Title: A rapid, accurate approach to inferring pedigrees in endogamous populations

Authors: Cole M Williams,[1,2] Brooke A Scelza,[3] Michelle Daya,[1] Ethan M Lange,[1] Christopher R Gignoux,[1,*] Brenna M Henn[2,**]

Affiliations:

1. Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical School, Aurora, CO 80045.
2. Dept. of Anthropology, and UC Davis Genome Center, University of California, Davis, CA 95616.
3. Dept. of Anthropology, University of California, Los Angeles, CA 90095.

Correspondence: *chris.gignoux@cuanschutz.edu and **bmhenn@ucdavis.edu

Abstract

Accurate reconstruction of pedigrees from genetic data remains a challenging problem. Pedigree inference algorithms are often trained only on urban European-descent families, which are comparatively ‘outbred’ compared to many other global populations. Relationship categories can be difficult to distinguish (e.g. half-sibships versus avuncular) without external information. Furthermore, published software cannot accommodate endogamous populations where there may be reticulations within a pedigree or elevated haplotype sharing. We design a simple, rapid algorithm which initially uses only high-confidence first degree relationships to seed a machine learning step based on the number of identical by descent segments. Additionally, we define a new statistic to polarize individuals to ancestor versus descendant generation. We test our approach in a sample of 700 individuals from northern Namibia, sampled from an endogamous population. Due to a culture of concurrent relationships in this population, there is a high proportion of half-sibships. We accurately identify first through third degree relationships for all categories, including half-sibships, half-avuncular-ships etc. We further validate our approach in the *Barbados Asthma Genetics Study* (BAGS) dataset. Accurate reconstruction of pedigrees holds promise for tracing allele frequency trajectories, improved phasing and other population genomic questions.

1 Introduction

Geneticists have long relied on pedigrees to map disease-causing alleles, uncover modes of trait inheritance, and to construct recombination maps (Ott 1974; Griffiths et al. 1999; Kong et al. 2002; Hall et al. 1990, Thompson et al. 1981). Traditionally, pedigrees for biomedical studies are constructed using self-reported information from study participants. For example, the pedigrees in the San Antonio Mexican American Family Studies (SAMAFS) were constructed using self-reported relationships and verified by kinship inference algorithms (Hunt et al. 2005). As genomic datasets become exponentially larger, the proportion of individuals with close relatives in GWAS datasets or biobanks increases, necessitating the need to detect cryptic relatedness (Ramstetter et al. 2017; Henn et al. 2011). Thus, there is a need to construct pedigrees from genetic data alone. Particularly challenging is distinguishing pedigree relationships within the same degree of relatedness; for instance, grandparent-grandchild, avuncular, and half-siblings are all second degree relatives. If a second degree pair is present in a pedigree, it is crucial to the overall structure of the pedigree to accurately infer their pedigree relationship. Thus, construction of extended pedigrees depends on the ability to accurately infer pedigree relationships.

Most kinship inference programs use some identity by descent (IBD) metric based on allele frequencies or segments (Ramstetter et al. 2017). For IBD proportions, IBD1 and IBD2 measure the proportion of the genome that is identical by descent at one and two alleles, respectively. KING, for example, uses allele frequencies to estimate the proportion of the genome that is IBD1 and IBD2 and estimates the degree of relatedness of a pair, up to fourth degree (Manichaikal et al. 2010). Coefficients of relatedness (r) and kinship coefficients (Φ) refer to whole-genome proportions and are generally less informative than IBD1 and IBD2 values. In contrast, IBD segments, or tracts, identify specific shared haplotypes. A pair share a segment IBD if that segment passes unbroken by recombination to them from their common ancestor (Thompson 2013; Powell et al. 2010). Because IBD segments can be broken by meiotic recombination, the more meioses that separate a pair, the smaller the expected IBD segment. This additional time component imparts IBD segments with more statistical sensitivity than IBD proportions (Ramstetter et al. 2017; Huff et al. 2011; Li et al. 2014; Hill and Whyte 2013). IBD segments can be used to link distant relatives (up to 12th degree relatives in some cases) (Huff et al. 2011; Ramstetter et al. 2017). Importantly, IBD segments can also be used to distinguish pedigree relationships with the same kinship coefficient (Henn et al. 2012; Hill and Whyte 2013).

Grandparent-grandchildren, avuncular, and half-sibling relationships have the same Φ (1/8) and are classified as second degree relatives. Because second degree relatives provide the framework for larger, extended pedigrees beyond nuclear families, distinguishing these pedigree relationships is vital for accurate pedigree construction. Misclassification in pedigree structures can drastically change implications drawn from genetic data. This is especially problematic when the relative linking the individuals is absent from the dataset: for example, an uncle-niece pair where the uncle's full sibling (niece's parent) is missing.

Recent work has attempted to solve this problem. Li et al. developed ERSA 2.0, which takes advantage of the fact that pedigree relationships with the same Φ differ in the number of shared ancestors: grandparent-grandchild share zero ancestors, an avuncular pair shares two ancestors (the uncle/aunt's parents), and half-siblings share one ancestor (the shared parent) (2011). They use a Poisson distribution of expected IBD sharing patterns to distinguish the number of shared ancestors. Similarly, Hill and Whyte use the distribution of IBD segments throughout the genome to distinguish second degree relatives, which differ in the number, lengths and positions of shared segments (2013). The authors report that half-sibling pairs are more likely to have segments covering the ends of chromosomes and that lineal transmission of IBD segments (such as from grandparent to grandchild) differ from avuncular/half-sibling transmission and so can easily be distinguished (Hill and Whyte 2013). Still, considering the genomic properties of IBD segments has its limitations; Hill and Whyte report that considering number, length, and position of IBD segments only has 80% sensitivity. Indeed, considering the number of IBD segments has traditionally been used to distinguish pedigree relationships with the same Φ. Henn et al. used IBD segment number derived from simulations to distinguish avuncular and grandparent-grandchild relationships but did not consider half-siblings (2012). Staples et al. use a different approach in distinguishing relatives in their software PRIMUS (2014). PRIMUS uses IBD1 and IBD2 proportions to estimate the degree of relatedness and iteratively builds pedigrees starting from first degree relatives. PRIMUS builds every possible pedigree and ranks them by their likelihood; thus, PRIMUS can distinguish second degree relatives, but may not be able to without sufficient pedigree information. Most recently, Qiao and Sannerud proposed CREST, which takes advantage of other relatives present in the dataset in order to distinguish half-siblings from avuncular and grandparent-grandchildren (2019). The idea behind CREST is that half-siblings should be related equally to shared ancestors, whereas in a grandparent-grandchild pair, for example, the grandparent should share more IBD with any shared ancestor of the grandchild.

It is important to note that standard methods in the field often ignore endogamy and consanguinity in their discussions of IBD segments and relationship inference (Huff et al. 2011; Shem-Tov and Halperin 2014). Endogamy results in increased IBD sharing, such that the number, length, and position of IBD segments may not reflect the actual pedigree relationship (Hill and Whyte 2013). ERSA's authors, for example, predict reduced performance on endogamous datasets (Huff et al. 2011). The authors of CREST note its limitations for use in endogamous populations because the assumption that half-siblings are equally related to shared ancestors is not necessarily true. However, the performance of IBD-based relationship inference software on endogamous or founder populations has not been thoroughly studied, and so the degree to which pedigree reticulations affect relationship inference is unknown for many of these algorithms. Notably, this affects numerous populations worldwide, and would be very critical to studies of large structured populations, as the norm in multiple geographic regions such as the Near East, South Asia, and Latin America (Nakatsuka et al. 2017; Mooney et al. 2018). Therefore, there is a clear need for tools designed for realistic population settings worldwide that can accommodate elevated levels of relatedness. More so, most current algorithms are symmetric, in that they can identify pairs, but not give polarized assignments, e.g. the parent in a parent-offspring pair.

