Title

Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations

Authors

Michael D. Martin[1,2], Flora Jay[2,3], Sergi Castellano[4], Montgomery Slatkin[2]

[1] Dept. of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), 7041 Trondheim, Norway

[2] Center for Theoretical Evolutionary Genomics, Dept. of Integrative Biology, University of California Berkeley, 3040 Valley Life Sciences Building, Berkeley, California 94720, USA

[3] Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud, Paris-Saclay, France

[4] Dept. of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Corresponding author: Michael D. Martin
NTNU University Museum, Norwegian University of Science and Technology (NTNU)
7041 Trondheim, Norway
E-mail: mike.martin@ntnu.no
Fax: +47 73 59 21 45

Running title: Genomic relatedness from pedigree simulations

Keywords: Relatedness, Pedigree, Humans, Ancient DNA, Computer simulation, Genome, Genomics, Genetics, DNA, Single nucleotide, Polymorphism

Abstract

We develop and evaluate methods for inferring relatedness among individuals from low-coverage DNA sequences of their genomes, with particular emphasis on sequences obtained from fossil remains. We suggest the major factors complicating the determination of relatedness among ancient individuals are sequencing depth, the number of overlapping sites, the sequencing error rate, and the presence of contamination from present-day genetic sources. We develop a theoretical model that facilitates the exploration of these factors and their relative effects, via measurement of pairwise genetic distances, without calling genotypes, and determine the power to infer relatedness under various scenarios of varying sequencing depth, present-day contamination, and sequencing error. The model is validated by a simulation study as well as the analysis of aligned sequences from present-day human genomes. We then apply the method to the recently published genome sequences of ancient Europeans, developing a statistical treatment to determine confidence in assigned relatedness that is, in some cases, more precise than previously reported. As the majority of ancient specimens are from animals, this method would be applicable to investigate kinship in non-human remains. The developed software *grups* is implemented in Python and freely available.

Introduction

Genetic relatedness among individuals is a fundamental aspect of human society upon which many of our laws, traditions, and social structures are based. But as such a general concept, precise descriptions of genetic relatedness are elusive, and various definitions have been proposed. For our purposes, related individuals share at least one allele that is recently identical by descent (IBD). As members of a homogenous population are ultimately related by their ancient, common genetic ancestry, we examine only relationships whose genetic affinity is beyond the background relatedness of an entire population.

It follows from basic statistical genetics that closely related individuals share large fractions of their genomes IBD. For example, on average siblings are expected to share in common 50% of their genomes IBD. This degree of relatedness can be expressed as r, the coefficient of relationship (*e.g.* $r = 0.5$ in the case of sibling and parent-offspring relationships), which expresses the probability that, at a given locus, an allele randomly selected from each of two individuals will be shared by IBD. But this measure of relatedness is only a probability; due to the stochastic nature of meiotic recombination, the actual proportion of the genome that is shared between relatives varies across the genome and between pairs with the same kinship (Weir *et al*. 2006; Speed & Balding 2014).

Precise quantification of relatedness is useful in a variety of cases. For example, in the field of forensic genetics it is routine to employ microsatellite or single nucleotide polymorphism (SNP) markers to determine the relatedness of genetic samples for paternity testing and to search for matches to genetic profiles attained from crime scenes (Evett & Weir 1998; Weir *et al*. 2006). An accurate quantification of inter-sample

relatedness is also a prerequisite for most population genetic analyses, which often establish the probabilities of observing sampled genotypes under the assumption that individuals are genetically unrelated.

Due to rapid progress in next-generation sequencing (NGS) technologies, large quantities of genomic sequences can now be obtained quickly and relatively cheaply. There exist various methods to quantify individual relatedness using genomic sequences (*e.g.* see Wang [2011] for a non-exhaustive list). The simplest of these calculate genome-wide averages across a panel of single-SNP haploid or diploid genetic distances (Tal 2013) or allele-sharing coefficients (Pemberton *et al*. 2010; Speed & Balding 2014). More advanced methods identify IBD regions as shared haplotype segments within densely spaced, preferably unlinked genomic markers, and then infer relationships from the total proportion of IBD (Purcell *et al*. 2007; Browning & Browning 2007; Gusev *et al*. 2009; Kong *et al*. 2008; Browning & Browning 2010; Hill & White 2013). The most advanced methods can detect distant relationships up to the 9th degree, yielding probabilities of relatedness conditional on the total IBD, as in previous methods, as well as additional information including the number of IBD chromosomal segments, their lengths, and the genotypes they contain (*e.g.* Albrechtsen *et al*. 2009; Huff *et al*. 2011; Li *et al*. 2014). Useful as these methods are, most require high quality, high-depth sequences that are not always available, particularly for sequences obtained from fossils in which little endogenous DNA remains.

NGS approaches can also be applied to degraded, archival, and/or ancient DNA (aDNA) samples, especially since the advent of *in situ* hybridization capture technology (*e.g.* St John & Quinn 2008; Briggs *et al*. 2009; Bahcall 2013). Indeed, numerous recent studies reported low-coverage genomic sequences from ancient human specimens (*e.g.*

Lazaridis *et al*. 2013; Haak *et al*. 2015; Allentoft *et al*. 2015), and many more still are underway. This trend is due in part to advances that mitigate *post mortem* chemical damage of DNA and contamination from present-day DNA, both of which introduce errors when calling genotypes from reference-aligned sequences and have created problems in the field since its earliest days (Willerslev & Cooper 2005; Briggs *et al*. 2007; Ginolhac *et al*. 2011; Shapiro & Hofreiter 2012).

Accurate determination of relatedness within ancient human archaeological contexts could elucidate pre-historic family structure as well as social behaviors such as burial practices, insights that could not be determined without genetic data. But previous application of these relatedness determination methods to degraded/ancient samples has been limited, usually leading to the exclusion of related individuals from further analysis (Green *et al*. 2010; Lazaridis *et al*. 2015; Haak *et al*. 2015). Without a statistical model to account for the presence of contaminating DNA from present-day humans, precise inference of individual relatedness is complicated due to the unknown consequences of contamination on relatedness estimation, and often these contaminated samples are also excluded from further analysis (Hofreiter *et al*. 2001; Yang & Watt 2005).

These more recent datasets motivate the development of methods to determine relatedness within real archaeological contexts. In this paper, we seek to determine the limitations for discriminating close genetic relationships using very low-coverage NGS sequences. Although genome-wide average genetic distance is perhaps the crudest statistic for estimating relatedness, it can be determined directly from randomly sampled NGS sequences and may be a good option for low-coverage datasets.

We carried out simulations of simple pedigrees using present-day human genome sequences, allowing us to characterize the distribution of genetic distances for several familial relationships under realistic scenarios of contamination, sequencing error, and sequencing depth. We aimed specifically to determine the upper bounds of allowed sample contamination and lower bounds of NGS sequence depth necessary to infer the degree of relatedness between individual ancient humans. We validated the simulations first with theory and then with direct measurements of pairwise genetic distance from aligned NGS data. Finally, we used simulations to verify relatedness claims from previously published ancient human genomic sequences from European archaeological samples dated to between 5,311 and 1,780 BCE (Haak *et al*. 2015), determining the most likely coefficient of relationship and, in some cases, revealing relatedness that had gone unreported.

Materials & Methods

Genomic sequences for pedigree simulations

For pedigree simulations, we used 77,818,345 diallelic single nucleotide polymorphisms (SNPs) in the autosomes of the human reference genome (build 37) and tracked their transmission from parents to offspring. The occurrence of these SNPs in present-day populations and individuals were obtained from the phase 3 data release of the 1000 Genomes Project (G1K hereafter; Abecasis *et al*. 2010; Abecasis *et al*. 2012). Insertion/deletion variants were ignored. Phased genome sequences of 503 unrelated individuals from the European (EUR) super-population were included in pedigree simulations. Allele frequencies in the African (AFR) super-population were used for simulations of contamination from a different population.

Pedigree simulations using present-day human genomic sequences

Simulations were carried out using custom Python scripts. In each simulation replicate, we constructed a family pedigree starting with the unrelated EUR individuals. Each 'mating' of two individuals generated an offspring individual with 22 autosomal chromosomes produced by recombination of parent haplotype sequences randomly selected from all possible combinations of gametes. The probabilities of recombination events in intervals along each chromosome were non-uniform and determined using a genetic map (Kong *et al*. 2002; IHMC 2007). Offspring individuals were then 'mated' with other individuals to produce offspring in accordance with the desired pedigree (**Figure 1**). Replicates (n = 1000 in this study) of the same pedigree were generated through random selection of the unrelated initial mating couple. The per-generation

mutation rate was assumed to be small enough to ignore, and transmission of sex chromosomes was not simulated.

Simulation parameters and calculations of pairwise genetic distance

Input to the model includes seven major parameters that can be modified at run-time. (a) The source population (*e.g.*, the European super-population EUR) from which random genome sequences are selected for pedigree simulation. (b) The minimum allele frequency of chosen SNPs in the source population in order for a genome position to be included in the SNP panel. (c) The rate at which genome positions harboring known SNPs (within the set of all G1K individuals) are randomly selected, which is used to construct a SNP panel of particular size. (d) Individual-specific mean sequencing depths at targeted genome positions, modeled as the rate of a Poisson distribution. Only sites with simulated sequence depth ≥ 1 in both individuals can be considered in calculation of pairwise differences, thus modification of this parameter affects the number of sites where overlapping sequences enable an assessment to be made. (e) Individual-specific sequencing error rates, modeled as the rate at which the observed nucleotide is not correct, with the erroneous nucleotide chosen with equal probability from the three remaining possibilities. (f) Individual-specific rates of contamination by user-selected super-populations, simulated by randomly sampling alleles at a contamination rate c from a pool with the allele frequencies of the contaminant super-population. Contamination by a specified number of individuals randomly chosen from the contaminant super-population is also implemented. (g) A heterozygosity down-sampling parameter that randomly chooses SNPs for which to reduce the minor allele frequency in the pedigree population to zero, enabling simulation of pedigrees from

populations with mean heterozygosity lower than that of the simulation source population.

Each individual-specific parameter for sequencing error rate, contamination rate, and mean sequencing depth can also be expressed as a range. In this case, during each pedigree replicate, the model generates a user-selected number of parameter replicates. For each parameter replicate, the simulated values of these parameters are randomly selected from uniform distribution within a range input by the user. In our explorations of uncertainty in sequencing error, five random parameter replicates were generated for each pedigree simulation (n = 1000), generating distributions containing a total of 5000 replicates.

Once genetic data are simulated for a given scenario, the model computes pairwise differences between individuals with particular relationships and reports genetic distance as the mean number of mismatches at the randomly selected panel of variant positions. The distributions of simulated genetic distances were used to assess power to discriminate between different coefficients of relationship r under various scenarios. As the simulated genetic distances could not necessarily be assumed to conform to normal distributions, we chose to estimate overlap between distributions of simulated genetic distance for each pair of relationships with the Bhattacharyya coefficient. The possible values of this coefficient exist in the range 0 to 1 and indicate 0 to 100% overlap (Bhattacharyya 1947):

$$BC(v, s) = \sum_{i=1}^{n} \sqrt{v_i s_i}, \qquad \text{Eq. (1)}$$

where v and s are the distributions under comparison n is a chosen number of bins, and v_i and s_i are the number of samples falling within bin i. We chose a somewhat arbitrary number of uniform bins equal to the number of combined data points in the pair of distributions, divided by 10 (*e.g.* $n = 2000/10 = 200$). In these pairwise tests of relationship overlap, a cutoff value of BC ≤ 0.05 was used to determine significant separation of two relationships. For relationships with identical expected values of r, statistical testing and comparisons were performed conservatively on the relationship with the largest actual variance.

Theoretical expectations

The joint probability of genotype pairs at a diallelic locus within an outbred population are summarized by Slatkin (2008). Considering an ancient and a contaminating present-day population, the joint probabilities of sampling genotype pairs (one from each population) depend on the ancient population allele frequency p_A and the contaminating population allele frequency p_C. Applying Hardy-Weinberg expectations, we demonstrate the derivation of a formula for E[ ] (E[M]),the expected value of the probability M of observing mismatching nucleotides between single sequences sampled from two samples from the same population—each contaminated at known rates c_1 and c_2 by a contaminating population and subject to sequencing error q (Eq. 9, Supplementary Materials & Methods). When $q = 0$ and $c_1 = c_2 = 0$, the mean expected value of pairwise genetic distance assessed between identical twins or between two samples generated from the same individual, reduces to = (1 −) ($M_S = p_A (1 - p_A)$). Similarly, the expected value for parent–offspring and sibling–sibling relationships is $\frac{3}{2}$ (3/2 M_S). The expected value for grandparent–grandchild, avuncular (i.e. uncle–nephew), and

half-sibling relationships is $\frac{7}{4}$ $(7/4\ M_S)$, and between cousins it is $\frac{15}{8}$ $(15/8\ M_S)$. Finally, under these assumptions, the expected value of genetic pairwise distance assessed between unrelated individuals is 2^{-} $(2M_S)$.

Direct observations of genetic distance from published genomic sequences

Binary sequence Alignment Map (BAM) files from the G1K phase 3 data release were obtained from the G1K data repository (www.1000genomes.org/data), and likely PCR duplicate sequences were removed using the MarkDuplicates function implemented in Picard tools version 1.130 (http://broadinstitute.github.io/picard). BAM files from Lazaridis *et al*. (2013) and Haak *et al*. (2015) were obtained from the European Nucleotide Archive (accession numbers PRJEB6272 and PRJEB8448). We used MapDamage2.0 version 2.02 (Jónsson *et al*. 2013) to mitigate residual DNA damage by processing each BAM file using default settings. This software fits a position-specific aDNA damage model from Briggs *et al*. (2007) to aligned genomic sequences, recalibrating base quality scores so that they more accurately represent each base's probability of being erroneous.

For each pairwise comparison, SAMtools version 0.1.19 (Li *et al*. 2009) was used to convert aligned sequences to pileup format, excluding sequences with Phred-scaled mapping quality scores < 30 as well as bases quality scores < 30. The pileup file was passed to a custom Python script that calculated the mean number of pairwise differences observed at a provided panel of target genomic positions, without need for genotype calling, by randomly selecting one nucleotide from the observed nucleotides of each individual. In tabulating these counts, deletions and insertions in NGS sequences were removed from the pool of observations before random sampling. When running

the script in transition-filtration mode, transitions were removed, as were sites where an individual carried an allele unobserved in the called genotypes of G1K present-day human populations. Self-comparisons within a single individual further necessitate that at least 2 sequences must be observed at a given position. To mitigate possible biases in self-comparisons of low-coverage individuals, a randomly drawn sequence was never compared against itself. Replicates of direct observations of genetic distance, each generated by randomly sampling from available sequences, were used to determine a distribution of observations, enabling the calculation of a variance about the mean value.

