makes first-degree relationships straightforward to classify from genomic data. In contrast, several distinct pedigree configurations produce the same expected sharing for second-degree relatives (25% IBD), including half-siblings, grandparent–grandchild, avuncular (niece/nephew–uncle/aunt), and double first-cousin pairs, and the genome-wide patterns of IBD sharing are similar across these relationship classes [22, 5]. Among these, half-sibling and avuncular pairs are particularly important to distinguish in biobank data because grandparent–grandchild and double first cousin pairs are rare in cohorts recruited from a single adult generation.

Accurately distinguishing half-sibling from avuncular relationships is important for several domains of genomic research. Most directly, the two relationships imply different pedigree structures, so distinguishing them improves pedigree reconstruction in large cohorts and other analyses that rely on accurate family structure. Such pedigree reconstruction is particularly important in forensic genetics and missing-person

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

investigations, where lineage inference depends on the correct family configuration [22, 23]. It also has methodological implications for genomic analyses: half-siblings are more effective phase anchors than avuncular pairs because they share a parent directly, linking haplotypes across chromosomes more effectively than avuncular sharing, which is fragmented by an additional generation of recombination [24]. Finally, distinguishing these relationships can improve the interpretation of family-based analyses and clinical genetics. Half-siblings share more similar childhood environments and parental effects than avuncular pairs, and misclassification can therefore bias estimates of shared variance and genetic risk in pedigrees [20, 21, 25].

Current approaches to distinguish half-sibling from avuncular pairs primarily rely on the distribution of IBD segments; half-siblings inherit shared genetic segments directly from their common parent, whereas segments shared by niece/nephew-avuncular pairs have undergone an additional meiosis, typically leading to a higher count of shorter segments [6]. Despite these theoretical differences, empirical distributions of segment length and count exhibit substantial overlap due to the stochastic nature of recombination [19]. Furthermore, while demographic data such as age differences can provide heuristic support, they are often insufficient for high-confidence classification and such data can sometimes be unavailable.

In this paper, we propose a novel computational framework that utilizes across-chromosome phasing to resolve the two most common types of second-degree relationships in biobank data. By deriving a 2-dimensional feature vector, we capture the specific inheritance logic of parental haplotypes. We demonstrate that modeling these features with a multivariate Gaussian mixture model (GMM), fitted via the Expectation-Maximization (EM) algorithm [26], provides near-perfect separation between half-sibling and avuncular pairs.

Finally, we show that these identified relationships can be used as phase anchors, establishing a feedback loop that improves the performance of across-chromosome phasing. This approach offers a scalable, genotype-only solution that eliminates the dependency on explicit pedigree or phenotypic data.

2 Methods

2.1 UK Biobank sample and ancestry definition

The method was developed and evaluated using data from the UK Biobank [12], focusing on individuals of European descent. This choice was motivated by the fact that individuals of European descent constitute the overwhelming majority of the UK Biobank cohort and therefore by the poorer performance of across-chromosome phase estimation in non-European samples [12]. The approach we describe should work equally well in non-European samples of similar size to the UK Biobank. Genetic principal components (PCs) were computed on a linkage-disequilibrium–pruned set of common autosomal SNPs [6]. 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 [12].

2.2 Detection of 2nd-degree relatives

We calculated the standard measures of diploid SNP similarity [2], $\hat{\pi}^{(f,i)}$ and $\hat{\pi}_{\mathrm{IBD2}}(f,i)$, as defined in Equations 1 and 2.

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

In this equation, f denotes the focal individual, i indexes the non-focal individuals, p_k is the minor allele frequency at SNP k, x_{kf} and x_{ki} 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.

$$\hat{\pi}_{\mathrm{IBD2}}(f,i) = \frac{1}{m}\sum_{k=1}^{m}\frac{d_{kf}d_{ki}}{4p_k^2(1-p_k)^2} \tag{2}$$

where d_{kf} represents the dominance deviation for the focal individual f at SNP k (and similarly d_{ki} for individual i), defined conditionally based on the diploid genotype $x_{kf} \in \{0,1,2\}$:

$$d_{kf} = \begin{cases} -2p_k^2 & \text{if } x_{kf}=0 \\ 2p_k(1-p_k) & \text{if } x_{kf}=1 \\ -2(1-p_k)^2 & \text{if } x_{kf}=2 \end{cases} \tag{3}$$

All UK Biobank participants were ages 40-69 at recruitment [12], so true grandparent–grandchild pairs are expected to be rare. Our candidate 2nd-degree relative set is assumed to be dominated by half-sibling and avuncular pairs.

