Popli *et al. Genome Biology* (2023) 24:10
https://doi.org/10.1186/s13059-023-02847-7

Genome Biology

SOFTWARE

Open Access

KIN: a method to infer relatedness from low-coverage ancient DNA

Divyaratan Popli*, Stéphane Peyrégne and Benjamin M. Peter*

*Correspondence: divyaratan_popli@eva.mpg.de; benjamin_peter@eva.mpg.de

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Abstract

Genetic kinship of ancient individuals can provide insights into their culture and social hierarchy, and is relevant for downstream genetic analyses. However, estimating relatedness from ancient DNA is difficult due to low-coverage, ascertainment bias, or contamination from various sources. Here, we present KIN, a method to estimate the relatedness of a pair of individuals from the identical-by-descent segments they share. KIN accurately classifies up to 3rd-degree relatives using at least 0.05x sequence coverage and differentiates siblings from parent-child pairs. It incorporates additional models to adjust for contamination and detect inbreeding, which improves classification accuracy.

Introduction

Why study relatedness?

Identifying related individuals is a common task in genetic studies. Relatedness is of direct interest in, e.g., DNA forensics, where familial search can aid in solving criminal cases, and to identify unknown deceased persons [1, 2]. Genetic paternity tests have an important application in resolving family relation, e.g., in establishing relationship between an individual applying for immigration and the claimed relatives [3]. It is also an essential step in population genetics and association studies, where samples are typically assumed to be independent random draws from the population. For animal and plant breeders and conservation biologists, reconstructing pedigrees and finding related individuals is important to avoid inbreeding and ensure diversity [4–6].

In ancient DNA studies, relatedness can be used to identify bones and teeth belonging to the same individual. Given adequate familiarity with the subject, relatedness can provide an understanding of an ancient society's social structures, mobility and inheritance rules [7–9].

Approaches to estimate relatedness from high-coverage data

Commonly, pairs of related individuals are identified by looking for parts of the genome that are identical by descent (IBD), i.e., inherited from a recent common ancestor. Due to the laws of Mendelian segregation, each parent will share exactly one set of chromosomes IBD with their offspring, while a grandparent will on average share a quarter of their genome with a grandchild due to recombination. Along the genomes of a pair of diploid individuals, there are three IBD states possible at any given position: the individuals share zero, one, or two chromosomes IBD. The genome-wide proportions of these states (usually referred to as k_0, k_1, k_2, so that $k_0 + k_1 + k_2 = 1$) can be used to infer the degree and nature of relatedness for a pair of individuals. For example, a pair of siblings are expected to have all three possible IBD states with proportions of 0.25, 0.5, 0.25, respectively (Fig. 1). These IBD probabilities can directly be used to categorize their relatedness as shown in Table 1. One can also use these probabilities to estimate the coefficient of relatedness r, which is defined as the proportion of the genome that is IBD. In the absence of inbreeding, this would be calculated as $r = k_1/2 + k_2$.

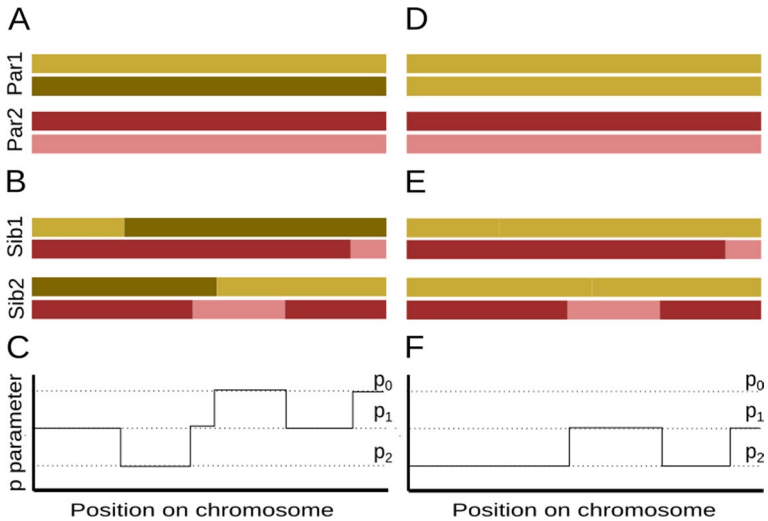

Fig. 1 IBD sharing between siblings without and with runs of homozygosity (ROH). **A** Schematic of chromosomes for two parents (Par1, Par2). **B** Schematics of recombinant chromosomes of two children (Sib1, Sib2). **C** Expected differences between Sib1 and Sib2 (p) along the chromosome. In both cases, p can take values of p_0, p_1 or p_2, which are the expected proportions of differences in IBD states 0,1 and 2, respectively. **D**–**F** Same as **A**–**C**, except parent 1 is assumed to be homozygous

Table 1 IBD sharing probabilities for different relations in absence of inbreeding

Relatedness	k_0	k_1	k_2
Unrelated	1	0	0
3rd degree	0.75	0.25	0
2nd degree	0.5	0.5	0
Siblings	0.25	0.5	0.25
Parent–child	0	1	0
Identical/twins	0	0	1

However, since it is not possible to directly observe IBD segments, a common approach is to first identify segments of the genome that are identical by state (IBS) and to use population allele frequencies obtained from an out-of-sample reference panel to calculate the probability of IBD given IBS. There are several methods that incorporate reference panel allele frequencies, phase information, recombination maps, or genotype calls to co-estimate IBD and the relatedness coefficient [10–20].

Approaches that address problems with ancient DNA data

One major issue with applying the above-mentioned methods to ancient DNA data is that the sequence coverage is typically low, making it difficult to obtain accurate genotype calls. Several methods surmount this problem by using genotype likelihoods [21, 22]. In this way, it is possible to account for the uncertainty in genotype calls by summing over all possible genotypes, weighted by their genotype likelihoods. However, these approaches typically still require at least 2x coverage, since genotype likelihoods may be imprecise at lower coverages [23]. Ancient DNA analyses often face additional challenges such as the unavailability of reference panels to estimate population allele frequencies, contamination with present-day DNA [24], and an ascertainment bias caused by DNA capture approaches [25, 26].

Several methods have been proposed to estimate relatedness without a reference panel from ancient DNA, but they require either > 4x coverage [27], or a large sample size to get an estimate of allele frequencies in the population from which the sampled individuals originate [28]. A second issue is contamination. If the contamination stems from another population, contaminated data will look more dissimilar to other individuals from the analyzed population, and hence relatedness will be underestimated. In addition, some analyzed genomes may have long runs of homozygosity (ROH), for example due to a small population size, or recent inbreeding. Long ROH cause related individuals to seem genetically more similar to each other but do not affect the genetic distance between unrelated individuals.

Moreover, ancient DNA is commonly captured with a SNP array that enriches for informative variants. Particularly, methods based on the fraction of sites in different IBS states are sensitive to the ascertainment bias caused by this non-random selection of targeted sites [27].

READ [29] addresses several of the issues encountered in the analysis of ancient DNA. In particular, the lack of genotype calls is dealt with by randomly sampling alleles from each individual. A string of these alleles at each position (called pseudo-haploids) are then compared to other individuals to estimate average pairwise genetic distances, which in turn are used to infer relatedness. However, READ only infers the degree of relatedness, and only up to second degree.

How our method works

Here, we present KIN (Kinship INference), a hidden Markov model (HMM)-based approach to estimate genetic kinship and IBD from low-coverage ancient DNA data. KIN can detect up to 3rd-degree relatives and differentiates between siblings and parent–child relationships. KIN is also able to take into account ROH and contamination and is not sensitive to SNP ascertainment. We validate the performance of KIN using simulations and

show that we are able to infer relatedness in real data from two datasets: a group of Neandertals and a group of Bronze age individuals.

Results

Algorithm

To infer relatedness, KIN fits one HMM for each pair of individuals and for each possible relatedness. The KIN-HMM infers IBD sharing between a pair of low-coverage individuals, optionally taking ROH tracts and contamination estimates in each individual into account. The best-fit model is then assigned as the inferred relatedness. If the locations of ROH tracts are unknown, we provide another HMM (ROH-HMM) to coarsely estimate the location of ROH for samples with sufficient coverage ($\geq 0.1x$). Our method is available on https://github.com/DivyaratanPopli/Kinship_Inference along with a python package (KINgaroo) to generate the input files for the models directly from bam files.

