Received: 25 November 2023 | Revised: 19 March 2024 | Accepted: 28 March 2024

DOI: 10.1111/1755-0998.13960

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

# Benchmarking kinship estimation tools for ancient genomes using pedigree simulations

Şevval Aktürk[1] | Igor Mapelli[1] | Merve N. Güler[1] | Kanat Gürün[1] | Büşra Katırcıoğlu[1] | Kıvılcım Başak Vural[1] | Ekin Sağlıcan[2] | Mehmet Çetin[1] | Reyhan Yaka[1,3,4] | Elif Sürer[5] | Gözde Atağ[1] | Sevim Seda Çokoğlu[1] | Arda Sevkar[6] | N. Ezgi Altınışık[6] | Dilek Koptekin[2] | Mehmet Somel[1]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey

[2]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[3]Centre for Palaeogenetics, Stockholm, Sweden

[4]Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden

[5]Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[6]Department of Anthropology, Hacettepe University, Ankara, Turkey

**Correspondence**
Şevval Aktürk, Merve N. Güler, and Mehmet Somel, Department of Biological Sciences, Middle East Technical University, Ankara 06800, Turkey.
Email: sevvalakturk96@gmail.com; merveglr2626@gmail.com and somel.mehmet@googlemail.com

**Handling Editor:** Michael M. Hansen

## Abstract

There is growing interest in uncovering genetic kinship patterns in past societies using low-coverage palaeogenomes. Here, we benchmark four tools for kinship estimation with such data: lcMLkin, NgsRelate, KIN, and READ, which differ in their input, IBD estimation methods, and statistical approaches. We used pedigree and ancient genome sequence simulations to evaluate these tools when only a limited number (1 to 50 K, with minor allele frequency ≥0.01) of shared SNPs are available. The performance of all four tools was comparable using ≥20 K SNPs. We found that first-degree related pairs can be accurately classified even with 1 K SNPs, with 85% $F_1$ scores using READ and 96% using NgsRelate or lcMLkin. Distinguishing third-degree relatives from unrelated pairs or second-degree relatives was also possible with high accuracy ($F_1 > 90\%$) with 5 K SNPs using NgsRelate and lcMLkin, while READ and KIN showed lower success (69 and 79% respectively). Meanwhile, noise in population allele frequencies and inbreeding (first-cousin mating) led to deviations in kinship coefficients, with different sensitivities across tools. We conclude that using multiple tools in parallel might be an effective approach to achieve robust estimates on ultra-low-coverage genomes.

**KEYWORDS**
ancient DNA, inbreeding, kinship coefficient estimation, low coverage, pedigree simulation

## 1 | INTRODUCTION

The use of palaeogenomes for inferring genetic kin relations in ancient human populations is growing at an accelerating pace. These studies have unraveled diverse types of social relations of past human societies, from the composition of households (Ning et al., 2021; Yaka et al., 2021) or burial treatment of mass murder victims (Schroeder et al., 2019) to matrilineal (Kennett et al., 2017) or patrilineal traditions studied in graves (Fowler et al., 2022; Mittnik et al., 2019; Rivollat et al., 2023; Sánchez-Quinto et al., 2019). However, determining the degree of kinship using single nucleotide polymorphism (SNP) data from low-coverage genomes is fraught with difficulties, mainly arising from data scarcity. Most published palaeogenomes are below 1× coverage and

Şevval Aktürk, Igor Mapelli and Merve N. Güler: Equal contribution.

thus do not allow reliable diploid genotyping, required by popular kinship estimation tools such as KING (Manichaikul et al., 2010). Although imputation has recently been shown to produce reliable diploid genotypes using shotgun genomes >0.5× (Martiniano et al., 2017; Sousa da Mota et al., 2023), a substantial fraction of palaeogenomes still do not reach this threshold; e.g. in the AADR repository (v54.1.p1) (Mallick et al., 2024), out of 2041 published shotgun genomes with reported coverage from their original source, 916 (45%) have coverage <0.5×.

A number of solutions fine-tuned for performance on low-coverage ancient DNA (aDNA) data have been published over the last few years. These algorithms use pseudo-haploid genotypes (e.g. Kuhn et al., 2018), genotype likelihoods (e.g. Hanghøj et al., 2019; Lipatov et al., 2015; Žegarac et al., 2021) or read information (e.g. Popli et al., 2023) instead of diploid calls. These methods also differ in (a) how they normalize the pairwise mismatch values between two genomes to infer the kinship degree and (b) whether they use method-of-moment estimators or probabilistic approaches. The most widely cited tool, READ (Kuhn et al., 2018), compares the rate of average mismatch (P0) between a genome pair with the median (or maximum) P0 of a large enough sample from the same population, assuming this median estimate represents the expected P0 of an unrelated pair. This is similar to the pairwise mismatch rate (PMR) calculation by Kennett and colleagues (Kennett et al., 2017). Two other commonly used tools, lcMLkin (v2) (Lipatov et al., 2015; Žegarac et al., 2021) and NgsRelate (v2) (Hanghøj et al., 2019), use genotype likelihoods and population allele frequency estimates to infer the kinship degree between pairs within a likelihood framework. The TKGWV2 (Fernandes et al., 2021) algorithm also uses population allele frequencies within a method-of-moments framework. Finally, the recently published method, KIN (Popli et al., 2023), uses a likelihood-based framework as well as a Hidden Markov Model (HMM) to infer segments of identity-by-descent (IBD) between pairs of individuals. KIN also uses the average mismatch in a sample for normalizing P0 rates for inferring identity-by-descent (IBD), akin to READ.

Although each of these methods is being widely used by the palaeogenomics community, their relative accuracy and performances have not been systematically investigated. One recent exception is a study by Marsh et al. (2023), who compared these methods using real ancient and modern-day genomic datasets. The authors lacked knowledge of real relationships but studied how consistency among estimates was affected by downsampling high-coverage genomes, reporting that READ, PMR, and TKGWV2 were less affected by low coverage than lcMLkin and NgsRelate. However, this study was limited by the lack of a ground truth set of relationships.

Beyond palaeogenomes, inferring genetic relatedness from low-coverage genomes can also be a challenge faced by conservation programmes, as well as evolutionary and ecological studies of wild populations. Such wildlife studies frequently rely on accurate pedigree information (Galla et al., 2022; Oliehoek et al., 2006; Pemberton, 2008), with motivations ranging from minimizing inbreeding to investigating heritability patterns. Identifying close genetic kin is also relevant for population genomics studies, where relatives are removed to ensure independence among samples.

Similar to human studies, relatedness estimation in wildlife samples has been traditionally performed using microsatellites (STRs), which can be powerful and cost-effective (e.g. Godoy et al., 2022; Koch et al., 2008; Moran et al., 2021; O'Reilly & Kozfkay, 2014). However, genome-wide SNP data can also be an effective, if not more precise, alternative to using STRs in wildlife studies (Galla et al., 2020). Importantly, many wildlife genetics studies rely on fecal DNA samples, where DNA is usually available in degraded form, broken into short fragments and/or in too low amounts to allow reliable diploid genotyping for a large number of samples (e.g. Pinho et al., 2014). Genome-wide SNP data obtained through either shotgun sequencing or capture sequencing may be a useful alternative in such cases (Bérénos et al., 2014; de Flamingh et al., 2023; Galla et al., 2020). Moreover, many such genomic datasets will be of low and heterogeneous coverage, with limited numbers of SNPs per individual. The relatedness estimation methods discussed in this work are therefore directly applicable to such data.

Here, we compare the performances of four algorithms designed for kinship estimation with sparse SNP data, lcMLkin, NgsRelate, READ, and KIN, using ancient-like genomic data from pedigree simulations to distinguish close kin (first- to third-degree relatives) and non-kin. We test the effects of ultra-low coverages (using down to 1000 SNPs per pair), inbreeding, and noise in allele frequency estimates (i.e. random fluctuations around the true allele frequency values that mimic biological or random technical variation in the real data). We chose READ, lcMLkin, and NgsRelate as these are among the most widely used algorithms on low-coverage genomes (Table 1). Meanwhile, we chose KIN along with NgsRelate as these algorithms are designed to separate genetic correlations due to direct kinship or inbreeding. Importantly, READ and KIN use sample-based normalization, while lcMLkin and NgsRelate use population allele frequencies to infer IBD.

## 2 | MATERIALS AND METHODS

### 2.1 | Pedigree simulations

We simulated ancient genome data representing pairs of individuals with known relationships. We first created 600 founder genotype data from scratch using 8,677,101 SNPs with minor allele frequency (MAF >0.01) from $n=112$ Tuscany (TSI) genomes from the 1000 Genomes Project v3 (Auton et al., 2015) and randomly creating diploid genotypes (Appendix S1). Note that our approach eliminates any background relatedness among founders as well as any homozygosity tracts within founder genomes; this is not realistic but simplifies the interpretation of the kinship estimation results. We repeated the creation of founder data 12 times (runs), each producing different sets of founders.

We then employed Ped-sim (v1.3) (Caballero et al., 2019) to simulate pedigrees using this founder pool, producing genotypes from pedigrees of various relationship degrees and types separately,

TABLE 1 Different methods and the number of publications using them for kinship estimation.

| Software | Study | Number of publications using the software | Input data type | Expected dissimilarity calculation approach | Relatedness classification |
|---|---|---|---|---|---|
| NgsRelate | Korneliussen and Moltke (2015) | 47 | Genotype likelihoods | Population allele frequencies | Up to third degree |
| NgsRelate v2 | Hanghøj et al. (2019) | 57 | Genotype likelihoods | Population allele frequencies | Up to third degree |
| lcMLkin | Lipatov et al. (2015) | 49 | Genotype likelihoods | Population allele frequencies | Up to third degree |
| lcMLkin v2 | Žegarac et al. (2021) | 1 | Genotype likelihoods | Population allele frequencies | Up to third degree |
| READ | Kuhn et al. (2018) | 128 | Pseudo-haploid genotypes | Average dissimilarity within the sample | Up to second degree |
| TKGWV2 | Fernandes et al. (2021) | 6 | Pseudo-haploid genotypes | Population allele frequencies | Up to second degree |
| KIN | Popli et al. (2023) | 3 | Aligned reads (BAM files) | Average dissimilarity within the sample | Up to third degree |

*Note*: The data were collected by revising literature citing the named articles in Google Scholar (retrieved 4 November 2023) and filtering for publications (including journal publications and pre-prints, but excluding academic theses) that directly used the software (Table S7). Expected dissimilarity calculation approach: all methods use an expected dissimilarity estimate representing dissimilarity between an unrelated pair to normalize the observed pairwise dissimilarity values and estimate the kinship coefficient.

including first-, second and third-degree relatedness without inbreeding, as well as first-degree and second-degree relatedness with first-cousin mating (Figure 1). Within Ped-sim, we used a linearly interpolated sex-specific recombination map (Bhérer et al., 2017) with the '-m' option and crossover interference model (Housworth & Stahl, 2003) using the '--intf' option; we also kept track of founder sexes (Appendix S1). We thus simulated $n = 72$ pedigrees composed of first-degree, $n = 96$ second-degree, and $n = 96$ third-degree related pairs. The founders of each pedigree and simulated individuals from distinct pedigrees were treated as 'unrelated'. From these simulated pedigrees, we randomly chose $n = 48$ pairs for each relationship type (Table 2) (Appendix S1).

