

University of Szeged
Albert Szent-Györgyi Medical School
Doctoral School of Multidisciplinary Medical Science

A novel method to determine kinship relationships up to 4th degree

Ph.D. Thesis

Nyerki Emil

Supervisor:

Dr. Maróti Zoltán, senior research fellow

Albert Szent-Györgyi Health Center

Department of Pediatrics and Pediatric Health Center

Genetical Diagnostical Laboratory

Szeged

2024

1. List of publications

1.1. List of papers directly related to the subject of the thesis:

- I. **Emil Nyerki**, Tibor Kalmár, Oszkár Schütz, M. Lima Rui, Endre Neparácski, Tibor Török, Zoltán Maróti. *Correctkin: An Optimized Method to Infer Relatedness up to the 4th Degree from Low-Coverage Ancient Human Genomes*. Genome Biology 24, no. 1 (2023). <https://doi.org/10.1186/s13059-023-02882-4>. **IF:12.3 D1**
- II. Gergely I.B. Varga, Lilla Alida Kristóf, Kitti Maár, Luca Kis, Oszkár Schütz, Orsolya Váradi, Bence Kovács, Alexandra Gînguță, Balázs Tihanyi, Péter L. Nagy, Zoltán Maróti, **Emil Nyerki**, Tibor Török, Endre Neparácski. *The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy*, Journal of Genetics and Genomics, Volume 50, Issue 1, (2023) **IF:5,9 Q1**

1.2. Other publications

- III. Maár, Kitti, Gergely I. B. Varga, Bence Kovács, Oszkár Schütz, Zoltán Maróti, Tibor Kalmár, **Emil Nyerki**, István Nagy, Dóra Latinovics, Balázs Tihanyi, and et al. 2021. *Maternal Lineages from 10–11th Century Commoner Cemeteries of the Carpathian Basin*, Genes 12, no. 3: 460. **IF: 4,141**
- IV. Zoltán Maróti, Endre Neparácski, Oszkár Schütz, Kitti Maár, Gergely I.B. Varga, Bence Kovács, Tibor Kalmár, **Emil Nyerki**, István Nagy, Dóra Latinovics, Balázs Tihanyi, Antónia Marcsik, György Pálfi, Zsolt Bernert, Zsolt Gallina, Ciprián Horváth, Sándor Varga, László Költő, István Raskó, Péter L. Nagy, Csilla Balogh, Albert Zink, Frank Maixner, Anders Götherström, Robert George, Csaba Szalontai, Gergely Szenthe, Erwin Gáll, Attila P. Kiss, Bence Gulyás, Bernadett Ny. Kovacsóczy, Szilárd Sándor Gál, Péter Tomka, Tibor Török, *The genetic origin of Huns, Avars, and conquering Hungarians*, Current Biology, Volume 32, Issue 13, 2022, **IF:9,1**
- V. Gînguță A, Kovács B, Schütz O, Tihanyi B, **Nyerki E**, Maár K, Maróti Z, Varga GIB, Băcăuț-Crișan D, Keresztes T, Török T, Neparácski E. *Genetic identification of members of the prominent Báthory aristocratic family*. iScience. 2023 Sep 14;26(10):107911, **IF: 5.08**
- VI. Tibor Szabó, Melinda Magyar, Kata Hajdú, Márta Dorogi, **Emil Nyerki**, Tünde Tóth, Mónika Lingvay, Győző Garab, Klára Hernádi, László Nagy, *Structural and*

Functional Hierarchy in Photosynthetic Energy Conversion—from Molecules to Nanostructures, Nanoscale Res Lett 10, 458 (2015) **IF:2,58**

- VII. Tibor Szabó, **Emil Nyerki**, Tünde Tóth, Richárd Csekő, Melinda Magyar, Endre Horváth, Klára Hernádi, Balázs Endrődi, Csaba Visy, László Forró, László Nagy, *Generating photocurrent by nanocomposites based on photosynthetic reaction centre protein*, Physica Status Solidi b, 2015, **IF=1,66**

Total cumulative impact factor: 40,761

2. Table of content

1.	List of publications.....	1
1.1.	List of papers directly related to the subject of the thesis:	1
1.2.	Other publications	1
2.	Table of content	3
3.	List of abbreviations.....	5
3.1.	List of population abbreviations.....	5
4.	Introduction	7
4.1.	Ancient DNA.....	7
4.2.	Kinship analysis	10
5.	Objective	13
6.	Materials and Methods	14
6.1.	Used Data	14
6.2.	New bioinformatics tools	14
6.3.	Principal component analysis.....	15
6.4.	Simulating the effect of low coverage from fully typed modern datasets.....	15
6.5.	Simulation of aDNA-related genotyping errors	16
6.6.	Uncorrected kinship coefficient estimation.....	17
7.	Results	18
7.1.	Effect of random pseudo-haploidization on PCA	18
7.2.	Effect of random pseudo-haploidization on kinship coefficient	19
7.3.	The effect of overlapping marker fraction on the kinship coefficient calculation ...	20
7.4.	The effect of genotyping errors on the corrected kinship coefficient	21
7.5.	The effect of reference population selection on kinship analysis	23
7.6.	Effect of reference population selection on kinship analysis in a complex admixed family with multiple ethnic relations	24
7.7.	Statistical validation, assessment of technical errors	26