Model description

The goal of KIN-HMM is to infer how two individuals are related via the patterns of shared IBD states along a pair of genomes. For this purpose, we subdivide the genomes of a pair of individuals into L large genomic windows (typically of size 10 Mb) and infer the pattern of IBD-sharing for each relatedness case we consider (unrelated, $5th$ degree, $4th$ degree, $3rd$ degree, grandparent–grandchild, avuncular, half-siblings, parent–child, siblings, and identical). We then compare the likelihood between all models, and classify each pair to the model with the highest likelihood. We also return the most likely locations of IBD tracts (using the standard Viterbi algorithm), and the IBD state posterior probabilities.

The details of the likelihood computation are given in the "Log likelihood of the KIN-HMM" section. As is the case for any HMM, KIN-HMM requires a set of emission (the "Emission probability" section) and transition probability matrices. We assume the transition matrix for each relatedness case is fixed. Our emissions include a vector of parameters (δ) that describe the variance in the data for each IBD state that we infer from the data (the "Emission probability" section).

Input of KIN-HMM

The inputs of our algorithm are (i) the number of overlapping sites for the wth window N_w for which both samples have at least one read available, (ii) the number of pairwise differences D_w at these sites, and (iii) the probability of ROH in windows, by default obtained from ROH-HMM described in the "ROH estimation model" section.

For high-coverage data, D_w can be directly obtained by comparing genotypes, but for low-coverage samples, D_w needs to be estimated from the sequencing data. The simplest approach is to randomly sample a read from each position [30–32]. However, such an approach may result in loss of data, and hence we estimate D_w by implicitly summing over all possible samplings:

$$D_w = \sum_{s=1}^{N_w} v_i(s)(1 - v_j(s)) + (1 - v_i(s))v_j(s) \quad (1)$$

Here, $v_i(s)$ and $v_j(s)$ are the proportions of reads carrying the derived allele at SNP index s for individuals i and j, respectively. Throughout, we will use bold-face notation to refer to the vector (or matrix) collecting all the terms, e.g. $\mathbf{D} = (D_1, D_2, \ldots D_L)$.

Log likelihood of the KIN-HMM

The KIN-HMM uses **D** and **N** to classify each window into three hidden states $Z_w \in (0, 1, 2)$, reflecting zero, one, or two shared chromosomes IBD, respectively. To take ROH into account, we define the variable $H_w \in (0, 1, 2)$ that designates that zero, one, or both individuals are homozygous in window w. Since H_w is unobserved, in practice, we use the estimates from ROH-HMM: $h_{wj} = P(H_w = j)$.

There are three additional model parameters, $\boldsymbol{\pi}, \mathbf{A}$ and $\boldsymbol{\delta}$. The transition matrix **A** gives the probability of moving from state i to state j, given by a_{ij}, and is fixed and estimated from simulations for each relatedness case (the "Simulations" section). The initial probabilities $\boldsymbol{\pi}$ give the probabilities of being in each state Z_0 (at the beginning of each chromosome), which we set to the stationary distribution of the transition matrix for simplicity. The over-dispersion parameter $\boldsymbol{\delta}$ takes into account that SNPs in each window vary in their allele frequencies (see next section). For compactness of notation, we group the fixed parameters: $\theta = (\mathbf{N}, \mathbf{A}, \boldsymbol{\pi})$.

Thus, the complete data likelihood for the HMM is

$$\begin{aligned}\log P(\mathbf{D}, \mathbf{Z}|\mathbf{H}, \theta, \boldsymbol{\delta}) &= \log P(\mathbf{D}|\mathbf{Z}, \mathbf{H}, \theta, \boldsymbol{\delta}) + \log P(\mathbf{Z}|\theta) \\ &= \sum_w \log P(D_w|Z_w, H_w, N_w, \boldsymbol{\delta}) + \sum_w \log P(Z_w|Z_{w-1}, \boldsymbol{A}) + \log P(Z_0|\boldsymbol{\pi}).\end{aligned} \tag{2}$$

Here, **Z** is not dependent on **H** and $\boldsymbol{\delta}$.

Emission probability

Using this setup, we can isolate the emissions $P(D_w|Z_w, H_w, \theta, \boldsymbol{\delta})$ from Eq. (2) and optimize them for $\boldsymbol{\delta}$. The simplest model is to assume that sites in each window are equally distributed and independent. In this case, we could use the binomial likelihood:

$$P(D_w|Z_w, H_w, N_w) \sim \text{Binom}[D_w; p(Z_w, H_w), N_w],$$

where p is the proportion of differences expected for a particular IBD and ROH state. If the two individuals are unrelated in a particular window (i.e. $Z_w = 0$), then the expected proportion of pairwise differences depends solely on the population history, and we denote this proportion with p_0. If the two individuals share one or even both copies of the genome IBD, we would expect the proportion of differences to be reduced to $p_1 = \frac{3}{4}p_0$, and $p_2 = \frac{1}{2}p_0$, respectively, since either one or two of the four possible comparisons will be between identical chromosomes [29]. Thus, $p(Z_w = i, H_w = 0) = p_i$.

The proportion of differences between unrelated individuals p_0 is an important parameter. We follow READ [29] and estimate p_0 as the median of differences for all possible pairs of individuals, which works well if the majority of individuals in the sample are unrelated.

The presence of long tracts of homozygosity resulting from recent inbreeding adds an additional complication, as the number of shared chromosomes may be overestimated

[33]. For example, when considering two bones from the same individual, we would expect the entire genome to have a pairwise difference of p_2, because two out of the four compared chromosomes are identical copies. However, in inbred regions, all four chromosomes will be identical, and so the expected pairwise differences are zero (p_4 in Eq. (3)). Note, however, that p_0 does not depend on the presence of ROH, since all comparisons are between unrelated chromosomes even if both individuals are homozygous at a particular locus.

Taken together, we can summarize $\boldsymbol{p}$ in the following matrix, where rows give the state of Z_w, and columns of H_w:

$$p(Z_w, H_w) = \begin{bmatrix} p_0 & p_0 & p_0 \\ p_1 & p_2 & p_4 \\ p_2 & p_4 & p_4 \end{bmatrix} \tag{3}$$

As explained above, we would expect p_4 to be zero. However, as we do all our calculations in large windows, the start/end positions of windows may not coincide with that of ROH tracts, and we found that we obtain better results by setting $p_4 = \frac{p_2}{2}$, to take into account that many windows will only partially have four comparisons between identical chromosomes.

The effect of these considerations is that even though we have nine possible combinations of Z_w and H_w for each window, there are actually only four unique p-parameters p_i with $i \in (0,1,2,4)$.

Beta binomial model

We empirically find that the data often has considerably higher variance than would be expected from a binomial model (Additional file 1: Fig. S2). We take this into account by adding an over-dispersion parameter $\boldsymbol{\delta}$. Just like $p(Z_w, H_w)$, $\delta(Z_w, H_w)$ depends on the number of chromosomes compared, and so each of the four p_i has a corresponding δ_i parameter.

Taken together, our emission probabilities are

$$\begin{aligned} P(D_w|Z_w, H_w, N_w, \boldsymbol{\delta}) \sim & BB[D_w; p_i, \delta_i, N_w] \\ = & \binom{N_w}{D_w} \frac{B(D_w + p_i\delta_i, N_w - D_w + \delta_i(1-p_i))}{B(p_i\delta_i, (1-p_i)\delta_i)}, \end{aligned} \tag{4}$$

where i is fully determined by the combination of Z_w and H_w (see Eq. (3)).

This parameterization of the beta distribution in terms of expected value p and over-dispersion δ is also called the Balding-Nichols model [34] and is distinct from the more common parameterization in terms of α and β. We use this equation even if preprocessing steps (see Eq. (1) and the "Contamination correction" section) result in non-integer D_w and N_w, in which case we approximate the binomial coefficient using Gamma functions.

Estimation of δ

We estimate the δ-parameters using an expectation-maximization (EM) algorithm [35].

Initialization

The value of δ_i is unknown to start with, and we set it to a random value between 0 and 1000.

Expectation step

In the t-th iteration, we calculate the posterior probability of each IBD state in each window $\gamma_{wi}^{(t)} = P(Z_w = i|D_w, H_w, \theta_i, \delta_i^{(t)})$ using the forward-backward algorithm, where $\delta_i^{(t)}$ is the current estimate for δ for a given IBD state.