Determination of relatedness from direct observations of genetic distance

Pedigree simulations were conducted to assist in determining the relatedness of individuals using direct observations of pairwise genetic distance from their aligned sequences. A pileup file containing all sequences was supplied to the model, and the calculation of genetic distance in both the simulated and observed data was carried out only at sites where sequences passed the same user-selected filters (possibly including base quality, sequence depth, transitions, allele frequency, etc.). This step enables calculation of genetic distance at precisely the same positions in both simulations and direct observations, generating distributions of genetic distance that were specific to the pairwise comparisons being conducted. Contamination was not parameterized in our simulations as the aDNA samples considered were all characterized by contamination rates < 2%. In order to reduce the impact of aDNA damage, in all analyses with ancient samples, sequences at a particular target site were excluded unless they matched an allele known from the G1K panel.

We assigned a most likely coefficient of relationship *r* to each BAM-based distribution of observed pairwise genetic distance by calculating the probability of making those observations given the simulated genetic distances for each relationship. For each simulated test relationship, we first conducted one-sample Kolmogorov-Smirnov tests (Conover 1971), using a critical p-value of 0.01 to determine if simulation replicates could be approximated as a normal distribution. Although in theory skewness increases with more distant relatedness and none of these relationships should conform to true normal distributions (Hill & Weir 2011), in our tests a normal distribution was never rejected. Thus for convenience we proceeded to assign a relationship-specific z-score to the mean of our direct observations by assuming it was drawn from a normal distribution with mean and standard deviation equal to the empirical mean and standard deviation of simulated genetic distances for each relationship. The most likely coefficient of relationship among those considered was identified by the smallest z-score (absolute value).

We used odds ratios (ORs) to assess confidence in our choice of the most likely coefficient of relationship. Using normal densities as proxies for the density functions of the simulated genetic distances, the probability p_{Obs} of the mean BAM-based direct observation within each relationship was calculated as the one-tailed probability of making an observation further from the mean of that relationship's probability distribution. Then, an OR was calculated as the odds of the most likely relationship ($p_{Obs,R1}/(1-p_{Obs,R1})$) divided by the odds of the second-most likely relationship ($p_{Obs,R2}/(1-p_{Obs,R2})$). The most likely relationship was considered confidently determined if OR > 100. Otherwise, a new OR was calculated between the most likely relationship and the third-most likely relationship, and so on, until an OR > 100 was obtained.

Results

Confirmation of theoretical expectations of pairwise genetic distance

We performed 100 simulations of a simple pedigree (**Figure 1**) so that the genetic distance between differently related individuals could be characterized and compared. We investigated the following relationships between individuals who were assumed to be outbred: (A) unrelated, (B) parent–offspring, (C) siblings, (D) avuncular (*e.g.* uncle–nephew), (E) grandparent–grandoffspring, (F) half-siblings, (G) first cousins, and (H) self. In a practical case described later, simulations were compared to real genome sequences, and a ninth relationship was examined.

The simulations confirmed our theoretical expectations for the number of pairwise differences per site under scenarios with varying sequencing error rate (0% to 10%), sample-specific mean sequencing depth (0.1X to 10.0X), and contamination rate by present-day humans (0% to 75%; **Figure S1**). Our simulations generally confirm theory and previous observations of non-equal variance of genetic distances between individuals whose relationship has the same expected value (Hill & Weir 2011; Speed & Balding 2014). For example, although sibling and parent–offspring relationships have equal expected values for the proportion of the genome IBD, replicates of the sibling relationship achieve a wider range of genetic distances. Despite their identical expected values, a higher variance was also observed in the grandparent–grandchild relationships in comparison with half-sibling or avuncular relationships.

Effects of the number of target SNPs and their allele frequencies

One of our primary interests was to determine the number of SNP positions necessary to determine relatedness coefficients using SNP capture approaches. To this end, we

performed simulations assessing pairwise distances using randomly selected SNPs ranging in number from 3M to 3k. With the resulting genetic distances, we used the estimated overlap of each pair of seven test relationships to approximate the separability of the relationship pair. For a hypothetical genetic distance observed between relatives under the simulation parameters, this overlap estimates the probability that the observation could be incorrectly assigned to an overlapping coefficient of relationship rather than to the correct one, or that it could not be assigned confidently to either of the two values of *r*. The simulations confirm our expectation that assays with larger numbers of SNPs have more power to discriminate among close relationships (**Figure S2 a-d**). These tests show that whereas a panel of 3k randomly selected SNPs is adequate only for reliably discriminating (5% level, *i.e*. BC<5%) siblings from completely unrelated individuals or identical individuals from any other relationship, 3M randomly selected SNPs, sequenced to a depth of 10X, are capable of distinguishing all tested relatedness values except $r = 0.25$ from $r = 0.125$.

A majority of human genetic variation exists at very low frequency within populations, and as such, selection of the SNP panel likely has important consequences for relationship determination. Rather than selecting SNPs randomly, assaying only variants at some minimum frequency should grant greater discriminatory power. To demonstrate, we replicated the previous simulations while requiring the randomly selected SNPs to have a minor allele frequency (MAF) ≥ 5% in the EUR super-population. Our simulations confirm that the power to discriminate among relationships is higher when assaying SNPs at higher frequencies in the population of interest (**Figure S2 e-h**).

Effects of mean* sequencing *depth

To determine how our ability to discriminate between close relationships using genetic distance is influenced by mean sequencing depth, we performed simulations while varying the mean depth parameter from 10.00X to 0.05X (**Figure 2**). For our panel of 300k SNPs, there is a strong loss of relatedness separability at depths lower than 0.5X. These simulations demonstrate that reduction of sequencing depth reduces the discriminatory power of pairwise genetic distance by effectively decreasing the number of overlapping sites that achieve the minimum sequence depth in both individuals under comparison. However, we show that $r = 1.00$ can be distinguished from all other tested values even down to 0.05X sequencing depth.

Effects of sequencing error

Our simulations demonstrate that error rates in sequence assessment (assignment of incorrect bases to DNA sequences) in the range of 0% to 10% have very little influence on the power to discriminate between close relationships when using a panel of 300k SNPs with MAF ≥ 5% (**Figure S3a-d**). Sequencing error increases the per-site genetic distance between individuals, but the effects were still quite small, with a slightly stronger influence on discrimination when using panels of 300k SNPs that were not filtered for a minimum MAF (**Figure S3e-h**). Even in the comparison of samples with vastly different error rates, sequencing error did not greatly change the power to discriminate among different values of *r* (**Figure S4**).

We also investigated how not knowing the exact value of the sequencing error parameter might impact power in relatedness discrimination. To capture this uncertainty effectively requires integration over the range of possible values of the error rate during simulation. In our simulations, small, realistic ranges of uncertainty produce

distribution overlaps on the order of those seen in simulation with realistic exact error rates around 0.1% (**Figure S5**). However, larger uncertainty (0 to 10%) in the error parameter yields distributions of genetic distance that overlap far more than in simulations with an exact error rate of 10%. Thus we conclude that for discriminating between possible values of *r*, to some degree certainty in the sequencing error parameter is more important than the relative value of the actual sequencing error rate (**Figure S3, Figure S5**).

Effects of contamination by populations and individuals

Next we performed simulations designed to demonstrate the effects of contamination of a sample by DNA from a foreign, present-day population. Our results show that sample contamination has a strong effect on the power of pairwise genetic distance to discriminate between possible relatedness coefficients (**Figure 3a–d**). As contamination approaches 100%, pairwise genetic distances between individuals of any relationship approach the mean genetic distance between unrelated individuals in the contaminant population. The major effect of increasing the contamination rate of all samples is to increase the genetic distances between related individuals and to decrease the separability of distributions (for several values of *r*). However, the effect was weak even at moderate levels of contamination. Indeed, assaying 300k sites above 5% frequency in EUR produces enough relative difference in genetic distance to discriminate at the 0.01% level all but one pair of simulated values of *r* even in the presence of 50% contamination by a foreign population. Relatedness discrimination was also inversely related to the extent of contamination when samples had unequal rates of

contamination, and the effect scaled with the total fraction of contamination within the two samples (**Figure S6a–c**).

Our derived theoretical expectations ignore relatedness in the contaminating population. In the most likely scenarios, a single individual (molecular biologist or archaeologist) would contaminate each sample, which might produce a very different signature of pairwise relatedness. We investigated this by performing simulations in which the reads from each ancient individual in the comparison are contaminated by the genotypes of a single, random contaminating individual from the AFR super-population. We observe that low rates of contamination by a single individual resemble equal rates of contamination by population allele frequencies (**Figure 3e–g, Figure S6d–f**). However, at very high rates of contamination (≥ 75%), contamination by a single individual further reduces the separability of distributions of genetic distance by skewing them toward shorter genetic distances. The underlying cause for this is that contamination by a single present-day individual increases the probability of sampling the same contaminant allele in the ancient individuals under comparison.

Confirmation of the method using aligned sequences

To test our method on real genomic sequences, we used publicly available low-coverage aligned Illumina sequences published by the G1K. For example, we assessed all pairwise and self-comparisons in a subset of six individuals from the Tuscan (TSI) population. This subset contained a known pair of siblings and otherwise unrelated individuals. Mean sequence depth for the seven genomes ranged from 4X to 8X. Where overlapping sequence data permitted, we examined 77,818,345 sites known to harbor a SNP variant in present-day human populations. A mean of 58.3M SNP sites could be assessed in

these pairwise comparisons. Pairwise comparison of the sibling pair (NA20526/NA20792) yielded the expected value of approximately 150% of the within-individual comparisons (**Figure S7**).

To further validate observations of pairwise genetic distance within the Tuscan population, we conducted simulations of simple pedigrees using randomly selected EUR genomes and reproducing the observed sequence depths. The simulations show that the genetic distances observed directly from the G1K aligned sequence data fall within the distributions of genetic distance from simulations (**Figure S8**, **Figure S9**). In the majority of cases, ORs for relatedness $r = 0$ versus $r = 0.125$ were $> 10^3$. ORs for $r = 1$ in simulated self-comparisons were very highly significant (ORs $> 10^{29}$).

Our method could discern the proper degree of relatedness in the known sibling relationship of individuals NA20526 and NA20792, assigning the observed genetic distance to a simulated distribution for $r = 0.50$ with a highly significant OR $> 10^{13}$. The only comparison that could not be assigned a particular value of r with OR $> 10^2$ was that of NA20526/NA20511; the placement of this comparison within $r = 0.00$ could not be distinguished from the simulations of $r = 0.125$ (OR = 73).

Applications to SNP capture in genomes from archaeological samples: Case 1

We obtained BAM files from Haak *et al*. (2015), which made available aligned sequences for numerous sets of individuals from the same archaeological site/horizon and sometimes dated to narrow time intervals (**Table 1**). Sequences from a panel of 380,000 SNPs previously ascertained in present-day human populations were captured and enriched in these samples. Close relationships have been reported in the Haak *et al*. (2015) in some cases, although the exact relationships have not been determined. We

observed elevated A/G and C/T mismatches likely related to residual deamination from aDNA damage in aligned sequences for these ancient individuals, so we analyzed only known transversion SNPs.

Initially we assessed genetic distances between individuals from the Esperstedt (ESP) site in Germany. These samples dated to ~2,500 BCE and consisted of four individuals sequenced to depths ranging from 0.5X to 4.0X at the target SNP sites and low estimated nuclear genome contamination rates in the range 0.3–3.5%. Using a minimum depth of two sequences per site, the number of genomic sites where pairwise comparisons could be assessed ranged from 3,000 to 49,000. The self-comparison and pairwise comparison values are consistent between different Esperstedt individuals, excepting ESP2/ESP29 and ESP3/ESP3 (**Figure 4**). While all other individuals appear unrelated, genetic distance alone allows us to predict a parent-offspring or sibling-sibling relationship between ESP2 and ESP29. Haak *et al*. (2015) used genetic distances to report that these individuals "form a small group and appear to be genetically closely related," but simulations are necessary to determine the precise degree of relatedness with high confidence.

We performed simulations of simple pedigrees to compare distributions of genetic distance between individuals of different coefficients of relationship *r* with those directly observed from aligned sequences. In these simulations, sequencing error and present-day contamination were ignored because our previous simulations showed them to have only minor effects under realistic values of these parameters. Relatedness among the four Esperstedt individuals could often not be resolved beyond the level of unrelated or third-order relatives (*e.g.* first cousins), which was expected from our preliminary study given the number of SNPs available (**Figure 5**). A relationship with *r* =

0.50 between ESP2 and ESP29 was confirmed, and the assignment was highly significant with OR > 10^9. Self-comparison of individual ESP2 placed this individual outside the distribution of mean genetic distances for r = 1.00, possibly indicating extreme recent inbreeding in this individual's ancestry. To investigate this possibility, we added an inbred individual (the offspring of siblings) to the pedigree (**Figure S10**). The self-comparison of individual ESP3 was assigned to this inbred self-relationship distribution with a highly significant OR > 10^{13} over the outbred distribution for r = 1.00.

Applications to SNP capture in genomes from archaeological samples: Case 2

Next we assessed relatedness in ancient remains from the Els Trocs cave site in Spain. These samples had direct dates ranging from 5,311 to 5,066 BCE and consisted of five individuals. Haak *et al*. (2015) noted one pair (Troc3/Troc4) to be "close relatives." Mean sequence depth of target SNP sites was quite low for some individuals and ranged from 0.1X to 30.8X. These samples also had very low estimated nuclear genome contamination rates in the range 0.0–0.8%. Although genetic distances between most individual pairs were close to the theoretical value of $2M_S$ for unrelated individuals (**Figure S11**), the number of overlapping sites achieving a minimum depth of two sequences was quite low, for many pairs at around 1,000 sites. Indeed, the Troc1/Troc4 pair had only 420 overlapping sites. To maximize the number of available sites, we elected to examine pairwise genetic distance at sites with a minimum depth of one sequence, although this precluded all within-individual comparisons. This increased the number of overlapping sites to a mean number of 28,042, and mean values of pairwise genetic distances between most sample pairs were increased (**Figure S11**).