7.8.	Kinship analysis of ancient samples with known relations using kinship coefficient correction.....	28
7.9.	Kinship analysis of ancient samples from the AADR 1240K dataset.....	30
7.10.	Genetic validation of St.Ladiuslaus	32
8.	Discussion	33
9.	Conclusion.....	39
10.	Acknowledgements	41
	References	42
	Annex	1

3. List of abbreviations

1KG: 1000 Genomes Project Phase 3

AADR: Allen Ancient DNA Resource

aDNA: Ancient DNA

IBD: identity-by-descent

IBS: identity-by-state

NGS: New generation sequencing

PCA: Principal component analysis

READ: Relationship Estimation from Ancient DNA

RPSH: random pseudo-haploidization

SNP: single nucleotide polymorphism

PCs: principal components

VCF: variant call format

GL: genotype likelihood

3.1. List of population abbreviations

AFR: African superpopulation

CWC: Corded Ware Culture

CDX: Dai Chinese

CEU: Utah residents with Northern and Western European ancestry

CHB: Han Chinese in Beijing

CHS: Southern Han Chinese

EAS: East Asian superpopulation

EUR: European superpopulation

FIN: Finnish population

GBR: British population

HAN: Han population

IBS: Iberian population

JPT: Japanese population

KHV: Kihn Vietnamese

TSI: Toscani population

YRI: Yoruba population

4. Introduction

Kinship analysis is a crucial tool in both archaeogenetics and forensic sciences for determining relationships between individuals based on genetic information. In forensic science, kinship analysis aids in criminal investigations, associating individuals with objects or locations, and identifying missing persons ¹. Utilizing single nucleotide polymorphism (SNP) markers and high-throughput sequencing techniques, kinship panels have been developed to confirm genetic genealogy leads and conduct general forensic applications, emphasizing the significance of kinship analysis in forensic genetics ². Techniques such as pairwise kinship analysis based on chromosomal sharing using high-density SNPs have been developed to enhance kinship analysis in forensic genetics, highlighting the importance of genetic markers in determining relationships ^{3,4}.

In archaeogenetics, kinship analysis is essential for understanding ancient populations, migration patterns, and genetic relationships. The use of Y-chromosomal markers and microchip panels has proven valuable in kinship analysis for ancient remains, shedding light on the presence or absence of kinship among individuals from the past ⁵. Additionally, the development of new genetic panels and SNP markers has enabled researchers to explore genetic diversity, haplotypic structure of ancient populations, aiding in biogeographical ancestry inference and kinship testing ⁶.

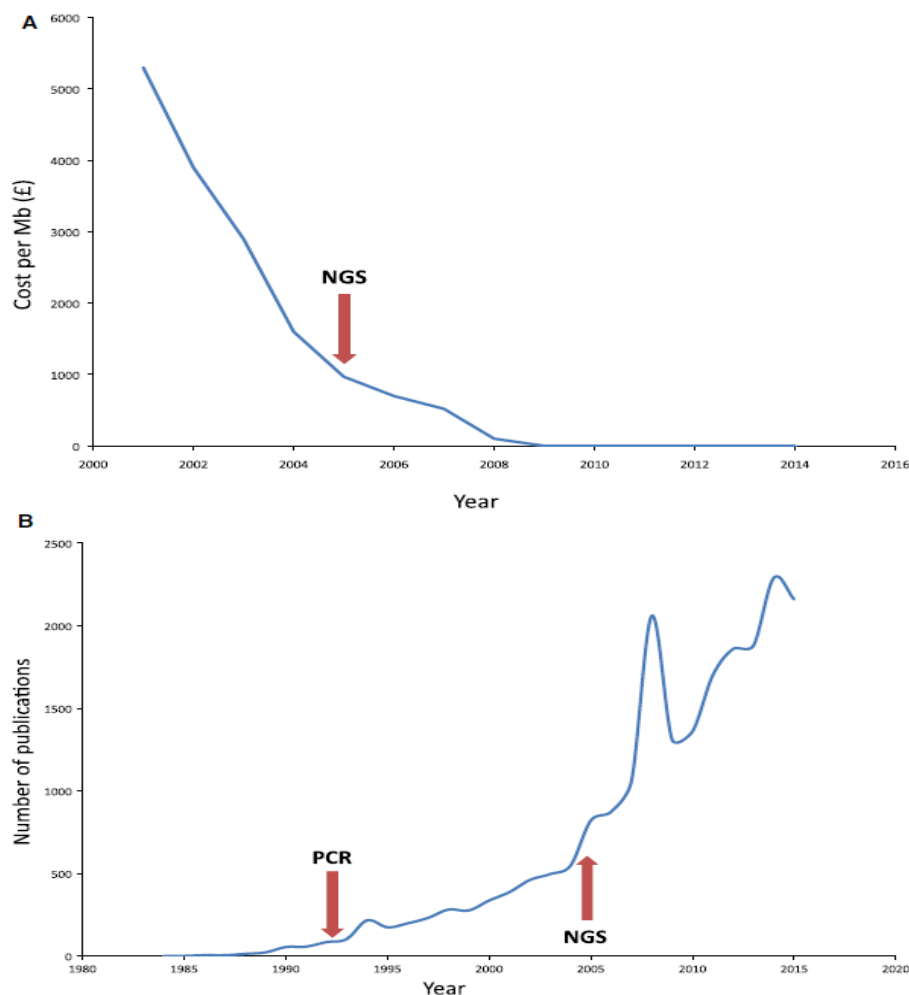
Beside the understanding the ancient population, kinship analysis is highly recommended during population genetic analysis, but it is also a precondition for most population genetic analyses to exclude close relatives from datasets (e.g., ADMIXTURE, principal component analysis (PCA)).