Maximization step

The only free parameters we estimate in the maximization step are the over-dispersion parameters δ_i. We do this optimization using a cost function, which is the log-emission probability weighted by the posterior probabilities of the hidden states γ_{wj} and optionally the ROH state-probabilities $h_{w\omega}$ obtained from the ROH-HMM.

$$\begin{aligned}\mathcal{C} =& \mathbb{E}[\log P(D_w|Z_w, H_w, \theta, \delta^{(t-1)})] \\ =& \sum_{w=1}^{L}\sum_{j=0}^{2}\sum_{\omega=0}^{2} \log P(D_w|N_w, \delta(j,\omega)^{(t-1)}, p(j,\omega))h_{w\omega}\gamma_{wj}\end{aligned} \tag{5}$$

Using Eq. (3), we simplify this by grouping all the terms that have the same number of pairwise comparisons between identical chromosomes, which would result in the same p_i and δ_i, i.e.

$$\begin{aligned}g_{w0} =& \gamma_{w0} \\ g_{w1} =& \gamma_{w1}h_{w0} \\ g_{w2} =& \gamma_{w2}h_{w0} + \gamma_{w1}h_{w1} \\ g_{w3} =& 0 \\ g_{w4} =& \gamma_{w2}h_{w1} + \gamma_{w2}h_{w2} + \gamma_{w1}h_{w2},\end{aligned} \tag{6}$$

where g_{w3} is always 0 because there is no case that leads to only three comparisons between identical chromosomes.

So, we can rewrite Eq. (5) as:

$$\mathcal{C} = \sum_{i=0}^{4}\sum_{w=1}^{L} \log P(D_w|N_w, \delta_i, p_i)g_{wi} \tag{7}$$

The cost function (Eq. (7)) has one independent term for each i and so we can separate them and estimate each δ_i independently using the minimize_scalar algorithm implemented in `scipy.optimize` [36].

We constrain the optimization space of the δ_i, as unconstrained optimization could result in some confounding of cases. We know that different cases of relatedness have different numbers of IBD states possible. For example, siblings may have all three IBD states present while a parent–child pair has only $Z_w = 1$. However, the parent–child model could fit data generated under the sibling model by assigning it a very high δ_1, which would reduce the performance (for example, see Additional file 1: Fig. S5, S6). We

avoid this problem by constraining the δ such that the beta distributions for the different i overlap by at most one standard deviation.

Model comparison

To infer the most likely relatedness case, we run our model on all relatedness cases mentioned in the "Model description" section and compare the resulting likelihoods. We output the relatedness corresponding to the maximum likelihood model. Since we compare models where the parameters are not subset of each other, standard likelihood-ratio theory for nested models cannot be used to obtain confidence intervals. Instead, we use the log-likelihood ratio between the two best models as a statistic to assess the confidence in our classifications and use simulations to obtain critical values (Additional file 1: Fig. S8).

Grouping of cases

Particularly for low-quality data, we may not be able to distinguish all cases. Thus, we group $4th$- and $5th$-degree relatives together with unrelated. Similarly, we group half-siblings, avuncular and grandparent-grandchild to $2nd$-degree relatives in the final results. We report the final pairwise classification in the following categories: unrelated, third degree, second degree, parent–child, siblings, identical individuals.

Critical values

To investigate the limits of our method, we plotted the true-positive and false-positive rates for classification of different relatedness in control simulations (without contamination, ROH and ascertainment bias, see the "Simulations" section) when we use a particular difference in log-likelihood as a cutoff (Additional file 1: Fig. S8). The figure shows that for all relatedness cases except for $3rd$-degree, the false positive rate is below 5% even when simply selecting the model with the highest likelihood. We observe that for $3rd$-degree relatives, using a cutoff of 1.0 brings down the false positive rate close to 5% for all coverages except 0.05x. Thus, we recommended using a cutoff of 1.0 for all cases where ROH and contamination are not a concern.

Example case of siblings

In Fig. 2, we show the inferred IBD fragments from different KIN-HMMs when they are applied to simulated data from a pair of siblings. The models for identical, parent–child, or unrelated relationships allow for just one IBD state, resulting in a flat line and low likelihood for this data. The other three cases all allow for different IBD states, but the siblings-model predictions match true IBD states the most, as reflected by the highest log-likelihood and the close correspondence of the inferred and true IBD states.

ROH estimation model

Our HMM to detect ROH tracts works similarly to the KIN-HMM described above, but in this case, we only consider one individual at a time and only consider positions covered by at least two reads. For each site, we calculate the proportion of reads that carry different alleles and sum them up in windows along the genome. We call the vector with the number of differences Δ, and the vector with the number of sites with at least

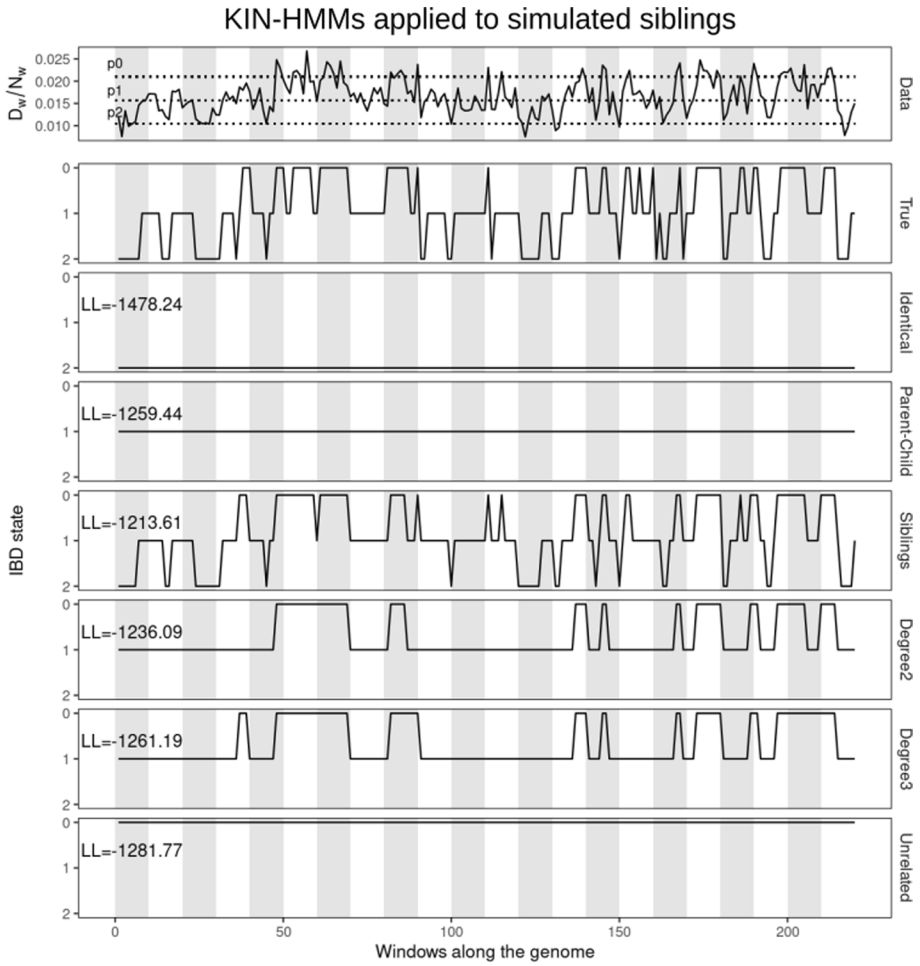

Fig. 2 Comparison of pairwise difference data and inferred IBD fragments. The top panel shows the proportion of differences in each window along the genome for a pair of simulated siblings. Dashed lines represent p_0, p_1 and p_2 estimates. The second panel shows the true IBD state for each window. The remaining panels show the IBD states predicted by particular relatedness models. The log-likelihood value for each model is shown on upper left corner of the panel. Light and shaded backgrounds represent distinct chromosomes

two reads **M**. Our model has two possible hidden states: homozygous state ($Y_w = 4$), and non-homozygous state ($Y_w = 2$). As above, we collect the hidden states in a vector $\mathbf{Y} = (Y_1, \ldots Y_w, \ldots, Y_L)$. The complete data likelihood for the model in this case is then:

$$\begin{aligned} logP(\Delta, \mathbf{Y}|\Theta, \boldsymbol{\delta}) =& logP(\Delta|\mathbf{Y}, \Theta, \boldsymbol{\delta}) + logP(\mathbf{Y}|\Theta) \\ =& [\sum_w logP(\Delta_w|Y_w, M_w, \boldsymbol{\delta}) + \sum_w logP(Y_w|Y_{w-1}, \boldsymbol{A})] + P(Y_0|\boldsymbol{\pi}), \end{aligned} \quad (8)$$

where Θ is a vector of initial probabilities ($\boldsymbol{\pi}$), transition matrix ($\boldsymbol{A}$) and $\boldsymbol{M}$.