Our initial simulations showed poor agreement with the observed genetic distances due to a strong bias toward higher values, resulting in all individuals being at least as related as first cousins (**Figure S12**). We calculated the mean error between the mean simulated genetic distance of unrelated individuals and the observed genetic distances of putatively unrelated individuals with approximately the same value for genetic distance. As we saw increased error in pairwise comparisons with less overlapping sequence, we calculated the mean error weighted by the number of overlapping sites assessed in each pairwise comparison. The mean weighted error was 7.04%. This discrepancy is likely due to ascertainment bias in our SNP panel, which almost certainly includes sites harboring variants that were maintained at frequencies different in the ancient Els Trocs population than in present-day EUR populations. Because of this ascertainment bias, the simulations relied on a SNP panel with a higher rate of heterozygosity in present-day Europeans than in the population of interest. In this case, our 7.04% error rate actually gives us some measure of differentiation between the Els Trocs population and the EUR super-population at the overlapping sites.

Thus we conducted additional simulations in which EUR allele frequencies were reduced to 0% at a randomly chosen subsample of 7.04% of target SNP positions and found these simulations agreed better with the observed genetic distances (**Figure S13**). These simulations facilitated confident assignment of the observed pairwise genetic distances to the $r = 0.00$ distribution in most relationships. Troc4 had the lowest mean sequence depth (0.12X) of the samples considered in this study, and as a result separability of $r = 0.00$ and $r = 0.125$ distributions was often low for this individual. However, the Troc3/Troc4 comparison was assigned to the $r = 0.50$ distribution with

high confidence (OR > 10^5), and Troc3/Troc7 were assigned to the $r = 0.125$ distribution with high confidence (OR > 10^3). For Troc4/Troc7, $r = 0.125$ is most likely, although neither the ORs for $r = 0.125$ versus $r = 0.00$ nor $r = 0.25$ were significant at the level of 10^2.

The coefficients of relationship assigned within this particular subset of Els Trocs samples make it possible to narrow down the possible pedigrees to those compatible with the available data. Troc3 was a male, Troc4 and Troc7 were females, and the three carried different mitogenome haplotypes (**Table 1**). Because they do not share mitogenome haplotypes, one can rule out that Troc4 was the mother of Troc3. For the same reason, Troc3 and Troc4 could not have been full siblings. Although $r = 0.125$ (a first-cousins relationship) is most likely for Troc3/Troc7 and Troc4/Troc7, these relationships are not compatible given that Troc3 and Troc4 were not siblings. If Troc3 was the father of Troc4, then a first-cousins relationship between Troc3 and Troc7 would have made Troc7 and Troc4 more distantly related, but not entirely unrelated, which is consistent with the entirety of the data (**Figure S14**).

Discussion

We set out to determine the feasibility of using pairwise genetic distance to characterize genetic relatedness from low-depth next-generation genome sequences. We have shown that while in some cases relatedness can be determined in the absence of population allele frequencies, estimates of allele frequencies allow for more precise determination of relatedness. For ancient human populations, however, estimated allele frequencies are not generally available.

Implications for relatedness studies in ancient samples

We find that while contamination from present-day sources reduces the power to discriminate relatedness among individuals sampled, sequencing error does not. However, sequencing error due to deamination (DNA damage, as in aDNA datasets) would be expected to increase the similarity of heavily damaged samples, albeit in a way that could be parameterized (*e.g.* Jónsson *et al*. 2013). We have also shown our method is able to estimate relatedness even for inbred individuals, as long as sequences from at least one known outbred individual is available to determine the ancient population's 'true' within-individual genetic distance.

Our simulations demonstrate the importance of a wisely chosen SNP panel. A study that uses SNP loci with a higher frequency in a given test population would have more power to discriminate coefficients of relationship. This is seen in our analysis of sequences from Haak *et al*. (2015). Even with relatively low sequencing depth, this panel of only 300k SNPs was sufficient to determine most coefficients of relationship in two sets of samples. Although the SNP panel was compatible with the ca. 4,000-year old (Middle Neolithic) individuals from Esperstedt, Germany, ascertainment bias was

apparent in our application of this panel to the ca. 7,200-year old (Early Neolithic) samples from Els Trocs, Spain. Multiple major genetic turnovers occurred throughout Europe's Neolithic period (8,000–7,000ya), which helps to explain the different magnitudes of divergence of Els Trocs and Esperstedt allele frequencies from the present-day populations used for simulation (Haak *et al*. 2015).

Our pedigree simulations used present-day EUR genomes with down-sampled heterozygosity to reconcile the observations of genetic distance among the Els Trocs individuals with corresponding simulations. But in this there is an inherent assumption that certain individuals are unrelated. An equally promising alternative approach would be to calculate the average error (as the distance from the mean observed value to the mean of the simulated distribution) across all self-comparisons, but this would require limiting the analysis to sites with minimum depth of two sequences. Since it requires only the assumption that no individuals are inbred, this approach is preferable in most cases. However, when few sequences are available, as was the case for some Els Trocs individuals, an analysis of this type may not be possible without excluding certain individuals. Otherwise, increasing the number of sequenced individuals from ancient populations would help to identify individuals with unusual pairwise genetic distances. Until then, comparisons with the remains of individuals from sites nearby in time or space are advised. These additional individuals could be assumed to be from the same population, but not the same family.

Potential applications

Happily, even under the condition of considerable sequencing error and high contamination, pairwise sequence differences are powerful enough to discriminate

identical or unrelated biological samples within low-coverage genome sequences with a sufficient number of SNPs. This assumes the three major parameters of our model are relatively well characterized. The contamination rate can be estimated from haploid (mitochondrial and Y chromosome) sequences using simple approaches such as rates of heterozygosity and mismatch to the consensus sequence, but more complex, likelihood-based estimators are often applied to the nuclear genome (Meyer *et al*. 2012). Estimation of sequencing error is more nuanced, although a profile of new Illumina platforms' sequencing error has been characterized with a mean value of 0.18–0.30%, suggesting a relatively narrow range within these values for sequences from present-day genomes (Ross *et al*. 2013). For aDNA data, damage models like mapDamage (Jónsson *et a*l. 2013) can be used to either characterize the error rate contributed by DNA damage, or—as in our approach—simply to correct it before sequence analysis.

In our work with SNPs identified in capture sequences from ancient humans, we found that determination of self-relatedness was possible even with low sequence coverage. This opens the possibility of high-throughput screening of mixed archaeological samples such as bone fragments in order to cobble together those fragments belonging to the same individual or to determine the minimum number of individuals represented. Our approach would be especially useful for sorting specimens heavily contaminated with genetic material from archaeological excavators, museum personnel, or other handlers.

This work also has implications for population genetic studies of archaeological human populations, as these studies should ideally take into account individual relatedness when calculating population genetic statistics. By identifying the degree of relatedness in ancient individuals already published and analyzed in a population

genomic context, we have demonstrated that this is a possible outcome for future population genomic studies. This point is even more relevant in light of our observations of inbreeding within these ancient populations. Following on the recent report of multiple Neanderthal exomes obtained through targeted sequence capture (Castellano *et al*. 2014), we suggest that our method could be applied to low-coverage sequences from remains of Neanderthal individuals with a close association in both time and space. Relatedness determination within samples such as these, which could represent families or clan groups, would extend the potential for fascinating insights into ancient social and family structures of our closest relatives.

Lastly, our framework may one be useful in forensic genomic scenarios—as in the aftermath of a disaster (Brenner & Weir 2003)—where genomic sequences from degraded and/or mixed samples from the same site could be quickly screened to determine if they are genetically identical, or to test for relatedness to potential family members from whom corresponding genomic sequence data has been obtained.

Future directions

Agreement of our simulations with observed genetic distances depends on the degree of differentiation between the observed population and the population(s) used for simulations, which complicates comparison of pedigree simulations to populations with unknown allele frequency spectra. Thus future work to simulate genetic distances within an unknown (ancient) population should attempt to estimate its allele frequency spectrum so that appropriate sites can be targeted.

As it assesses genetic distance only where sequences overlap a panel of sites known to harbor variants at high frequency in the population of interest, our method

does not use sequences off these targets that may still be informative about pairwise relatedness. Assessment of pairwise genetic distance at all sites with overlapping sequences would be a desirable alternative, as this would not rely on assumed allele frequencies and would use all available genetic information, which is scarce in low-depth sequencing studies. These genetic distances could then be simulated using estimates of contamination, sequencing error, and heterozygosity. But at ~0.1%, the Illumina sequencing error rate, which varies between samples and sequencing runs, is comparable to the per-base heterozygosity of the human genome (Nakamura *et al*. 2011; Schirmer *et al*. 2015). Thus if all overlapping sites are assessed in pairwise comparisons, sequencing errors threaten to eclipse true genetic differences between individuals. For our purposes, the use of all genomic positions where there is overlapping sequence data would be possible if an accurate model of sequencing error were used to parameterize the sequencing error rate.

Other recent work (Korneliussen & Moltke 2015; Lipatov *et al*. 2015) also estimates relatedness from low-coverage sequences, building on methods implemented in PLINK (Purcell *et al*. 2007) and related software, but with the inclusion of genotype likelihoods based on per-base sequencing error probabilities. The simulation results are promising in that these methods, which rely on allele frequencies estimated in a known population, enable more accurate determination of relatedness than previous methods that utilize called genotypes. For Lipatov *et al*. (2015), this was true even with population divergence up to F_{ST}=0.1 from assumed allele frequencies. However, the use of genotype likelihoods limits these types of analysis to individuals sequenced to mean depth ≥ 2X, and Lipatov *et al*. (2015) report that their method performs poorly with admixed and inbred individuals. Methods of this kind, though, do have added benefits in

that they are able to distinguish parent-offspring and sibling-sibling relationships, which typically cannot be achieved using only genetic distance. Thus while our method accounts for sample contamination and is applicable to any overlapping sequencing data regardless of depth, the method of Lipatov *et al*. (2015) would likely be superior for relatedness studies with negligible present-day human contamination and in which the study population's allele frequencies are not strongly divergent from reference populations.

Acknowledgements

We offer our sincerest thanks to Eric Durand, Fernando Racimo, Rebekah Rogers, Melinda Yang, Amy Ko, and other members of the Center for Theoretical Evolutionary Genomics for their useful comments and suggestions during the development of the methods presented here. We thank Vanessa Bieker for helping to test software compatibility. This research was supported in part by a grant from the US NIH R01-GM40282 to M. Slatkin.

Data Accessibility

The scripts developed to implement this study's methods together are called *grups* (Genetic Relatedness Using Pedigree Simulations) and are available at: github.com/sameoldmike/grups.

Author Contributions

MDM, FJ, and MS developed the methods presented here. MDM performed the analyses. MDM, FJ, SC, and MS wrote the paper.

References

Abecasis, G. R. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073.

Abecasis, G. R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65.

Albrechtsen A, Korneliussen TS, Moltke I, Hansen TO, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* **33**, 266–274.

Allentoft, M. E. *et al.* (2015) Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172.

Bhattacharyya A (1947) On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35**, 99–109.

Brenner CH, Weir BS (2003) Issues and strategies in the DNA identification of World Trade Center victims. *Theor. Popul. Biol.* **63**, 173–178.

Briggs, A. W. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 14616–14621.

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**, 1084–1097.

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet* **86**, 526–539.

Castellano, S. *et al.* (2014) Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6666–6671.

Conover WJ (1971) *Practical Nonparametric Statistics*. New York: John Wiley & Sons. pp. 309–314.

Evett IW, Weir BS (1998) Interpreting DNA evidence: Statistical genetics for forensics scientists. Sinauer Associates Inc., Sunderland, MA.

Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27,** 2153–2155.

Gusev A. *et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19,** 318–326.

Haak, W. *et al.* (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211.

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8,** 299–309.

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research* **93**, 47–64.

Hill WG, White IMS (2013) Identification of pedigree relationship from genome sharing. *G3* **3**, 1553–1571.

Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S (2001) Ancient DNA. *Nature Reviews Genetics* **2**, 353–359.

Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, *et al*. (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research* **21**, 768–774.

International HapMap Consortium [IHMC] (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) MapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29,** 1682–1684.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics* **31**, 241–247.

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

Lazaridis I, *et al.* (2013) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413.

Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079.

Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, *et al*. (2014) Relationship Estimation from Whole-Genome Sequence Data. *PLoS Genetics* **10**, e1004144.

Lipatov M, Komal S, Patro R, Veeramah K. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *bioRxiv*, dx.doi.org/10.1101/023374.

Meyer M, *et al*. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226.

Pemberton TJ, Wang C, Li JZ, Rosenberg NA (2010) Inference of unexpected genetic relatedness among individuals in HapMap phase III. *American Journal of Human Genetics* **87**, 457–464.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome

association and population-based linkage analyses. A*merican Journal of Human Genetics* **81**, 559–575.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB (2013) Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51.

Slatkin M (2008) Exchangeable models of complex inherited diseases. *Genetics* **179**, 2253–2261.

Speed D, Balding DJ (2014) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16**, 33–44.

Shapiro B, Hofreiter M (2012) *Ancient DNA: Methods and Protocols*. New York: Humana Press. pp. 247.

Tal O (2013) Two complementary perspectives on inter-individual genetic distance. *BioSystems* **111,** 18–36.

Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources* **11**, 141–145.

Weir BS, Anderson AD, Hepler AB. 2006. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics* **7**, 771–780.

Willerslev E, Cooper A (2005) Ancient DNA. *Proc. Biol. Sci.* **272**, 3–16.

Yang DY, Watt K (2005) Contamination controls when preparing archaeological remains for ancient DNA analysis. *Journal of Archaeological Science* **32**, 331–336.

Tables

Table 1. Provenance of Haak *et al*. (2015) archaeological samples utilized in this study. ETS = Els Trocs, Spain. EG = Esperstedt, Germany.