For the pedigree simulations with inbreeding, first-degree and second-degree pedigrees (parent–offspring and grandparent–grandchild relationships) were simulated in the presence of first-cousin mating (i.e. the parents of an offspring or a grandchild are first cousins respectively). We also used $n = 48$ pairs for each relationship type (Table 2).

## 2.2 | Ancient DNA sequence data simulation and pre-processing

To create realistic ancient-like genotypes, we first simulated Illumina short read data based on each simulated individual's genotype, with aDNA-like postmortem damage and sequencing error introduced using the gargammel software (Renaud et al., 2017) (Appendix S1). We limited the generated data to randomly chosen 200,000 autosomal SNPs. We generated ancient read data with 5× depth of coverage per individual, without any present-day human or microbial contamination.

We then processed the gargammel-simulated read data following the same procedure as applied to ancient genome sequencing libraries in our group and other research teams (e.g. Altınışık et al., 2022; Koptekin et al., 2023; Yaka et al., 2021). First, we removed the adapters from the simulated ancient reads and then merged the paired end reads (Schubert et al., 2016). The reads were then mapped to the human reference genome (hs37d5) using the bwa software 'samse' function (v0.7.15) (Li & Durbin, 2009) with the '-aln' option, and parameters are set to '-l 16,500', '-n 0.01' and '-o 2'. Third, we removed the duplicate reads with identical starting and end positions using FilterUniqueSAMCons.py script (Kircher, 2012). We also eliminated the reads with a minimum of 10% mismatches to the human reference genome. Finally, the remaining reads were trimmed 10 bps from both ends to remove the PMD-related C-to-T and G-to-A substitutions using the bamUtil software with the 'trimBAM' option (Jun et al., 2015).

## 2.3 | Genotyping and downsampling

We randomly downsampled the BAM files of all simulated individuals from 5× to 1× coverage using Picard Tools DownsampleSam (2.25.4) (Broad Institute, 2019). Because we aim to study the performance

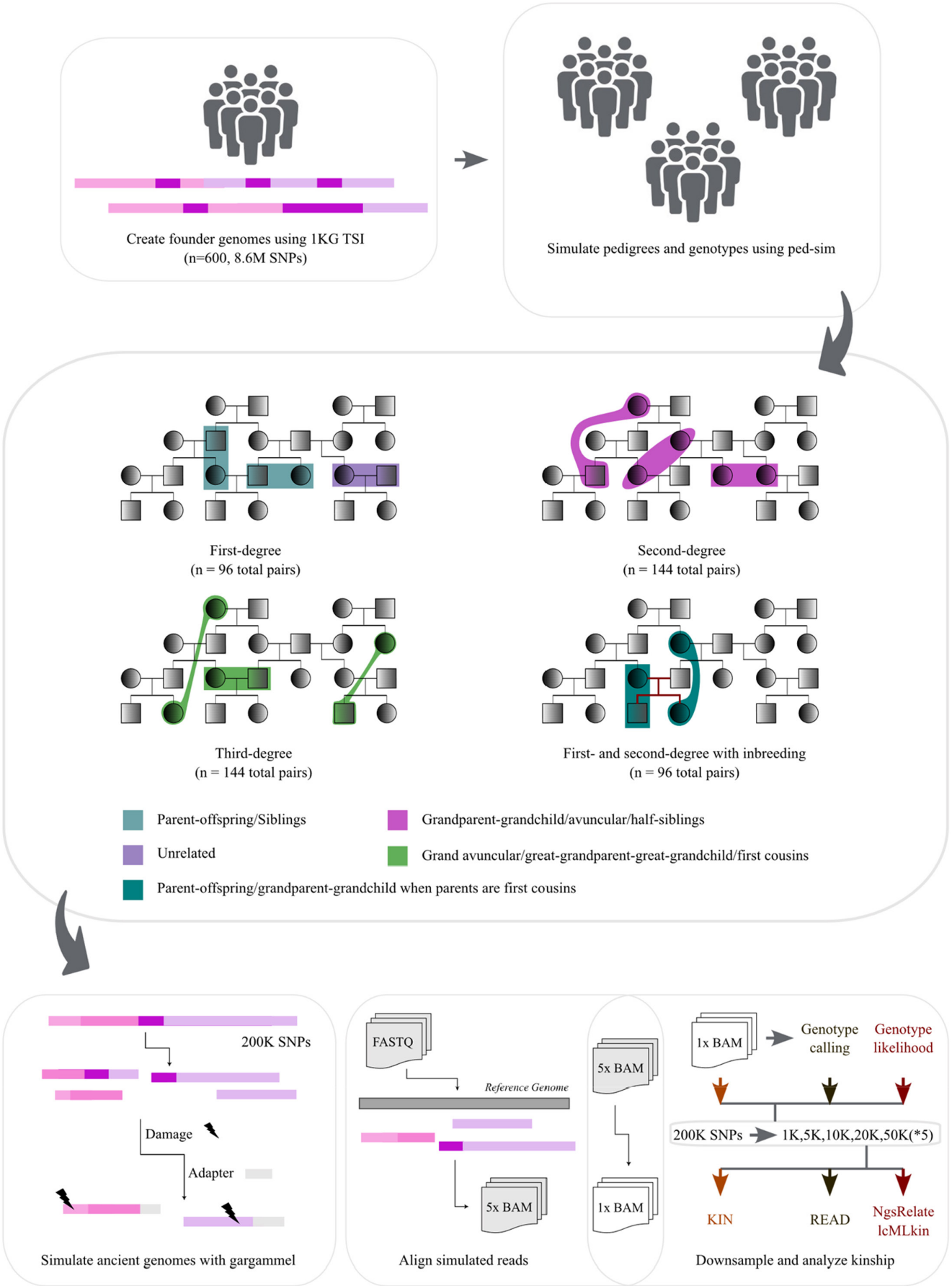

**FIGURE 1** Primary simulations and analysis workflow. We created 600 synthetic founder genomes using 1000 Genomes Project v3 Tuscany (TSI) samples. We used these founder genomes to create pedigrees with Ped-sim and human genetic maps, from which we chose sets of related pairs of different types, with $n = 48$ pairs created for each relationship type (two types for first degree and three types each for second and third degree) (Table 2). We also created parent–offspring and grandparent–grandchild pairs, where the offspring was the child of first cousins. We sub-sampled these genotypes to 200 K SNPs and created aDNA-like sequencing read data using the gargammel tool. The reads were then aligned to the reference genome to produce 5× BAM files, which were further downsampled to 1× coverage. We called pseudo-haploid genotypes or calculated genotype likelihoods (GL) for the same 200 K SNPs and downsampled these to 1–50 K subsets, each SNP counts downsampled randomly five times. The genotypes, GL, or BAM files were input into the four kinship estimation tools.

**TABLE 2** The relationships used for palaeogenomic data simulation.

| Relationship | Degree | Number of sex combinations | Number of individuals | Number of pairs |
|---|---|---|---|---|
| Parent–offspring | First | 4 | 72 | 48 |
| Siblings | First | 3 | 96 | 48 |
| Half-siblings | Second | 6 | 96 | 48 |
| Grandparent–grandchild | Second | 4 | 72 | 48 |
| Avuncular | Second | 8 | 96 | 48 |
| First cousins | Third | 10 | 96 | 48 |
| Great-grandparent–great-grandchild | Third | 8 | 72 | 48 |
| Grand avuncular | Third | 16 | 96 | 48 |
| Parent–offspring (inb) | First | 8 | 72 | 48 |
| Grandparent–grandchild (inb) | Second | 4 | 72 | 48 |

*Note*: Number of sex combinations: the count of distinct configurations of individuals' sexes within the same pedigree for each simulation run. Number of pairs: the number of independently simulated pairs for each type of relationship. 'inb': pairs where inbreeding was simulated with the child or grandchild being the offspring of a first-cousin mating (Figure 1).

of the kinship coefficient ($\theta$) estimation on low-depth ancient data, most of our analyses involve sub-samples of the 1× data (only one read per SNP). We used the 5× data only to test noise in population allele frequencies.

We next performed pseudo-haploid genotyping (Skoglund et al., 2012) from simulated 1× ancient genomes using the SAMtools (v.1.9) 'mpileup' function (Danecek et al., 2021), followed by running pileupCaller (v1.4.0.5) with the '--randomHaploid' parameter (Schiffels, n.d.). We used the 200 K autosomal SNPs we selected earlier to generate text pileup files for all BAM files. Mapping quality and base quality filters were set to >30 in SAMtools (v.1.9) mpileup. The output pileup files were given as input to pileupCaller software to produce pseudo-haploid genotype data by randomly sampling one read at each SNP. The output files were then converted to binary PLINK files using ADMIXTOOLS *convertf* package (Patterson et al., 2012) with parameter '-p' and then to transposed ped/fam format using PLINK (v1.9) (Purcell et al., 2007). Last, we retained only non-missing genotype calls for each pair of individuals using PLINK (v1.9) with the option '--geno 0' (note that missing SNPs are removed only for the analysed pair). This reduced the number of SNPs from 200 K to an average of 77 K for 1× depth of coverage.

We randomly chose subsets of 1, 5, 10, 20, and 50 K SNPs shared between each simulated pair five times to explore the lower limits of using ancient genomes for genetic relatedness estimation. This allowed us to study how much kinship coefficient estimates vary depending on the set of variants used for the analysis. We note that the term replicate, used for the downstream analysis, refers to this repeated downsampling ($n = 5$).

## 2.4 | Simulations with background relatedness

In addition to the primary dataset we generated above using synthetic founders based on 1000 Genomes TSI, we created another founder dataset comprising 250 individuals with background relatedness (i.e. due to drift). For this, we employed the msprime engine (Baumdicker et al., 2022; Kelleher et al., 2016) in the mode of 'HomSap' from the stdpopsim library (Adrion et al., 2020; Lauterbur et al., 2023) to simulate the genetic data of these founder individuals. We utilized the 'HapMapII-GRCh37' (Frazer et al., 2007) with the '-g' option as the recombination map. We simulated the 500 haploid genomes descended from the Linearbandkeramik (LBK) population, which can be described as early European Neolithic populations of Anatolian descent (Kılınç et al., 2016), of the multi-population model of ancient Eurasia model (Kamm et al., 2020), with the '-d AncientEurasia-9K19 0 500' option. Subsequently, we transformed the succinct tree sequence output generated by the stdpopsim software into VCF using the tskit library (Kelleher, 2018) 'vcf' command with the '--ploidy 2' option. We then narrowed our analysis to 200 K randomly selected SNP positions through a customized bash script. These selected positions were further used to extract reference bases from the

human reference genome (hs37d5) using the 'getfasta' command of BEDtools (v2.27.1) (Quinlan & Hall, 2010). We estimated the transition:transversion rate from the 1000 Genomes Dataset v3 TSI population to assign alternative alleles to the retrieved reference positions. With this information, we stochastically generated alternative alleles for each position in our dataset, employing a customized R script. This approach was instrumental in replicating genetic variation according to the observed rates within the TSI population, offering a realistic distribution of allele frequencies within our simulated dataset. The rest of the pipeline, comprising pedigree simulation, ancient sequence simulation, pre-processing, genotyping, and downsampling, was identical to that used to create our primary dataset.