Since the source of ROH may not be known, we estimate both transitions and emissions. We calculate the emissions using a beta-binomial likelihood, and fix the mean of the distributions corresponding to $Y_w = 4$ and $Y_w = 2$ at expected proportion of differences in a homozygous tract (p_4) and expected proportion of differences in a non-homozygous tract (p_2), respectively. The expectation step outputs the posterior

probability Γ of being in state $Y_w = 4$ or the state $Y_w = 2$ in each window. The maximization step for emissions is analogous to that in the KIN model (Eq. 5), and the optimization step here is done with the following cost function:

$$\mathcal{C} = \sum_{w=1}^{L} \sum_{i} P(\Delta_w|\delta_i, p_i)\Gamma_{wi}, \tag{9}$$

where $P(\Delta_w|\delta_i, p_i)$ is a beta-binomial probability with mean p and over-dispersion parameter δ similar to Eq. 4, and i can take values 2 and 4 corresponding to the hidden states Y_w.

To estimate transitions, we initialize the transition matrix with the value 0.2 for the off-diagonal entries, and update it using the standard Baum-Welch update step [37]. Similar to the KIN-HMM, we avoid fitting issues by forcing all windows whose proportion of differences is larger than p_2 to be in the non-homozygous state (Additional file 1: Fig. S7).

Contamination correction

Contamination by DNA from present-day people is a common feature of human ancient DNA datasets [38]. To address this issue, we developed a heuristic that adjusts both D_w and N_w to minimize the influence of contamination on the relatedness inference.

We assume that contamination rates in both individuals are known and small ($< 5\%$), and set $C_{ij} = C_i + C_j$, where C_i, C_j are the contamination estimates from the two individuals. We also assume the divergence ϕ between our target population and the putative contaminant popoulation is known. With probability C_{ij}, a comparison between two random reads from the pair of tested individual will contain a contaminant read, and thus contain a difference with probability ϕ, and with probability $1 - C_{ij}$ it will be between endogenous ones. The comparisons between two contaminant reads are ignored, since we assume C_{ij} to be small.

We estimate the expected number of differences from comparison of endogenous reads $\mathbb{E}[D_w^{'}]$, and the total number of sites with overlapping endogenous reads $\mathbb{E}[N_w^{'}]$.

For any particular comparison showing a difference D (not to be confused with the number of differences D_w), we calculate the probability of the event E that it is between endogenous reads as

$$P(E|D) = \frac{P(D|E)P(E)}{P(D|E)P(E) + P(D|\neg E)P(\neg E)}. \tag{10}$$

Then, by linearity of expectation, we obtain our estimator for the expected number of endogenous comparisons with a difference as

$$\mathbb{E}[D_w^{'}] = D_w P(E|D) = D_w \times \frac{P(D|E)P(E)}{P(D|E)P(E) + P(D|\neg E)P(\neg E)} \tag{11}$$

.

Of these terms, $P(E) = 1 - C_{ij} = 1 - P(\neg E)$, and $P(D|\neg E) = \phi$.

For $P(D|E)$, we use an estimator based on the genome-wide average:

$$P(D|E) = \rho = \frac{\frac{\sum_w D_w}{\sum_w N_w} - C_{ij}\phi}{1 - C_{ij}}. \tag{12}$$

Taken together,

$$\mathbb{E}[D_w^{'}] = D_w \frac{\rho(1 - C_{ij})}{\rho(1 - C_{ij}) + C_{ij}\phi} \tag{13}$$

Analogous considerations lead to the expected number of endogenous comparisons that yield no difference:

$$\mathbb{E}[S_w^{'}] = S_w P(E|\neg D) = S_w \frac{(1 - \rho)(1 - C_{ij})}{(1 - \rho)(1 - C_{ij}) + C_{ij}(1 - \phi)}. \tag{14}$$

Here, $S_w = N_w - D_w$. Hence, we set $\mathbb{E}[N_w^{'}] = \mathbb{E}[S_w^{'}] + \mathbb{E}[D_w^{'}]$. We do a similar contamination correction for the input of ROH-HMM.

Evaluation with simulations

We first tested the performance of KIN with simulated pedigrees. We performed coalescent simulations to generate 8 unrelated diploid genomes and artificially mated them to form pedigrees of 17 individuals with relationships up to 5*th* degree (Additional file 1: Fig. S3). To evaluate the effect of sequence coverage on the performance of KIN, we generated artificial pileups at each polymorphic site for each individual following a Poisson distribution with 6 different average depths varying between 4*x* and 0.03*x* (see the "Materials and methods" section). To mimic ROH, for some pedigrees, we picked a single allele at heterozygous sites in some regions as determined by a Markov chain so that on an average about 17% of the genome is ROH. We also created versions with contamination, by introducing alleles from distantly related individuals, and ascertainment bias, by selecting polymorphic sites identified in a subset of the individuals (see the "Materials and methods" section). We created 60 pedigrees for each combination of average coverage and scenarios of presence/absence of ROH, contamination, and ascertainment bias, totalling 2880 pedigrees.

ROH detection

In Fig. 3, we present an example of data and inference of the ROH-HMM for simulations with and without ROH, potentially with ascertainment bias and contamination. In all cases, we find that the inferred ROH closely matches the simulations, but the confidence in the classification tends to increase with the simulated coverage.

A systematic evaluation of the performance of the ROH-HMM is given in Table 2: for the purpose of this analysis, we classified all windows that were at least 20% homozygous as ROH and the remainder as non-ROH. Likewise, we classified all windows with a posterior probability of ROH of at least 20% as ROH. We used these cases to compute sensitivity (Se) and specificity (Sp). In the control case (simulations with no ascertainment,

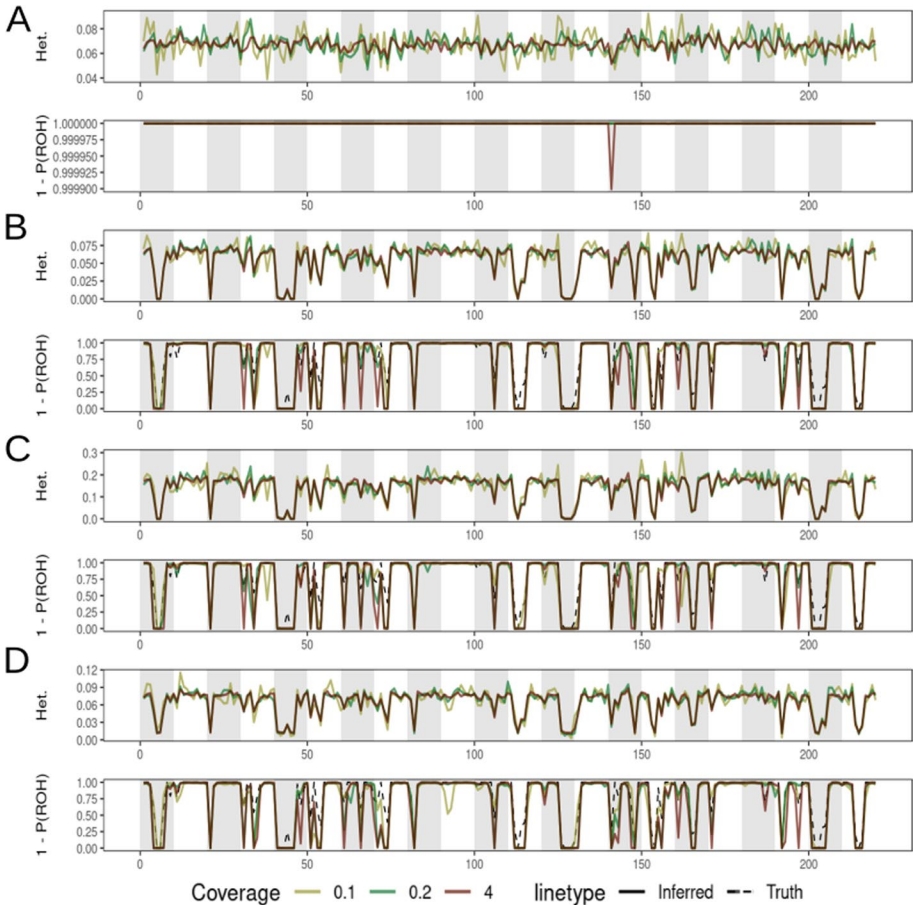

Fig. 3 Estimation of ROH probabilities along the genome in simulations. The top row in each panel shows the proportion of differences in a simulated individual along the genome. In the bottom row, the dotted line shows the proportion of each window not in ROH, and the solid lines show the estimated probability of not being in the ROH state. **A** Simulation with no ascertainment, contamination, or ROH. **B** Simulation with ROH. **C** Simulation with ROH and ascertainment. **D** Simulation with ROH and contamination

Table 2 Model performance for ROH prediction. Here, we test ROH-HMM in four different cases of simulations: control (without ROH, ascertainment bias or contamination), R (with ROH), RA (with ROH and ascertainment bias), RC (with ROH and contamination)

Coverage	Control (Sp)	R (Se)	R (Sp)	RA (Se)	RA (Sp)	RC (Se)	RC (Sp)
4x	> 0.99	0.97	0.98	0.97	0.98	0.97	0.97
0.5x	> 0.99	0.96	0.98	0.96	0.98	0.96	0.97
0.2x	> 0.99	0.92	0.98	0.91	0.98	0.93	0.97
0.1x	> 0.99	0.85	0.98	0.83	0.97	0.82	0.97
0.05x	0.99	0.65	0.96	0.55	0.96	0.57	0.96

contamination, or ROH) we see that the specificity remains ≥ 0.99 for all coverages. In the cases where we simulate ROH, we find that sensitivity decreases from 97% at 4x coverage to less than 65% at 0.05x coverage, but specificity remains high at above 0.96 for all cases, suggesting that the number of erroneously called ROH segments is low.