4.1. Ancient DNA

Ancient DNA (aDNA) and archaeogenetics have significantly advanced our understanding of human history and evolution. By combining genome-wide data, isotopic evidence, and archaeological and anthropological data, researchers have been able to shed light on the complexity of social status, inheritance rules, and mobility over ages ⁷. The extraction, sequencing, and analysis of aDNA fragments have transformed our concepts of the history of human and animal species ⁸. Furthermore, archaeogenetics has been applied to many problems on a regional scale, providing information on the genetic origins of domestic

animals and their wild progenitors and congeners, sheds new light on the genetic origins of domesticates and the domestication process itself ⁹.

The start of aDNA studies started in 1984 when Higuchi et al. ¹⁰ successfully extracted DNA from a museum specimen of a Quagga (*Equus quagga quagga*). DNA derived from remains dating 150 years was shown to have adequate size and quality, facilitating the determination of phylogenetic relationships within this species. In 1985 Svante Pääbo (who got his Nobel Prize for archaeogenetic research) extracted DNA from an Egyptian human mummy, which is the first human aDNA sample. ¹¹ During the 1990s, the methodological improvements focused mainly on DNA extraction protocols, rather than sequencing. After the introduction of NGS sequencing, the cost of sequencing decreased and the number of publications on aDNA increased significantly, as shown in Figure 1.



1. Figure A. Sequencing cost per Mb. B number of publications about aDNA. Figure from Linderholm,2016 ¹²

Post-mortem damage, endogenous contamination, and exogenous contamination are the main difficulties in archaeogenetics, in addition to the low coverage of the sequences, that researchers must address to ensure the authenticity and reproducibility of their findings.

Post-mortem damage refers to the degradation of DNA over time after the death of an organism. This degradation can lead to fragmentation and chemical modifications of DNA molecules, making it difficult to obtain intact genetic material from ancient samples⁹. Factors such as environmental conditions, temperature, and humidity play an important role in DNA preservation, and tropical regions pose a particular challenge due to accelerated DNA decay¹³. In addition, chemical processes can cause DNA damage, including cytosine deamination, further complicating the analysis of ancient DNA¹⁴.

Exogenous contamination involves the presence of DNA from microorganisms or other sources within the archaeological sample itself. This contamination can arise from bacteria, fungi, or other organisms that have infiltrated the postmortem sample, leading to the mixing of exogenous DNA with the ancient genetic material of interest⁹. The presence of exogeneous contaminants can skew genetic results and compromise the accuracy of population studies based on ancient DNA analysis.

Endogenous contamination, on the other hand, occurs when modern DNA from researchers, laboratory equipment or the surrounding environment contaminates the ancient sample during excavation, handling, or analysis. Even trace amounts of modern DNA can significantly impact the results of genetic studies, leading to false interpretations of population histories and genetic relationships⁹. To mitigate endogenous contamination, strict protocols for sample collection, processing, and analysis are essential in archaeogenetic research. Researchers in the field of archaeogenetics must employ rigorous methods to address these challenges and ensure the reliability of their findings.

Techniques such as ancient DNA extraction, library preparation, and bioinformatic analysis have been developed to minimise the impact of postmortem damage and contamination on genetic data obtained from archaeological samples¹⁵. By implementing stringent quality control measures and utilising advanced sequencing technologies, researchers can overcome the difficulties posed by postmortem damage, endogenous contamination, and exogenous contamination in archaeogenetics to reconstruct accurate genetic profiles of ancient populations and unravel their evolutionary histories.

4.2. Kinship analysis

Genetic kinship analysis is a method used to determine the degree of relatedness between individuals based on genetic information. This analysis is essential in various fields, such as population genetics, bioarchaeology, forensic genetics, and epidemiology. Traditional methods of kinship analysis have evolved with advancements in technology, particularly with the advent of next-generation sequencing (NGS) techniques.