## 2.5 | Genetic relatedness estimation using READ, NgsRelate, lcMLkin and KIN

### 2.5.1 | READ

This non-parametric genetic relatedness estimation tool relies on the proportion of mismatching sites between pseudo-haploid genomes, i.e. the pairwise mismatch rate (P0) (Appendix S1). We ran READ with pseudo-haploid genotype data of the simulated individual pairs using default parameters. We combined all READ results for all $n=48$ pairs of each eight relationship types into eight sets, each combined with unrelated pairs (~2000–4000) from different pedigrees of this type (Appendix S1). As these sets are mainly composed of unrelated individuals, we used their median P0 value for normalization (~0.24). The kinship coefficient ($\theta$) estimate for each related and unrelated pair was calculated using the formula:

$$\theta = 1 - \left(P0_{\text{pair}} / P0_{\text{median}}\right)$$

These $\theta$ estimates can be negative when a pair shares fewer alleles IBS than the ones of the average unrelated pair (Konovalov & Heg, 2008), suggesting a non-kin relationship. Thus, we set the negative $\theta$ estimates to 0.

### 2.5.2 | NgsRelate

NgsRelate v2 (Hanghøj et al., 2019) (hereon NgsRelate) is a maximum likelihood-based method estimating Jacquard coefficients ($J_1$, $J_2$, ..., $J_9$) given genotype likelihoods (GL) and population allele frequencies. To calculate the GLs for each individual separately from the gargammel-produced BAM files, we used the ANGSD program (Korneliussen et al., 2014) with the '--gl 2' option. We limited GL calculation to the chosen 200 K autosomal SNPs (MAF >0.01) using the '-sites' parameter. The beagle text output file of ANGSD (--doGlf 2) was manipulated to generate a GL file containing only two individuals with their shared SNPs. We eliminated pairwise missing SNPs by keeping only sites with GL values not equal to 0.33 for three genotype states (major/major, major/minor, and minor/minor) for both individuals with a custom R script. Next, we randomly downsampled the shared SNPs between every pair of individuals to 1–50 K, five times each, using an in-house bash script. Then, every pair's GL files with five different SNP subsets were converted to binary GL file format. The background allele frequency files for corresponding SNPs were prepared using their MAF of the 1000 Genomes TSI sample with $n=112$ individuals. The MAF threshold of NgsRelate was set to 0 with the option '-l'. The output file produced by NgsRelate for each pair includes a $\theta$ value corresponding to the kinship coefficient estimate.

### 2.5.3 | NgsRelate with alternative background allele frequencies

With NgsRelate, we also conducted trials with alternative background MAF. This analysis was restricted to the two first-degree relatedness categories, parent–offspring ($n=48$) and siblings ($n=48$); we reasoned these effects would be consistent across different relatedness types. We ran the ANGSD program with the above-mentioned parameters on the BAM files using chosen 200 K autosomal SNPs on the BAM files and we processed the resulting GL file to obtain pairwise GL files without missing SNPs. We then used three alternative background MAF calculations: (1a) MAFs from the 1000 Genomes TSI population ($n=112$) as in the original analyses. (1b) MAFs calculated from gargammel-produced 5× coverage BAM files of the same individuals used in this analysis: 72 individuals comprising the 48 parent–offspring and 96 individuals comprising the 48 sibling pairs. For this, we ran the ANGSD program with the same parameters on the 5× coverage BAMs and obtained MAFs for both relatedness categories separately. (1c) MAFs estimated from gargammel-produced 1× coverage BAM files of the same individuals.

We also used modified MAFs in three ways: (2a) No noise. (2b) Adding a low level of random noise (i.e. random variation). Here, we introduced random noise to the original MAFs from the TSI, as well as MAFs called from 1× and 5× genomes (as described earlier), while ensuring the resulting values remained within the valid range of 0 to 0.5. For this, we first transformed the MAF values with the logit function $\text{logit}(p) = \log(p/(1-p))$. This transformation aims to stretch the original allele frequencies to the entire real number space, making them amenable to adding random noise. Then, we generated the noise-added allele frequency values following a Gaussian distribution with a mean based on the logit-transformed MAF values and a standard deviation of 0.5. Then, we applied the expit function, $\text{expit}(p) = 1/(1+\exp(-p))$, to the random values to transform them back to the 0 to 1 interval. Finally, we adjusted the MAF values to ensure they fell within the valid range of 0 to 0.5. This adjustment involved subtracting any values that exceeded 0.5 from 1. (2c) Adding a high level of random noise. Here, we repeated the same steps as in (2b) but added Gaussian noise with a standard deviation of 1.

We manipulated the resulting GL and MAF files for each pair to have five replicates of 1, 5, 10, 20, and 50 K shared autosomal SNPs between pairs of samples. We then ran NgsRelate with the parameters described earlier for each pair of parent–offspring and sibling categories with these nine different background MAF values (Table 3).

### 2.5.4 | lcMLkin

lcMLkin (Lipatov et al., 2015) is another maximum-likelihood-based software detecting genetic relatedness. Unlike NgsRelate, it assumes a non-inbred population and estimates Cotterman coefficients using GL and population allele frequencies. We prepared input VCF files for each pair to run lcMLkin (v2.1) (Altınışık, 2023) implemented for Python3. We used BCFtools mpileup and call commands (Li, 2011) to estimate the genotype likelihoods of each individual using BAM files for the 200 K SNP set with the mapping and base quality filter parameters '-q10' and '-Q13', respectively. These thresholds were selected based on the default filters of ANGSD to estimate GLs for NgsRelate analysis. lcMLkin requires the genotype data of the selected background population for allele frequency estimation. These genotype data were provided in *PLINK* format (bed/bim/fam) with the argument '-p'. We prepared these genotype data using the 200 K selected autosomal SNPs (MAF >0.01). We changed the default allele frequency thresholds integrated into the lcMLkin python script from minimum 0.05 and maximum 0.95 to minimum 0.01 and maximum 0.99. We filtered out missing (non-shared) SNPs from VCF files using an in-house bash script to collect only overlapping SNPs between each simulated pair. After that, we randomly selected 1, 5, 10, 20, and 50 K shared SNPs between pairs of samples, independently five times each, and generated downsampled VCF files using BCFtools view (Li, 2011) with the '-R' parameter. As the linkage disequilibrium (LD) pruning application of lcMLkin removes closely linked SNPs from the relatedness analysis, we modified the program script such that downsampled SNPs are not pruned by LD. This was done for simplicity to ensure we use the same number of SNPs in each trial and across different software. Also, with ≤50 K SNPs across the genome, the linkage between neighboring SNPs will be minimal. The relatedness coefficient ($r$) is represented with the 'PI_HAT' estimate in the output files of lcMLkin. We calculated the kinship coefficient value as $\theta = r/2$.

### 2.5.5 | KIN

KIN (Popli et al., 2023) has been recently developed to estimate relatedness up to the third degree and differentiate between parent–offspring and sibling pairs using Hidden Markov Models (HMM). The algorithm uses P0 estimates (like READ) in genomic windows calculated directly from BAM files, and further estimates possible ROH and IBD tracks using HMM (Appendix S1). As KIN does not run with only two individuals and because we wanted to test one pair at a time to control for the shared SNP counts between individuals, we first grouped our BAM files into triplets for each relationship type, including one pair of BAM files to be analysed and one BAM file of a randomly chosen simulated individual. We determined the read depth of each site at the predefined 200 K SNPs for each triplet using SAMtools (v1.9) (Danecek et al., 2021) 'depth' with the '-q 30' and '-Q 30' options. Then, we removed sites that do not contain at

**TABLE 3** The overview of the datasets utilized, MAF sources incorporated, pedigree types employed, and the corresponding SNP counts investigated for kinship estimation performance across the four tools evaluated in this study.

| Software | Dataset with or w/o background relatedness | Pedigree type (w/o inbreeding) | MAF source | Noise application | No. of SNPs inspected | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 K | 5 K | 10 K | 20 K | 50 K |
| READ | w/o | All | NA | NA | √ | √ | √ | √ | √ |
| | With | Parent-offspring and siblings | NA | NA | √ | X | X | √ | X |
| NgsRelate v2 | w/o | All | TSI | Without noise | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | TSI | With noise (SD = 0.5) | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | TSI | With noise (SD = 1) | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 1× BAMs | Without noise | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 1× BAMs | With noise (SD = 0.5) | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 1× BAMs | With noise (SD = 1) | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 5× BAMs | Without noise | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 5× BAMs | With noise (SD = 0.5) | √ | √ | √ | √ | √ |
| | w/o | Parent-offspring and siblings | 5× BAMs | With noise (SD = 1) | √ | √ | √ | √ | √ |
| | With | Parent-offspring and siblings | TSI | None | √ | X | X | √ | X |
| lcMLkin v2 | w/o | All | TSI | None | √ | √ | √ | √ | √ |
| KIN | w/o | All | NA | NA | X | √ | √ | √ | √ |

*Note*: The pedigrees with inbreeding are not shown. NA denotes tools that do not use MAF information. None denotes tools without the application of noise, though they employ MAF information.

least one read shared between a pair of individuals using a custom bash script.

We randomly downsampled the remaining sites to 1, 5, 10, 20, and 50 K, independently five times each, for each pair and gave these downsampled SNP lists as input with the '--bed' argument to run the KINgaroo algorithm, a python package to generate ROH estimates and input files for KIN. We ran KINgaroo with default parameters without contamination correction (using the '--cnt 0' option) and without indexing and sorting of BAM files (using the '--s 0' option) for each triplet separately 25 times ($n = 5$ SNP counts $\times$ $n = 5$ replicates).

We separately collected pairwise mismatch values (P0) of pairs for each relationship type (found in 'p_all.csv' file under the 'hmm_parameters' directory created by KINgaroo) and calculated their median P0 values for each SNP count and replicate, corresponding to a P0 value of an average unrelated pair. To apply normalization for kinship estimation with these median values (~0.24), we manually changed the text files of P0, 'p_0.txt' created by KINgaroo under the 'hmm_parameters' directory. We then ran KIN with input files separately for each triplet using default parameters. The KIN output file includes the Jacquard coefficients ($k_0$, $k_1$, and $k_2$) for each pair analysed. We calculated the kinship coefficient using these estimates as $\theta = (k_1 / 4 + k_2 / 2)$.

We note that KIN gave sporadic errors when analysing pairs with 1 K SNPs and when using data from grandparent–grandchild pairs under first-cousin mating (Appendix S1).

## 2.6 | Classification of kinship coefficient estimates

We categorized each simulated pair into one of four relationship categories, i.e. first-, second-, third-degree related or unrelated, using their $\theta$ estimates. Here, we used two assessment criteria. The first criterion was the arithmetic mean (average) of the theoretical kinship coefficient values. The arithmetic mean of two expected values $\theta_1$ and $\theta_2$ would be $(\theta_1 + \theta_2) / 2$, i.e. the mid-point of expected kinship coefficient values of two relatedness degrees (Table S6). READ and TKGWV2 also use this mid-point cut-off approach. The alternative criterion we explored was the geometric mean. The geometric mean of two expected values $\theta_1$ and $\theta_2$ would be $\sqrt{\theta_1 \times \theta_2}$, which is always smaller than the arithmetic mean. As $\theta$ values decrease with lower degrees of relatedness in a non-linear fashion (see Figures S1–S4), we asked if using the geometric mean may improve the accuracy of relatedness type classification. The cut-offs used are shown in Table S6. As zero values cannot be tolerated while calculating the geometric mean, we applied a modified geometric mean for third-degree cut-off using the *splicejam* (v0.0.63.900) package in R (Ward, 2023).