Individual ID	Alternate ID	Sampling location	Estimated nuclear genome contam. rate (%)	Date (cal BCE)	Sex	Mean seq. depth at targeted autosomal SNP sites (X)	mtDNA haplotype	Y haplotype
Troc1	I0409	ETS	0.0	5311-5218	F	0.80	J1c3	
Troc3	I0410	ETS	0.8	5178-5066	M	3.47	pre-T2c1d2	R1b1
Troc4	I0411	ETS	0.4	5177-5068	F	0.12	K1a2a	F*
Troc5	I0412	ETS	0.6	5310-5206	M	30.82	N1a1a1	12a1b1
Troc7	I0413	ETS	0.0	5303-5204	F	3.49	V	
ESP2	I0114	EG	0.3	2131-1979	M	1.14	I3a	I2a2
ESP3	I0115	EG	2.8	1931-1780	F	0.55	U5a1	
ESP4	I0116	EG	3.5	2118-1961	M	4.15	W3a1	I2c2
ESP29	I0117	EG	2.6	2199-2064	F	2.32	I3a	

Figures

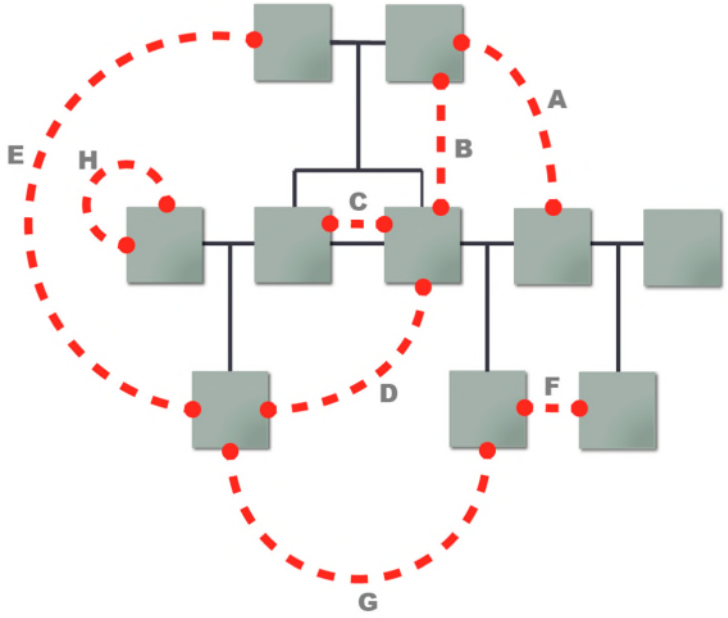

Figure 1. Diagram of the family pedigree used for simulations and quantification of genetic distance between various relationships. Solid connecting lines indicate haploid parental contributions to offspring. Dashed connecting lines indicate the following genetic relationships simulated throughout the study: A) Unrelated. B) Parent–offspring. C) Siblings. D) Avuncular (uncle–nephew). E) Grandparent–grandoffspring. F) Half-siblings. G) First cousins. H) Self (or equivalently, identical twins). All individuals are unrelated unless otherwise indicated.

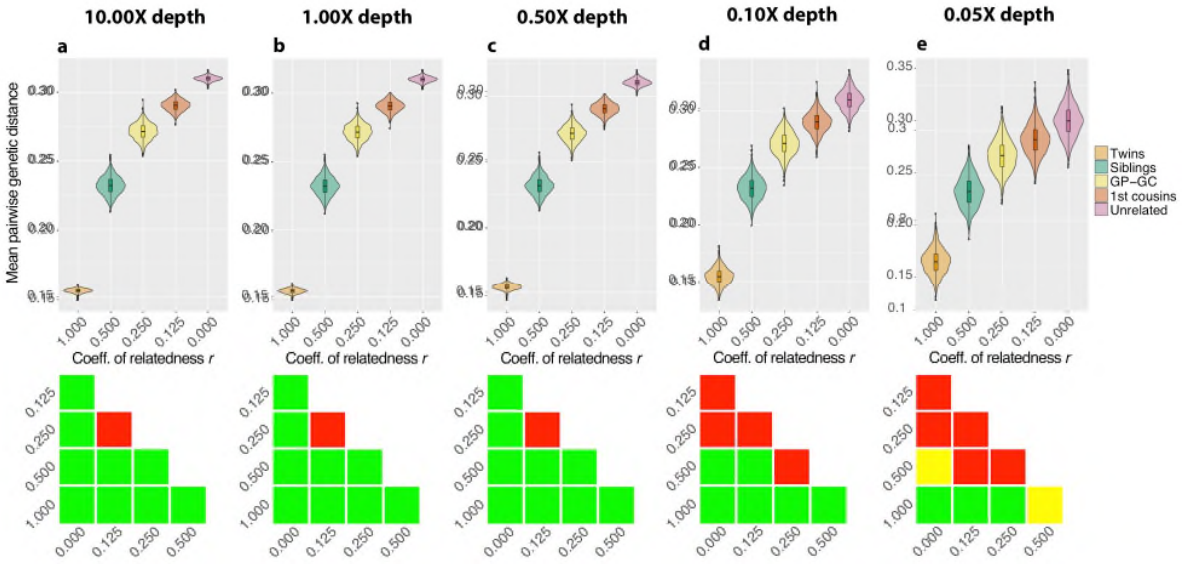

Figure 2. Pedigree simulations demonstrate the influence of mean sequence depth on power to discriminate relatedness using pairwise distances. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, contamination rate = 0%, sequencing error rate = 0.0%, mean sequence depth ranging from 10X to 0.01X. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. T, twins.

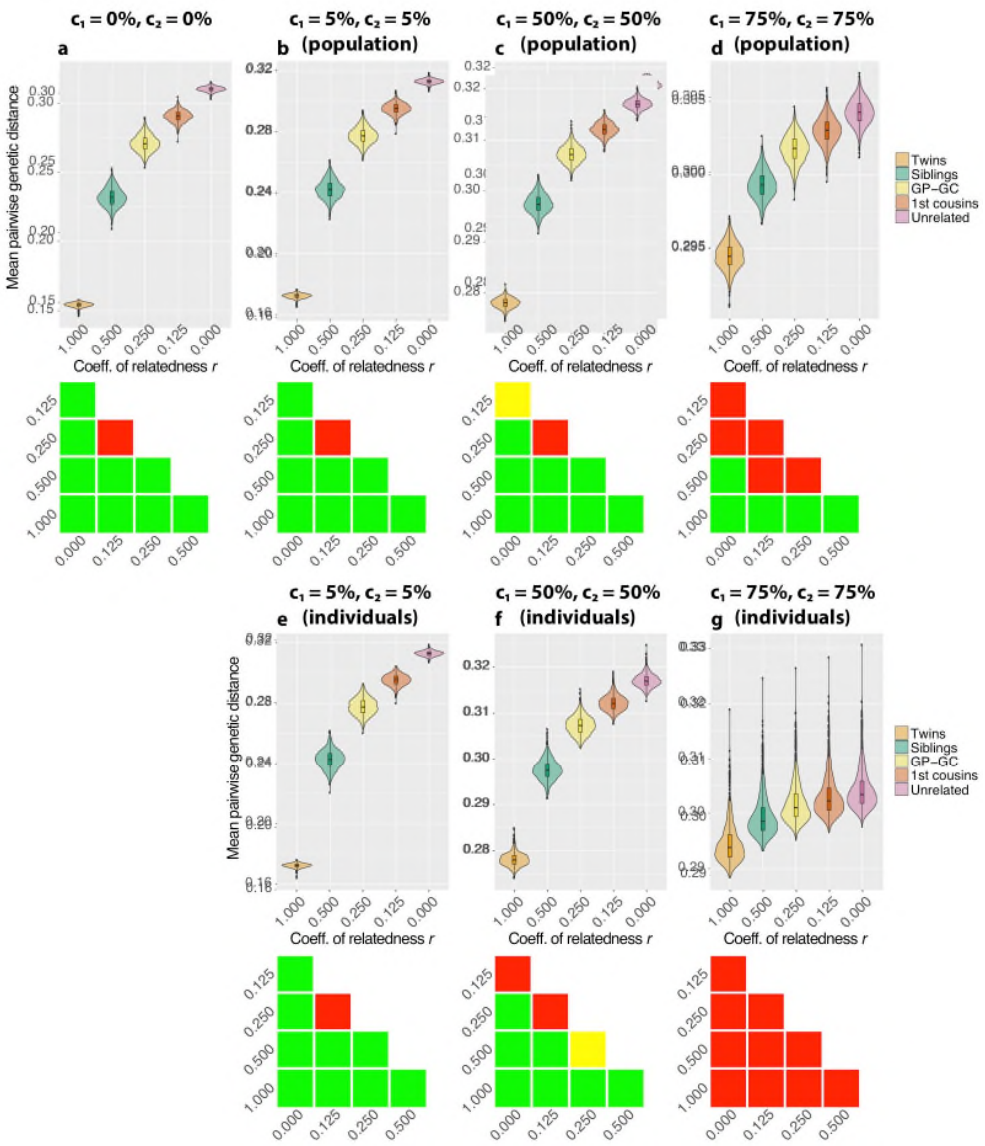

Figure 3. Pedigree simulations demonstrate the influence of sample contamination by a foreign population on power to discriminate relatedness using pairwise distances. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, mean sequence depth = 10X, sequencing error = 0%. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. c_1 and c_2 describe the contamination rates for each individual in the pairwise comparison. In panel **a**, simulations were performed without contamination. In panels **b** – **d**, each sample's contaminant sequences were drawn from the allele frequencies of the AFR super-population. In panels **e** – **g**, each sample's contaminant sequences were drawn from a single, unique AFR individual.

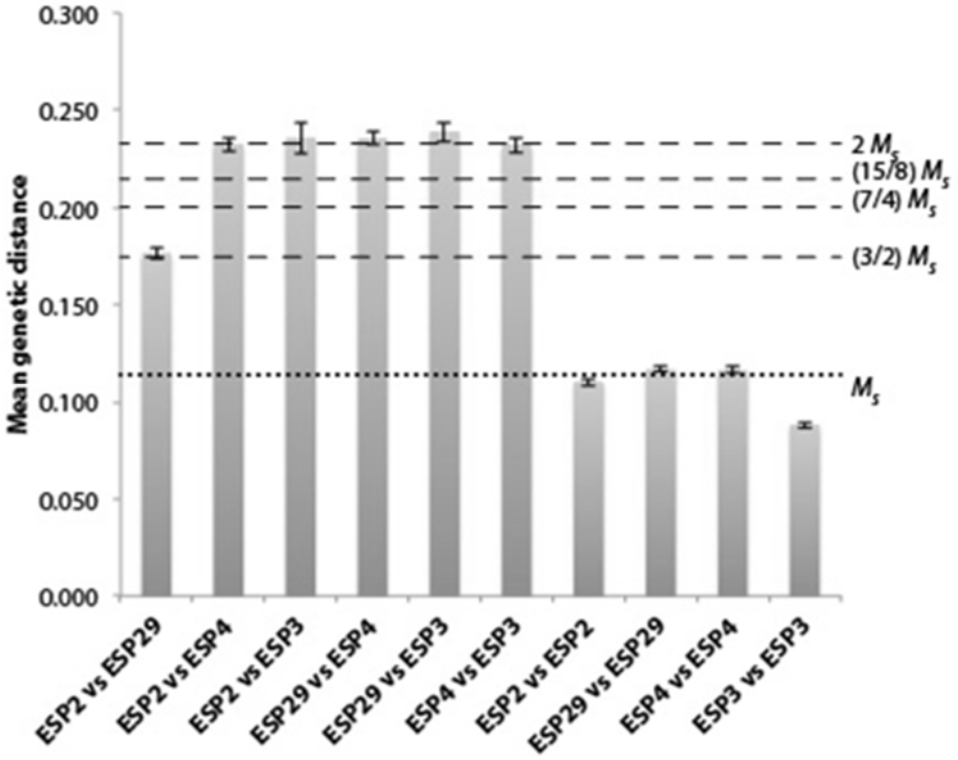

Figure 4. Pairwise genetic distances observed between ancient human individuals from Esperstedt, Germany dated to 2,199-1,780 cal BCE. A mean of 19,586 target genomic positions with overlapping sequence data in both individuals were examined in pairwise comparisons. Whiskers indicate ± 2 SD (standard deviation) around the mean of 20 replicates of pairwise genetic distance generated by randomly sampling from available sequences. Positions were included in pairwise comparisons only if they had a minimum sequence depth of 2 in both individuals. Sequences were excluded if they did not support known transversion alleles. indicates the mean of self-comparisons of the three putatively outbred individuals ESP2, ESP29, and ESP4.

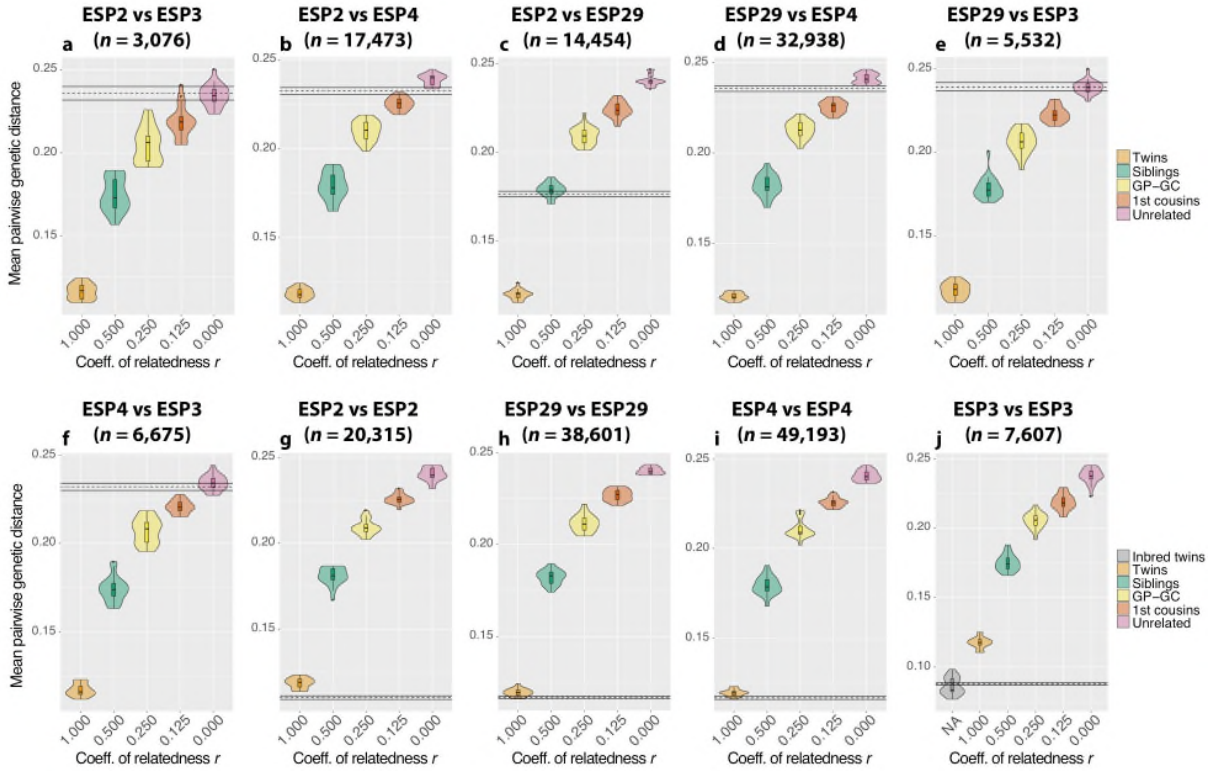

Figure 5. Results of pedigree simulations corresponding to pairwise comparison of aligned sequence data from ancient human individuals excavated from an archaeological site in Esperstedt, Germany. The horizontal black lines indicate the mean (±SD) of 100 replicate observations of genetic distance generated by randomly sampling from available aligned sequences. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: transversion SNP positions passing depth and quality filters including a minimum depth of 2 in aligned sequences in both individuals, sequencing error = 0%, contamination rate = 0%. GP–GC, grandparent-grandchild relationship. OR, odds ratio against second most likely coefficient of relationship. **a**, $r = 0.00$ or $r = 0.125$, OR > 10^3. **b**, $r = 0.00$ or $r = 0.125$, OR > 10^2. **c**) $r = 0.50$, OR > 10^9. **d**, $r = 0.00$ or $r = 0.125$, OR > 10^4. **e**, $r = 0.00$, OR > 10^4. **f**, $r = 0.00$, OR > 10^2. **g**, $r = 1.00$, OR > 10^{40}. **h**, $r = 1.00$, OR > 10^{33}. **i**, $r = 1.00$, OR > 10^{37}. **j**, Self-related (inbred offspring of siblings), OR > 10^{13}.