Here, we introduce PONDEROSA, **P**arent **O**ffspri**N**g pe**D**igree inf**E**rence **RO**bu**S**t to endog**A**my, an algorithm for accurately recovering family-level relationships from genome-wide data. We focus on the case of second-degree relatives, however our framework is easily extended across arbitrary relatedness categories. We begin by creating a supervised set of second-degree relative classes identified through parent-offspring transitivity. Parent-offspring relationships are largely unambiguous, even under human endogamy, given that 100% of the genome should be at least IBD1 between pairs. We introduce the concept of a haplotype-score ratio (HSR), a statistic based on the long-range phasing that is possible in endogamous populations. This long-range phasing information in the context of elevated IBD has previously been leveraged as the basis for novel phasing algorithms (Loh et al. 2016). This information, along with the distribution of IBD segment lengths, then can be used as input features in a machine learning classifier. By providing this information in a supervised machine learning context, we can account for switch errors and elevated background relatedness in a data-driven manner specific to the population of interest. We demonstrate the utility of HSR, along with sex-specific recombination rates, to identify paternal and maternal half-sibs in a highly endogamous population from northern Namibia. We also describe the use of HSR to polarize specific pedigree relationships, accurately recovering the younger or older generation of individuals in parent-offspring, avuncular, and grandparent-grandchildren pairs.

2 Materials and Methods

2.1 Algorithm overview

The first phase of PONDEROSA generates a list of high-confidence relationships—full siblings (FS), grandparent-grandchildren (GP/GC), avuncular (AV), paternal half-siblings (PHS), maternal half-siblings (MHS), cousins (CO), half-cousins (HC), and half-avuncular (HV)—using the assumption that parent-offspring (PO) relationships are the only pairs that can be inferred with full confidence under arbitrary demographic histories (Figure 1). In essence, PONDEROSA traces parent-offspring lineages to find these relationships. Two machine learning classifiers are trained using summary statistics based on these high-confidence relationships. The first classifier utilizes IBD1 and IBD2 values to predict second degree relatives and solve ambiguous sibships (i.e. individuals that share one parent but are missing a second parent and can be either full siblings or half-siblings). PONDEROSA then analyzes the phased identity by descent segments and computes the number of shared IBD segments (*n*) and HSR of each known and putative relative. HSR is a summary statistic that characterizes the haplotype state of shared IBD segments and is used to distinguish half-siblings from avuncular and grandparent-grandchild. Using *n* and HSR of known second degree relatives, a second classifier is trained and used to predict the pedigree relationship of putative second-degree relatives. Using the HSR, the generation of GP/GC and AV pairs is inferred, or whatever relationship pairs the researcher is interested in evaluating in their dataset. Both classifiers are a linear discriminant analysis (LDA) classifier from the Scikit-learn Python package (Pedregosa et al. 2011).

2.2 PONDEROSA implementation

PONDEROSA takes as input IBD segment estimates (either in GERMLINE or iLASH format), as well as pairwise IBD1 and IBD2 values in a KING-formatted file (to define parent-offspring pairs) and a PLINK-formatted .fam file (to define known paths through the pedigree) (Gusev et al. 2009; Shemirani and Belbin 2019; Manichaikul et al. 2010). The user can also supply reported age data (to constrict possible pedigree relationships) and a PLINK .ped file (to more accurately merge IBD segments). PONDEROSA is written in Python 3.7 to take advantage of the scikit.learn() library for classification, although the user need only use it on the command line. A full implementation of PONDEROSA, including source code, is available at: github.com/williamscole/PONDEROSA.

2.3 Generation of high-confidence relationships

PONDEROSA assumes that only parent-offspring can be inferred with perfect confidence because they should be at least IBD1 across the entire genome. Lineal relationships can be inferred by tracing parent-offspring lineages and transitivity; for example, an individual's grandparent is their parent's parent. All of these specific relationships can be inferred with near perfect confidence. A similar approach is described in DRUID in order to initialize the comparison of full siblings to avuncular relationships (Ramstetter et al. 2018). It is important to note that the distinction between half-siblings (or any second degree relative) and FS is straightforward in outbred populations, given that FS have expected IBD2 of 0.25, but is more ambiguous in endogamous populations, as second degree relatives can have inflated IBD2 values. Because of this, PONDEROSA initially defines FS as pairs of individuals sharing both parents (meaning both parents must be present in the dataset) and not by their IBD2 values. Once PONDEROSA has generated FS and second degree relative pairs, it trains the first LDA classifier with their IBD1 and IBD2 values in order to resolve ambiguous sibships, where a pair shares one parent but are both missing a second parent.

2.4 Differences in IBD segment transmission across relationship pairs

A GP/GC pair are expected to share fewer segments but of longer length than an AV pair (Henn et al. 2012). This is because a grandparent and grandchild are separated by two meioses, whereas an aunt and niece are separated by three. A complication arises when using haploid IBD estimation because—depending on where the recombination events occur within the chromosome—an IBD detection algorithm may output either one or two IBD segments. IBD segments can be observed in an individual that would appear to be mosaics in an ancestor. For example, a chromosome transmitted from parent to offspring is a recombinant of the maternal

and paternal chromosome of the parent. This chromosome is a single IBD segment even though it is a recombinant of segments from the parent. Haploid IBD estimation however would consider them to be several segments. Thus, it is disingenuous to calculate the total number of segments by simply adding up the number of IBD segments. PONDEROSA calculates *n* by stitching together IBD segments that overlap or are within 1 cM and have, at most, one discordant homozygote. PONDEROSA does offer flexibility, allowing the user to change the maximum cM gap and the maximum number of discordant homozygotes. We also find that Himba MHS share more IBD segments than PHS, consistent with previous findings and the higher recombination rate in females (Caballero et al. 2019; Kong et al. 2002). Thus, *n* can be used to distinguish MHS from PHS in addition to AV from GP/GC.
Half-siblings differ from AV and GP/GC pairs in the distribution of IBD segments across parental chromosomes (Figure 2). In a GP/GC pair, the GC should inherit all IBD segments shared with the GP from one parent (the child of the GP) and thus all IBD shared with the GP should fall on the same parental chromosome. However, those same IBD segments can be found on both maternal and paternal chromosomes of the GP. Similarly, across all IBD shared in an avuncular pair, all segments should fall on only one parental chromosome of the niece/nephew but can fall on both parental chromosomes in the aunt/uncle. However, both individuals in a half-sibling pair should share all IBD on the same parental chromosome corresponding to their shared parent.

2.5 Haplotype score ratio

We define a new statistic called the haplotype score ratio (HSR) that summarizes pairwise IBD sharing across parental haplotypes. For individuals *i* and *j* we first calculate a haplotype score (*h*) for each as follows:
where is the set of IBD segment lengths on the 0 haplotype of chromosome n and is the set of IBD segment lengths on the 1 haplotype of chromosome n.
The HSR is simply the ratio of h_i and h_j:

For a member of a half-sibling pair, a niece/nephew, or grandchild, the expected *h* is 1, since all IBD segments should be on the same haplotype for each chromosome. For an aunt/uncle or grandparent, *h* should be less than one, but no more than 1/2 as PONDEROSA always picks the larger of the two sums for each chromosome. The HSR of a pair relates the individual haplotype scores by taking a ratio of them, where the denominator is the larger of the two haplotype scores. Thus, because each individual in a half-sibling pair has an expected *h* of one, the HSR should be one. However, for a GP/GC or AV pair, only the grandchild or niece/nephew, respectively, should have *h* close to one; the grandparent or uncle/aunt should have *h* less than one, so the HSR for a GP/GC or AV pair should be significantly less than 1.