Kinship analysis in archaeogenetics can provide important data for a better understanding of familial relationships, social structures, and migration patterns of ancient populations. By analysing ancient DNA samples from archaeological sites, researchers can infer kinship relationships between individuals buried together or within the same community¹⁶. Furthermore, kinship analysis in archaeogenetics has been instrumental in studying the genetic relationship and population structure of ancient civilisations.

Identity-by-state (IBS) is a concept in genetic analysis that focuses on the sharing of genetic markers between individuals, regardless of whether the markers are inherited from a common ancestor. IBS sharing is based on genetic similarity due to population-relatedness rather than familial relatedness. It is a fundamental aspect of genetic analysis, particularly in kinship studies, forensic genetics, and population genetics. Distinguishing between IBS and identity-by-descent (IBD) sharing is crucial in accurately estimating kinship coefficients and determining the degree of relatedness between individuals.¹⁷

The calculation of IBS sharing involves examining genetic markers, such as single nucleotide polymorphisms (SNPs), to identify regions where individuals share identical alleles. The kinship coefficient based on IBS sharing can be mathematically expressed as follows:

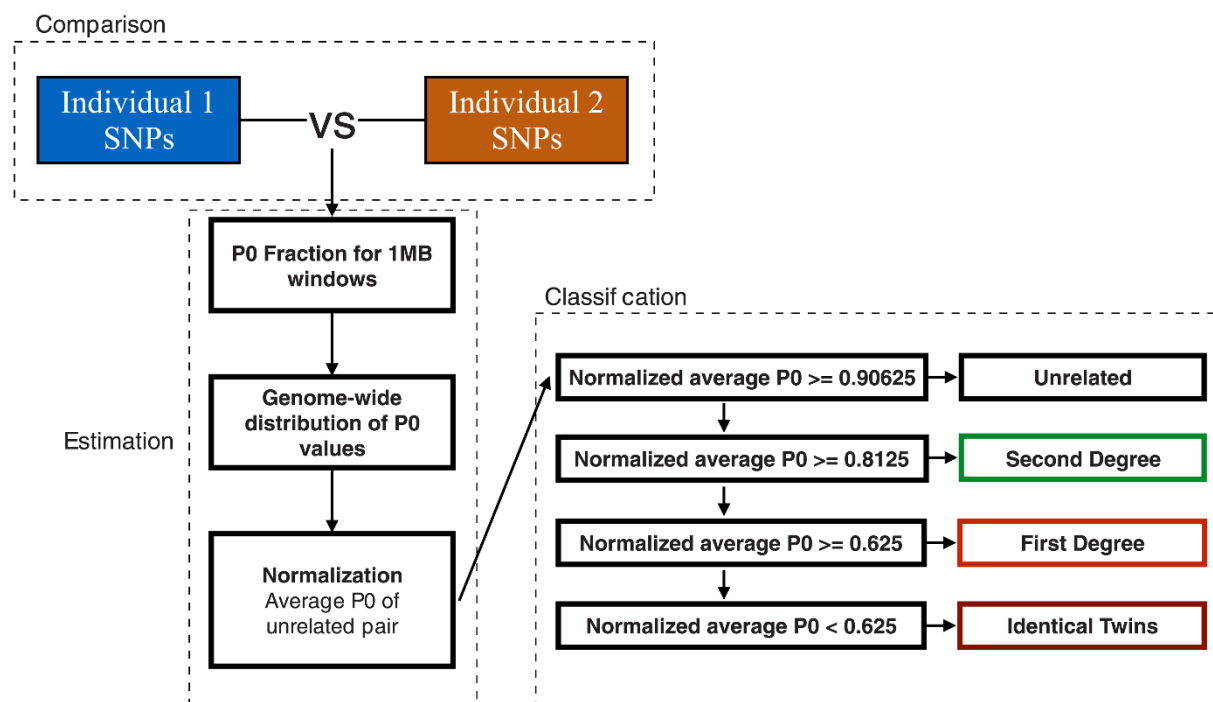
$$\phi = F1 + 0.5F2$$

Where F1 represents the probability that two alleles at a specific locus are IBD, and F2 represents the probability that two alleles at a specific locus are IBS due to a more distant common ancestor. This equation allows for the quantification of the genetic relatedness between individuals by considering both recent and more distant shared ancestry.^{18,19}

In archaeogenetics there are multiple difficulties to distinguish. The post-mortem damage and fragmented sequences are the main problems.²⁰ Moreover, the lack of complete pedigrees and genealogical records in ancient populations poses a significant obstacle in determining IBD accurately. Unlike modern populations where detailed family histories are available, ancient

populations may lack comprehensive genealogical information, making it challenging to infer true IBD segments and relationships between individuals.²¹ Another problem could be that the sample size of the experiments are low, because of the price of the highly degraded DNA sequences.^{22,23}

Multiple methods are used in the field of archaeogenetics to determine the kinship relationship. One of the most widely used is READ (Relations Estimated from Ancient DNA)²⁴, which is originally made for archaic samples, and can work from as low as 0.1x coverage, the algorithm can detect up to 2nd degree relationships. READ takes as input TPED/TFAM files containing pseudo-haploid genotypes of a population. READ divides the genome into nonoverlapping 1MB windows and calculates the proportion of nonmatching alleles (P0) for each pair of individuals within each window. Before classifying the relationship between pairs, P0 is normalised using expected values derived from unrelated individuals within the population or between populations with similar diversity. READ classifies pairs of individuals as unrelated, second-degree relatives, first-degree relatives, or identical individuals based on the normalised proportion of shared alleles. The tool produces the degree of relationship that is best suited, based on the average P0 point estimates, with low false positive rates observed. It provides users with graphical summaries of classification results, including uncertainties and certainty expressed as multiples of the standard error of the mean (Z).²⁴ Figure 2 shows the workflow of the software.