Supplementary Materials & Methods

Theoretical expectations

At a diallelic locus with alleles A and a, the joint probability of genotype pairs within an outbred population depend on p, the frequency of allele a, θ, the probability that exactly one allele is inherited IBD, and γ, the probability that exactly two alleles are inherited IBD. Following from Slatkin (2008) and previous work, these are:

$$\begin{aligned}
&\Pr(\mathrm{AA},\mathrm{AA}) = (1-\theta-\gamma)p^4 + \theta p^3 + \gamma p^2) \\
&\Pr(\mathrm{AA},\mathrm{Aa}) = \Pr(\mathrm{Aa},\mathrm{AA}) = (1-\theta-\gamma)2p^3(1-p) + \theta p^2(1-p) \\
&\Pr(\mathrm{AA},\mathrm{aa}) = \Pr(\mathrm{aa},\mathrm{AA}) = (1-\theta-\gamma)p^2(1-p)^2 \\
&\Pr(\mathrm{Aa},\mathrm{Aa}) = (1-\theta-\gamma)4p^2(1-p)^2 + \theta p(1-p) + \gamma 2p(1-p) \\
&\Pr(\mathrm{Aa},\mathrm{aa}) = \Pr(\mathrm{aa},\mathrm{Aa}) = (1-\theta-\gamma)2p(1-p)^3 + \theta p(1-p)^2 \\
&\Pr(\mathrm{aa},\mathrm{aa}) = (1-\theta-\gamma)(1-p)^4 + \theta(1-p)^3 + \gamma(1-p)^2
\end{aligned} \qquad \text{Eq. (2)}$$

At a diallelic locus with alleles A and a, considering an ancient and a contaminating present-day population, the joint probabilities of sampling genotype pairs (one from each population) depend on the ancient population allele frequency p_A and the contaminating population allele frequency p_C. Thus, from Hardy-Weinberg expectations we have:

$$\begin{aligned}
&\Pr(\mathrm{AA},\mathrm{AA}) = {p_A}^2{p_C}^2 \\
&\Pr(\mathrm{AA},\mathrm{Aa}) = 2{p_A}^2\mathrm{p}_C(1-\mathrm{p}_C) \\
&\Pr(\mathrm{AA},\mathrm{aa}) = {p_A}^2(1-\mathrm{p}_C)^2 \\
&\Pr(\mathrm{Aa},\mathrm{AA}) = 2\mathrm{p}_A(1-\mathrm{p}_A){p_C}^2 \\
&\Pr(\mathrm{Aa},\mathrm{Aa}) = 4\mathrm{p}_A(1-\mathrm{p}_A)\mathrm{p}_C(1-\mathrm{p}_C)
\end{aligned} \qquad \text{Eq. (3)}$$

$$\Pr(\mathrm{Aa}, \mathrm{aa}) = 2\mathrm{p}_A(1-\mathrm{p}_A)(1-\mathrm{p}_C)^2$$

$$\Pr(\mathrm{aa}, \mathrm{AA}) = (1-\mathrm{p}_A)^2 {p_C}^2$$

$$\Pr(\mathrm{aa}, \mathrm{Aa}) = 2(1-\mathrm{p}_A)^2\mathrm{p}_C(1-\mathrm{p}_C)$$

$$\Pr(\mathrm{aa}, \mathrm{aa}) = (1-\mathrm{p}_A)^2(1-\mathrm{p}_C)^2$$

Next we consider the probability of observing a mismatch when randomly drawing a single sequence from each of two diploid individuals in a pairwise comparison. At a single site, the probability *M* of observing mismatching nucleotides drawn from correctly mapped sequences of each individual depends only on the true genotypes and the sequencing error rate ε (equivalent to the probability of drawing an erroneous nucleotide), which here we assume to be equal for both individuals. For example, if both individuals have the AA genotype, the probability of observing mismatching alleles is the joint probability of observing the true allele of each individual multiplied by the probability of the observed alleles being different, summed over all four possible states of correctness:

$$M_{\mathrm{AA,AA}} = \varepsilon(1-\varepsilon)(1) \;+\; (1-\varepsilon)\varepsilon(1) \;+\; \varepsilon^2(2/3) \;+\; (1-\varepsilon)(1-\varepsilon)(0) \qquad \text{Eq. (4)}$$

It follows from this reasoning that the mismatch probabilities for all joint genotypes are:

$$M_{\mathrm{AA,AA}} = M_{\mathrm{aa,aa}} = \frac{2}{3}\varepsilon(3-2\varepsilon)$$

$$M_{\mathrm{AA,Aa}} = M_{\mathrm{Aa,AA}} = M_{\mathrm{aa,Aa}} = \frac{1}{18}(9+12\varepsilon-8\varepsilon^2) \qquad \text{Eq. (5)}$$

$$M_{\mathrm{aa,AA}} = M_{\mathrm{aa,AA}} = \frac{1}{9}(9-6\varepsilon+4\varepsilon^2)$$

$$M_{\mathrm{Aa,Aa}} = M_{\mathrm{Aa,aa}} = \frac{1}{18}(9 + 12\varepsilon - 8\varepsilon^2)$$

Thus the expected value of *M* observed when comparing single sequences sampled from two individuals from ancient and contaminating populations is obtained by multiplying each joint genotype frequency (Eq. 3) by its corresponding mismatch probability (Eq. 5), obtaining:

$$\mathrm{E_{AC}}[M] = \frac{1}{9}[p_C(3 - 4\varepsilon)^2 - p_A(-1 + 2\,p_C)(3 - 4\varepsilon)^2 + (18\varepsilon - 12\varepsilon^2)] \qquad \text{Eq. (6)}$$

It follows that the expected value of *M* observed when comparing two sequences both sampled from a large contaminating population is given by:

$$\mathrm{E_{CC}}[M] = \frac{1}{9}[p_C(3 - 4\varepsilon)^2 - p_C(-1 + 2\,p_C)(3 - 4\varepsilon)^2 + (18\varepsilon - 12\varepsilon^2)] \qquad \text{Eq. (7)}$$

Similarly, the expected value of *M* observed when comparing two putatively related individuals from the ancient population is derived from Eqns. (2) and (5), depends on θ, γ, q, and p_A, and is given by:

$$\mathrm{E_{AA}}[M] = \frac{1}{18}[36\varepsilon - 24\varepsilon^2 - \mathrm{p}_A(3 - 4\varepsilon)^2(-4 + 2\,\gamma + \theta) + \mathrm{p}_A{}^2(3 - 4\varepsilon)^2(-4 + 2\,\gamma + \theta)] \qquad \text{Eq. (8)}$$

Finally, we obtain the expected value of *M* observed between single sequences sampled from two genomic datasets from the same population—each contaminated at known rates c_1 and c_2 by a contaminating population—by multiplying the independent

probability c of sampling a contaminated sequence by the weight (expected value of M) of each scenario:

$$\begin{aligned}\mathrm{E}[M] = (1-c_1)(1-c_2)\mathrm{E}_{\mathrm{AA}}[M] + c_1(1-c_2)\mathrm{E}_{\mathrm{AC}}[M] \\ + (1-c_1)c_2\mathrm{E}_{\mathrm{AC}}[M] + c_1c_2\mathrm{E}_{\mathrm{CC}}[M]\end{aligned} \qquad \text{Eq. (9)}$$

In a pairwise comparison of aligned genomic sequences from two individuals, the average number of expected pairwise differences per site $\bar{M}$ is obtained by calculating the sum of this value divided by the number of assessed overlapping sites N:

$$\bar{M} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathrm{E}[M]_i \qquad \text{Eq. (10)}$$

When $q = 0$ and $c_1 = c_2 = 0$, the mean expected value of pairwise genetic distance assessed between identical twins, between two samples generated from the same individual, or within the same individual, reduces to $M_S = p_A\,(1-p_A)$. Similarly, the expected value for parent–offspring and sibling–sibling relationships is $\frac{3}{2}M_S$. The expected value for grandparent–grandoffspring and uncle–nephew relationships is $\frac{7}{4}M_S$, and between cousins it is $\frac{15}{8}M_S$. Finally, under these assumptions, the expected value of genetic pairwise distance assessed between unrelated individuals is $2\bar{M}_S$.

Supplementary Figures

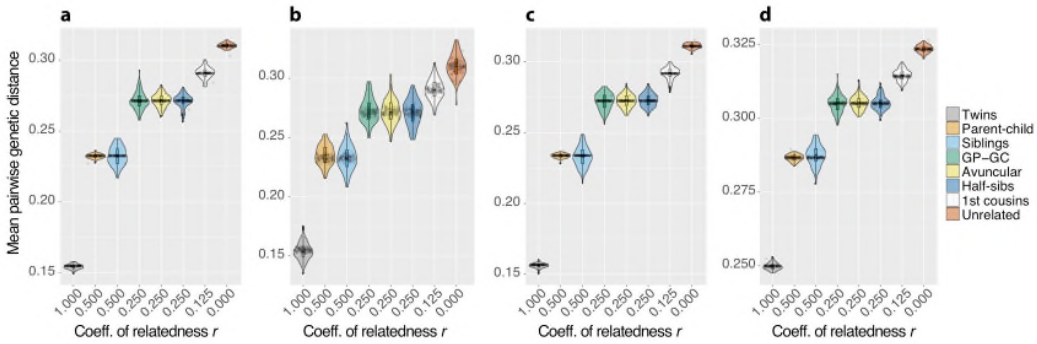

Figure S1. Theoretical expectations of pairwise genetic distance between relatives are confirmed by pedigree simulations (n = 100) under various scenarios of contamination, sequence depth, and sequencing error. Open circles indicate the theoretically expected mean value of pairwise genetic distance of each replicate for each simulated relationship (blue = unrelated, pink = first cousins, red = grandparent–grandoffspring and avuncular, purple = siblings and parent–offspring, green = self-related). The height of the black box within each violin signifies the interquartile range of all simulation replicates around their mean. Box whiskers extend to 1.5 times the interquartile range. Simulations were initialized with random, unrelated EUR individuals and carried out under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, contamination by AFR allele frequencies. GP–GC, grandparent-grandchild relationship. **a**, mean sequence depth = 10X, sequencing error = 0%, c_1 = c_2 = 0%. **b**, mean sequence depth = 0.1X, sequencing error = 0%, c_1 = c_2 = 0%. **c**, mean sequence depth = 10X, sequencing error = 0.1%, c_1 = c_2 = 0%. **d**, mean sequence depth = 10X, sequencing error = 0%, c_1 = 5%, c_2 = 50%.