2.6 Data generation

We test the performance of PONDEROSA on 678 Himba individuals genotyped on Illumina H3Africa and MEGAex arrays (Scelza et al. 2019; Scelza et al. 2020). The Himba are an endogamous cattle pastoralist population from northern Namibia/ southern Angola numbering approximately 50,000 individuals (Scelza 2011). The Himba diverged from the Herero population in the mid-19th century, possibly accompanied by a population bottleneck, and are distinguished today by a variety of cultural norms (Bollig 1997, Malan 1995). DNA from Himba families was initially collected to address hypotheses about the frequency of extra-pair paternity. Thus, the study design resulted in a large number of half-sibling relationships. We relied on genetic inference of parent-offspring rather the self-reported assertions. After merging the SNP array datasets and extensive QC, we subsequently analyzed 486,000 autosomal SNPs. We also test the performance of PONDEROSA using data from BAGS. In BAGS, families were ascertained through asthmatic probands from Barbados through referrals from local clinics/hospitals, and their nuclear and extended family members were recruited into a family-based asthma genetics study [PMID: 8921368]. As previously described, samples were genotyped on the Illumina HumanHap650Y BeadChip and the African Diaspora Power Chip [PMID: 19910028], yielding 974,505 autosomal SNPs for 684 individuals after merging and QC [PMID: 19910028, 30787307]. The data were phased with SHAPEIT without the **–duohmm** flag, which would constrain possible haplotypes (O’Connell et al. 2014). KING v2.1.5 was used to find parent-offspring pairs and IBD proportions (IBD1 and IBD2) (Manichaikul et al. 2010). We estimated IBD segments using GERMLINE’s **–haploid** flag to preserve phasing (Gusev et al. 2009).

3 Results

3.1 Algorithm

PONDEROSA initially identifies 1551 and 387 Himba and BAGS (respectively) second degree pairs and infers an additional 404 and 130 (respectively) second degree pairs (Table 1). The Himba data has a higher density of genotyped parent-offspring, so 65% of these pairs were connected before PONDEROSA, compared to only 28% of pairs in BAGS. PONDEROSA’s runtime in the Himba was 66s and 18s in BAGS. The haplotype score ratio computation, the rate-limiting step of PONDEROSA, depends on the number of close relatives in the dataset as opposed to the number of individuals. We anticipate that poor phase quality and/or IBD calling that results in an increased number of IBD segments will increase runtime due to the increased time needed for haplotype score ratio calculation. Additionally, the runtime should depend more on the number of close relatives in the dataset than the overall size of the dataset.

3.2 Benchmarking PONDEROSA

To compute the performance of PONDEROSA, ERSA v2.1, and KING v2.1.5, we took a subset of pairs identified through parent-offspring transitivity and compared their relatedness assignments. In assessing the machine learning classification of PONDEROSA, we performed leave-one-out cross-validation.

The subset used included first through fourth degree relatives. We removed third and fourth degree pairs whose IBD1 values were outside 2 SD of the mean IBD1 value for their respective relative class in order to remove pairs that may be related through two parents (e.g. identified first cousins who are also putative second degree relatives will have elevated IBD1 values compared to strictly first cousin pairs). Note that when PONDEROSA trains the LDA classifier, it performs a similar step of removing outlier pairs. We also removed fourth degree pairs that KING and ERSA report as more distantly related than fourth degree so as not to penalize them for underestimating relatedness (PONDEROSA only infers relatedness up fourth degree related and cannot underestimate relatedness for fourth degree relatives).

3.3 Assigning degrees of relatedness

Here we analyze PONDEROSA's ability to assign degrees of relatedness in the Himba compared to KING and ERSA; we do not include its performance in BAGS because BAGS does not have elevated IBD sharing and PONDEROSA, ERSA, and KING all performed well assigning degrees of relatedness.

The first LDA (Supp. Figure 1A) in PONDEROSA is used to assign degrees of relatedness in order to find putative second degree relatives and resolve ambiguous sibships (i.e. distinguishing full-siblings from half-siblings). PONDEROSA outperforms KING and ERSA in assigning degrees of relatedness across first through fourth degree relatives in the Himba with high sensitivity and specificity (Figure 3A). Importantly, PONDEROSA has >97.5% sensitivity and specificity in assigning second degree relatives (Supp. Figure 2). KING has high sensitivity (99.5%) in assigning second degree relatives but low specificity (84%). This is expected because the Himba share more IBD than an outbred population, so existing algorithms should overestimate relatedness. We compared the relatedness assignments of PONDEROSA, ERSA, and KING in order to assess their performance and whether they under-or overestimate relatedness (Figure 4). KING's low specificity is the result of assigning third degree relatives as second degree related; however, its sensitivity remains high because second degree relatives need a high IBD2 value for KING to infer them as FS. KING infers 8 half-siblings as FS, all of which have IBD2 values >0.08 (expected for half-siblings is zero; expected for FS is 0.25). ERSA has a low sensitivity (67%) in assigning Himba second degree relatives but high specificity (99.5%). ERSA's low sensitivity, compared to KING, is because it tends to underestimate relatedness for second degree relatives in the Himba; it maintains high specificity because it neither under-nor overestimates relatedness for third degree relatives: unlike KING, ERSA does not commonly assign third degree relatives as second degree related. Ramstetter et al. report similar findings: ERSA underestimates relatedness for second degree relatives but outperforms KING for third degree relatives. PONDEROSA rarely misclassifies first through fourth degree relatives, but when it does, it underestimates relatedness. For instance, it reports 2% of second degree relatives as third degree and 4.8% of third degree relatives as fourth degree.

3.4 Classifying 2nd degree relatives

PONDEROSA uses an LDA classifier to classify putative second degree relatives as either PHS, MHS, AV, or GP/GC (Figure 5). We calculated the sensitivity and specificity of PONDEROSA and ERSA in assigning the pedigree relationship of Himba second degree relatives, excluding pairs that ERSA inferred as third degree related or as FS (Supp. Figure 3). PONDEROSA greatly outperforms ERSA across HS, GP/GC, and AV (Figure 3B). ERSA is particularly poor in assigning AV pairs, classifying all AV pairs as HS (Supp. Figure 4). Note that ERSA does not distinguish MHS from PHS, so all PHS and MHS pairs were aggregated as HS and PONDEROSA was not penalized for misclassifying the parental sex. To understand PONDEROSA's performance in distinguishing PHS from MHS, we reran the analysis stratifying HS pairs by parental sex. When PONDEROSA correctly assigns an HS pair as HS, it correctly assigns the sex of the shared parent 92.5% of the time.

Next, we looked at the most common pedigree relationship assignments of PONDEROSA for

Himba PHS, MHS, GP/GC, and AV to assess whether outlying HSR or *n* were driving PONDEROSA's misclassifications (Figure 6A). For PHS and MHS, the most common misclassification was MHS and PHS, respectively. This suggests that HSR values for half-siblings are close to expectation "1" and PONDEROSA performs well in distinguishing half-siblings from AV and GP/GC, but that the overlap in distributions of *n* for PHS and MHS drive their misclassification. However, for AV and GP/GC pairs, the most common misclassification was MHS and PHS, respectively, suggesting that PONDEROSA's ability to distinguish AV and GP/GC from HS is limited by high HSR values in some AV and GP/GC pairs. Because GP/GC pairs have the lowest sensitivity due to high HSRs, we next stratified them by sex: PGM (paternal grandmother), PGF (paternal grandfather), MGM (maternal grandmother), and PGM (paternal grandmother) and analyzed their HSR and *n*. If the high HSR of GP/GC pairs was driven by phase error, we would expect these high HSR pairs to be represented equally across all four GP/GC types. Alternatively, we hypothesized that high HSR values were driven by sex differences in the recombination map and that the lower recombination rate in males would elevate HSRs in grandfather-grandchildren. Grandfather-grandchildren do have elevated HSR relative to grandmother-grandchildren and it does not matter whether it is a PGF or MGF (Supp Figure 5A). On the other hand, both the sex of the grandparent and the linking parent affects *n*: PGF share the fewest IBD segments and MGM share the most (Supp. Figure 5B). These differences in HSR and *n* affect PONDEROSA's relationship prediction (Supp. Figure 6). For example, PGF and MGF are most often misclassified as PHS because of their higher HSR. MGM are misclassified most often as AV because of their high *n*. MGF have lower HSR and *n* and therefore have the highest sensitivity, as a GP/GC pair is expected to have the lowest HSR and *n* out of all second degree relatives.