2. Figure Workflow of READ software²⁴

The other kinship estimation method commonly used for low coverage aDNA is lcMLkin. Similar to other recent methods aiming to infer population genetic parameters from low coverage data by utilizing genotype likelihoods instead of assuming a single best genotype lcMLkin, can infer biological relatedness down to 3rd degree relatives from simulated data, even when coverage is as low as 2x in both individuals examined.²⁵ lcMLkin starts the calculation from variant call format (VCF), where the samples have genotype likelihood (GL) fields. From this field it calculates IBD/IBS probability described in the publication. Lastly they define K as a 3-tuple of coefficient (k_0, k_1, k_2) and calculates the kinship coefficient.

5. Objective

PC-Relate algorithm accounts for population structure (ancestry) among sample individuals through the use of principal components (PCs) to estimate the kinship coefficient. Consequently, while the reference data has to be present in the analysed data set, the best reference does not have to be explicitly provided for each potential related individuals. The original algorithm has very good accuracy using fully genotyped modern data, however it is not suitable for the analysis of partially genotyped haploid data from ancient DNA sequences. During my research my task was to test our hypothesis that the raw kinship coefficient estimation of PC-Relate algorithm from pseudo-haploid partial marker data can be corrected based on the fraction of the genotyped markers in the samples. We used the PCAngsd version 0.99 implementation of the algorithm. During my research I participated in the following tasks.

1. In the initial examination, we sought to understand whether the commonly employed random pseudo-haploidization significantly influences the outcomes of Principal Component Analysis (PCA). Subsequent investigations involved simulations to assess the impact of random pseudo-haploidization (RPsH) on calculations of kinship degrees.
2. Archaic samples have a huge variability of genome coverage due to the generally low (variable) endogenous DNA content. In the second study we utilised simulations by downscaling known genome coverage data to investigate how the partially genotyped marker fraction affects the estimated kinship coefficient.
3. In the third investigation, the aim was to simulate the random errors and PMD error's effect on the corrected kinship coefficient.
4. Our aim was to test, whether the lack of exact reference population (often the case in case of ancient data) hugely invalidate our results. Furthermore, we wanted to test that in admixed relatives the un-admixed source populations can be used as reference.
5. Validation of the improved algorithm compared to READ. Validation of the algorithm on known relatives of the AADR data set.
6. The archaeogenomic validation of the Saint Ladislaus' relic

6. Materials and Methods

6.1. Used Data

In all analyses conducted, the genome coordinates of the 1240K SNP set from Allen Ancient DNA Resource (AADR) ²⁶ were used. For marker overlap simulations, two distinct fully-typed modern datasets were employed: the 1000 Genomes Project Phase 3 data ²⁷, and a sizable admixed Cabo Verdean-Hungarian family of known pedigree with first- (siblings), second- (half siblings), and fifth-degree relatives sourced from an anonymised clinical biobank. Variants within joint VCF (Variant Call Format) files underwent filtration based on 1240K SNP coordinates and were subsequently imported into plink 1.9 binary format ²⁸.

To examine the impact of genome coverage on the estimated kinship coefficients derived from genuine ancient data, unpublished 1240K genotype data of a documented medieval parent-offspring pair were used. This dataset comprises low-coverage, partially typed pseudo-haploid genotype data derived from 3+2 distinct library preparations sourced from various biological samples of the individuals under analysis. The unpublished medieval datasets have been deposited in the PLINK 1240K binary format, as detailed in this manuscript, and are available via Zenodo ²⁹.

The public AADR V42.4 1240K dataset ²⁶ served as validation of methodology in a diverse range of ancient individuals. Specifically, we included ancient samples with more than 100K genotyped markers (N=2810), while excluding those before 8000BC (N=216) due to their scarcity and inadequate representation as a reference population (³⁰ Additional file 7: Figure S4). Furthermore, to prevent the analysis of individuals lacking sufficient or appropriate reference populations, samples were restricted based on their geographical coordinates (Longitude -12 – 120 and latitude 28 – 65), resulting in the exclusion of 458 individuals (³⁰ Additional file 7: Figure S5). Following filtration, the dataset comprised 2136 ancient individuals (³⁰ Additional file 6: Table S5).