2.3 Ground truth labeling

To obtain labeled pairs 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 $\hat{\pi}_{\mathrm{IBD2}}$ sharing. From these relationships, the following labels could be assigned:

- **Half-siblings**: a pair of individuals sharing one parent.
- **Avuncular pairs**: a pair in which one individual is inferred to be a child of the other individual's sibling.

This procedure led to the identification of 395 avuncular pairs but only 5 half-siblings. To increase the number of half-sibling pairs, we used 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 half-siblings:

- **Unrelated first cousins**: pairs in which each individual ($half\text{-}sib_1$ and $half\text{-}sib_2$) has a first cousin (FC_1 and FC_2) who is unrelated to the other member of the pair (Figure 1). This configuration is incompatible with an avuncular relationship. In an avuncular pair (A, N), any first cousin of A would be a first cousin of N's parent (and a cousin once-removed of N) and therefore would share $\hat{\pi} \approx 0.0625$ with N, precluding unrelatedness (Figure 2). Thus, pairs exhibiting this pattern cannot be avuncular and are classified as half-siblings.
- **Shared first cousin**: pairs in which two individuals ($half\text{-}sib_1$ and $half\text{-}sib_2$) share a first cousin (FC) as illustrated in Figure 3. Such pairs cannot be avuncular, because, as illustrated in Figure 2, an avuncular individual's first cousin (CA) would share $\hat{\pi} \approx 0.0625$ with N rather than $\approx .125$. Therefore, these pairs are classified as half-sibling.

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 half-siblings, 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 when $\hat{\pi} < 0.025$.

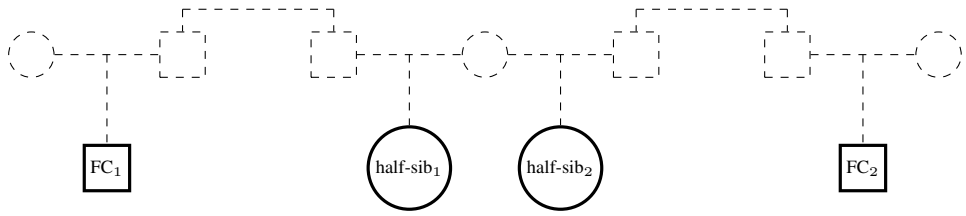

Figure 1: Example pedigree with two half-siblings (half-sib$_1$, half-sib$_2$) sharing a mother and each having a paternal first cousin (FC$_1$, FC$_2$), illustrating a configuration that is incompatible with an avuncular relationship between half-sib$_1$ and half-sib$_2$.

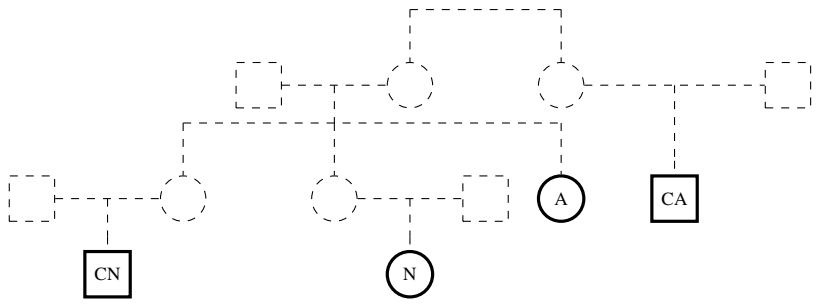

Figure 2: Example pedigree showing that a first cousin (CA) of an avuncular relative A must also be related to the niece/nephew N.

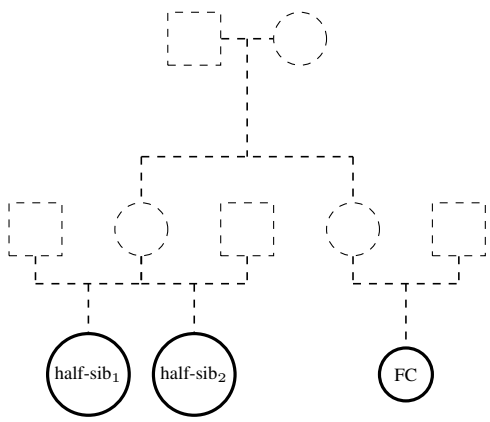

Figure 3: Refined pedigree showing two half-siblings ($half\text{-}sib_1$, $half\text{-}sib_2$) sharing a mother and a maternal first cousin (FC).