PONDEROSA also performs well for inferring BAGS second degree pairs (Figure 6B; see Supp. Figure 1B for LDA classifier). In the Himba dataset, all relative pairs analyzed are connected through individuals that are genotyped (e.g. the shared mother of an MHS pair is genotyped). However, because the BAGS pedigrees have already been constructed, the dataset also contains relative pairs that are connected through an ungenotyped parent (referred to as a dummy parent). First, we analyzed PONDEROSA's performance across all second degree pairs (i.e. relatives connected by genotyped parents or dummy parents). PONDEROSA correctly identifies >90% of AV and GP/GC pairs (better performance for these categories than with the Himba) but has low performance for HS pairs. Note that there are comparatively few HS pairs (5 PHS and 93 MHS pairs), which likely results in their poor performance.

Next, we specifically analyzed the 120 second degree pairs connected through a dummy parent. These are pairs whose pedigree relationship was determined from questionnaire data but that would otherwise be unknown. Of these 120 second degree pairs, PONDEROSA inferred the correct relationship of 101 pairs (84%). For an additional 16 pairs, the reported relationship was the second most-likely relationship as inferred by PONDEROSA. Therefore, the reported relationship was among the two most likely relationship categories in 97.5% of these pairs.

3.5 Orienting pairs in a pedigree using haplotype scores

Even if a pair's pedigree relationship can be accurately assigned, intergenerational pairs pose a challenge in orienting the individuals in a pedigree (e.g. which, in an AV pair, is the uncle/aunt). Age data can be used but is problematic when it is incomplete, inaccurate or used for orienting non-lineal relationships. For example, an uncle can be younger than his niece. Breaking down the HSR into its individual haplotype scores allows PONDEROSA to orient GP/GC and AV pairs in a pedigree. The genetically younger individual in the pair (i.e. the grandchild/niece/nephew) should have a higher h than the genetically older individual, as we expect all IBD to fall on the same parental chromosome. Using the same logic, individual haplotype scores can be used to orient PO. For instance, if KING finds an isolated PO pair and there is no age data, h can be used to determine which individual is the parent and which is the offspring.

We assessed the ability of the haplotype scores to correctly orient PO, AV, and GP/GC pairs by taking the ratio of the genetically older individual's h and the genetically younger individual's h in known Himba relative pairs (Figure 7). The genetically younger individual should have a higher h than the genetically older individual, so this ratio should be less than one. The generational assignment of individuals in pairs that are less than one are correct. Haplotype scores correctly orient >98% of GP/GC and AV pairs, and 99.5% of PO pairs. Note that most incorrect assignments occur when the difference in haplotype scores is close to zero, suggesting there may be a cutoff where the difference in h is too small to accurately orient a pair. The ability to orient PO pairs depends on the phase quality and thus the dataset; in the Himba, the mean difference in h is 0.22 but is lower in BAGS (0.17).

4 Discussion

PONDEROSA offers a new approach for inferring pedigrees by initially identifying high-confidence relationships already present in the dataset and using them as a training set for a machine learning classifier. The machine learning steps of PONDEROSA are key to its application in diverse human populations. PONDEROSA trains a linear discriminant analysis (LDA) classifier with the IBD1 and IBD2 values of high-confidence second degree relatives from a given population dataset to identify unconnected second degree relatives. This reduces the danger of assigning a second degree relationship to actual third degree relatives. We then use a second LDA classifier to infer the pedigree relationship of second degree relatives. The second classifier uses two data points: the number of IBD segments shared (a commonly used metric) and a novel metric we call a haplotype score ratio (HSR). These IBD and HSR parameters are computed from the actual data such that deviations from an idealized population or technology-specific biases are incorporated into the segment lengths distributions for a relationship pair.

The HSR reflects the how many IBD segments are shared on a single parental chromosome. In essence, the segment score measures how much a second degree pair "looks like" a half-sibling pair. Half-sibling pairs are unique in that they should share all their segments on the chromosomes they inherit from their shared parent. An HSR closer to one is evidence of a half-sibling relationship. Additionally, the HSR can be broken down into individual haplotype scores, which can be used to assign individuals in a pair to their older or younger pedigree generations. This allows PONDEROSA to estimate which individual in a GP/GC or AV pair, for example, is the grandchild or niece/nephew, respectively, without using age data. This is particularly useful for inferred AV pairs, because a niece can be older than her uncle, for example, so age can be unreliable. Lastly, PONDEROSA has the ability to estimate the sex of the shared parent of two half-siblings, taking advantage of the higher recombination rate in females, which leads to maternal half-siblings sharing more IBD segments than paternal halfsiblings. Note that currently PONDEROSA does not use information from sex chromosomes or mtDNA, but we acknowledge that such data should improve PONDEROSA's performance and its integration should be implemented in future versions.

We test our approach on two datasets with different population histories: the endogamous Himba of northern Namibia and an African-descent population from Barbados (BAGS). These datasets are realistic scenarios for PONDEROSA's best use: both have dense family structure present with many PO and FS pairs but also have unresolved close relative pairs.

Our results with the Himba show the importance of our machine learning step in inferring degrees of relatedness: 20% of Himba third degree relatives are inferred as second degree relatives by KING. Considering these pairs as second degree would drastically change the pedigree structure and the implications made from them. Additionally, PONDEROSA has high accuracy in inferring PHS, MHS, and AV. The accuracy can remain high with few training pairs per category, e.g. accuracy increases only marginally as the number of training pairs increases above 10 per relationship category. For comparison, there are between 300 and 400 training pairs per relationship in the Himba dataset we use. These findings show that PONDEROSA will retain its accuracy even in datasets with few training relationships. PONDEROSA also performs well in the BAGS dataset for AV and GP/GC pairs; it has reduced performance for PHS and MHS, likely because the training size is comparatively small.

PONDEROSA's reduced accuracy in inferring GP/GC in the Himba is due to an elevated HSR, which is particularly high in grandfather-grandchild pairs. We hypothesize that this is due to the lower recombination rate in males. Because a GP/GC may share only a single IBD segment on a chromosome, if there is not a recombination event in the region of the segment in the grandparent, it will fall on the same parental chromosome. This will drive the HSR closer to one since this makes the pair indistinguishable from half-siblings at that chromosome. This does not affect AV relationships to the same extent, most likely because AV are more likely to share more than one IBD segment on an average chromosome. However, PONDEROSA's accuracy for GP/GC approaches the 80% accuracy reported by Hill and Whyte's method (2013) and is less of a concern because age data is most useful for inferring GP/GC, as half-siblings—particularly maternal half-siblings—are not likely to have large age gaps.

The challenge of summarizing IBD sharing across parental haplotypes is that phase cannot be maintained between chromosomes. Using population-based phasing (as opposed to trio-based), it is difficult to determine whether two IBD segments on different chromosomes are of the same parental origin. This is why the HSR equation chooses the larger of the two sums for each chromosome (where each sum is the sum of total IBD on each haplotype state), which in turn skews the HSR closer to 1. If phase quality is low, the HSR calculated will not reflect the actual parental state of the IBD segments. In this case, PONDEROSA will struggle distinguishing HS from AV and GP/GC, but the calculation of n (the number of shared IBD segments) is more robust to phase quality. Therefore, even with poorly phased data, PONDEROSA can reduce the number of possible second degree relationships to two. For example, pairs with low n are most likely to be GP/GC or PHS. Additional phenotypic information (e.g. age, existing pedigree structure) can be used to narrow down to one possible relationship. In datasets from populations where half-siblings are markedly rare, n alone may be enough to infer the relationship, or consider an age cohort where AV or GP/GC pairs are rare. We have found that PONDEROSA may work best in endogamous populations, possibly because the lower haplotype diversity allows for better quality phasing. However, our results with BAGS suggests that endogamy may not be a prerequisite for high phase quality.