6.2. New bioinformatics tools

To facilitate seamless importation, manipulation, and analysis of genotype data within our proposed workflow, we developed essential tools, including the following:

- importHaploCall: Designed to import pseudo-haploid genotype calls from ANGSD.
- pseudoHaplo: to perform RPsH using a diploid dataset
- markerOverlap: Calculating the pairwise marker overlap fraction matrix.

- filterRelates: Correcting the kinship coefficient and filtering relatives based on error models and/or hard kinship coefficient thresholds.

For controlled studies on the impact of partially genotyped markers and comparison with analyses of fully genotyped modern samples, we utilised:

- depleteMarkers: Simulating the desired marker overlap fraction between selected samples.
- depleteIndivs: Simulating a random cohort of partially genotyped samples.

These tools are compatible with the major genotype data formats (PLINK, EIGENSTRAT, PACKEDANCESTRYMAP). Detailed usage instructions and command examples are provided in the publication ³⁰ Additional file 8: Note S1. The tools are accessible via Zenodo and the GitHub repository (<https://github.com/zmaroti/correctKin>).

To calculate the pairwise overlapping marker fraction between samples, the method defines it as the number of markers typed in both samples divided by the total number of markers in the dataset. Utilizing the "pseudoHaplo" tool, 100 randomly pseudo-haploidized datasets were generated from the fully typed modern diploid dataset using different random seeds. Throughout the presented examples, the "markerOverlap" tool was employed to compute the pairwise overlapping marker fraction matrix of samples utilized for kinship coefficient correction.

6.3. Principal component analysis

The British(GBR), Toscani (TSI), Iberian (IBS) and Finnish (FIN) populations were selected from the 1000 Genomes Project (1KG) dataset, which totalled 404 samples. Subsequently, the diploid dataset was randomized using three different seeds. Smartpca ^{31,32} analysis was performed on both the original diploid dataset and the three randomly pseudohaploidized datasets, incorporating the "inbreed: YES" option. Visualisation of individuals on the PC1 and PC2 axes was performed using R (version 4.0.5) ³³ and the ggplot2 R package (version 3.3.5) ³⁴.

6.4. Simulating the effect of low coverage from fully typed modern datasets

To systematically examine the impact of coverage and the resultant lower genotyping percentages on kinship coefficient calculations, we employed "depleteMarkers" to randomly deplete markers from a fully typed dataset (in PLINK or EIGENSTRAT format). This process

allowed us to achieve the desired percentage of marker overlap between two samples. Using this tool, overlapping marker fractions were simulated within selected samples in a range of 5 to 100%, with increments of 5 percent.

To evaluate the technical errors associated with low or variable coverage data throughout the workflow, a set of 1020 fully typed diploid Eurasian samples from diverse populations were curated from the 1000 Genomes Project phase 3 dataset. Utilizing "depleteIndivs," a random, partially typed sample cohort was generated, with marker counts ranging between 100,000 and the complete set of 1,150,639 markers. Populations were the followings: IBS, GBR, FIN, TSI, Utah residents with Northern and Western European Ancestry (CEU), Dai Chinese (CDX), Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), Japanese (JPT) and Khin Vietnamese (KHV)

Subsequently, from the partially typed diploid dataset, a pseudo-haploidized dataset was generated using the "pseudoHaplo" tool. Kinship analysis was performed using PCAngsd, and estimated kinship coefficients were corrected based on the marker overlap fraction of sample pairs within the partially genotyped datasets. These results were compared with those obtained from the original fully typed diploid dataset for analysis.

6.5. Simulation of aDNA-related genotyping errors

The PLINK and EIGENSTRAT data formats were designed for biallelic markers. There are only four possible allelic states (homozygote major allele, homozygote minor allele, heterozygote major / minor and missing), thus any other nucleotide that is different from the minor or major allele cannot be represented, and the allelic state of samples with invalid alleles is set to the "missing" state at such marker positions.

Based on the data format restriction, the three typical aDNA-related genotype errors can be simulated in the following ways for the pseudo-haploid PLINK dataset:

- Post mortem damage; if the conversion of C->T or G->A conversion leads to nucleotides different from the minor or major allele, the state is set to 'missing'; otherwise, if the minor and major alleles are C/T, T/C, G/A or A/G, the minor and major homozygote states are flipped.
- Exogenous (non-human DNA) contamination; since the exogenous DNA consists mainly of DNA from microorganisms (usually in ancestral state), it leads to excessive

homozygote major allele, thus random subsets of markers are set to the homozygote major allele state.

- Endogenous (human DNA) contamination; Random subsets of markers are set to the state of the same markers genotyped from another sample (theoretically, the largest number of SNPs is expected to be flipped in case the population has the largest FST from the test individuals or practically if the test is contaminated with sample from a very different population).

In most population genetic analyses, highly contaminated samples are excluded. In the comprehensive AADR ancient dataset, the following criteria are used to mark bad quality sequences:

- ANGSD X contamination (applicable only to males) 0.02–0.05="QUESTIONABLE", >0.05="QUESTIONABLE_CRITICAL" or "FAIL".
- mtcontam <0.8 is "QUESTIONABLE_CRITICAL", 0.8-0.95 is "QUESTIONABLE", and 0.95–0.98 is recorded, but "PASS", gets overridden by ANGSD X contamination.