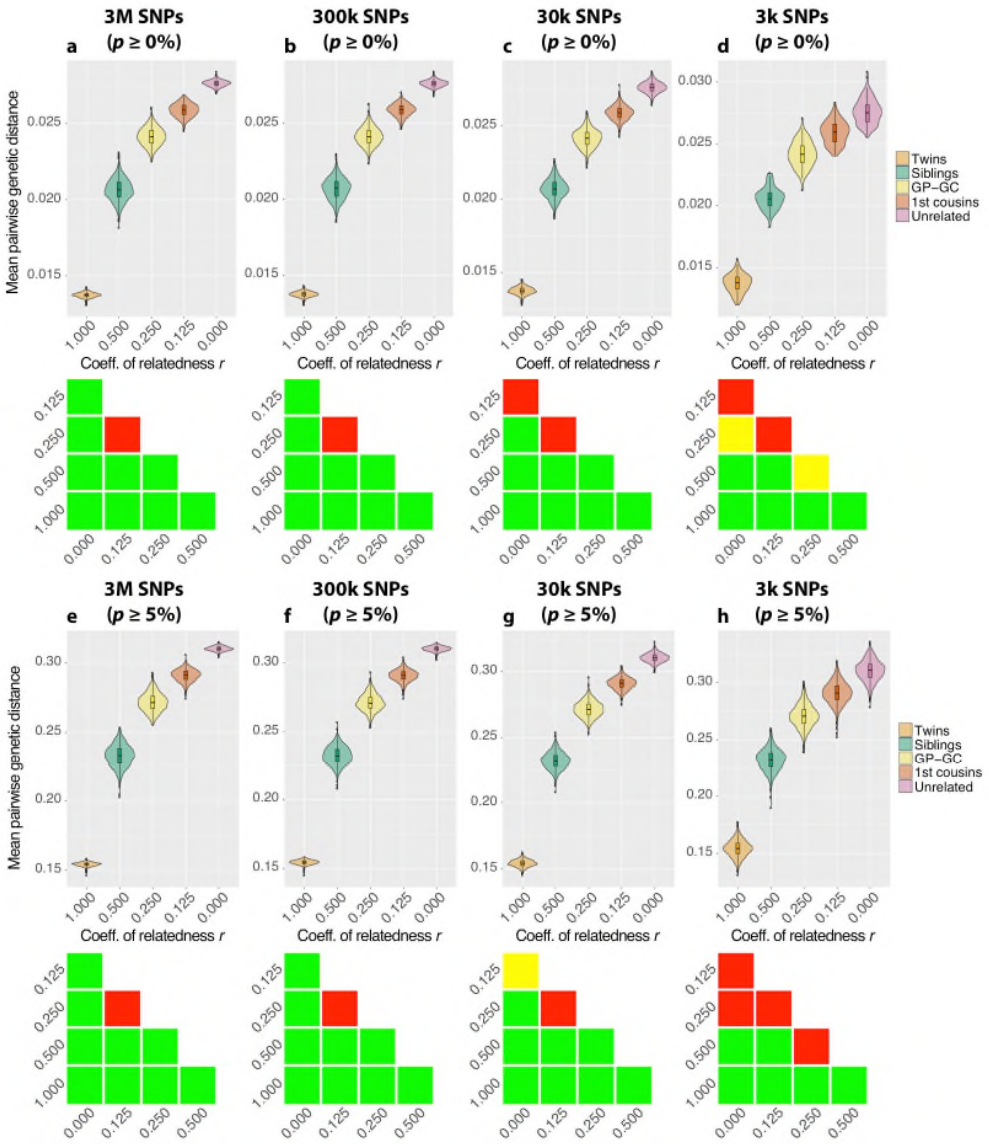

Figure S2. Pedigree simulations demonstrate the influence of the number of targeted SNP loci on power to discriminate relatedness using pairwise distances. Violin plots represent the distribution of pairwise differences per site in 1000 replicates of each analyzed relationship. The height of the black box within each violin signifies the interquartile range of all simulation replicates around their mean. Box whiskers extend to 1.5 times the interquartile range. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: mean sequence depth = 10X, sequencing error = 0%, contamination rate = 0%, number of SNPs ranging from 3M to 3k. Sites were retained according to the European minor allele frequency *p*: **a** – **d**, $p \geq 0\%$. **e** – **h**, $p \geq 5\%$. T, twins.

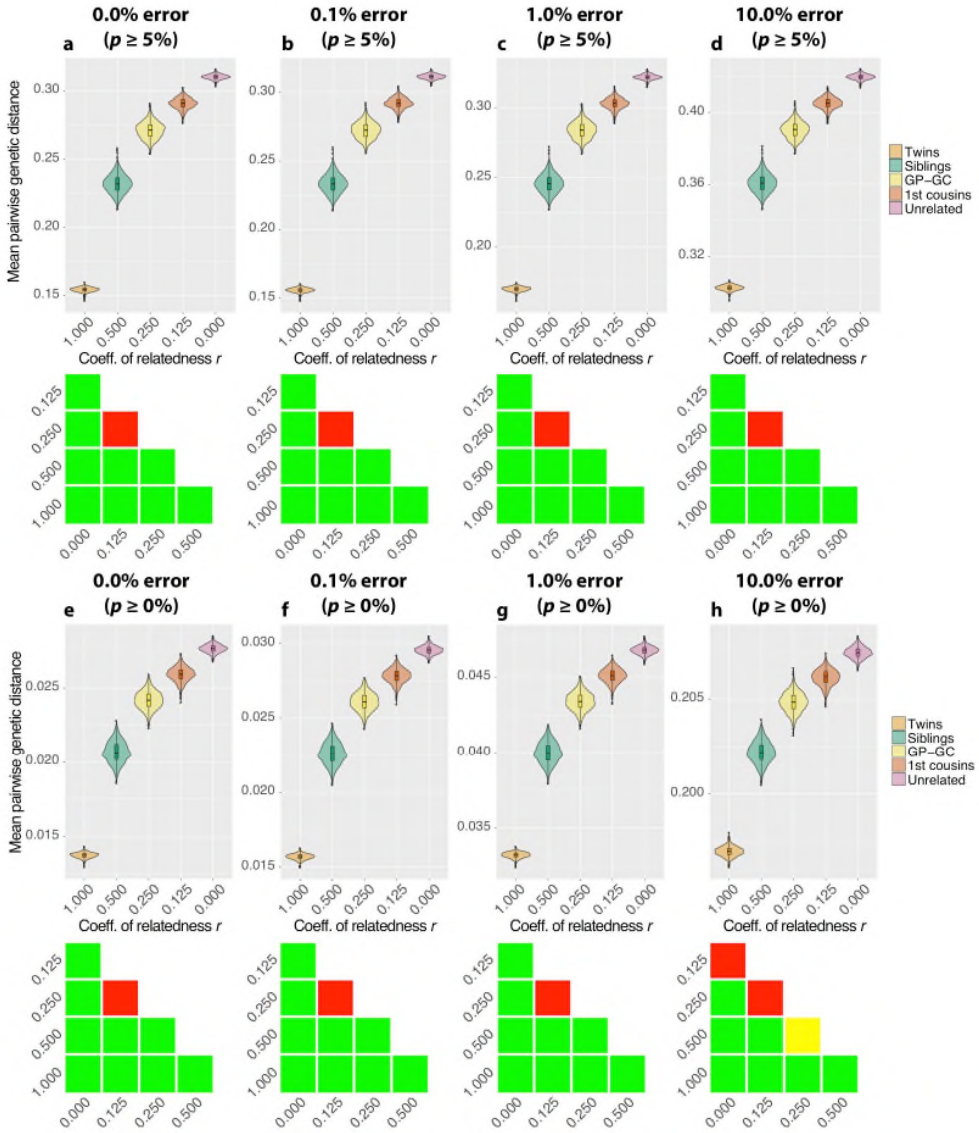

Figure S3. Pedigree simulations demonstrate the minor influence of sequencing error on relationship discrimination using pairwise distances. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: 300k random SNP sites, 0% contamination rate, 10.0X mean sequence depth, sequencing error rate ranging from 0.0% to 10.0%. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. **a–d**, EUR allele frequency $p \geq 5\%$. **a–d**, EUR allele frequency $p \geq 0\%$ (no filtering of allele frequency).

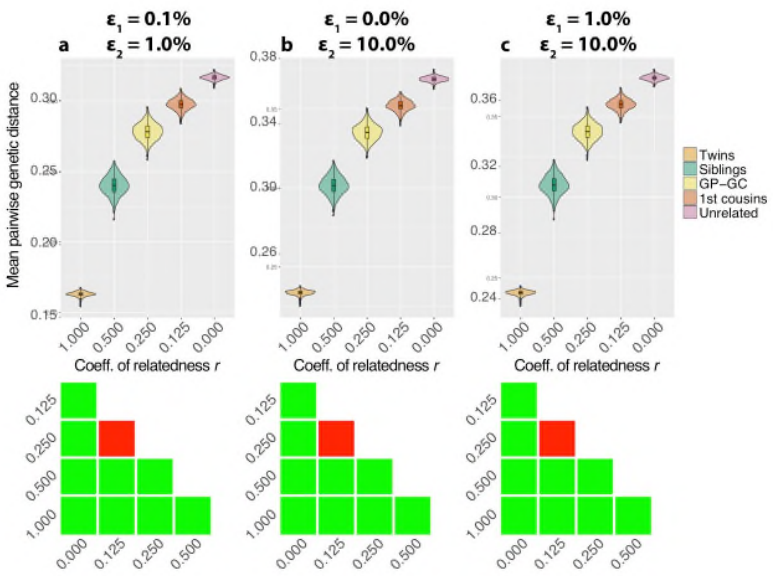

Figure S4. Pedigree simulations demonstrate only a minor influence of unequal rates of sequencing error on relationship discrimination using pairwise distances. Simulations were performed with random EUR individuals from under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, mean sequence depth = 10X, sequencing error ranging from 0% to 10%. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. ε_1 and ε_2 indicate error rates in sequences from the two individuals under comparison.

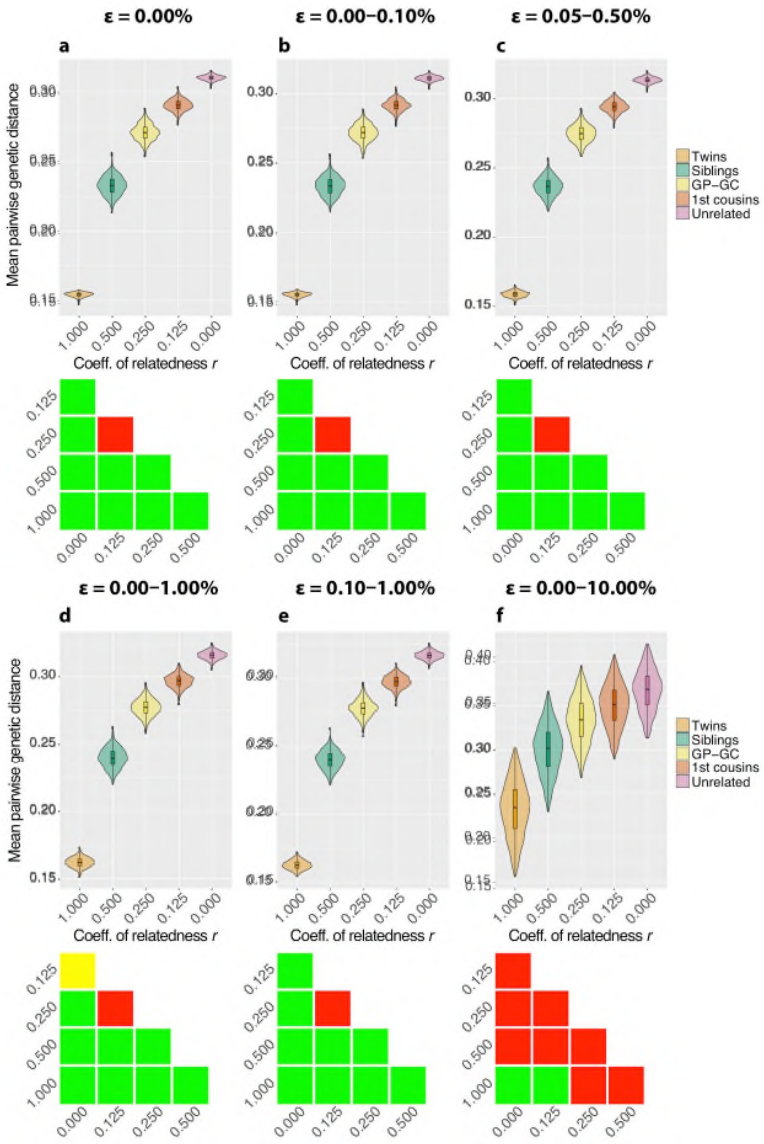

Figure S5. Pedigree simulations demonstrate the influence of uncertainty in the sequencing error parameter on relationship discrimination using pairwise distances. Simulations were performed with random EUR individuals from under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, mean sequence depth = 10X, sequencing error ranging from 0% to 10%. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. ε indicates the error rate in sequences from both individuals under comparison.

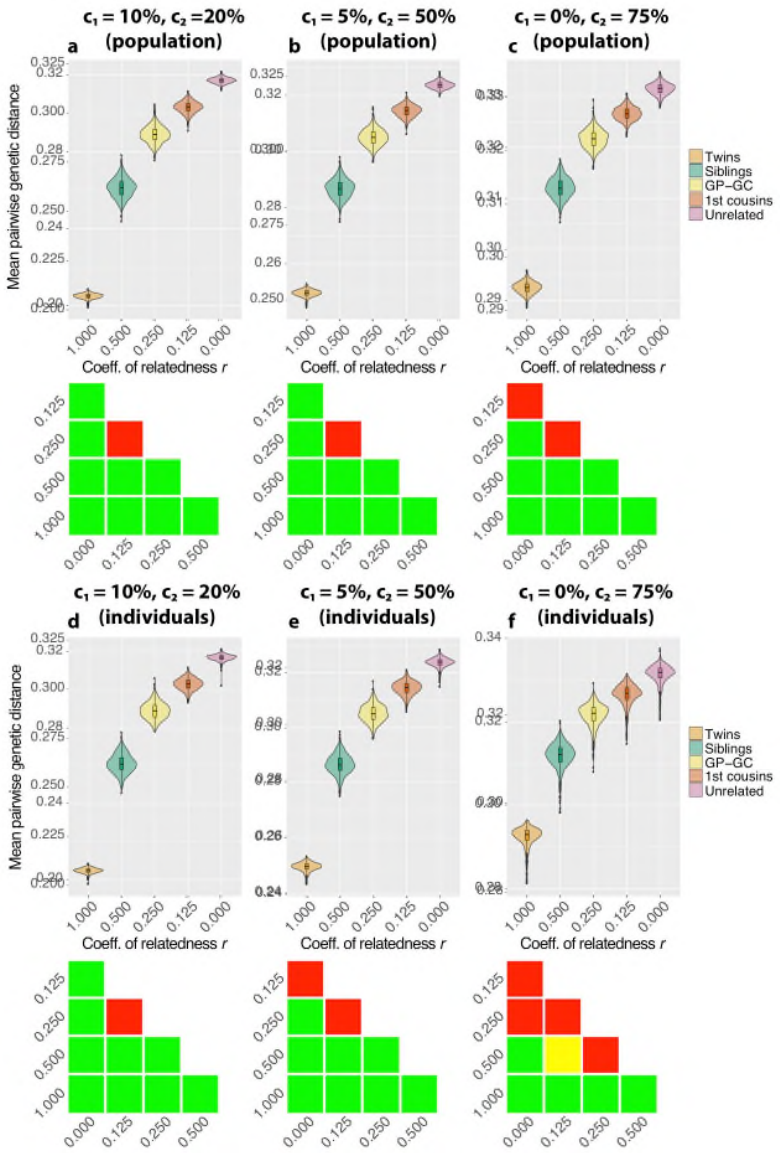

Figure S6. Pedigree simulations demonstrate the influence of unequal rates of sample contamination by a foreign population on relationship discrimination using pairwise distances. Simulations were performed with random EUR individuals from under the following parameters: 300k random SNP sites with EUR allele frequency ≥ 5%, mean sequence depth = 10X, sequencing error = 0%. GP–GC, grandparent-grandchild relationship. Heatmaps below each violin plot illustrate overlap (BC, Bhattacharyya coefficient) in simulations of each relationship pair. Green, BC < 1%. Yellow, BC 1–5%. Red, BC > 5%. In panels a-c, contaminant sequences were drawn from AFR allele frequencies. In panels d-f, each sample's contaminant sequences were drawn from a single, unique AFR individual. c_1 and c_2 indicate contamination rates in sequences from the two individuals under comparison.