Our results show the power in considering the haplotype state of IBD segments in inferring pedigree relationships and sex-specific recombination; haplotype state is not considered in—at least currently—any publicly available kinship inference algorithms. Advances in phasing will only improve the performance of PONDEROSA. For example, Tourdot and Zhang (2019) provide a framework for whole chromosome phasing using NGS. As genetic databases increase in size, there is an increasing need to construct extended pedigrees with members of the databases without reliance on self-reported relationships, which may be either absent, incomplete, or inaccurate. PONDEROSA offers a solution for finding second degree relatives and building extended pedigrees. Building pedigrees out of large genetic datasets will enable the study of disease and recombination across a range of populations.

5 Description of Supplemental Data

Supplemental Data includes 7 figures.

6 Declaration of Interests

None declared.

7 Acknowledgments

This work was funded by the NSF (BCS-1534682 to B.A.S.).

8 Web Resources

PONDEROSA is available at https://github.com/williamscole/PONDEROSA.

9 Data and Code Availability

Code availability: PONDEROSA is available at https://github.com/williamscole/PONDEROSA.

Data availability: dbGaP: phs001995.v1.p1 (Himba) and dbGaP: phs001143.v3.p1 (BAGS).

10 References

Astle W, Balding DJ. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statist. Sci*., 24(4), pp. 451-471.

Bollig, M., Schulte, A. Environmental Change and Pastoral Perceptions: Degradation and Indigenous Knowledge in Two African Pastoral Communities. Human Ecology 27, 493–514 (1999). https://doi.org/10.1023/A:1018783725398

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.,* 81(5), pp. 1084-1097.

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.,* 12(10), pp. 703-714

Caballero M, Seidman DN, Qiao Y, Sannerud J, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Carmi S, Williams AL. 2019. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PloS Genet. 15(12): e1007979*.

Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9, pp. 179-181.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. *Am J Hum Genet*, 67(5), pp. 1219-1231.

Griffiths AJF, Gelbart WM, Miller JH, et al. Modern Genetic Analysis. New York: W. H. Freeman; 1999. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21248/

Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*., 19(2), pp. 318-326.

Hall JM, Lee MK. Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988), pp. 1684-1689.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramchandran S, Hon L, Bribin A, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA*. 2011 Mar 29;108(13):5154-62.

Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, Mountain JL. 2012. Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PloS One,* 7(4).

Hill WG, White IM. 2014. Identification of pedigree relationship from genome sharing. *G3,* 3(9), pp. 1553-1571.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, et al. 2011. The landscape of recombination in African Americans. *Nature*, 476(7359), pp. 170-175.

Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, Tuohy TM, Neklason DW, Burt RW, Guthery Sl, Woodward SR, Jorde LB. 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res*, 21, pp. 768-774.

Hunt KJ, Lehman DM, Arya R, Fowler S, Leach RJ, Harald HH, Almasy L, Blangero J, Dyer TD, Duggirala R, Stern MP. 2005. Genome-Wide Linkage Analyses of Type 2 Diabetes in Mexican Americans. *Diabetes*. 54(9), pp. 2655-2662.

Kayser M, de Knijff P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet.,* 12(3), pp. 179-192.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Hallbeck B, Masson G, Shlien A, et al. 2002. *Nat Genet.,* 31(3), pp. 241-247.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, et al. 2008. *Nat Genet.*, 40(9), 1068-1075.

Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, Witherspoon D, Guthery SL, Mauldin DE, Jorde LB. 2014. Relationship Estimation from Whole-Genome Sequence Data. *PloS Genetics*, 10(1).

Loh PR, Palamara PF, Price AL. 2016. Fast and accurate long-range phasing in a UK Biobank cohot. *Nat Genet*. 48(7): 811-816.

Malan G. 1995. Cooperative breeding and delayed dispersal in the Pale chanting goshawk, Melierax canorus. University of Cape Town.

Manichaikul A, Mychaeleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), pp. 2867-2873.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670), pp. 581-584.

Mooney JA, Huber CD, Service S, Sul JH, Marsden CD, Zhang Z, Sabatti C, Ruiz-Linares A, Bedoya G, Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes, Freimer N, Lohmueller KE. 2018. Understanding the Hidden Complexity of Latin American Population Isolates. *Am J Hum Genet*. 103(5): 707-726.

Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S, et al. 2017. The promise of disease gene discoveery in South Asia. *Nat Genet*. 49(9): 1403-1407.

Newman DL, Abney M, McPeek MS, Ober C, Cox NJ. The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet*. 2001;69(5):1146–1148.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, et al. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PloS Genet*. 10(4).

Ott J. 1974. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet*. 26, pp. 588-597.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR* 12, pp. 2825-2830, 2011.

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet*. 11(11), pp. 800-805.

Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Mezey JG, Williams AL. 2017. Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics*. 207(1), pp. 75-82.

Ramstetter MD, Shenoy SA, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Mezey JG, Williams AL. 2018. Inferring Identical-by-Descent Sharing of Sample Ancestors Promotes High-Resolution Relative Detection. *Am J Hum Genet*. 2018 Jul 5;103(1):30-44.

Scelza BA. 2011. Female choice and extra-pair paternity in a traditional human populationBiol. Lett.7889–891. http://doi.org/10.1098/rsbl.2011.0478

Scelza BA, Prall SP, Swinford N, Gopalan S, Atkinson E, McElreath R, Sheehamas J, Henn BM. 2020. High rate of extrapair paternity in a human population demonstrates diversity in human reproductive strategies. *Sci Adv.; 6 :* eaay6195 19 February 2020

Shem-Tov D, Halperin E. 2014. Historical Pedigree Reconstruction from Extant Populations Using PArtitioning of RElatives (PREPARE). *PLoS Comput Biol*. 10(6): e1003610.

Shemirani R, Belbin GM, Avery CL, Kenny EE, Gignoux CR, Ambite JL. 2019. Rapid detection of identity-by-descent tracts for mega-scale datasets. *BioRxiv. doi: https://doi.org/10.1101/749507*.

Staples J, Qiao D, Cho MH, Silverman EK, University of Washington Center for Mendelian Genomics, Nickerson DA, Below JE. 2014. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet*. 95(5), pp. 553-564.

Thompson EA. 1981. Pedigree analysis of Hodgkin's disease in a Newfoundland genealogy. *Ann Hum Genet*. 45(3), pp. 279-292.

Thompson EA. 2013. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*. 194(2), pp. 301-326.

Tourdot RW, Zhang CZ. 2019. Whole Chromosome Haplotype Phasing from Long-Range Sequencing. *bioRxiv.* doi: https://doi.org/10.1101/629337.

Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*. 1(3).

11 Figures Titles and Legends

Figure 1

The three phases of PONDEROSA.

An overview of the three phases of PONDEROSA and the relationships inferred in each.

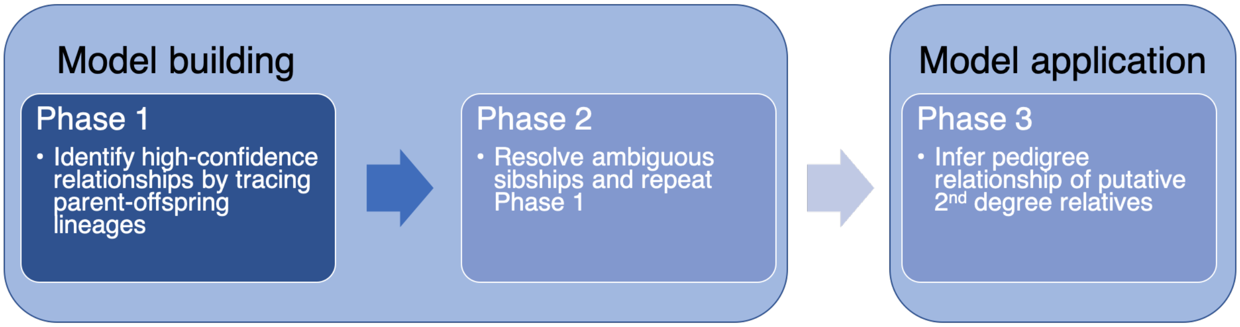

Model building
Phase 1
• Identify high-confidence relationships by tracing parent-offspring lineages
Phase 2
• Resolve ambiguous sibships and repeat Phase 1
Model application
Phase 3
• Infer pedigree relationship of putative 2nd degree relatives

Figure 2

IBD sharing patterns of 2nd degree relatives.
For both individuals in a half-sibling pair (upper left), all IBD should fall on the same parental chromosome. This is only true for one of the individuals in a grandparent-grandchild (lower left) or avuncular pair (right) (the grandchild and niece/nephew, respectively) because the IBD can be on both the maternal and paternal of the grandparent/aunt/uncle.