## 2.7 | Classification and accuracy

We created a confusion matrix using either the arithmetic or geometric mean criteria for each relationship type. We used the 'confusionMatrix' function of the R *caret* (v3.5) package (Kuhn, 2008). To maintain the balance between classes in confusion matrix calculation, we randomly selected only 96 second- and third-degree related and unrelated pairs using the 'sample' function of R without replacement. We used the same number of each relationship type for second and third-degree pairs ($n = 32$ each). After that, we prepared four different datasets for the tools, consisting of classified estimates based on either the arithmetic or geometric mean and their actual classes.

The classification metrics we used were the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), precision, and the *F* score ($F_1$). To understand how often the four software correctly identified genetic relatedness, we also determined the relative frequency of both true and false predictions for each class and SNP count. Additionally, we categorized false predictions according to their inferred classes.

## 2.8 | Statistical tests on kinship coefficient estimates

### 2.8.1 | Linear mixed effect model

We used a linear mixed effect model to study the effect of software choice and SNP count on $\theta$ estimates for each relationship type. The fixed effects were (a) the type of genetic relatedness estimation tools we used, i.e. READ, NgsRelate, KIN, and lcMLkin, and (b) SNP counts shared between simulated individuals (5, 10, 20, and 50 K). The pair of individuals was included as a random effect.

We used the 'lmer' function in the R *lmerTest* package (Bates et al., 2015) with the R code: lmer($\theta$ ~ Software + SNPCount + (1| pairs)). We repeated the analysis with the estimates for each relationship type separately. We used the R base function 'summary' on the lmer object to visualize *p* values of pairwise mean $\theta$ difference among software and SNP counts, using lcMLkin and 50 K SNPs as the baseline. To ensure data independence, if multiple pairs included the same individual (which happened among parent–offspring, grandparent–grandchild, and great-grandparent–great-grandchild pairs), we chose only one of the pairs. In this way, we kept only 24 pairs for these three relatedness types.

We further tested the effect of software and shared SNP counts on $\theta$ estimates using repeated measures ANOVA with the 'aov' function in R (R Core Team, 2022). We integrated the pair of individuals as an error term to represent individual differences while identifying within-group variabilities. We repeated the analysis with the $\theta$ estimates for each relationship type, SNP count, and replicate separately. As mentioned earlier, we chose only one of the pairs from parent–offspring, grandparent–grandchild, and great-grandparent–great-grandchild pairs to maintain data independence. We thus kept only 24 pairs for these three relatedness types.

Additionally, we applied the same linear mixed effect model using as a response variable the absolute residuals, i.e. the absolute differences between the $\theta$ estimate of a pair and theoretical $\theta$ value,

AMD = $|\theta_{expected} - \theta|$. This way, we investigated the possible deviations from the theoretical values while accounting for the variance between pairs.

#### 2.8.2 | Levene's test

We performed the Levene's test to explore the homogeneity of variances between the kinship coefficient estimates of the tools using the 'leveneTest' function in the R *car* package (Fox & Weisberg, 2011). We first divided the estimates from READ, NgsRelate, lcMLkin, and KIN into groups based on SNP counts and replicates. Then, we applied Levene's test separately to each group.

## 3 | RESULTS

### 3.1 | Similar performance among tools at ≥20 K SNPs

We studied the performance of lcMLkin, NgsRelate, READ, and KIN on genomic data from simulated first- to third-degree relatives and unrelated pairs using various shared SNP numbers from 1 to 50 K, without background relatedness or inbreeding (Section 2). $\theta$ distributions across all studied pairs and replicates (Figures S1–S4), the mean $\theta$ estimates (Figure 2), as well as correct kinship degree assignment rates (Figure 3) were similar among the four tools using downsampled sets of either 50 or 20 K SNPs. The variance in $\theta$ tended to be negatively correlated with the SNP count, and in the analyses of first-degree pairs, all $\theta$ estimates had higher variance between siblings than between parent–offspring.

We found that identifying first-degree relatives is possible with ≥5 K SNPs with all four tools using this dataset with high reliability (≥97.5% correct assignment). Even with 1 K SNPs, READ could assign first-degree pairs correctly with a frequency of 85%, and NgsRelate and lcMLkin at a frequency of >96% (Figure 3). NgsRelate and lcMLkin achieved acceptable performance levels with as few as 1 K SNPs for distinguishing between second- versus third-degree kin and third-degree kin versus unrelated pairs (Figure 3). In contrast, READ and KIN required ≥10 K SNPs to achieve >80% correct assignment for these classes.

### 3.2 | Bias and variation in $\theta$ estimates among the four tools

We found that $\theta$ estimates from all tools display slight biases, but their level and directions depend on the relationship type and tool. One consistent trend was underestimating $\theta$ in first-degree relationships and grandparent–grandchild pairs and overestimating $\theta$ among unrelated pairs (Figure 2). We tested the effect of software choice and SNP count on $\theta$ estimates with a linear mixed effect model (Table S1) and with repeated measures ANOVA separately for different SNP counts (Table S2), which supported the observation of slight but significant differences in estimation among tools, especially in third-degree relationship types. We further compared the absolute mean differences between observed and expected $\theta$ (residuals) with the same linear mixed effect approach. Testing all eight kinship types separately, and for each type, at least one pair of software showed significant differences in the magnitude of residuals (at *t*-test $p < .05$) (Table S3).

We next studied whether variance among $\theta$ estimates (as opposed to bias) significantly differs among tools. We ran Levene's test for variance differences, comparing estimates among the four tools for each relatedness type and SNP count separately (Table S4). This revealed significant differences in $\theta$ variances among the tools, especially with ≤10 K SNPs (72/90 of comparisons with $p < .05$), which is consistent with their variable classification performance at low SNP counts (Figure 3). The only exceptions were grandparent–grandchild and great-grandparent–great-grandchild pairs, for which variances were similar among tools.

### 3.3 | Higher classification accuracy with NgsRelate and lcMLkin than other tools

Next, we calculated standard accuracy metrics to represent the four tools' classification performances (Figure 4). All tools had high (>98%) $F_1$ accuracy values for first-degree relatives down to 5 K SNPs. Even using 1 K SNPs, READ had $F_1$ 86%, while NgsRelate and lcMLkin had $F_1$ 96% (Table S5). For second-degree relatives at 5 K SNPs, lcMLkin and NgsRelate had $F_1$ values of 93 and 94%, respectively, while READ $F_1$ was only 83% and that of KIN was 88%, similar to values reported by Popli and colleagues (Popli et al., 2023). We found similarly compromised assignments for third-degree related pairs using READ and KIN at 5 K SNPs (69–79%) compared to lcMLkin and NgsRelate (91–93%). We also note that second- and third-degree relative estimations never reached 100% accuracy, even at 50 K SNPs.

### 3.4 | Using geometric versus arithmetic mean thresholds

Because $\theta$ and kinship degrees are not linearly correlated (e.g. see Figure S1), we asked if the geometric mean may provide a more suitable threshold (Section 2). We ran the classification of the same pairs using the same $\theta$ estimates from all four tools using the geometric mean as the threshold. We found slightly higher true positive rates using the geometric mean over the arithmetic mean for all categories except third-degree relatives (Figure S5). Overall, the differences between the thresholds appear too modest to entail a change in assignment strategy.

### 3.5 | Noise in population allele frequency leads to over- or underestimation of $\theta$

Higher Gaussian noise in background allele frequencies led to systematic overestimation of $\theta$ (>0.25) for all 96 pairs that

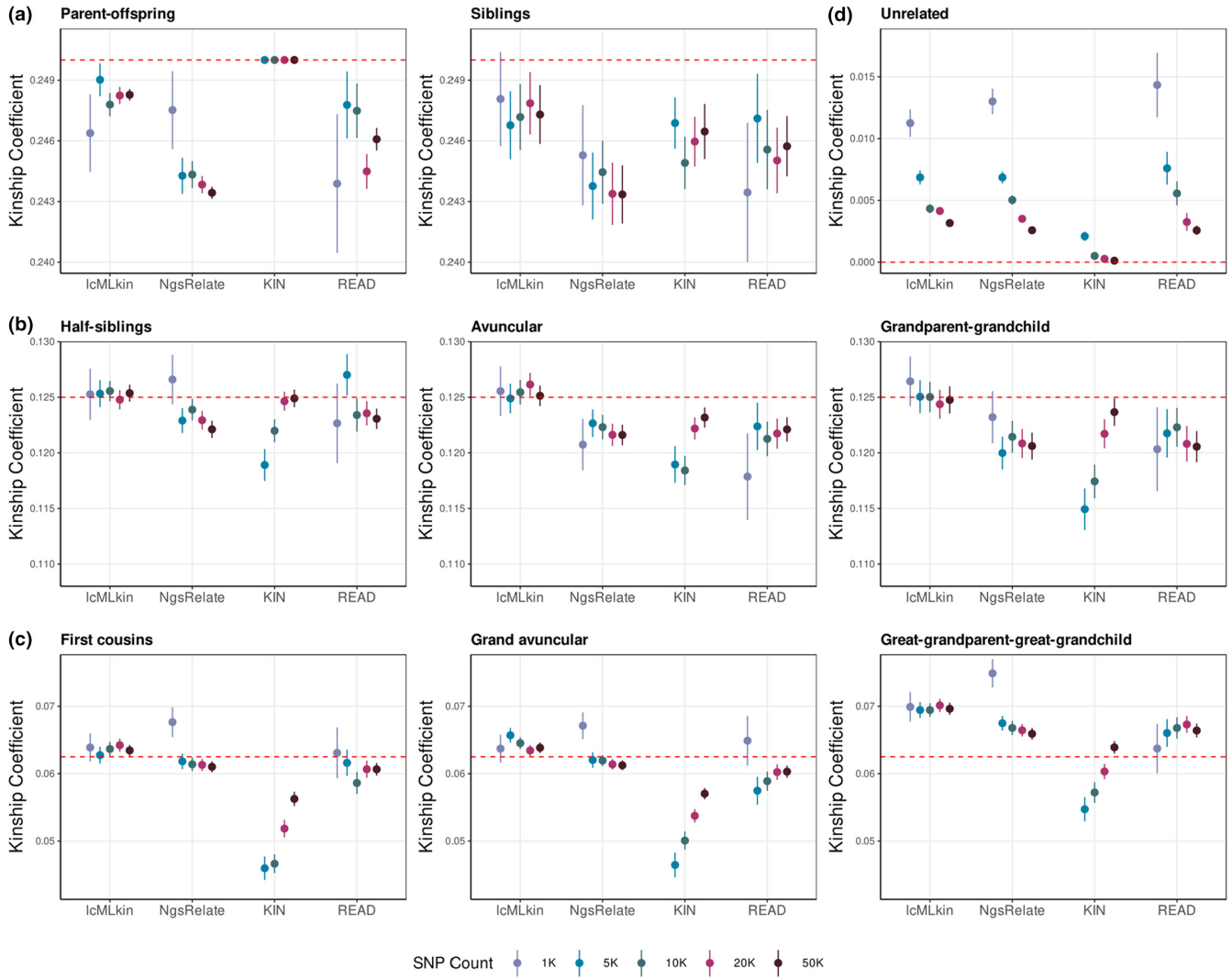


FIGURE 2 The mean $\theta$ estimates across different tools and SNP counts for (a) first-degree pairs, (b) second-degree pairs, (c) third-degree pairs, and (d) unrelated pairs, using all pairs ($n=48$) and replicates ($n=5$ per pair). Results for each overlapping SNP count are described with distinctive colours. The points show the mean and the vertical lines show $\pm 1$ standard error, estimated using all pairs ($n=48$) and replicates ($n=5$ per pair). The red dashed line represents the theoretical $\theta$ value for the corresponding relatedness degree. The results reveal variable levels of bias, which are not necessarily correlated with SNP counts.

we analysed (48 siblings and 48 parent–offspring pairs) using NgsRelate (Figure 5a,b and Figures S7, S8). However, noise related to imprecise minor allele frequency estimation led to a slight but systematic underestimation of $\theta$, with 95% of parent–offspring pair comparisons ($n=48$ pairs $\times$ $n=5$ SNP counts $\times$ $n=5$ replicates) with $\theta<0.25$ and 76% sibling pair comparisons with $\theta<0.25$ (Figure 5a,b and Figures S7, S8). Indeed, the underestimation trend was mitigated when using allele frequencies estimated from 5× genomes instead (Figure 5a,b and Figures S7, S8).