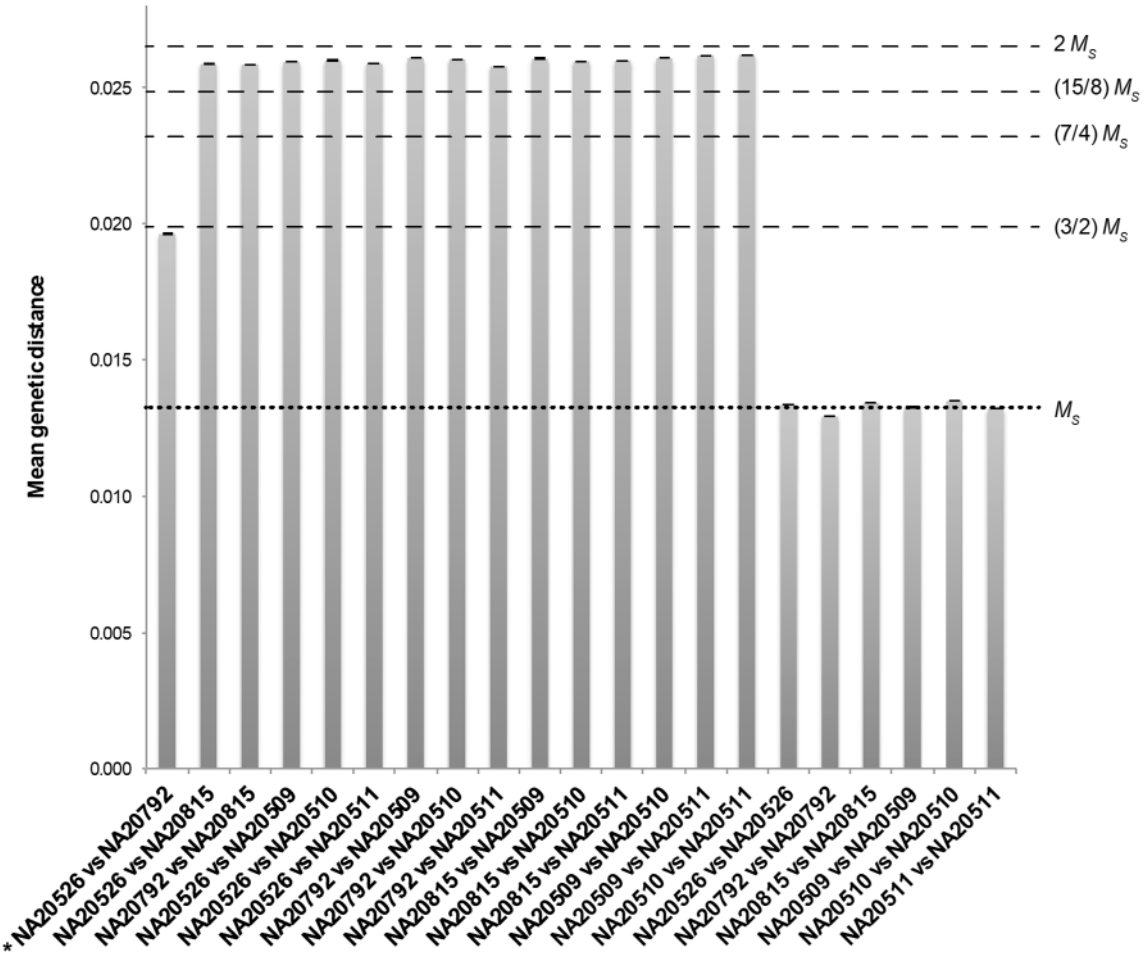

Figure S7. Pairwise genetic distances observed between individuals from the Tuscan (TSI) population. Error bars indicate ± 2 SD (standard deviation) around the mean from 20 replicates generated by randomly sampling from available sequences. Positions were considered in a pairwise comparison only if they had a minimum sequence depth of 2 in both individuals. * indicates the known sibling relationship between individuals NA20526 and NA20792.

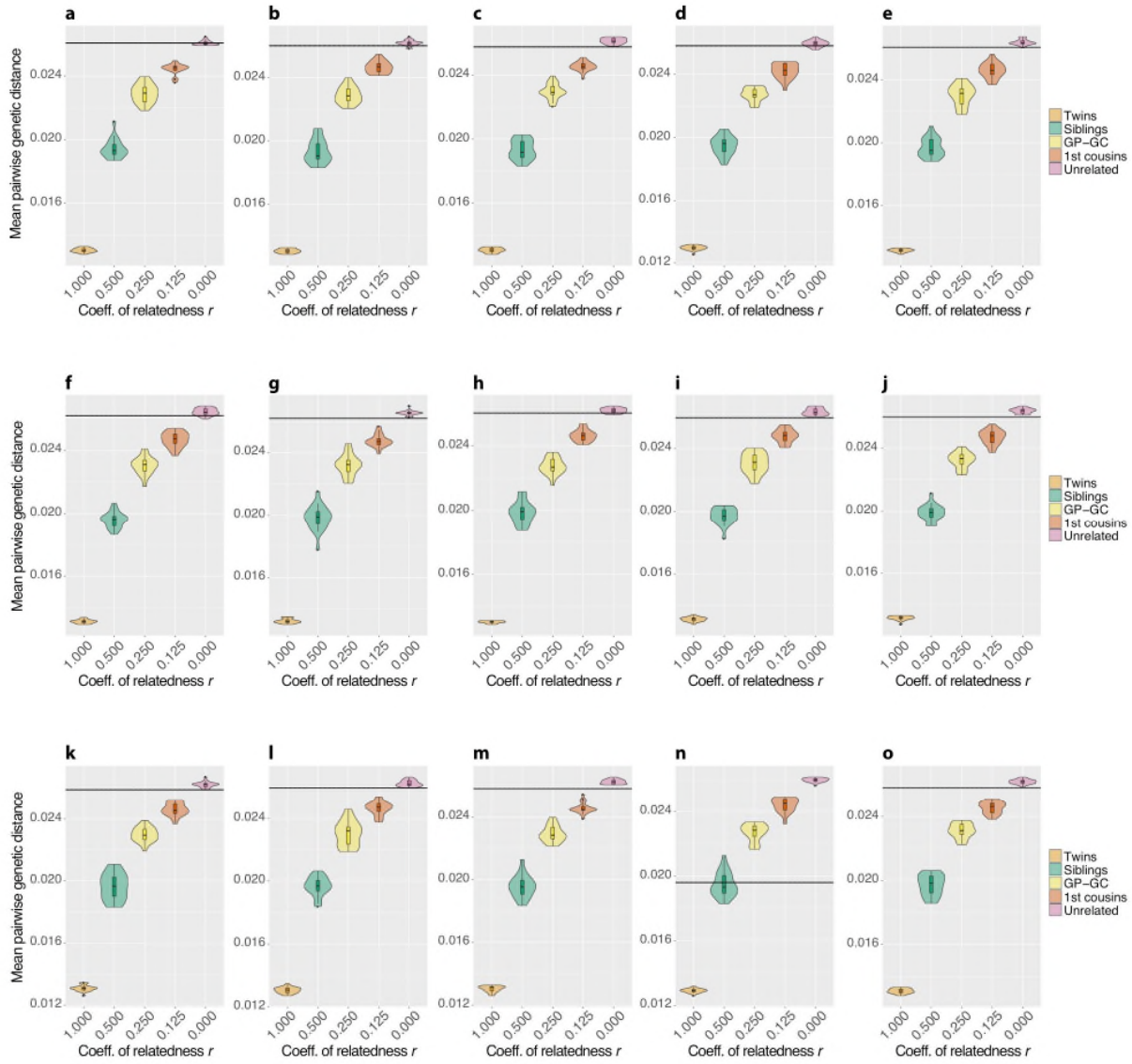

Figure S8. Results of pedigree simulations corresponding to pairwise comparison of aligned sequence data from the Tuscan population. Violin plots represent the distribution of pairwise differences per site in 100 replicates of each analyzed relationship. Simulations were performed with random EUR individuals under the following parameters: SNP positions passing depth and quality filters including a minimum sequence depth of 2 in aligned sequence data from each pairwise comparison, 0% sequencing error, 0%contamination rate. GP–GC, grandparent-grandchild relationship. OR, odds ratio against second most likely coefficient of relationship. **a**, NA20792 vs. NA20509 (r = 0, OR > 10^9). **b**, NA20792 vs. NA20510 (r = 0, OR > 10^4). **c**, NA20792 vs. NA20511 (r = 0, OR > 10^3). **d**, NA20792 vs. NA20815 (r = 0, OR > 10^3). **e**, NA20509 vs. NA20510 (r = 0 or 0.125, OR > 10^5). **f**, NA20509 vs. NA20511 (r = 0, OR > 10^3). **g**, NA20510 vs. NA20511 (r = 0 or 0.125, OR > 10^4). **h**, NA20815 vs. NA20509 (r = 0, OR > 10^4). **i**, NA20815 vs. NA20510 (r = 0 or 0.125, OR > 10^5). **j**, NA20815 vs. NA20511 (r = 0 or 0.125, OR > 10^5). **k**, NA20526 vs. NA20509 (r = 0 or 0.125, OR > 10^5). **l**, NA20526 vs. NA20510 (r = 0 or 0.125, OR > 10^4). **m**, NA20526 vs. NA20511 (r = 0 or 0.125, OR > 10^6). **n**, NA20526 vs. NA20792 (r = 0.50, OR > 10^{10}). **o**, NA20526 vs. NA20815 (r = 0 or 0.125, OR > 10^7).

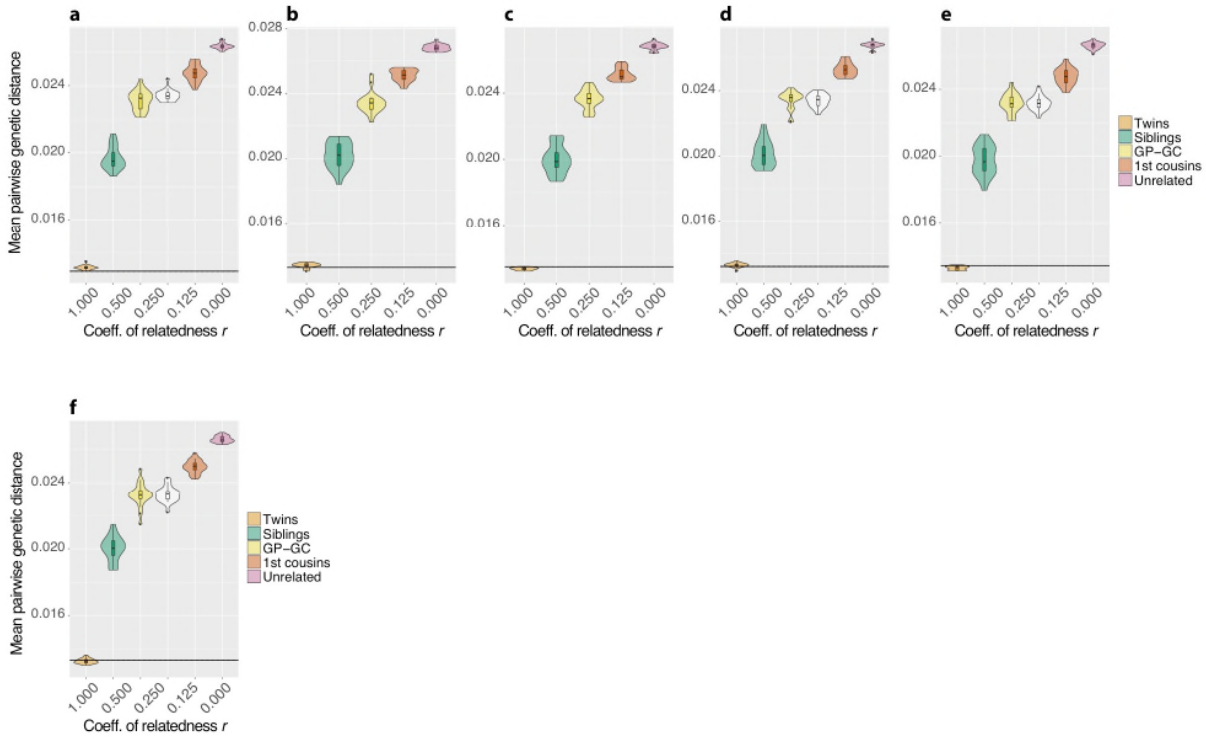

Figure S9. Results of pedigree simulations corresponding to self-comparison of aligned sequence data from Tuscan individuals. Violin plots represent the distribution of pairwise differences per site in 100 replicates of each analyzed relationship. Simulations were performed with random EUR individuals under the following parameters: SNP positions passing depth and quality filters including a minimum sequence depth of 2 in aligned sequence data from each pairwise comparison, 0% sequencing error, 0% contamination rate. GP–GC, grandparent-grandchild relationship. OR, odds ratio against second most likely coefficient of relationship. **a**, NA20792 vs. NA20792 (r = 1, OR > 10^{23}). **b**, NA20509 vs. NA20509 (r = 1, OR > 10^{17}). **c**, NA20510 vs. NA20510 (r = 1, OR > 10^{14}). **d**, NA20511 vs. NA20511 (r = 1, OR > 10^{18}). **e**, NA20815 vs. NA20815 (r = 1, OR > 10^{12}). **f**, NA20526 vs. NA20526 (r = 1, OR > 10^{19}).

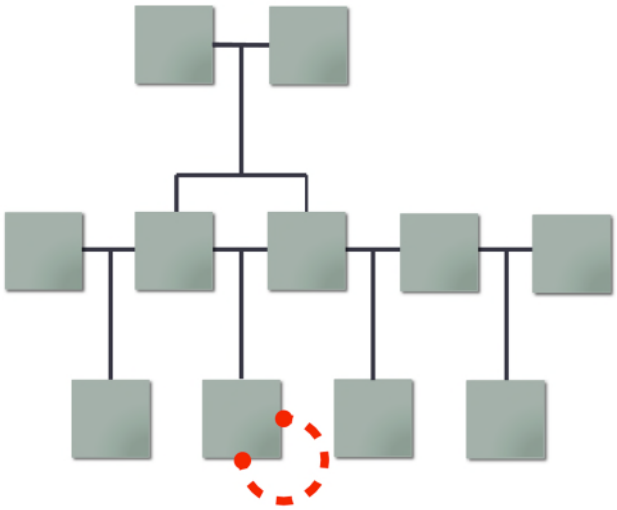

Figure S10. Diagram of the family pedigree used for simulations including an inbred individual that is offspring of siblings. Arrows indicate haploid parental contributions to offspring. The dashed line indicates the relationship of an inbred individual to itself. All individuals are unrelated unless otherwise indicated.