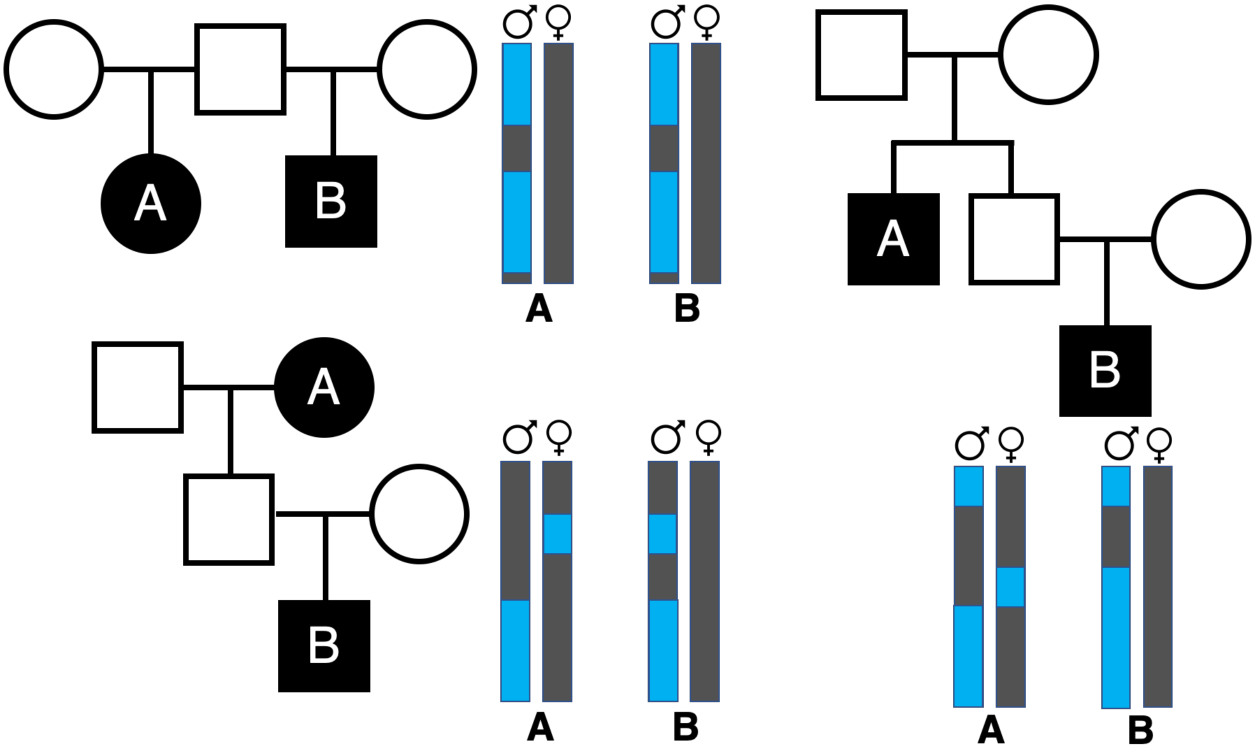


Figure 3

Area under the curve (AUC) of Himba relatedness assignments.
The AUC of PONDEROSA, ERSA, and KING in inferring degrees of relatedness in the Himba dataset. **(A)** PONDEROSA outperforms ERSA and KING across the first 4 degrees of relatedness. **(B)** PONDEROSA has greater accuracy in assigning Himba 2nd degree relatives across relationship categories, as compared to ERSA. The area under the curve (AUC) is not calculated for ERSA for paternal half-sibs (PHS) and maternal half-sibs (MHS) because ERSA does not distinguish between the two.

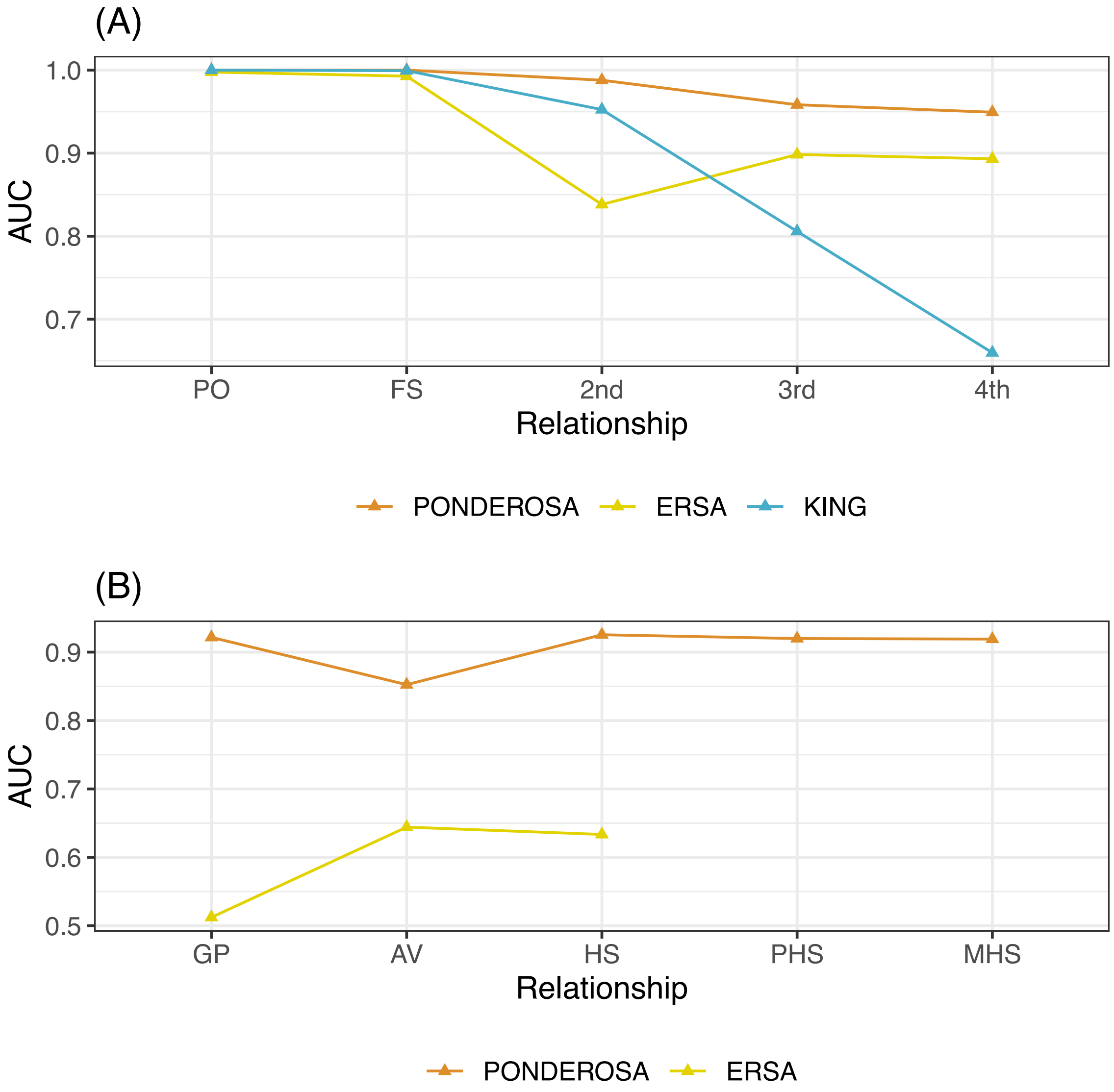


Figure 4

Classification of close relatives in an endogamous population.
We compare how three pedigree assignment software programs (PONDEROSA, ERSA, and KING) perform in terms of accurate assignment of Himba to parent offspring (PO), full siblings (FS), 2nd, 3rd, and 4th degree pairs. Each plot represents the proportion of a relationship type (rows) that is inferred as the true relationship type (columns). For example, the bottom right-hand plot shows the proportion of 4th degree pairs correctly inferred as 4th degree relatives.