## 3.6 | Background relatedness has a limited effect on kinship estimates

We studied the performance of READ and NgsRelate on genomes with background relatedness due to genetic drift, produced using population genetic simulations (Section 2). We found READ $\theta$ estimates were practically the same when genomes contained background relatedness compared to when they did not. Meanwhile, NgsRelate tended to underestimate $\theta$ with these genomes, albeit minimally (<0.025) (Figure 5c).

## 3.7 | The effect of inbreeding on $\theta$ estimates

Inbreeding, either through consanguinity or through small population size, can create distal IBD loops between pairs of individuals and elevate $\theta$ estimates beyond that expected from the proximal relationship (Figure 1). We tested the four tools first using parent–offspring simulations, where the parents of the offspring were the first cousins. Average $\theta$ values from READ, lcMLkin, and NgsRelate were 0.27–0.28, as expected (Figure 6a and Figure S9). KIN

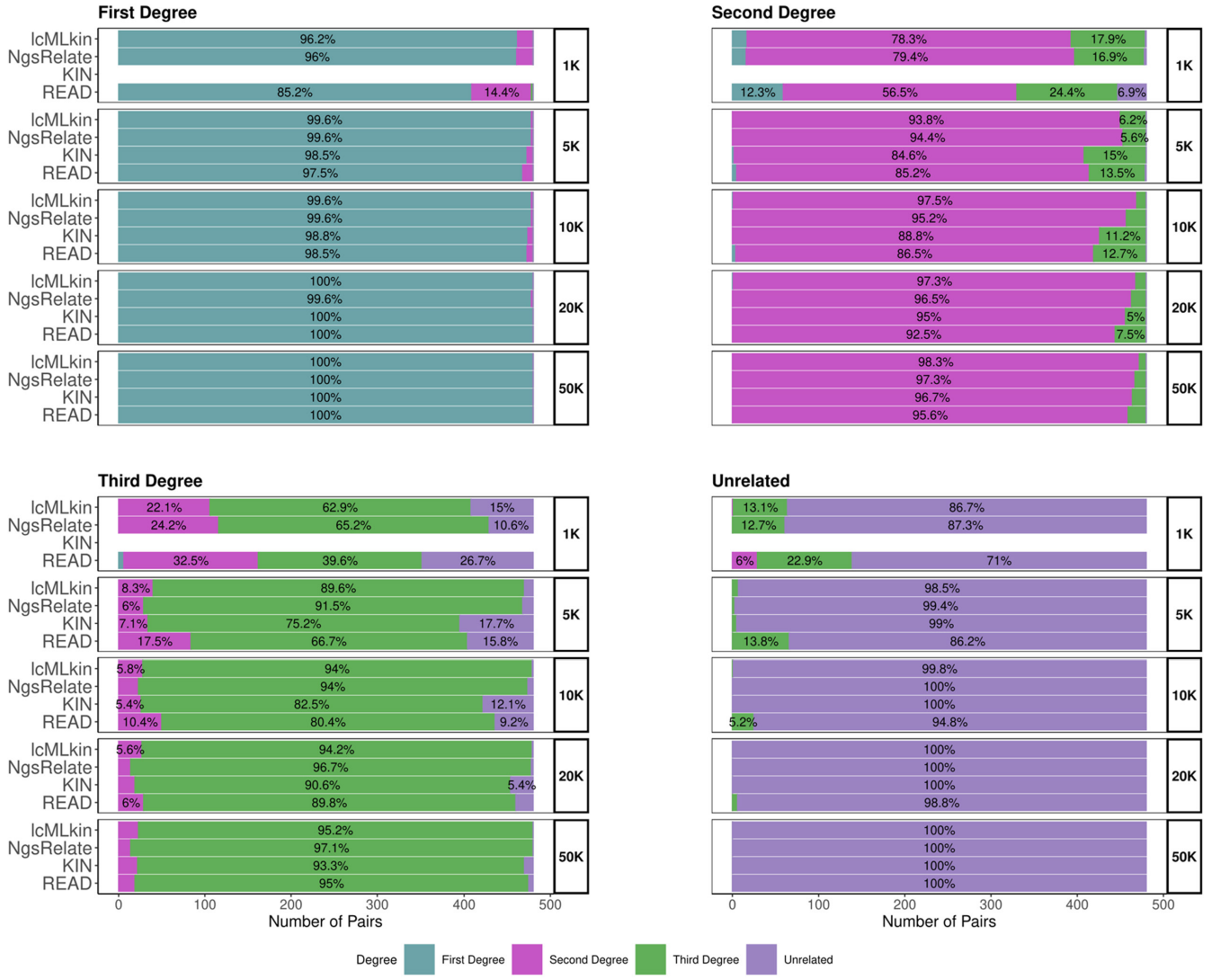


FIGURE 3 The relative frequency of pairs assigned to first-, second-, and third-degree related and unrelated categories by lcMLkin, NgsRelate, KIN, and READ. The kinship coefficient estimates from these tools were classified using the arithmetic mean of theoretical kinship coefficients. Colours refer to the assigned relatedness degree. The frequencies of pairs assigned to each category are indicated as percentages inside the bars (only for categories with frequency >5%). The results indicate similar performance of all tools at and above 20 K SNPs and better performances of lcMLkin and NgsRelate at low SNP counts.

estimates were all 0.25 (except for a single pair using 50 K SNPs). For NgsRelate, we also calculated a modified $\theta$ version, $\hat{\theta} = J_7 / 2 + J_8 / 4$, which is expected to reflect proximal IBD sharing without IBD due to distal loops. These $\hat{\theta}$ estimates were slightly but systematically lower than what would be expected from proximal loops (~0.24 using ≥5 K SNPs).

For grandparent–grandchild pairs, with the grandchild being the offspring of first cousins, READ, lcMLkin, and NgsRelate $\theta$ values were higher than expected from proximal loops (Figure 6b and Figure S10). This time, NgsRelate $\hat{\theta}$ values were also overestimated, but at a lower degree than the earlier three $\theta$ estimates. KIN did not perform with this dataset.

NgsRelate also estimates the individual inbreeding coefficient, *F*. This should be 0.0625 for first-cousin mating. The NgsRelate mean *F* estimates for the child were 0.075 for 1 K SNPs, but 0.051–0.055 for ≥5 K SNPs in the parent–offspring dataset; likewise, mean *F* was 0.068 for 1 K SNPs, but 0.041–0.048 for ≥5 K SNPs in the grandparent–grandchild dataset, suggesting that NgsRelate tends to over- or underestimate *F* in different settings.

## 4 | DISCUSSION

Our benchmarking using simulated genomes revealed a number of interesting observations on the four tools tested on sparse and low-coverage SNP data. First, all four tools, lcMLkin, NgsRelate, KIN and READ, perform well and are consistent with each other down to 20 K shared SNPs, even in separating third-degree and unrelated pairs (Figure 3). This SNP count lower limit corresponds to two genomes, each with ~0.1× coverage genotyped on a ~1 million SNP panel

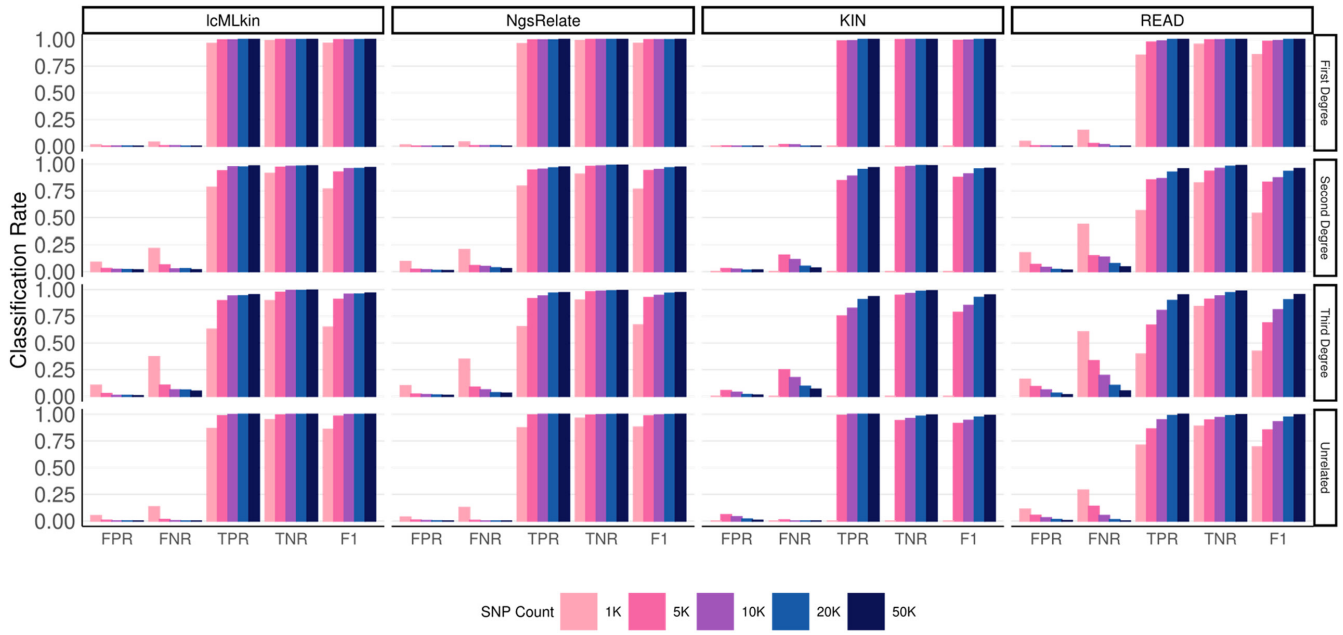


**FIGURE 4** Classification performance of the four tools using the primary dataset. FPR, false positive rate; FNR, false negative rate; TPR, true positive rate; TNR, true negative rate and $F_1$, accuracy. The classification was performed using $n = 48$ pairs x 5 replicates for each kinship type ($n = 96$ for first-, $n = 96$ for second-, and $n = 96$ for third-degree related and $n = 96$ for unrelated), generated using the primary dataset (no inbreeding, perfect background allele frequencies, and no background relatedness) and using the arithmetic mean to classify kinship coefficient estimates. Note that we randomly sub-sampled $n = 96$ pairs for second- and third-degree related categories with each relationship type represented equally ($n = 32$) to ensure balance. The colours represent the count of SNPs shared between individuals.