2.4 Across-chromosome phasing

All UK Biobank individuals were initially phased within chromosomes using Shapeit2 [12]. To extend phase across chromosomes, we implemented a method that assigns each haplotype to a consistent parental origin by leveraging genome-wide patterns of haplotype sharing [1].

The central idea of cross-chromosome phasing is that, although chromosomes are initially phased independently, haplotypes inherited from the same parent will exhibit systematically higher genetic similarity to the same set of external individuals (i.e., distant relatives) across the genome. This occurs because a common ancestor of the focal and non-focal individuals may transmit IBD segments on multiple chromosomes, and—absent inbreeding—these segments must be inherited through the same parental lineage in the focal individual. As a result, haplotypes from that parent will show elevated SNP similarity to the same non-focal individuals across multiple windows, producing correlated similarity profiles, whereas haplotypes from different parents will not. To capture this signal, the genome is partitioned into windows defined by recombination hotspots. Within each window, each haplotype of a focal individual is compared to both haplotypes of all non-focal individuals in a reference sample. For each comparison, we retain the maximum SNP similarity, which reduces sensitivity to switch errors in the reference panel. This produces, for each focal haplotype and each genomic window, a vector of similarity scores to the reference individuals.

Across windows, haplotypes originating from the same parent will show correlated patterns of similarity to the reference sample, whereas haplotypes from different parents will not. The algorithm therefore computes correlations of these similarity profiles across windows and iteratively groups haplotypes into two clusters corresponding to the two parental homologues (p_1 and p_2). By aggregating information across the genome in this manner, the method establishes a consistent parental assignment across all autosomes without requiring pedigree information or explicit IBD segment detection. Hereafter, we use *phased* to refer to such across-chromosome phased data.

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

For each candidate second-degree pair, we compute four genome-wide haplotype–haplotype similarity values ($\hat{\pi}^{\text{hh}}$), corresponding to all pairwise comparisons between the two phased haplotypes of each individual. Whereas the standard diploid $\hat{\pi}$ (Eq. 1) measures similarity between diploid genotypes, the haploid version evaluates similarity between individual haplotypes, yielding four values per pair.

For haplotype a of individual i and haplotype b of individual j, with $a, b \in \{1, 2\}$, we define the haploid SNP similarity as:

$$\hat{\pi}^{\text{hh}}(h_{ai}, h_{bj}) = \frac{1}{m} \sum_k \frac{(h_{kai} - p_k)(h_{kbj} - p_k)}{p_k(1 - p_k)}, \tag{4}$$

where h_{kai} and h_{kbj} denote the haploid genotypes (coded as 0 or 1) at SNP k for haplotypes a and b of individuals i and j, respectively, p_k is the reference allele frequency, and m is the total number of SNPs across the genome.

These values form a 2×2 matrix of haplotype–haplotype similarities for each pair (Table 1). Because haplotypes are unlabeled with respect to

parental origin, the ordering of haplotypes (1 and 2) is chosen post hoc to facilitate interpretation. Specifically, we relabel rows and columns such that $\hat{\pi}^{\text{hh}}_{\text{Max}}$ is the largest of the four entries, and among the two remaining entries in the same row or column, $\hat{\pi}^{\text{hh}}_{\text{R_Max}} > \hat{\pi}^{\text{hh}}_{\text{R_Min}}$. The final entry, $\hat{\pi}^{\text{hh}}_{\text{Opp}}$, is defined as the value diagonally opposite $\hat{\pi}^{\text{hh}}_{\text{Max}}$.

Table 1. Haplotype–haplotype $\hat{\pi}^{\text{hh}}$ values for a given pair of individuals. Haplotypes are indexed as $1, 2$ for individuals i and j, and are relabeled post hoc so that $\hat{\pi}^{\text{hh}}_{\text{Max}}$ is the largest entry.