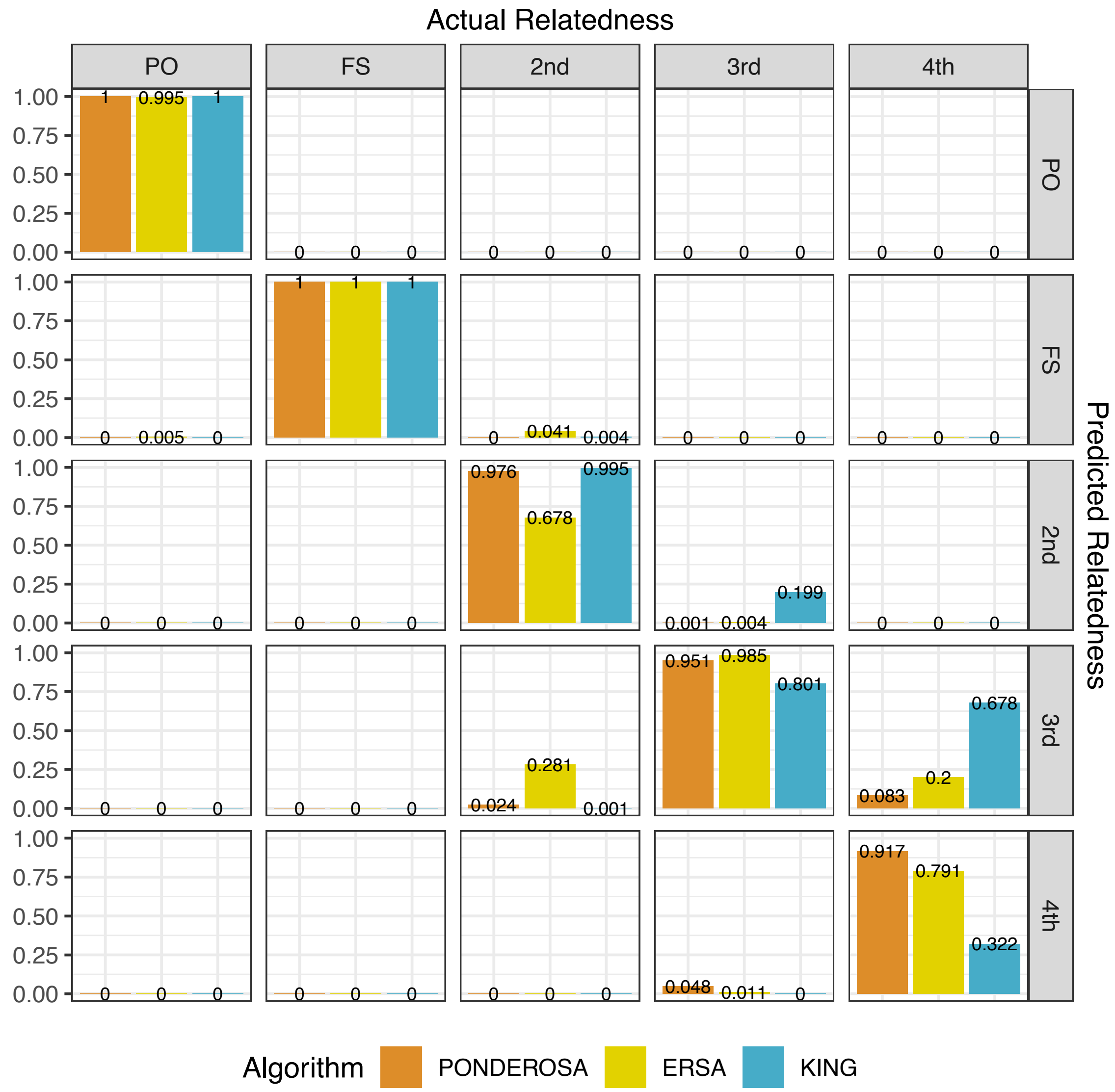


Figure 5

Classification of Himba 2nd degree relatives using LDA classifier.
A graphical representation of the linear discriminant analysis (LDA) classifier from scikit-learn used to infer the pedigree relationship of 2nd degree relatives. Each point is a Himba training pair; their haplotype score ratio (HSR) is plotted on the x-axis and the number of IBD segments on the y-axis. Points marked *x* have been incorrectly inferred by the classifier. The black lines separate regions of the plot belonging to each relative class.