IBD prediction

We investigated the accuracy of IBD state prediction along the genome by counting the number of genomic windows where we correctly predict IBD state for different relatedness and coverages (Fig. 4). The accuracy for coverages of 0.1x or higher is consistent and varies between 1.0 and 0.78, depending on relatedness. However, the accuracy decreases at lower coverages for most relatedness cases and ranges from 1.0 to 0.48 at 0.03x. The exceptions are for identical individuals and parent–child pairs, where the accuracy is nearly perfect for all investigated coverages, even at 0.03x coverage. We note that the accuracy of IBD prediction is lowest for siblings, followed by 2nd-degree and 3rd-degree relatives. This is because IBD in siblings is more variable and therefore harder to predict. A pair of 3rd-degree relatives, for example, are expected to only share 12.5% of their genome IBD, with the rest being unrelated. Therefore, even a naive classifier that classified everything as unrelated would have an accuracy of 87.5%. In contrast, siblings are expected to share one chromosome IBD for half their genomes, and both chromosomes for another 25%. This higher variability makes predictions harder.

We find that adding contamination, ascertainment bias in our simulations has little effect on IBD prediction. Adding ROH to the simulations reduces the IBD prediction in two cases: the average accuracy for second-degree relatives decreases from 0.89 to 0.85 and for siblings from 0.81 to 0.70 (Additional file 1: Fig. S9). We see this adverse effect of ROH in case of siblings, although we do not introduce long ROH directly, perhaps because ROH in the parents have an effect (see the "Materials and methods" section). They may cause a difficulty in differentiating between different combinations of IBD (Z_w) and ROH states (H_w). However, this does not affect the power to identify siblings in presence of ROH even at 0.05x (see Fig. 5).

Relatedness classification

We evaluated the classification accuracy of KIN (cutoff: 1 log-likelihood unit) and compared it to that of READ (cutoff: 1 standard deviation)) (Fig. 5). We first describe the results for relatedness cases detectable by READ, viz. identical, 1*st* degree, 2*nd* degree, and unrelated. In this case, we show that both methods have similar performance for low-coverage shotgun data ("control"-case) and for ascertained data. The true positive rate is above 0.97 for both KIN and READ, while the false negative rate

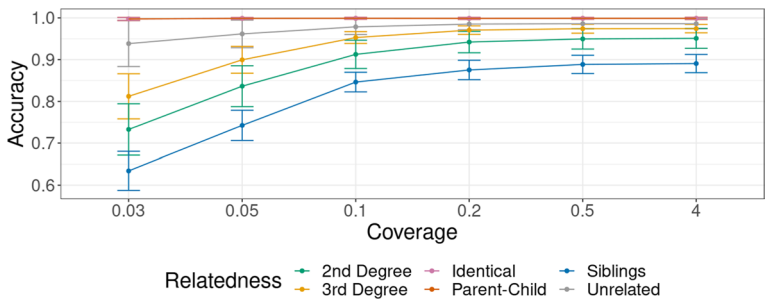

Fig. 4 Evaluation of IBD estimation at different coverages. *y*-axis shows per-window performance accuracy calculated over 60 simulations with each relatedness case and six different coverages shown on *x*-axis. The error bars are drawn at 1 standard deviation from the mean. Here, accuracy corresponding to relatedness cases for parent–child and identical individuals is always 1 and overlaps with each other. Accuracy is defined as the proportion of correctly predicted IBD states, when compared to the central position of the window

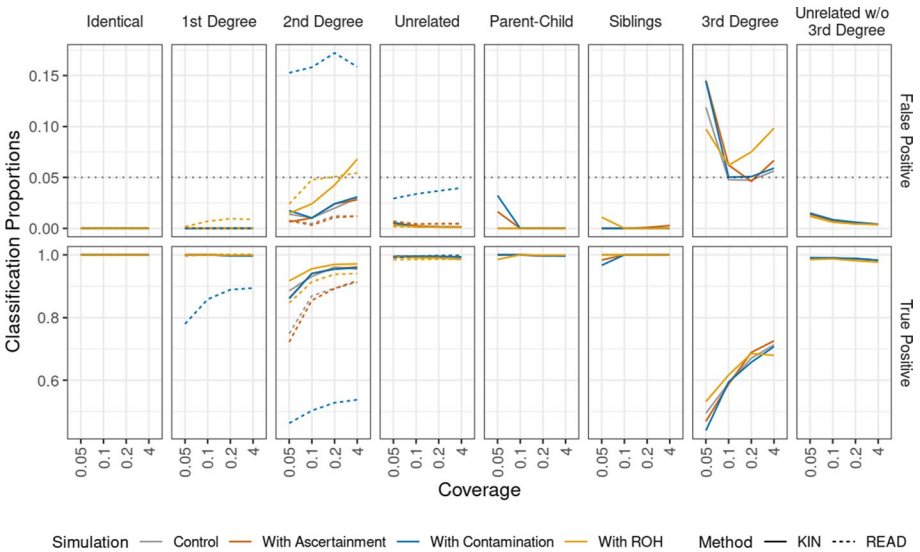

Fig. 5 Comparison of KIN with READ using simulations with different coverages, and different cases of ascertainment, contamination and ROH. "Unrelated" label here refers to KIN performance results when all unrelated, fifth degree, fourth degree, and third degree pairs are labeled as unrelated (for fair comparison with READ). "Unrelated w/o 3*rd* degree" refers to the performance results when 3*rd* degree is classified separately from the unrelated individuals

is below 0.02. One exception is 2*nd*-degree relatives, where KIN has higher power than READ, and as coverage decreases from 4x to 0.05x, KIN's true positive rate decreases from 0.95 to 0.89, compared to range of 0.91 to 0.75 for READ. The false positive rate in this case remains below 0.03 for both methods.

To investigate the impact of contamination, we performed simulations where we added up to 3% contamination to some individuals. We find that READ is strongly impacted by contamination as the true positive rate is in the range 0.89 to 0.78 and 0.54 to 0.46 for 1st- and 2nd-degree relatives, respectively, and the false positive rate reaches up to 0.04 and 0.17 for unrelated individuals and 2nd-degree relatives, respectively. In comparison, we also ran KIN, giving it the simulated contamination amounts for each individual. We find that the correction implemented in KIN is sufficient to remove this bias, and the true and false positive rates remain as in the control (Fig. 5). We also performed an analysis where we misspecified the contamination, and find that the effect of modest misspecification is small (Additional file 1: Fig. S13).

For simulations with ROH, we find that KIN also outperforms READ for 1st- and 2nd-degree relatives, although both the true and false positive rates increased for both methods compared to the control for 2nd-degree relatives. The increase in false positives is likely due to ROH making related individuals more similar. Finally, we also find that KIN has good power to detect relatedness cases that are not detectable by READ, i.e. parent–child, siblings and, to a lower level, 3rd-degree relatives (Fig. 5).

Application to real data

Chagyrskaya and Okladnikov Neandertals

To test KIN on real ancient data, we applied it to a Neandertal dataset from Chagyrskaya and Okladnikov Caves in Siberia, Russia [39–41]. This dataset contains genetic data from a total of 16 skeletal remains that likely belong to contemporary Neandertals who occupied the Chagyrskaya and Okladnikov caves between 59 and 51 kya and at least 44 kya respectively [41]. DNA extracted from each of these remains were captured with an array targeting variable sites identified in high-coverage Neandertal and Denisovan genomes and common variations in Africans [41]. This genetic data has low-to-intermediate depth of coverage ranging from 0.01x to 12.34x, with 8 samples at $< 1x$ coverage. Some of these specimens showed signs of long ROH and DNA contamination from modern humans as well as hyenas [41]. We focused our analysis on the variable sites in two high-coverage Neandertal genomes: Altai Neandertal (Denisova 5) [32] and Vindija 33.19 [42] as done by the authors for the relatedness analysis [41]. Our results for pairwise relatedness for these individuals are shown in Fig. 6. We found three specimens from the same individual (Chagyrskaya13-Chagyrskaya19-Chagyrskaya1141), a parent–child pair (Chagyrskaya07 and Chagyrskaya17), and a pair of 2*nd*-degree relatives (Chagyrskaya01/Chagyrskaya60). Further, we identify Chagyrskaya17 and Chagyrskaya60 as 3rd-degree relatives (Additional file 2: Table S1).