Consequently, the 1240K v42.2 AADR dataset (n=3589 ancient samples) 157 is marked CRITICAL/FAIL (>5% error rate), while the mean contamination rate X of all ancient samples is 1.28%.

The simulation comprised three distinct error types examined individually, together with a mixed scenario where all three error types were introduced equally, resulting in a consistent total error rate. Throughout, a maximum total genotyping error rate of 5% was maintained, adhering to the CRITICAL / FAIL tag threshold as per the AADR criteria. Consequently, each sample was subjected to random genotyping errors ranging from 0% to 5%, leading to an overall genotype error rate of approximately 2.5% across the dataset. This rate represents approximately twice the genotyping error rate observed in the experimental AADR ancient DNA dataset. Across each simulation, 100 iterations were performed using different random seeds, with the mean and standard deviation of the corrected kinship coefficients calculated accordingly.

6.6. Uncorrected kinship coefficient estimation

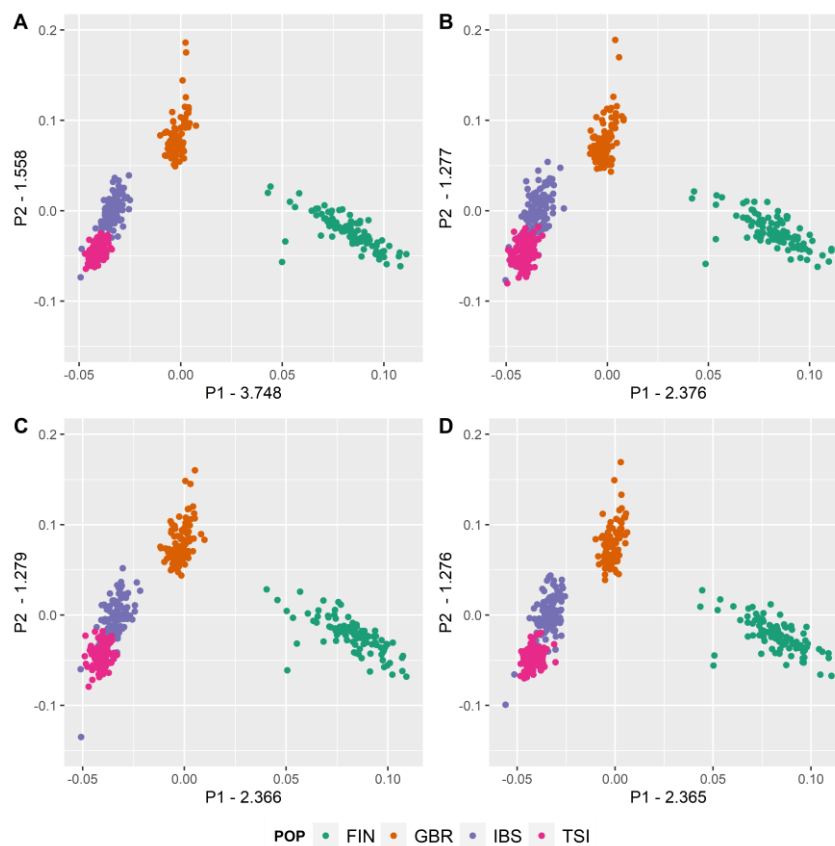
The estimation of the kinship coefficient was performed by the PCAnsd³⁵ software (version 0.99) of the ANGSD package³⁶ that implements a fast parallelised kinship calculation from the PLINK or EIGENSTRAT format based on the PC-Relate algorithm³⁷ with the parameters '-inbreed 1 -kinship'.

7. Results

7.1. Effect of random pseudo-haploidization on PCA

In our first experiment, we wanted to know that random pseudo-haploidization (RPsH) influences the result of the principal component analysis (PCA). RPsH is commonly used in the field of archaeogenetics, because the genomic coverage of the samples is usually low, between 0,1x-5x, so reliable diploid call is not possible. With RPsH, the heterogeneous call is randomly changed to homozygote reference or homozygote alternative. Although it is a method used every day in the field, no articles have tested whether it has any effect on PCA.

To test the effect, 404 samples of 4 European populations from the 1KG database were selected: British (GBR), Toscani (TSI), Iberian (IBS), and Finnish (FIN). The samples were randomised with three different seeds. After randomisation PCA analysis was performed with the smartpca software, with the original diploid and the 3 random pseudo-haploidized datasets.



3. Figure Result of Principal component analysis of diploid (A) and three random pseudo-haploidized dataset³⁰

As seen in Figure 3. the RPsH does not significantly alter the result of the PCA.

7.2. Effect of random pseudo-haploidization on kinship coefficient

To examine the effect of RPsH on the kinship coefficient, 509 individuals were selected from five different populations of 1KG. The populations were the following: FIN, GBR, TSI, and CEU for European, and Han Chinese (HAN) for East Asian.