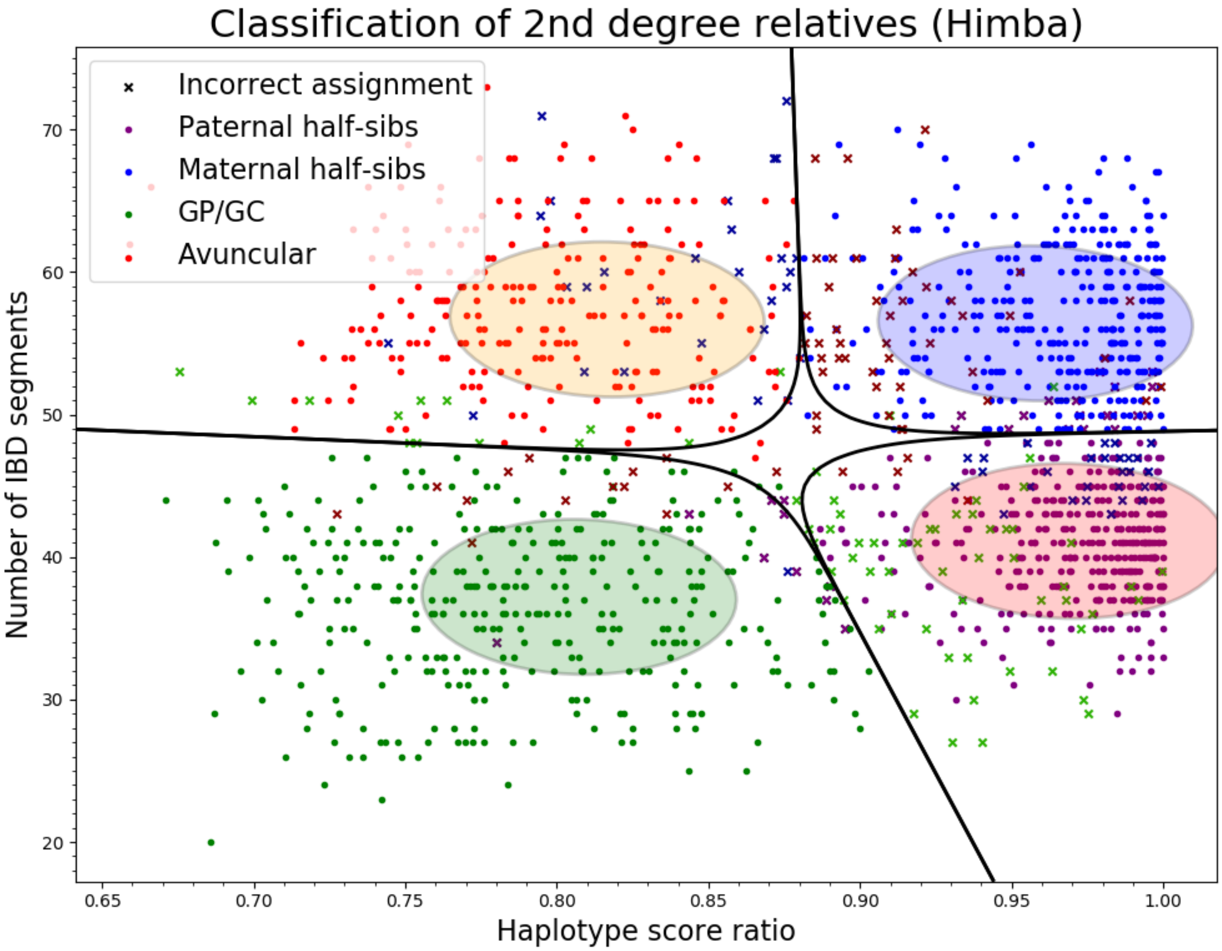


Figure 6

PONDEROSA classification of 2nd degree relatives from two African Datasets. Each plot represents the proportion of a relationship type from Phase 1 and 2 (column) that is classified into the relationship category listed by row. A) In the Himba dataset, ≥80% of all pairs in each 2nd degree category are correctly assigned. Misassignments are shown off-diagonal, e.g. 6.9% of true Himba maternal half-sibs (MHS) are classified as paternal half-sibs (PHS). B) In the BAGS dataset, >90% of grandparental and avuncular pairs are correctly assigned. However, half-siblings are less likely to be accurately identified, potentially due to lower sample size, e.g. 32.1% of true maternal half-sibs (MHS) are classified as avuncular (AV).

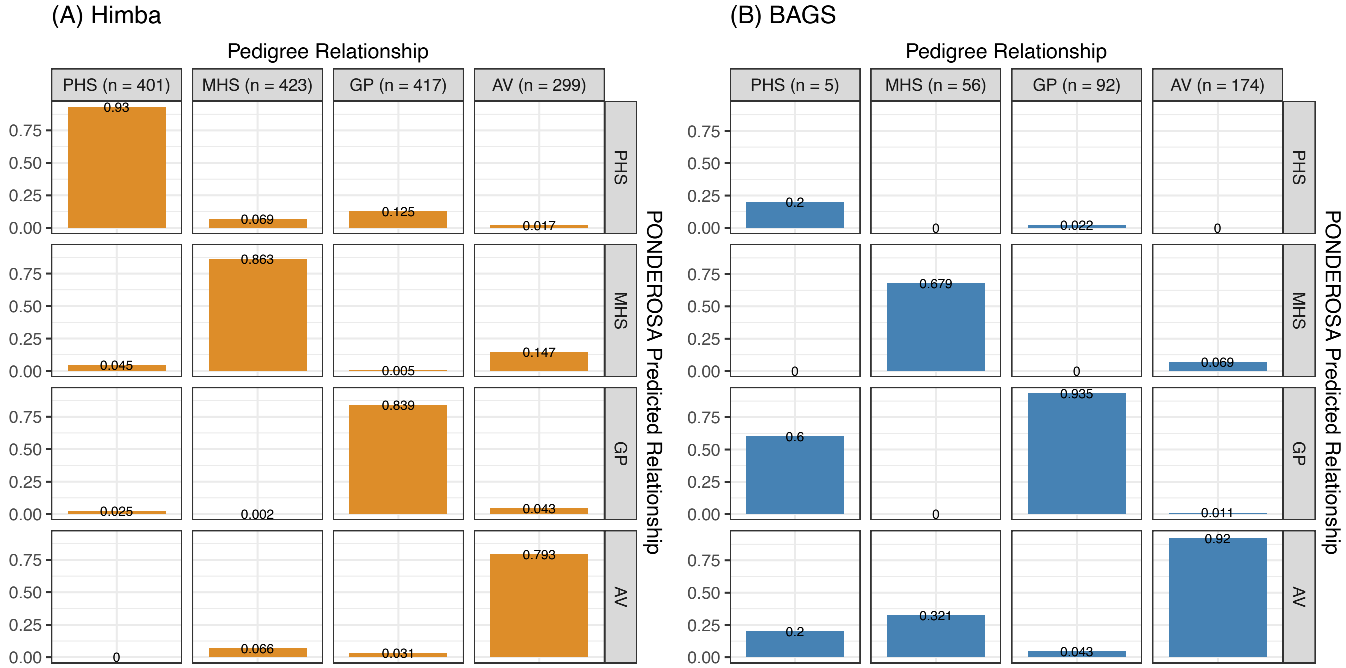
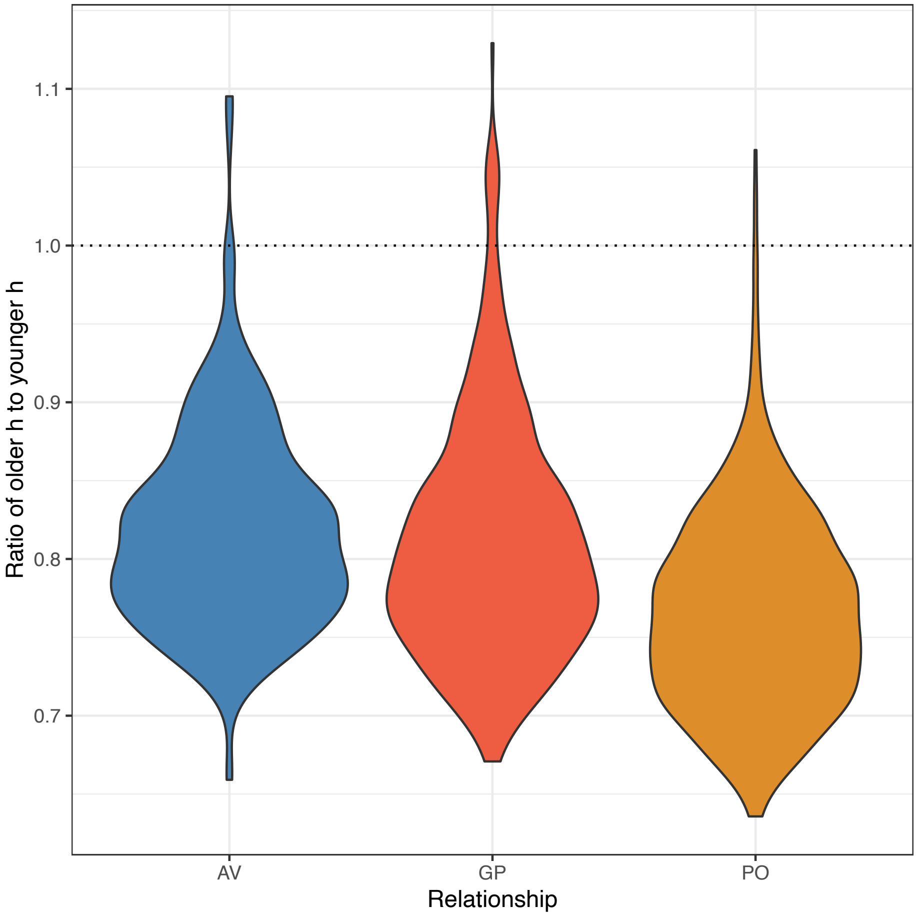


Figure 7

Assignment of individuals within a pair to older or younger generation.
For all AV, GP/GC, and PO Himba pairs, the ratio of the older individual's haplotype score (i.e. the parent/grandparent/uncle/aunt) to the younger individual's haplotype score is plotted on the y-axis. The haplotype score of the older individual is expected to be less than that of the younger individual, and thus the ratio should be less than one. More than >98% of GP/GC and AV pairs fall below one and 99.5% of PO pairs fall below one, demonstrating the accuracy in using haplotype scores to predict the orientation of inter-generational relative pairs.



12 Tables

Table 1

Relatives classified in each PONDEROSA step

	PO		FS		MHS		PHS		AV		GP/GC	
	Himba	BAGS	Himba	BAGS	Himba	BAGS	Himba	BAGS	Himba	BAGS	Himba	BAGS
Phase 1	655	693	106	59	312	35	399	5	146	14	417	92
Phase 2	655	693	178	196	424	56	411	5	299	234	417	92
Phase 3	655	693	178	196	476	72	563	6	413	322	503	117