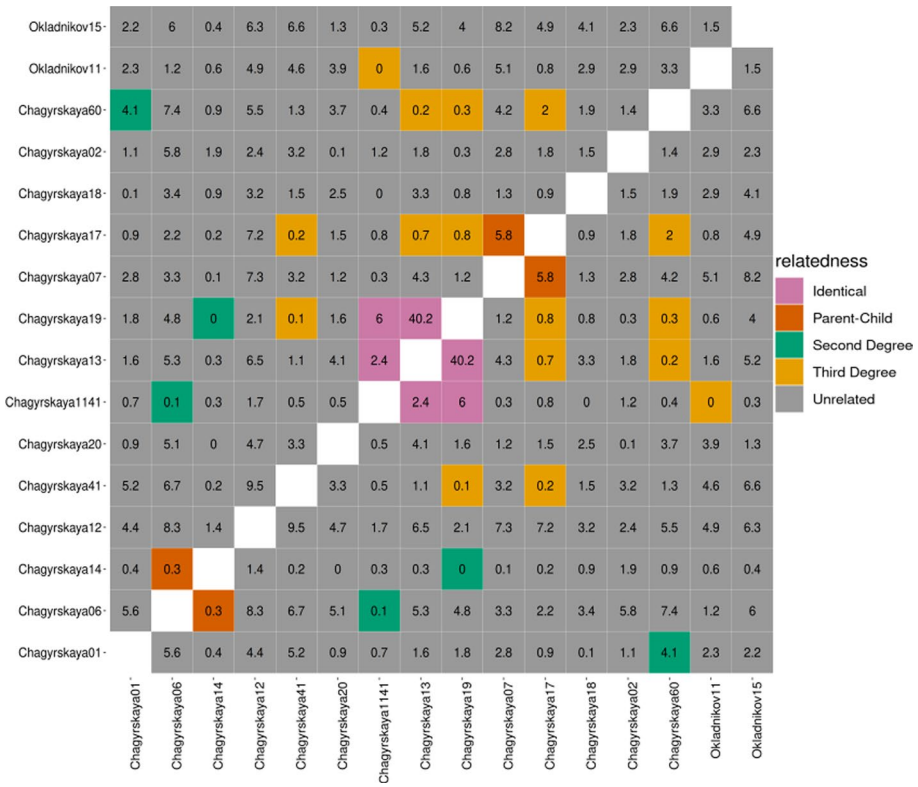

Fig. 6 Application of KIN to Neandertal remains from Chagyrskaya and Okladnikov Caves. The color of a square represents the relatedness, while the number within denotes log-likelihood ratio (ΔLL) between the two maximum likelihood models

Our estimates are consistent with those obtained using READ, except that READ is unable to detect 3rd-degree relationships, and does not distinguish between sibling and parent–child relationships (Additional file 1: Fig. S4). We also find that both KIN and READ classify Chagyrskaya06 and Chagyrskaya14 as parent–child with low confidence but from the morphology they likely stem from the same individual. We believe that the low confidence mis-classification may be due to uncorrected non-human contamination present in these libraries (1% and 2.9% respectively [41]), biasing the estimated differences between the individuals to higher values.

We compared the IBD estimates obtained using KIN to those from lcMLkin for all pairs for whom READ results (with s.d. > 1) and KIN results ($\Delta LL > 1$) match (Additional file 1: Fig. S10). We find that lcMLkin creates four different clusters based on the coefficient of relatedness (r) and the proportion of the genome in the unrelated state (k_0) for different relatedness cases (identical, parent–child, $2nd$ degree and unrelated). However, the location of these clusters strongly deviates from the expected values, likely due to the low coverage and ROH. Remains from the same individual, for example, are expected to be at $k_0 = 0$ and $r = 1$ but are at $k_0 \approx 0.3$, $r \approx 0.6$ for lcMLkin (Additional file 1: Fig. S10).

Ancient modern humans

We further applied KIN to a genome-wide dataset of 118 ancient individuals from the Lech Valley [8]. We compared our relatedness estimates to those obtained by the authors using READ and lcMLkin. We found that KIN was able to confidently classify 85% of the 6903 possible comparisons ($\Delta LL > 1$), READ 94% ($s.d. > 1$) and lcMLkin 53% of pairs ([8]).

Only for 28 pairs, KIN has confident classifications that differ from those obtained using READ (Additional file 2: Table S2, Additional file 1: Fig. S11). Twenty of these comparisons, KIN predicts to be third-degree relationships, and READ concordantly classifies them as unrelated (as it only infers first- and second-degree relationships). lcMLkin predicts $3rd$ - $5th$ degree in 14 of these cases, unrelated in 1 case (see Additional file 1: Fig. S11), second degree in 1 case (see Additional file 1: Fig. S11), and does not have enough data for classification in 4 cases. For 10 additional pairs, lcMLkin predicts $3rd$–$5th$ degree, but KIN infers them to be unrelated. There are a total of 83 pairs where READ obtains a confident call ($s.d. > 1$), but KIN does not ($\Delta LL < 1$). In 80 of these cases, READ classifies them as unrelated, KIN classifies $3rd$ degree, while lcMLkin predicts $3rd$–$5th$ degree. The remaining cases READ classifies unrelated, while KIN and lcMLkin call a $2nd$-degree relationship. Thus, the vast majority of differences can be explained by READ not considering 3rd-degree relationships.

For less than third-degree relations, only eight cases are classified differently between KIN and READ. lcMLkin matches KIN's prediction in two cases, and does not have enough data in 5 cases. In the last case, all three methods differ (Additional file 2: Table S3, Additional file 1: Fig. S11), and the true relatedness is unresolved in this case. Finally, there are three disagreements between KIN and lcMLkin in classification of parent–child versus siblings (READ predicts first degree for all three pairs). We plotted the pairwise differences for these three pairs in Additional file 1: Fig. S12, and found that the proportion of differences along the genome aligns with the prediction of KIN.

Discussion

Here, we present a new method called KIN to estimate genetic kinship and the location of IBD tracts from low-coverage data, in presence of long ROH, contamination, as well as ascertainment bias. Our method utilizes a set of HMMs to estimate IBD tracts and uses them to classify each pair of individuals into a possible relatedness case, along with a measure of classification confidence (Additional file 1: Fig. S1). We evaluated the method performance of KIN, and compared it to that of READ using simulated pedigrees. Finally, we show applications of KIN on two ancient datasets.

For detecting ROH, there is only one method available that works with low coverage ancient DNA [43]. This software can infer ROH at coverages $\geq 0.3x$ but requires a large reference panel. Instead, our method detects long ROH based on just the expected heterozygosity p_0, which can be estimated from a small number of unrelated individuals. On simulations, we find that our method reliably infers ROH regions with lengths on the order of 10 cM from samples sequenced to coverage $\geq 0.1x$. In our simulations, we find that adding long ROH ($\approx$ 17%) to simulations slightly improves the power of both KIN and READ, particularly at lower coverages. This is likely because ROH actually reduces the variance in differences depending on the relatedness and makes it easier to correctly classify relatedness. For READ, the presence of ROH causes a bias towards inferring closer relatedness cases, but the model we use in KIN reduces this bias (see Fig. 5).

We show that KIN reliably detects IBD tracts for $\geq 0.1x$ coverage even in the presence of contamination and ascertainment bias, and the accuracy for IBD detection reduces with lower coverages. Adding ROH to the simulations adversely affects IBD prediction in siblings (Additional file 1: Fig. S9), but this does not affect the power to correctly classify siblings (Fig. 5).

Contamination can affect the accuracy of relatedness inference for two reasons. For one, if a substantial fraction of samples is contaminated, then the estimation of p_0 becomes inaccurate, because the majority of pairwise differences will include at least one contaminated sample. The second issue is that contaminated samples will look less similar to other individuals and thus cause a bias towards inferring them to be less related to other individuals. The amount of contamination does not need to be large for this to be important; in our simulations, we find that even at contamination levels $\leq$ 3%, the performance of READ is substantially reduced. When contamination rates are correctly inferred, the correction we implemented leads to improved performance compared to naive methods such as READ, although they too could be ameliorated in a similar way [41]. However, in many cases, contamination estimates may be uncertain or inaccurate, and we show in Additional file 1: Fig. S13 that KIN's performance is robust to small deviations (small compared to average pairwise heterozygosity) in contamination estimates.