		Indiv. i	
		1	2
Indiv. j	1	$\hat{\pi}^{\text{hh}}_{\text{Max}}$	$\hat{\pi}^{\text{hh}}_{\text{R_Max}}$
	2	$\hat{\pi}^{\text{hh}}_{\text{R_Min}}$	$\hat{\pi}^{\text{hh}}_{\text{Opp}}$

When two individuals share a parent (as in half-siblings), the pair of haplotypes inherited from that shared parent is expected to yield $\hat{\pi}^{\text{hh}}_{\text{Max}} \approx 0.25$, while the remaining three haplotype pairs have expected values near 0. In contrast, for avuncular pairs, the two largest entries, $\hat{\pi}^{\text{hh}}_{\text{Max}}$ and $\hat{\pi}^{\text{hh}}_{\text{R_Max}}$, are both expected to be approximately 0.125. This reflects the fact that IBD segments transmitted to the niece/nephew reside on a single haplotype, whereas in the avuncular relative they are distributed across both haplotypes. The remaining two entries have expected value 0. The expected $\hat{\pi}^{\text{hh}}$ patterns for half-sibling and avuncular pairs are summarized in Table 2.

Table 2. Expected haplotype–haplotype $\hat{\pi}^{\text{hh}}$ values for half-sibling (left) and avuncular (right) pairs in perfectly across-chromosome phased data. Haplotypes are indexed as $1, 2$ for the two individuals, with ordering chosen so that the largest entry appears in the $(1, 1)$ position.

		Half-sib$_1$	
		1	2
Half-sib$_2$	1	0.25	0
	2	0	0

		Niece/Nephew	
		1	2
Avuncular	1	0.125	0
	2	0.125	0

For each candidate second-degree pair, we partition the genome into nine groups based on the genotype status—common homozygous (CH), heterozygous (Het), or rare homozygous (RH)—of the two individuals in the pair. Doubly heterozygous (Het-Het) loci increase the informativeness of phased comparisons, as such sites directly indicate whether the same allele is carried on a given haplotype in both individuals, thereby reducing noise from loci that do not inform haplotypic sharing.

Conversely, doubly common homozygous (CH-CH) and doubly rare homozygous (RH-RH) loci are independent of across-chromosome phasing and heterozygosity and can be used to standardize the $\hat{\pi}^{\text{hh}}$ values. Based on these two observations, we define two features:

$$\text{Max_Standardized} = \frac{\hat{\pi}^{\text{hh}}_{\text{Max}}|_{\text{HetHet}}}{\hat{\pi}^{\text{hh}}_{\text{Max}}|_{\text{CHCH}} + \hat{\pi}^{\text{hh}}_{\text{Max}}|_{\text{RHRH}}} \tag{5}$$

$$\text{Min_Standardized} = \frac{\min\left(\hat{\pi}^{\text{hh}}_{\text{R_Min}}|_{\text{HetHet}}, \hat{\pi}^{\text{hh}}_{\text{Opp}}|_{\text{HetHet}}\right)}{\hat{\pi}^{\text{hh}}_{\text{Max}}|_{\text{CHCH}} + \hat{\pi}^{\text{hh}}_{\text{Max}}|_{\text{RHRH}}} \tag{6}$$

From these, we derive a two-dimensional feature vector:

$$\mathbf{x} = (\text{Max_Standardized}, \text{Min_Standardized}). \tag{7}$$

2.6 Multivariate GMM and Posterior Probability $P(\text{half-sibling})$

To classify pairs, we fit a two-component multivariate normal mixture model to the 2-D feature vectors of the labeled ground-truth set using the Expectation-Maximization (EM) algorithm (Implemented in C and R, available at [27]). The joint density for a pair with feature vector $\mathbf{x}$ is modeled as:

$$f(\mathbf{x}|\boldsymbol{\Theta}) = \pi\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \tag{8}$$

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 $k \in \{\text{half-sibling, avuncular}\}$. For every pair with a complete 2-D feature vector, we compute the posterior probability $P(\text{half-sibling})$ as the probability of the half-sibling component given the observed vector $\mathbf{x}$:

$$P(\text{half-sibling}|\mathbf{x}) = \frac{\pi\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{half-sibling}}, \boldsymbol{\Sigma}_{\text{half-sibling}})}{\pi\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{half-sibling}}, \boldsymbol{\Sigma}_{\text{half-sibling}}) + (1-\pi)\phi(\mathbf{x}|\boldsymbol{\mu}_{\text{avuncular}}, \boldsymbol{\Sigma}_{\text{avuncular}})} \tag{9}$$

This calculated probability mathematically falls within the interval [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 identified two distinct clusters within the 459 ground-truth (GT) pairs based solely on the 2-D feature vector; critically, the relationship labels (half-sibling or avuncular) are not utilized by the algorithm during the cluster discovery phase [26].