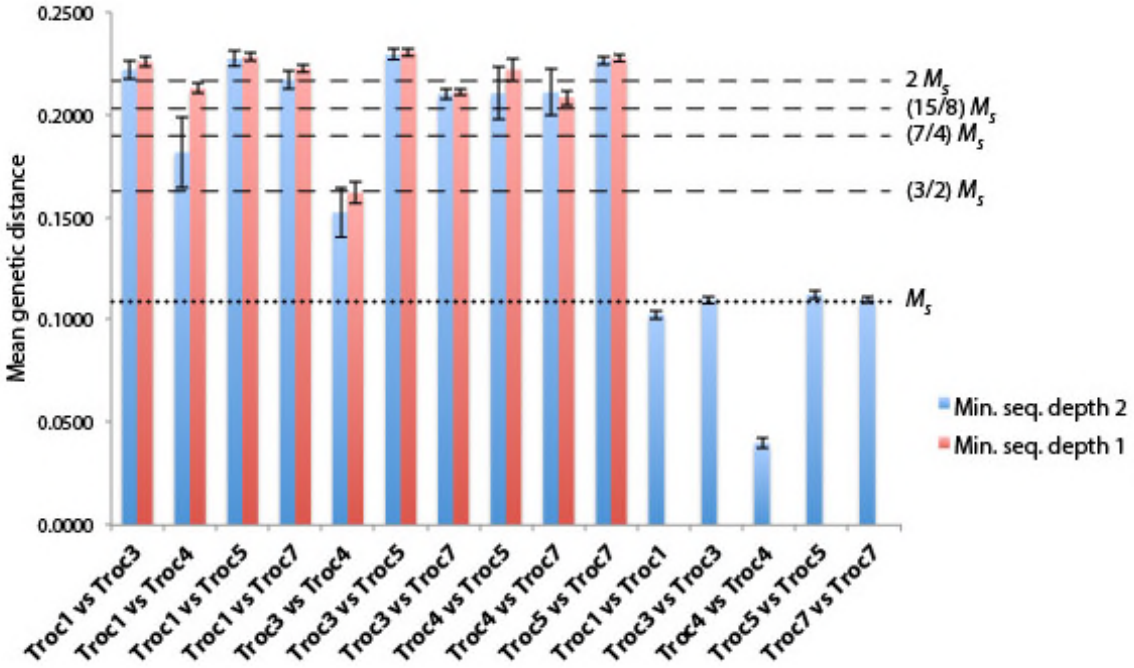

Figure S11. Pairwise genetic distances observed between ancient human individuals from the Els Trocs cave site in Spain. Samples were directly dated 5,311 to 5,066 cal BCE. Whiskers indicate ± 2 SD around the mean of 100 replicates generated by randomly sampling from available sequences. Sequences were considered only if they supported known transversion alleles. A mean of 22,699 target genomic positions with sequence depths ≥ 2 in both individuals were examined, although overlapping sites were as few as 420 in one case (Troc1 vs. Troc4). When targeting genomic positions with sequence depths ≥ 1 in both individuals, a mean of 28,042 positions passed filters, and the smallest number of overlapping sites assessed was 3,467 (Troc1 vs. Troc4). M_S indicates the mean of within-individual comparisons from the four putatively outbred individuals Troc1, Troc3, Troc5, and Troc7.

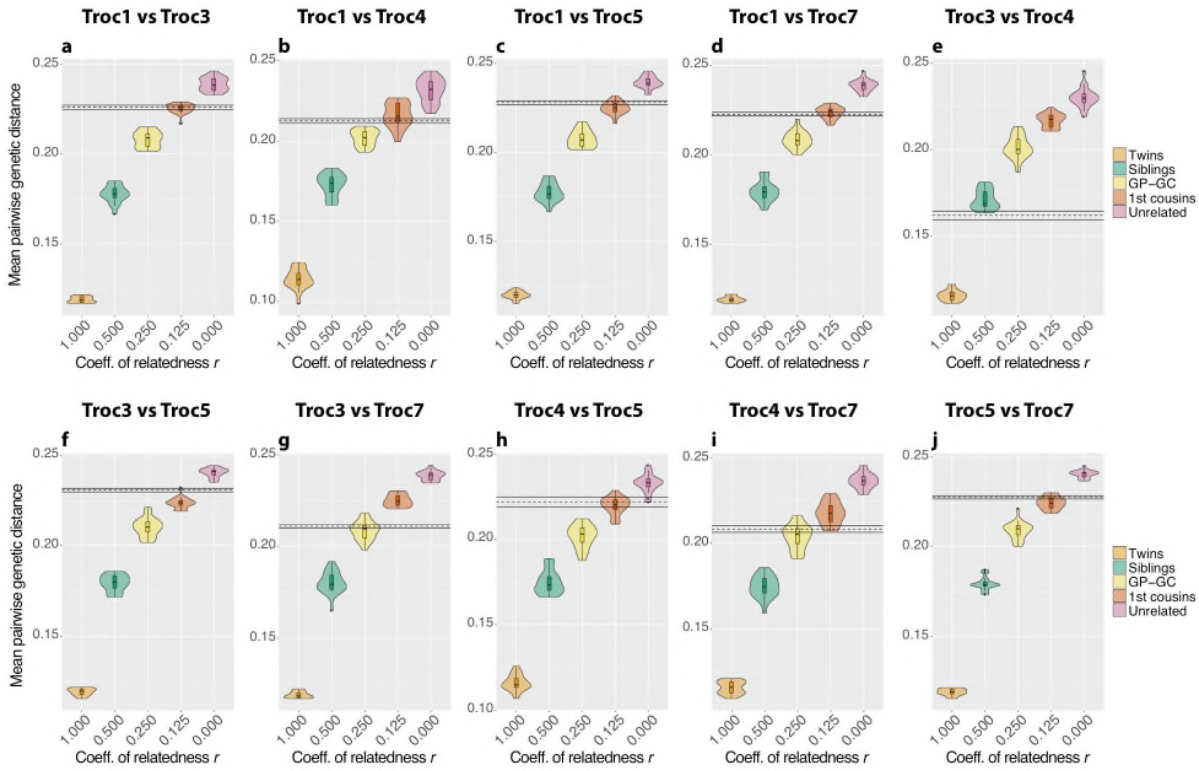

Figure S12. Results of pedigree simulations (without heterozygosity down-sampling) corresponding to pairwise comparison of aligned sequence data from ancient human individuals from the Els Trocs cave site in Spain. Violin plots represent the distribution of pairwise differences per site in 20 replicates of each analyzed relationship. Simulations were performed with random EUR individuals under the following parameters: transversion SNP positions passing depth and quality filters including a minimum sequence depth of 1 in aligned sequence data from each pairwise comparison, sequencing error = 0%, contamination rate = 0%. GP–GC, grandparent-grandchild relationship. **a**, Troc1 vs. Troc3. **b**, Troc1 vs. Troc4. **c**, Troc1 vs. Troc5. **d**, Troc1 vs. Troc7. **e**, Troc3 vs. Troc4. **f**, Troc3 vs. Troc5. **g**, Troc3 vs. Troc7. **h**, Troc4 vs. Troc5. **i**, Troc4 vs. Troc7. **j**, Troc5 vs. Troc7.

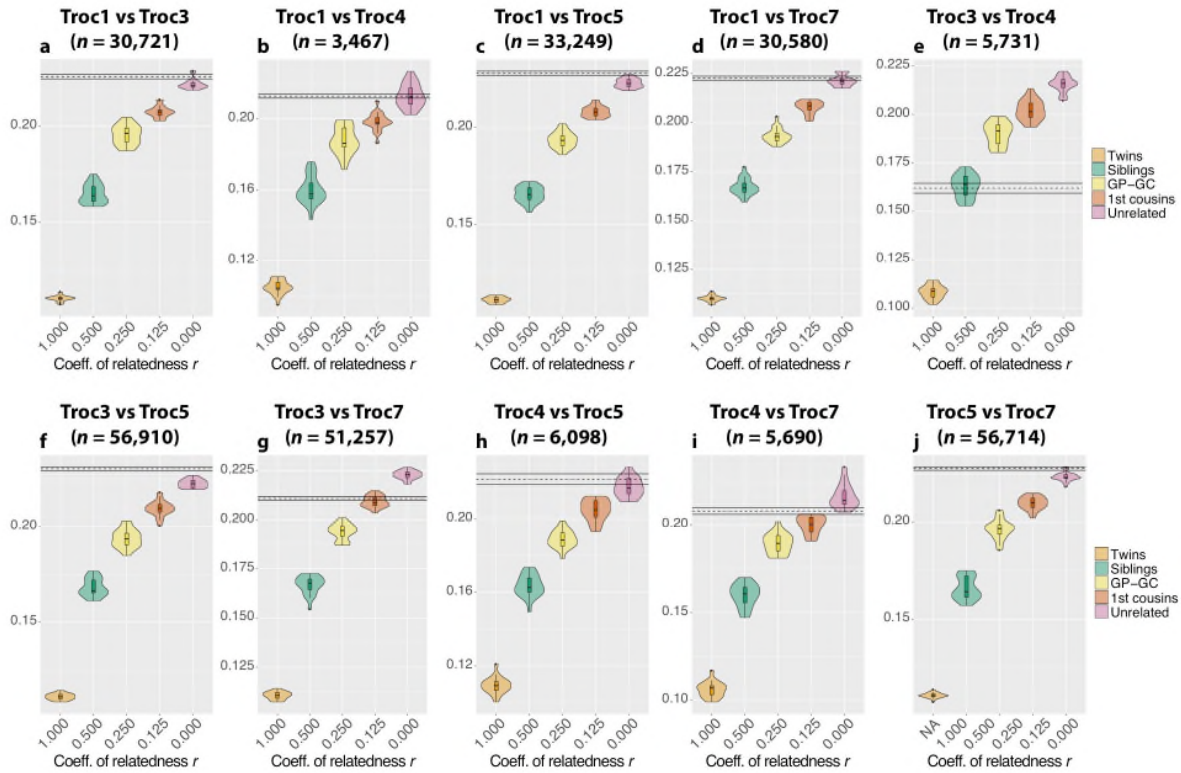

Figure S13. Results of pedigree simulations (with heterozygosity down-sampling) corresponding to pairwise comparison of aligned sequence data from archaeological human remains excavated from the Els Trocs cave site in Spain. The horizontal black lines indicate the mean (±SD) of 20 replicate observations of genetic distance generated by randomly sampling from available aligned sequence data. Simulations were initialized with random, unrelated EUR individuals and were carried out under the following parameters: transversion SNP positions passing depth and quality filters including a minimum sequence depth of 1 in aligned sequence data from each pairwise comparison, sequencing error = 0%, contamination rate = 0%, heterozygosity down-sampling rate = 7.04%. GP–GC, grandparent-grandchild relationship. OR, odds ratio against second most likely coefficient of relationship. **a**, $r = 0.00$, OR > 10^2 . **b**, $r = 0.00$, OR > 10^4. **c**, $r = 0.00$, OR > 10^{10}. **d**, $r = 0.00$, OR > 10^5. **e**, $r = 0.50$, OR > 10^6. **f**, $r = 0.00$, OR > 10^3. **g**, $r = 0.125$, OR > 10^4. **h**, $r = 0.00$, OR > 10^2. **i**, $r = 0.00$ or $r = 0.125$, OR > 10^2. **j**, $r = 0.00$, OR > 10^5.

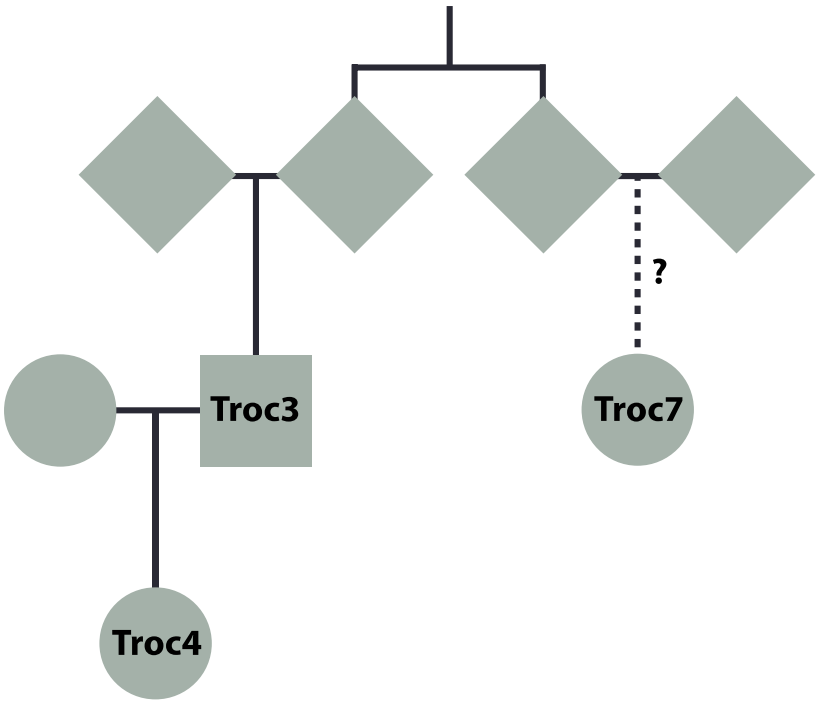

Figure S14. Possible relationship of three individuals from Els Trocs, Spain cave site. Squares represent males and circles represent females. Diamonds represent individuals of unknown sex. The available information leads to the most parsimonious pedigree in which Troc3 was the father of Troc4, Troc7 was the first cousin of Troc3, and Troc7 was the first cousin once removed of Troc4. However, the dashed line indicates that the very low number of overlapping nucleotide positions between Troc7 and Troc3 provides less confidence in Troc7's position in the pedigree.