The Lech Valley data has low contamination and no ROH. For pairwise comparisons with large numbers of overlapping sites ($>$ 10000), KIN, READ, and lcMLkin all mostly agree. However, KIN is able to differentiate between parent–child and siblings and identify second-degree relationship from just a few thousand polymorphic sites ($\approx$ 4000) overlapping between samples. KIN can also infer third-degree relation with $\approx$ 30,000 overlapping polymorphic sites. We show that when applied to Neandertal specimens from Chagyrskaya and Okladnikov Caves, KIN identifies a pair of 1*st*-degree relatives as parent–child, which is in agreement with the finding that the mtDNA haplotypes differ

between the samples (one sample is male and the other female) [41]. In addition, KIN identifies a pair of 3*rd*-degree relatives. In this case of a population with large amounts of ROH, we find that the inference by lcMLkin are heavily biased, but KIN's model takes ROH into account and both the coefficient of relatedness and k_0 are very close to what would be expected from the inference by both READ and KIN.

One limitation of our approach is that it assumes a single population. In case of a highly structured population, KIN may show inaccurate inference of p_0 causing inaccurate relatedness inference. Also, our method makes the assumption that the median pairwise genetic difference in the population reflects the population diversity p_0, which fails if almost all individuals in the dataset are related. We may get around this problem by using an estimate of p_0, calculated from a known pair of unrelated individuals from same population, or another population with similar diversity. We provide the user with an option to give an estimate of p_0. The current implementation of KIN is restricted to the six relatedness cases we expect to be the most common, but it might be feasible to extend it to other cases, such as double first cousins, using a corresponding IBD state transition matrix.

While we have focused on the application of KIN on ancient human samples, the model is not tied to this system. Assuming we know the recombination rate and hence can estimate the transition matrix (see the "Materials and methods" section), KIN can be widely applied to any diploid species. In addition, the output of KIN is a table which shows for each pair, the most likely model, and the second best guess, along with a confidence level represented by the log-likelihood ratio. This makes KIN easy to automatize for large datasets. To make application of KIN user-friendly, we provide a python package (KINgaroo) to create input files for KIN from processed bam files, while optionally estimating ROH and correcting for contamination estimates.

Conclusions

KIN is a useful tool to estimate relatedness and the location of IBD tracts from low-coverage ancient DNA samples in presence of long ROH, contamination, and ascertainment bias. This method is applicable to any diploid species and is easy to automatize for large datasets.

Materials and methods

Simulations

We use simulations both for estimating the transition matrices and for testing and validating our algorithm. All simulations are performed in a scenario mimicking the analysis of a Neandertal population contaminated by modern humans [40]. We simulate unrelated individuals using `msprime` [44], followed by an additional step where we simulate related individuals using a predetermined pedigree (Additional file 1: Fig. S3).

Simulating pedigrees

For our simulations of background diversity, we form a population (Pop1) with constant effective size of 10,000 and sample eight diploid individuals (each made up of two haploid individuals) from 2500 generations ago (Additional file 1: Fig. S3). For each individual, we simulate 22 chromosomes with length $L \approx 96$ Mb (same as chromosome 13) and

a recombination rate of $r = 10^{-8}$ per base pair per generation. We introduce mutations using an infinite sites model with rate $\mu = 10^{-8}$ per base pair per generation.

For the pedigree simulations, we first simulate a recombined set of chromosomes for either parent, and combine them to create the progeny. There are two different ways in which we generate recombination points. For the estimation of transition probabilities, we simulate recombination by first drawing the number of breakpoints as a Poisson random variable with parameter rL, and use a uniform distribution on $[1, L]$ to sample the positions of recombination points. For the testing of our method, we use Ped-sim [45] to simulate recombination points. This allows us to take into account sex-specific recombination rates and crossover interference, and thus is expected to give a more realistic recombination landscape. The pedigree simulations result in nine addition individuals, resulting in a total sample of 17 individuals.

Transition matrices

KIN requires a transition matrix for each relatedness case, which we estimate by counting the transition between IBD states for all pairs of individuals with that relatedness in a training set of 1000 simulations from our pedigree (Additional file 1: Fig. S3). For two cases, siblings and grandparent-grandchild, it is possible to write down the theoretical expectation of the transition matrix:

For grandparent-grandchild, rate matrix is

$$Q = \begin{bmatrix} -r & r \\ r & -r \end{bmatrix}$$

Similarly, for a pair of siblings, we calculate

$$Q = \begin{bmatrix} -4r & 4r & 0 \\ 2r & -4r & 2r \\ 0 & 4r & -4r \end{bmatrix}$$

The state space includes all IBD states in this case. For these two cases, we get the transition matrix in each case as e^{Qb}, where b is the window size.

Simulations for method evaluation

Apart from the related and unrelated individuals in Pop1, we simulated more haploid individuals in three other populations to create scenarios with ascertainment bias and contamination (Additional file 1: Fig. S3). We simulated two individuals to form an individual each from two other populations (Pop2, Pop3) with split time of 3500 and 4500 generations with Pop1 and sampling time of 2000 generations and 4000 generations ago respectively. We identified the sites that were polymorphic among these two individuals and used these sites to ascertain the genomes of individuals from Pop1. This scenario roughly models the ascertainment of the Chagyrskaya and Okladnikov Caves data. We tested the performance of our method in presence of long ROH ($\sim$ 17%), by simulating regions of homozygosity in unrelated individuals with a Markov chain using the transition matrix:

$$\mathbf{A} = \begin{bmatrix} 1 - 10/L & 10/L \\ 2/L & 1 - 2/L \end{bmatrix}$$

It is worth noting that we introduce long ROH in unrelated individuals, before artificially mating them to form pedigrees, which means that we do not directly introduce ROH in progeny, but it still affects relatedness inference among progeny as shown in Fig. 1. From the steps described above, we got genotypes of individuals in Pop1 in presence/absence of ROH and ascertainment. We further simulated five diploid individuals from Pop4 with split time of 20,000 generations with Pop1, sampled from the present time, as a source of contamination. For cases including contamination, we contaminated eight individuals with varying amounts between 0.5% to 3% contamination, while the remaining nine individuals did not have any contamination. We generated reads (derived/ancestral) for different genomic coverages ranging from 4x to 0.03x, assuming a Poisson distribution.

For testing, we replicate our simulation 60 times and create data from the same base simulations at varying levels of coverages and different scenarios: we have a control scenario (without ascertainment bias, ROH or contamination) and individual scenarios where we add ROH, SNP ascertainment and contamination. For the evaluation of the ROH-HMM, we also combine ROH with ascertainment or contamination.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02847-7.

Additional file 1: Supplementary figures. Figure S1. Overall schematic of the method showing the entire pipeline from processed bam files to final relatedness and IBD estimates. **Figure S2.** Comparison of fit with Beta-binomial and Binomial distributions. **Figure S3.** Overview of simulated dataset. **Figure S4.** Application of READ on Neandertal specimens from Chagyrskaya and Okladnikov Caves. **Figure S5.** Application of Identical KIN-HMM on a pair of identical individuals with low coverage (0.2x), and ROH tracts. **Figure S6.** Comparison of beta distributions estimated with the Identical KIN-HMM with and without variance constrained optimization of δ parameters. **Figure S7.** Comparison of Beta distributions estimated with ROH-HMM with and without constrained emissions. **Figure S8.** False positive and true positive rates as a function of cutoff on log-likelihood ratio. **Figure S9.** Comparison of IBD estimation in control simulations and in presence of contamination, ascertainment or ROH at different coverages. **Figure S10.** Comparison of IBD states estimated for Chagyrskaya specimens using lcMLkin and KIN. **Figure S11.** Plots showing proportion of differences in windows along the genome for some pairs of relatives for which there is contradiction among KIN, READ and lcMLkin. **Figure S12.** Plots showing proportion of differences in windows along the genome for the first-degree relatives for which KIN differs from lcMLkin. **Figure S13.** False positive and true positive rates as a function of incorrect contamination estimates provided to KIN.

Additional file 2: Supplementary Tables. Table S1. Relatedness and IBD estimates for Chagyrskaya and Okladnikov cave samples. **Table S2.** Relatedness estimates for Bronze age samples for which READ and KIN differ, and KIN has log likelihood ratio > 1. **Table S3.** Relatedness estimates for Bronze age samples for which lcMLkin and KIN differ, and KIN has log likelihood ratio > 1.

Additional file 3. Review history.