(Mallick et al., 2024); Mathieson et al., 2015), or each with ~0.06× genotyped on the 1000 Genomes v3 Africa diversity panel of ~5 million SNPs (Koptekin et al., 2023). Theoretically, this lower limit also applies to comparisons between a 1× genome and a 0.004× genome, using a 5 million SNP panel.

The variance in $\theta$ between replicates exhibited a negative correlation with SNP count (Figure 2 and Figures S1–S4), attributable to stochastic noise. As expected, $\theta$ estimates also displayed higher variance between siblings compared to parent–offspring pairs (Figure 2 and Figure S1), as IBD between siblings varies across the genome due to the randomness of recombination.

We also observed a number of systematic differences in performance among the tools. READ generally performs worse than the other three tools with these data in terms of higher variance in $\theta$ estimates and, hence, lower assignment accuracy (Figures 2 and 3 and Figures S1–S4). Meanwhile, KIN $\theta$ distributions have lower variance than the other tools but not improved accuracy, with higher degrees of misassignment than lcMLkin and NgsRelate (Figure 3). For instance, using 5 K SNPs, the correct assignment of first-degree relatives was 99.6% for both lcMLkin and NgsRelate, compared to 98.5% for KIN and 97.5% for READ. For third-degree relatives, using again 5 K SNPs, correct assignment rates were 91.5% for lcMLkin and 89.6% for NgsRelate, in contrast to 75.2% for KIN and 66.7% for READ. This difference may be expected, as lcMLkin and NgsRelate use more information (population allele frequencies per site) to normalize genomic mismatch rates.

Our comparisons of variance in $\theta$ estimates across tools using Levene's test also supported the earlier observations. We found significant differences among tools for nearly all relationship types below 10 K SNPs. Interestingly, grandparent–grandchild and great-grandparent–great-grandchild pairs were an exception to this pattern, such that all tools had comparable variances (Table S4). This observation may be attributed to fewer recombination events in these two kinship types (Qiao et al., 2021).

Beyond variance in $\theta$ estimates, average $\theta$ estimates were generally close to expected values under most conditions (Figure 2). Nevertheless, slight shifts from expected values can be noticed in Figures S1–S4 and Figure 2. The tools underestimated $\theta$ in first-degree relationships and grandparent–grandchild pairs but overestimated $\theta$ among unrelated pairs. Further, KIN diverged from the other tools in displaying the strongest downward bias for related pairs but the least upward bias for unrelated pairs. Except for KIN estimates, the observed biases were not strongly correlated with SNP counts. NgsRelate and lcMLkin appeared overall least biased, but not for all kinship types; e.g. for great-grandparent–great-grandchild pairs, READ estimates were closest to expectation. To summarize, we observed slight biases in the $\theta$ estimates by all tools, yet the magnitude and tendencies of these biases varied based on the type of relationship and the specific tool employed (Figure 2 and Table S1).

Similar trends emerged when analysing absolute mean differences from expectation (residuals) via a linear mixed effect model.

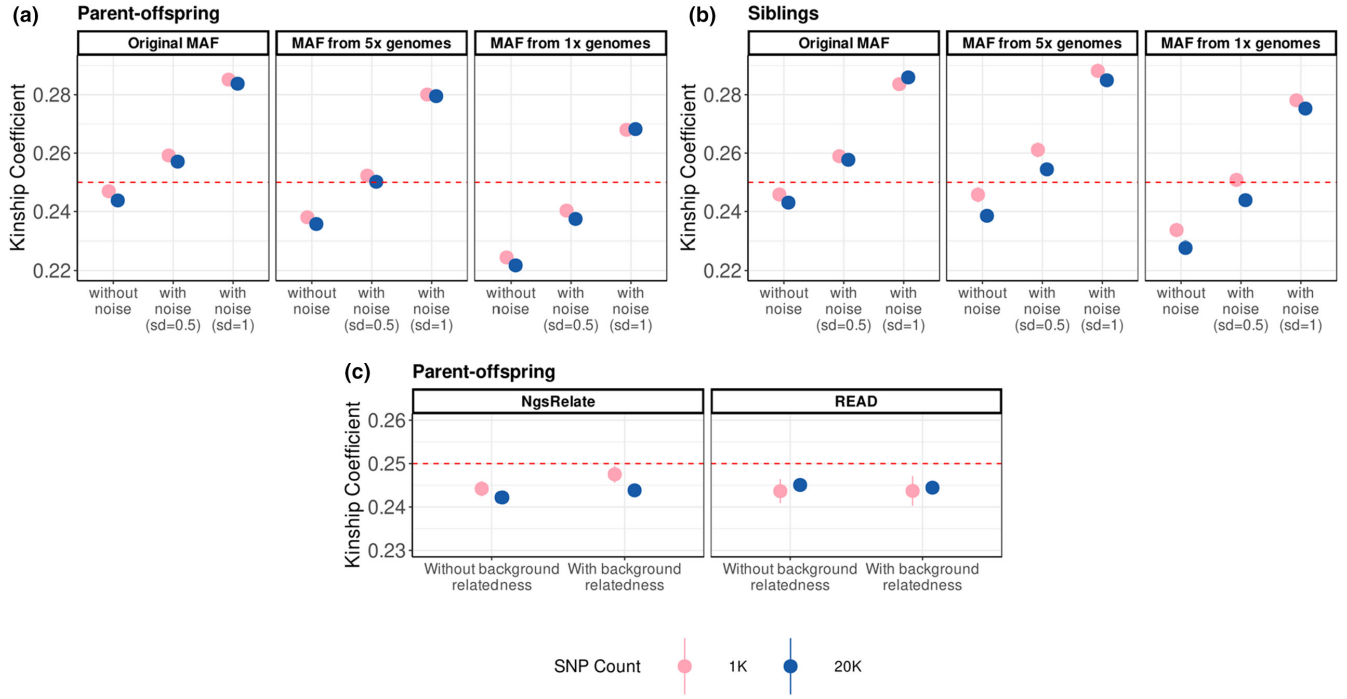


FIGURE 5 The effects of background allele frequency noise and background relatedness on $\theta$ estimations. (a) Parent–offspring and (b) sibling $\theta$ distributions under noise in allele frequencies, calculated using NgsRelate using $n=48$ pairs each, and 1 and 20 K SNPs. 'MAF without noise' indicates TSI allele frequencies (perfect information) or MAF from 5× and 1× genomes; 'MAF with noise (SD = 0.5)' and 'MAF with noise (SD = 1)' indicate cases where random Gaussian noise is added to allele frequencies; 'MAF from 5× genomes' and 'MAF from 1× genomes' indicate MAF called using genomes of the indicated coverage (Section 2). (c) Parent–offspring $\theta$ distributions without or with background relatedness using NgsRelate and READ. The points show the mean ($n=48$ pairs × $n=5$ replicates) and the vertical lines show ± one standard error (not visible in panels a and b) for 1 and 20 K SNPs. 'Without background relatedness': the main simulations where synthetic founders were created without background relatedness. 'With background relatedness': simulations where we produced founders using a coalescent simulator and realistic demographic model. The results indicate that Gaussian noise versus noise caused by imprecise population allele frequency estimates have opposing effects on $\theta$ estimates.

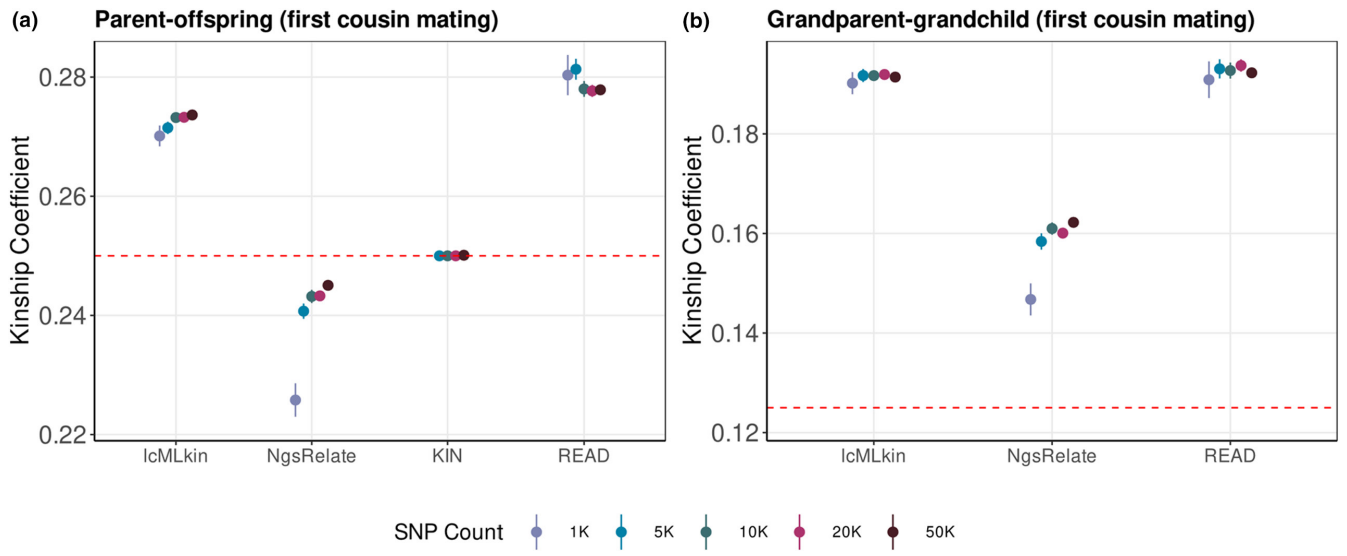


FIGURE 6 The mean $\theta$ estimates across different tools and SNP counts for (a) parent–offspring pairs (first-cousin mating) and (b) grandparent–grandchild pairs (first-cousin mating). Results for each overlapping SNP count are described with distinctive colours. The points show the mean and the vertical lines show ±1 standard error, estimated using all pairs ($n=48$) and replicates ($n=5$ per pair). The kinship coefficient from NgsRelate ($\widehat{\theta}$) was calculated ignoring the inbreeding-related Jacquard coefficients. The red dashed line represents the theoretical kinship coefficient value for the corresponding relatedness degree.

Across all eight kinship types examined individually, significant differences in residual magnitudes were detected between at least one pair of software (*t*-test $p < .05$) (Table S3). These trends, though, appear to have limited impact on classification accuracy: e.g. for siblings, NgsRelate displays the strongest downward bias in average $\theta$ estimates, but its classification accuracy is higher than both READ and KIN and is on a par with lcMLkin (Figure 3). As expected, SNP count also significantly affected residuals (i.e. variance), with larger residuals at lower SNP counts (Table S3).