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 half-sibling class) and to estimate the sensitivity and specificity of the approach for identifying half-sibling pairs (treating the half-sibling class as the positive condition) on this labeled set. This methodology ensures that the performance metrics reflect the model's ability to recover known biological structures without over-fitting to pre-defined labels.

The 2-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 $P(half-sibling)$ exhibits sharp bimodal peaks at 0 and 1, reflecting high model certainty in distinguishing between the two pedigree configurations.

Using a middle classification threshold of 0.5, the model achieves exceptional separation (Figure 4). All 395 ground-truth avuncular pairs were correctly classified ($TP = 395$), whereas 60 out of 61 ground-truth half-siblings were correctly identified ($TP = 60$), corresponding to a sensitivity of 98.4% and a specificity of 100%.

Beyond the ground-truth set, the model identified 513 previously unlabeled pairs ($P(half-sibling) > 0.5$), while 5,963 unlabeled pairs were predicted as avuncular (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.

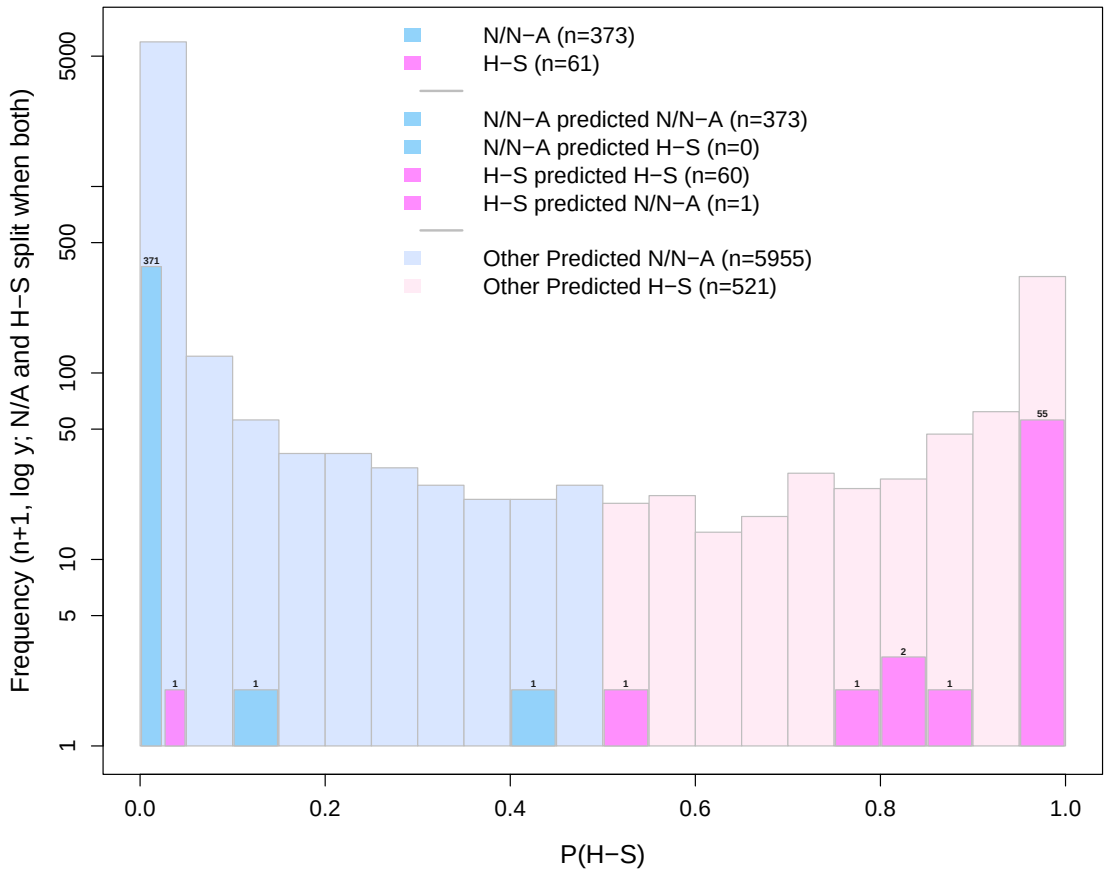

Figure 4: Distribution of the posterior probability $P(half - sibling)$ for ground-truth and unlabeled second-degree relative candidates. The dashed line indicates the threshold for half-sibling identification ($P(half - sibling) > 0.5$).