Acknowledgements

We thank Svante Pääbo, Janet Kelso, Harald Ringbauer, Johann Visagie, Zbigniew Jędrzejewski-Szmek, Laurits Skov, Leonardo N. M. Iasi, Alba Bossoms Mesa, Arev P. Sümer, Zuzana Hofmanová, Guido Alberto Gnecchi Ruscone, Ke Wang, and Luca Traverso for helpful comments and discussions. This work was supported by the Max Planck Society and the European Research Council (Grant No.: 694707) to Svante Pääbo.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 3.

Authors' contributions
Conceptualization (design of study): B.M.P.; software: D.P.; methodology—lead: D.P.; methodology—support: B.M.P., S.P.; formal analysis: D.P.; visualization—lead: D.P.; visualization—support: S.P.; data curation: D.P.; writing—lead: D.P.; writing—support: B.M.P., S.P.; supervision: B.M.P. The authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
An open-source implementation of KIN and KINgaroo in python along with a toy example dataset, and the scripts to generate our simulations are available on GitHub [46]. We have deposited the version of the software used in the manuscript, along with the above mentioned files on Zenodo [47]. The analyzed datasets were generated in previous studies, and are available in European Nucleotide Archive [48, 49].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 May 2022 Accepted: 4 January 2023
Published online: 17 January 2023

References

1. Murphy E. Law and policy oversight of familial searches in recreational genealogy databases. Forensic Sci Int. 2018;292:e5–9. https://doi.org/10.1016/j.forsciint.2018.08.027.
2. Ram N, Guerrini CJ, McGuire AL. Genealogy databases and the future of criminal investigation. Science. 2018;360(6393):1078–9. https://doi.org/10.1126/science.aau1083.
3. Egeland T, Mostad PF, Mevåg B, Stenersen M. Beyond traditional paternity and identification cases: selecting the most probable pedigree. Forensic Sci Int. 2000;110(1):47–59. https://doi.org/10.1016/S0379-0738(00)00147-X.
4. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389–97. https://doi.org/10.1534/genetics.107.081190.
5. Kardos M, Luikart G, Allendorf FW. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity. 2015;115(1):63–72. https://doi.org/10.1038/hdy.2015.17.
6. Oliehoek PA, Windig JJ, van Arendonk JAM, Bijma P. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics. 2006;173(1):483–96. https://doi.org/10.1534/genetics.105.049940.
7. Baca M, Doan K, Sobczyk M, Stankovic A, Węgleński P. Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. BMC Genet. 2012;13(1):30. https://doi.org/10.1186/1471-2156-13-30.
8. Mittnik A, Massy K, Knipper C, Wittenborn F, Friedrich R, Pfrengle S, et al. Kinship-based social inequality in Bronze Age Europe. Science. 2019;366(6466):731–4. https://doi.org/10.1126/science.aax6219.
9. Sikora M, Seguin-Orlando A, Sousa VC, Albrechtsen A, Korneliussen T, Ko A, et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. Science. 2017;358(6363):659–62. https://doi.org/10.1126/science.aao1807.
10. Boehnke M, Cox NJ. Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet. 1997;61(2):423–9. https://doi.org/10.1086/514862.
11. Browning BL, Browning SR. A Fast, Powerful method for detecting identity by descent. Am J Hum Genet. 2011;88(2):173–82. https://doi.org/10.1016/j.ajhg.2011.01.010.
12. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, et al. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 2009;19(2):318–26. https://doi.org/10.1101/gr.081398.108.
13. Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 2011;21(5):768–74. https://doi.org/10.1101/gr.115972.110.
14. Li H, Glusman G, Huff C, Caballero J, Roach JC. Accurate and robust prediction of genetic relationship from whole-genome sequences. PLoS ONE. 2014;9(2):e85437. https://doi.org/10.1371/journal.pone.0085437.
15. Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, et al. Relationship estimation from whole-genome sequence data. PLoS Genet. 2014;10(1):e1004144. https://doi.org/10.1371/journal.pgen.1004144.
16. Lynch M, Ritland K. Estimation of pairwise relatedness with molecular markers. Genetics. 1999;152(4):1753–66.
17. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73. https://doi.org/10.1093/bioinformatics/btq559.
18. Nyerki E, Kalmár T, Schütz O, Lima RM, Neparáczki E, Török T, et al. An optimized method to infer relatedness up to the 5th degree from low coverage ancient human genomes. Genetics. 2022. https://doi.org/10.1101/2022.02.11.480116.
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

20. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91(1):122–38. https://doi.org/10.1016/j.ajhg.2012.05.024.
21. Korneliussen TS, Moltke I. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics (Oxford, England). 2015;31(24):4009–11. https://doi.org/10.1093/bioinformatics/btv509.
22. Lipatov M, Sanjeev K, Patro R, Veeramah KR. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. bioRxiv. 2015;023374. https://doi.org/10.1101/023374.
23. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15(1):356. https://doi.org/10.1186/s12859-014-0356-4.
24. Peyrégne S, Prüfer K. Present-day DNA contamination in ancient DNA datasets. BioEssays. 2020;42(9):2000081. https://doi.org/10.1002/bies.202000081.
25. Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. Computational challenges in the analysis of ancient DNA. Genome Biol. 2010;11(5):R47. https://doi.org/10.1186/gb-2010-11-5-r47.
26. Vai S, Amorim CEG, Lari M, Caramelli D. Kinship determination in archeological contexts through DNA analysis. Front Ecol Evol. 2020;8:83. https://doi.org/10.3389/fevo.2020.00083.
27. Waples RK, Albrechtsen A, Moltke I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. Mol Ecol. 2019;28(1):35–48. https://doi.org/10.1111/mec.14954.
28. Theunert C, Racimo F, Slatkin M. Joint estimation of relatedness coefficients and allele frequencies from ancient samples. Genetics. 2017;206(2):1025–35. https://doi.org/10.1534/genetics.117.200600.
29. Kuhn JMM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. PLoS ONE. 2018;13(4):e0195491. https://doi.org/10.1371/journal.pone.0195491.
30. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft sequence of the Neandertal genome. Science (New York, NY). 2010;328(5979):710–22. https://doi.org/10.1126/science.1188021.
31. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207–11. https://doi.org/10.1038/nature14317.
32. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43–9. https://doi.org/10.1038/nature12886.
33. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;7(10):771–80. https://doi.org/10.1038/nrg1960.
34. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995;96(1):3–12. https://doi.org/10.1007/BF01441146.
35. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol). 1977;39(1):1–38.
36. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
37. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat. 1970;41(1):164–71. https://doi.org/10.1214/aoms/1177697196.
38. Peyrégne S, Prüfer K. Present-day DNA contamination in ancient DNA datasets. BioEssays. 2020;42(9):2000081. https://doi.org/10.1002/bies.202000081.
39. Kolobova KA, Roberts RG, Chabai VP, Jacobs Z, Krajcarz MT, Shalagina AV, et al. Archaeological evidence for two separate dispersals of Neanderthals into southern Siberia. Proc Natl Acad Sci. 2020;117(6):2879–85. https://doi.org/10.1073/pnas.1918047117.
40. Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, et al. A high-coverage Neandertal genome from Chagyrskaya Cave. Proc Natl Acad Sci. 2020;117(26):15132–6. https://doi.org/10.1073/pnas.2004944117.
41. Skov L, Peyrégne S, Popli D, Iasi LNM, Devièse T, Slon V, et al. Genetic insights into the social organization of Neanderthals. Nature. 2022;610(7932):519–25. https://doi.org/10.1038/s41586-022-05283-y.
42. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science (New York, NY). 2017;358(6363):655–8. https://doi.org/10.1126/science.aao1887.
43. Ringbauer H, Novembre J, Steinrücken M. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. Nat Commun. 2021;12(1):5425. https://doi.org/10.1038/s41467-021-25289-w.
44. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 2016;12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842.
45. Caballero M, Seidman DN, Qiao Y, Sannerud J, Dyer TD, Lehman DM, et al. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. PLoS Genet. 2019;15(12):e1007979. https://doi.org/10.1371/journal.pgen.1007979.
46. Popli D. 2022. https://github.com/DivyaratanPopli/Kinship_Inference/releases/tag/v3.1.2. Accessed 10 Sept 2022.
47. Popli D. https://doi.org/10.5281/zenodo.7067142.
48. Mittnik A, Massy K, Knipper C, Wittenborn F, Friedrich R, Pfrengle S, et al. Kinship-based social inequality in Bronze Age Europe. Datasets. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB34400. Released 10 Oct 2019.
49. Skov L, Peyrégne S, Popli D, Iasi LNM, Devièse T, Slon V, et al. Genetic insights into the social organization of Neanderthals. Datasets. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB55327. Released 19 Oct 2022.

Publisher's Note