When evaluating standard accuracy metrics (Figure 4), we found that all tools achieved high $F_1$ accuracy values for first-degree relationships, even with as few as 5 K SNPs. However, NgsRelate and lcMLkin consistently outperformed READ and KIN for relationships beyond the first degree, particularly at lower SNP counts (Table S5). This trend aligns with the observed higher variation in READ $\theta$ estimates and downward bias in KIN $\theta$ estimates.

As discussed earlier, READ and KIN display lower performance at low SNP counts than lcMLkin and NgsRelate. READ and KIN use the median mismatch rate in a sample of pairs for normalization, whereas lcMLkin and NgsRelate use population allele frequency estimates. We reasoned that using perfect knowledge of allele frequencies (frequencies used to create the founders) in our analysis may have favoured the performance of lcMLkin and NgsRelate. Indeed, Lipatov et al. (2015) tested imperfect allele frequencies by using the Balding–Nichols model with various $F_{ST}$ values (0.01, 0.05, and 0.1) at each SNP and observed overestimation of $\theta$ with increasing $F_{ST}$. Hence, we repeated NgsRelate with imperfect population allele frequencies in a subset of our data. Consistent with Lipatov et al. (2015), we found that higher random Gaussian noise led to systematic overestimation of $\theta$, which arises because inaccurate background allele frequencies inflate the impact of being identical-by-state (IBS) between any pair.

We then introduced another type of noise, imprecise minor allele frequencies, when running NgsRelate. Intriguingly, this led to an underestimation of $\theta$ for the majority of parent–offspring and sibling pairs (Figure 5). The reason for this underestimation trend could be related to the lower representation of relatively rare variants when estimating allele frequencies from low-coverage genomes (Figure S6). Overall, these results suggest that different sources of noise in population allele frequency estimates can compromise the performance of lcMLkin and NgsRelate. This would also be consistent with the results by Marsh et al. (2023), who reported low performance of the latter two tools on real genomic datasets.

We further asked if background relatedness among the founders, which would arise due to drift, may cause a shift in $\theta$ estimates. At least in our simulated scenario of European Neolithic ancestry with an effective population size of 250, the presence of background relatedness among founders did not substantially influence the accuracy or reliability of $\theta$ estimates produced by READ and NgsRelate using either 1 K or 20 K SNP sets (Figure 5c).

We mark that these results reflect the upper bounds of performance in real datasets for a number of reasons:

a. Most of our lcMLkin and NgsRelate analyses presented used perfect information on background allele frequencies, which may be slightly or highly unrealistic in real settings, depending on the dataset.
b. Our sets of sample pairs used for normalizing mismatch rates, used by READ and KIN, do not include population structure. Heterogeneous ancestries in a sample can lead to overestimation of $\theta$, as pointed out by Popli and colleagues (Popli et al., 2023).
c. Our primary genome simulation dataset lacks background relatedness among the founders, which would be present at variable degrees in real data and could confound estimates of proximal IBD. This involves results from all four tools. Our experiment with founders obtained from a realistic demographic model did not create a major shift in $\theta$ estimates. Still, we note that the effect depends on the effective population size, so that in bottlenecked populations $\theta$ estimates might be affected.
d. We did not include identical genomes or fourth-degree or more distant kin in the simulations. The presence of more variable classes would increase the chance of misidentification and would lower classification accuracy overall.

In our primary simulations, NgsRelate and lcMLkin were found to be more accurate than READ and KIN, with lower false positive and false negative rates, especially when using <20 K shared SNPs (Figure 3 and 4). The former tools both use genotype likelihoods and population allele frequencies. However, as our trials with noise-added or imperfectly estimated population allele frequencies reveal, this performance might be compromised in real-life applications. In fact, in our own experience, READ results appear highly robust and reproducible compared to those of other tools (e.g. Altınışık et al., 2022; Yaka et al., 2021).

Among the tools tested, KIN performs the most sophisticated estimation, which includes inference of both ROH and shared IBD segments using HMMs, calculating likelihoods for kinship degree assignment, and classifying parent–offspring and sibling pairs (we note that the recently released READv2 also distinguishes parent–offspring and siblings; Alaçamlı et al., 2024). KIN also differed from the other tools in estimating all simulated parent–offspring pairs' kinship coefficients as precisely as 0.25 due to the authors having constrained the parameter optimization space for this relationship type (Popli et al., 2023). However, the accuracy of KIN was not generally much superior to that of READ. We also note that we failed to run KIN on 1 K SNP datasets (due to sporadic errors likely due to convergence issues) and on one dataset that included inbreeding.

Marsh and colleagues (Marsh et al., 2023) recently tested the performance of kinship estimation software, including READ, NgsRelate, and lcMLkin (as well as TKGWV2 and PMR calculation), and using real high-coverage ancient human genomes from three different publications. Assuming the relatedness degree identified by the tools on the original high-coverage genomes as ground truth, they studied kinship estimates using downsampled versions of the same genomes (0.02×–2.1×). Interestingly, the authors found that the performance of genotype likelihood-based

methods (NgsRelate and lcMLkin) dropped starkly as the false negative rate increased. In contrast, the performances by READ, PMR, and TKGWV2 were relatively robust to low coverages. The reason for NgsRelate and lcMLkin performance being compromised in the Marsh et al. study might be sensitivity to noise in population allele frequencies.

Overall, these results suggest no single tool may be universally superior in estimating kinship levels with low-coverage genomes. Using multiple tools in parallel and interpreting the results in light of the superiorities and weaknesses of each tool and the particularities of each dataset (e.g. knowledge of allele frequencies, genetic structure within the sample, and the possibility of inbreeding) may be the most prudent and confident approach. Meanwhile, both the archaeogenomics community and wildlife geneticists may continue seeking novel and more powerful methods, such as combining the two alternative normalization approaches (population allele frequencies and the median mismatch in a sample) and using haplotype information (Ringbauer et al., 2024) to calculate more robust kinship coefficients.

## AUTHOR CONTRIBUTIONS

Ş.A., M.N.G., and M.S. designed the study; Ş.A., I.M., and M.N.G. produced the data with the support of K.B.V. Ş.A., M.N.G., and B.K. analysed data assisted by I.M., K.G., K.B.V., E.S., M.Ç., R.Y., E.S., G.A., S.S.Ç., A.S., N.E.A., D.K., M.S. Ş.A., M.N.G., B.K., and M.S. wrote the manuscript with contributions from all authors.

## ACKNOWLEDGEMENTS

We thank all members of the METU Biological Science CompEvo and of the Hacettepe Human_G groups, Torsten Günther, Gülşah Merve Kılınç, Aybar Can Acar, and Burçak Otlu for discussions and Divyaratan Popli and Douaa Zakaria for help. We acknowledge support from the European Research Council (ERC) Consolidator grant H2020 'NEOGENE' (no:772390 to M.S.) and H2020-WIDESPREAD-05-2020 TWINNING grant 'NEOMATRIX' (no: 952317 to M.S.), and by the 'TÜBİTAK-2210/A (The Scientific and Technological Research Council of Türkiye)' for M.N.G. and Ş.A.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

BAM files of simulated ancient sequence data from simulated pairs of related and unrelated individuals are openly available in Zenodo at http://doi.org/10.5281/zenodo.10070958 (Aktürk et al., 2023), http://doi.org/10.5281/zenodo.10079685 and http://doi.org/10.5281/zenodo.10079625. The source codes to generate ancient DNA sequences and run genetic relatedness tools are openly available in the GitHub repository under SevvalAkturk/Benchmarking_kinship (github.com) and to generate founder individuals as well as pedigrees under CompEvoMetu/adna_tools (github.com).

## ORCID

*Şevval Aktürk* https://orcid.org/0000-0003-4157-6551
*Igor Mapelli* https://orcid.org/0000-0001-9814-7884
*Merve N. Güler* https://orcid.org/0000-0001-7766-9333
*Kanat Gürün* https://orcid.org/0000-0002-0433-2593
*Büşra Katırcıoğlu* https://orcid.org/0009-0006-4095-9728
*Kıvılcım Başak Vural* https://orcid.org/0000-0003-3964-3065
*Ekin Sağlıcan* https://orcid.org/0000-0001-8646-1163
*Reyhan Yaka* https://orcid.org/0000-0002-9359-4391
*Elif Sürer* https://orcid.org/0000-0002-0738-6669
*Gözde Atağ* https://orcid.org/0000-0001-6173-3126
*Sevim Seda Çokoğlu* https://orcid.org/0000-0002-1055-3966
*Arda Sevkar* https://orcid.org/0000-0003-4573-6778
*N. Ezgi Altınışık* https://orcid.org/0000-0003-0653-4292
*Dilek Koptekin* https://orcid.org/0000-0003-2664-5774
*Mehmet Somel* https://orcid.org/0000-0002-3138-1307

## REFERENCES

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-del Vecchyo, D., … Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, *9*, e54967.

Aktürk, Ş., Mapelli, I., Güler, M. N., & Somel, M. (2023). Simulated Ancient Genomic Kinship Dataset: VCF and BAM (1×) Files for Related (including inbred) Pairs (1.0). *Zenodo*. https://doi.org/10.5281/zenodo.10070958

Alaçamlı, E., Naidoo, T., Aktürk, Ş., Güler, M. N., Mapelli, I., Vural, K. B., Somel, M., Malmström, H., & Günther, T. (2024). READv2: Advanced and user-friendly detection of biological relatedness in archaeogenomics. *BioRxiv*. https://doi.org/10.1101/2024.01.23.576660

Altınışık, E. (2023). lcMLkin v2.1. Github. https://github.com/altinisik/lcMLkin-v2.1

Altınışık, N. E., Kazancı, D. D., Aydoğan, A., Gemici, H. C., Erdal, Ö. D., Sarıaltun, S., Vural, K. B., Koptekin, D., Gürün, K., Sağlıcan, E., Fernandes, D., Çakan, G., Koruyucu, M. M., Lagerholm, V. K., Karamurat, C., Özkan, M., Kılınç, G. M., Sevkar, A., Sürer, E., … Somel, M. (2022). A genomic snapshot of demographic and cultural dynamism in upper Mesopotamia during the Neolithic transition. *Science Advances*, *8*, eabo3609.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., & McVean, G. A. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*, iyab229.

Bérénos, C., Ellis, P. A., Pilkington, J. G., & Pemberton, J. M. (2014). Estimating quantitative genetic parameters in wild populations: A comparison of pedigree and genomic approaches. *Molecular Ecology*, *23*, 3434–3451.

Bhérer, C., Campbell, C. L., & Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, *8*, 14994.

Broad Institute. (2019). *Picard Toolkit*. Github. https://broadinstitute.github.io/picard/

Caballero, M., Seidman, D. N., Qiao, Y., Sannerud, J., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Carmi, S., & Williams, A. L. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genetics*, *15*, e1007979.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008.

de Flamingh, A., Ishida, Y., Pečnerová, P., Vilchis, S., Siegismund, H. R., van Aarde, R. J., Malhi, R. S., & Roca, A. L. (2023). Combining methods for non-invasive fecal DNA enables whole genome and metagenomic analyses in wildlife biology. *Frontiers in Genetics*, *13*, 1021004.

Fernandes, D. M., Cheronet, O., Gelabert, P., & Pinhasi, R. (2021). TKGWV2: An ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. *Scientific Reports*, *11*, 21262.