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 haplotype assignment remains a key area for further refinement in biobank-scale pedigree reconstruction.

4 Application to Phasing

Beyond subtyping, the inferred half-sibling and avuncular relationships can be utilized as long-range phase anchors to improve both within- and across-chromosome phasing. Because half-siblings and nieces/nephews inherit shared segments from a single parental lineage, these regions provide critical constraints for parental haplotype assignment. In contrast, while an avuncular relative's shared regions cannot be cleanly attributed to a specific parent without additional information, they still offer informative links when integrated into the reference panel of a phasing algorithm.

We incorporated these relatives into the across-chromosome phasing framework described in Sapin et al. [1] by adjusting the $\hat{\pi}^{(f,i)}$ exclusion threshold from 0.1875 to 0.33, thereby explicitly including second-degree relatives as phase anchors. For ground-truth nieces/nephews where parental genotypes allowed for direct verification, this relationship-informed constraint measurably enhanced across-chromosome phasing accuracy. These results demonstrate that subtyping second-degree relatives provides a robust, genotype-only feedback loop for refining genomic reconstruction in biobank-scale datasets.

5 Conclusion and Future Work

This study demonstrates that a multivariate Gaussian mixture model (GMM) applied to haplotype-level sharing features provides a robust, scalable solution for distinguishing half-sibling from niece/nephew–avuncular pairs. By utilizing a two-dimensional feature 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 (100%) and sensitivity (98.4%) in distinguishing these relationships.

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. Improving the stability of across-chromosome haplotype assignment remains a key area for further refinement in biobank-scale pedigree reconstruction.

Beyond relationship subtyping, the inferred half-sibling and avuncular pairs serve as powerful phase anchors for genomic reconstruction. By incorporating these high-confidence relatives into the reference population, we established a feedback loop that measurably improves the accuracy of across-chromosome phasing. This approach offers a genotype-only solution that eliminates the dependency on explicit pedigree or phenotypic data, providing a scalable framework for large cohorts.

A critical direction for future work remains the extension of this haplotype-aware framework to resolve third-degree relationships, such as first-degree cousins (FDC) and half-avuncular pairs, who share approximately 12.5% of their genome. These configurations are theoretically identical in their expected haplotype sharing because both involve a pair of parents (one for each individual) being related to the first degree—full siblings in the case of first cousins, and a parent-offspring relationship in the case of half-avuncular pairs. Future research will focus on distinguishing the four "grand-haplotypes" corresponding to the four grandparents to resolve these ambiguities. By identifying which specific grandparental lineage a shared segment originates from, it may be possible to break the symmetry between these third-degree classes. In summary, this study provides a direct application of across-chromosome phasing with immediate implications for the construction of refined pedigrees and the control of cryptic relatedness in large-scale human populations.

In summary, this study provides a direct application of across-chromosome phasing with immediate implications for downstream genomic analyses. The ability to distinguish between distinct second-degree pedigree structures allows for the construction of refined pedigrees, better control of cryptic relatedness in association studies, and more accurate interpretation of shared genetic and environmental variance in large-scale human populations.

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]A Novel Method for Across-Chromosome Phasing without Relative Data Emmanuel Sapin, Kristen M. Kelly, Matthew C. Keller bioRxiv 2026.02.15.706064; doi: https://doi.org/10.64898/2026.02.15.706064

[2]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.

[3]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.

[4]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.

[5]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.

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

[7]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.

[8]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.

[9]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.

[10]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.

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

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

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

[14]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.

[15]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.

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

[17]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.

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

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

[20]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.

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

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

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

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

[25]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.

[26]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.

[27]E. Sapin, "N-N-A against half-sibling: 4-D Multivariate GMM for Pedigree Classification," GitHub repository, 2026. [Online]. Available: `https://github.com/emmanuelsapin/N-N-Aagainsthalf-sibling`