Fowler, C., Olalde, I., Cummings, V., Armit, I., Büster, L., Cuthbert, S., Rohland, N., Cheronet, O., Pinhasi, R., & Reich, D. (2022). A high-resolution picture of kinship practices in an early Neolithic tomb. *Nature*, *601*, 584–587.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Sage.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., & Gao, Y. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*, 851–861.

Galla, S. J., Brown, L., Couch-Lewis, Y., Cubrinovska, I., Eason, D., Gooley, R. M., Hamilton, J. A., Heath, J. A., Hauser, S. S., Latch, E. K., Matocq, M. D., Richardson, A., Wold, J. R., Hogg, C. J., Santure, A. W., & Steeves, T. E. (2022). The relevance of pedigrees in the conservation genomics era. *Molecular Ecology*, *31*, 41–54.

Galla, S. J., Moraga, R., Brown, L., Cleland, S., Hoeppner, M. P., Maloney, R. F., Richardson, A., Slater, L., Santure, A. W., & Steeves, T. E. (2020). A comparison of pedigree, genetic and genomic estimates of relatedness for informing pairing decisions in two critically endangered birds: Implications for conservation breeding programmes worldwide. *Evolutionary Applications*, *13*, 991–1008.

Godoy, I., Korsten, P., & Perry, S. E. (2022). Genetic, maternal, and environmental influences on sociality in a pedigreed primate population. *Heredity*, *129*, 203–214.

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*, giz034.

Housworth, E. A., & Stahl, F. W. (2003). Crossover interference in humans. *American Journal of Human Genetics*, *73*, 188–197.

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, *25*, 918–925.

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *115*, 1472–1487.

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*, e1004842.

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, *14*(11), e1006581.

Kennett, D. J., Plog, S., George, R. J., Culleton, B. J., Watson, A. S., Skoglund, P., Rohland, N., Mallick, S., Stewardson, K., Kistler, L., LeBlanc, S. A., Whiteley, P. M., Reich, D., & Perry, G. H. (2017). Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nature Communications*, *8*, 14115.

Kılınç, G. M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A. M., Bıçakçı, E., Baird, D., Dönertaş, H. M., Ghalichi, A., Yaka, R., Koptekin, D., Açan, S. C., Parvizi, P., Krzewińska, M., Daskalaki, E. A., Yüncü, E., Dağtaş, N. D., Fairbairn, A., Pearson, J., … Götherström, A. (2016). The demographic development of the first farmers in Anatolia. *Current Biology*, *26*, 2659–2666.

Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. *Methods in Molecular Biology*, *840*, 197–228.

Koch, M., Hadfield, J. D., Sefc, K. M., & Sturmbauer, C. (2008). Pedigree reconstruction in wild cichlid fish populations. *Molecular Ecology*, *17*, 4500–4511.

Konovalov, D. A., & Heg, D. (2008). TECHNICAL ADVANCES: A maximum-likelihood relatedness estimator allowing for negative relatedness values. *Molecular Ecology Resources*, *8*, 256–263.

Koptekin, D., Yüncü, E., Rodríguez-Varela, R., Altınışık, N. E., Psonis, N., Kashuba, N., Yorulmaz, S., George, R., Kazancı, D. D., Kaptan, D., Gürün, K., Vural, K. B., Gemici, H. C., Vassou, D., Daskalaki, E., Karamurat, C., Lagerholm, V. K., Erdal, Ö. D., Kırdök, E., … Somel, M. (2023). Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. *Current Biology*, *33*, 41–57.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356.

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*(24), 4009–4011.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kuhn, J. M. M., Jakobsson, M., & Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS One*, *13*, e0195491.

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B. C., Hasan, A. R., Huang, X., Iasi, L. N. M., Noskova, E., Obsteter, J., Pavinato, V. A. C., … Gronau, I. (2023). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife*, *12*, RP84874.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformation Oxford England*, *27*, 2987–2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Lipatov, M., Sanjeev, K., Patro, R., & Veeramah, K. R. (2015). Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *BioRxiv*. https://doi.org/10.1101/023374

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., & Reich, D. (2024). The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Scientific Data*, *11*(1), 182.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*, 2867–2873.

Marsh, W. A., Brace, S., & Barnes, I. (2023). Inferring biological kinship in ancient datasets: Comparing the response of ancient DNA-specific software packages to low coverage data. *BMC Genomics*, *24*, 111.

Martiniano, R., Cassidy, L. M., Ó'Maoldúin, R., McLaughlin, R., Silva, N. M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M. J., Serra, M., Burger, J., Parreira, R., Moran, E., Valera, A. C., Porfirio, E., Boaventura, R., Silva, A. M., & Bradley, D. G. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics*, *13*, e1006852.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. M. B., Carbonell, E., …

Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*, 499–503.
Mittnik, A., Massy, K., Knipper, C., Wittenborn, F., Friedrich, R., Pfrengle, S., Burri, M., Carlichi-Witjes, N., Deeg, H., Furtwängler, A., Harbeck, M., von Heyking, K., Kociumaka, C., Kucukkalipci, I., Lindauer, S., Metz, S., Staskiewicz, A., Thiel, A., Wahl, J., … Krause, J. (2019). Kinship-based social inequality in bronze age Europe. *Science*, *366*, 731–734.
Moran, B. M., Thomas, S. M., Judson, J. M., Navarro, A., Davis, H., Sidak-Loftis, L., Korody, M., Mace, M., Ralls, K., Callicrate, T., Ryder, O. A., Chemnick, L. G., & Steiner, C. C. (2021). Correcting parentage relationships in the endangered California condor: Improving mean kinship estimates for conservation management. *Ornithological Applications*, *123*, duab017.
Ning, C., Zhang, F., Cao, Y., Qin, L., Hudson, M. J., Gao, S., Ma, P., Li, W., Zhu, S., Li, C., Li, T., Xu, Y., Li, C., Robbeets, M., Zhang, H., & Cui, Y. (2021). Ancient genome analyses shed light on kinship organization and mating practice of late Neolithic society in China. *iScience*, *24*, 103352.
Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M., & Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, *173*, 483–496.
O'Reilly, P. T., & Kozfkay, C. C. (2014). Use of microsatellite data and pedigree information in the genetic management of two long-term salmon conservation programs. *Reviews in Fish Biology and Fisheries*, *24*, 819–848.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*, 1065–1093.
Pemberton, J. M. (2008). Wild pedigrees: The way forward. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 613–621.
Pinho, G. M., da Silva, A. G., Hrbek, T., Venticinque, E. M., & Farias, I. P. (2014). Kinship and social behavior of lowland tapirs (*Tapirus terrestris*) in a Central Amazon landscape. *PLoS One*, *9*, e92507.
Popli, D., Peyrégne, S., & Peter, B. M. (2023). KIN: A method to infer relatedness from low-coverage ancient DNA. *Genome Biology*, *24*, 10.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.
Qiao, Y., Sannerud, J. G., Basu-Roy, S., Hayward, C., & Williams, A. L. (2021). Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. *American Journal of Human Genetics*, *108*, 68–83.
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformation Oxford England*, *26*, 841–842.
R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Renaud, G., Hanghøj, K., Willerslev, E., & Orlando, L. (2017). gargammel: a sequence simulator for ancient DNA. *Bioinformatics*, *33*, 577–579.
Ringbauer, H., Huang, Y., Akbari, A., Mallick, S., Olalde, I., Patterson, N., & Reich, D. (2024). Accurate detection of identity-by-descent segments in human ancient DNA. *Nature Genetics*, *56*(1), 143–151.
Rivollat, M., Rohrlach, A. B., Ringbauer, H., Childebayeva, A., Mendisco, F., Barquera, R., Szolek, A., le Roy, M., Colleran, H., Tuke, J., Aron, F., Pemonge, M. H., Späth, E., Télouk, P., Rey, L., Goude, G., Balter, V., Krause, J., Rottier, S., … Haak, W. (2023). Extensive pedigrees reveal the social organization of a Neolithic community. *Nature*, *620*, 600–606.
Sánchez-Quinto, F., Malmström, H., Fraser, M., Girdland-Flink, L., Svensson, E. M., Simões, L. G., George, R., Hollfelder, N., Burenhult, G., Noble, G., Britton, K., Talamo, S., Curtis, N., Brzobohata, H., Sumberova, R., Götherström, A., Storå, J., & Jakobsson, M. (2019). Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proceedings of the National Academy of Sciences*, *116*, 9469–9474.
Schiffels, S. SequenceTools. GitHub. https://github.com/stschiff/sequenceTools.git/
Schroeder, H., Margaryan, A., Szmyt, M., Theulot, B., Włodarczak, P., Rasmussen, S., Gopalakrishnan, S., Szczepanek, A., Konopka, T., Jensen, T. Z. T., Witkowska, B., Wilk, S., Przybyła, M. M., Pospieszny, Ł., Sjögren, K. G., Belka, Z., Olsen, J., Kristiansen, K., Willerslev, E., … Allentoft, M. E. (2019). Unraveling ancestry, kinship, and violence in a late Neolithic mass grave. *Proceedings of the National Academy of Sciences*, *116*, 10705–10710.
Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*, 88.
Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M. T. P., Götherström, A., & Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, *336*, 466–469.
Sousa da Mota, B., Rubinacci, S., Cruz Dávalos, D. I., G. Amorim, C. E., Sikora, M., Johannsen, N. N., Szmyt, M. H., Włodarczak, P., Szczepanek, A., Przybyła, M. M., Schroeder, H., Allentoft, M. E., Willerslev, E., Malaspinas, A. S., & Delaneau, O. (2023). Imputation of ancient human genomes. *Nature Communications*, *14*, 3660.
Ward, J. (2023). Splicejam. Github. https://github.com/jmw86069/splicejam/
Yaka, R., Mapelli, I., Kaptan, D., Doğu, A., Chyleński, M., Erdal, Ö. D., Koptekin, D., Vural, K. B., Bayliss, A., Mazzucato, C., Fer, E., Çokoğlu, S. S., Lagerholm, V. K., Krzewińska, M., Karamurat, C., Gemici, H. C., Sevkar, A., Dağtaş, N. D., Kılınç, G. M., … Somel, M. (2021). Variable kinship patterns in Neolithic Anatolia revealed by ancient genomes. *Current Biology*, *31*, 2455–2468.
Žegarac, A., Winkelbach, L., Blöcher, J., Diekmann, Y., Krečković Gavrilović, M., Porčić, M., Stojković, B., Milašinović, L., Schreiber, M., Wegmann, D., Veeramah, K. R., Stefanović, S., & Burger, J. (2021). Ancient genomes provide insights into family structure and the heredity of social status in the early bronze age of southeastern Europe. *Scientific Reports*, *11*, 10072.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Aktürk, Ş., Mapelli, I., Güler, M. N., Gürün, K., Katırcıoğlu, B., Vural, K. B., Sağlıcan, E., Çetin, M., Yaka, R., Sürer, E., Atağ, G., Çokoğlu, S. S., Sevkar, A., Altınışık, N. E., Koptekin, D., & Somel, M. (2024). Benchmarking kinship estimation tools for ancient genomes using pedigree simulations. *Molecular Ecology Resources*, *24*, e13960. https://doi.org/10.1111/1755-0998.13960