In the experiment, 100 different pseudo-haploid datasets were generated from the original dataset using different random seeds. To exclusively study the effect of RPsH, sample duplicates were included from different sets of random seeds. Random individuals were selected from the GBR (HG00244.SG), FIN (HG00356.SG), CEU (NA12763.SG, NA12775.SG) and TSI (NA20798.SG) populations. This idealised experimental setup is equivalent to the monozygotic twin relation, while the maximal expected kinship coefficient (0.5) allows for a more sensitive analysis.

To study the effect of RPsH on true first / second order relatives, samples with known family relations from 1KG were selected. (I) HG00702-HG00657, a parent-child relation from Han population, where exactly 50% of genome is shared between the samples. (II) NA20526-NA20792 siblings from Toscani populations, where 50% of the genome comes from the same parent, but a different subset of markers was found due to recombination and segregation. (III) Second-order relatives HG00124-HG00119 of the British population, where statistically 25% of the genomes are shared. The kinship coefficients were calculated for each of the selected relatives (kin1/kin2) using their own reference population for the 100 different randomisations and calculated the mean and standard deviation of the estimated kinship coefficients.

ID1	ID2	POP	relatedness	expected kinship coeff.	diploid			haploid	
					estimated kinship coeff.	95% conf. interval lower	95% conf. interval upper	mean	SD
HG00244.SG	HG00244_DUP.SG	GBR	sample match	0,5	0,496	0,492	0,501	0,496	0,001
HG00356.SG	HG00356_DUP.SG	FIN	sample match	0,5	0,493	0,488	0,497	0,488	0,001
NA12763.SG	NA12763_DUP.SG	CEU	sample match	0,5	0,499	0,495	0,504	0,499	0,001
NA12775.SG	NA12775_DUP.SG	CEU	sample match	0,5	0,493	0,488	0,497	0,492	0,001
NA20798.SG	NA20798_DUP.SG	TSI	sample match	0,5	0,495	0,490	0,499	0,492	0,001
NA20526.SG	NA20792.SG	TSI	1 st	0,25	0,243	0,239	0,248	0,241	0,001
HG00657.SG	HG00702.SG	CHS	1 st	0,25	0,240	0,236	0,245	0,237	0,001
HG00119.SG	HG00124.SG	GBR	2 nd	0,125	0,159	0,155	0,164	0,158	0,001

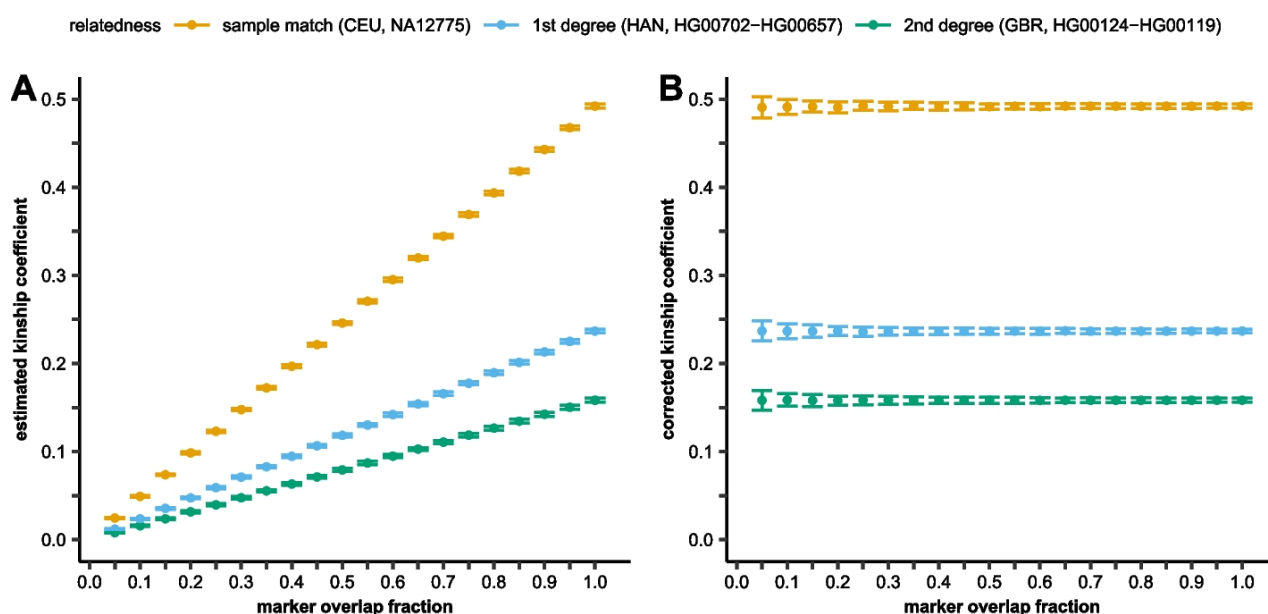
1. Table Effect of RPsH on kinship coefficient. Five artificial identical twin and 3 true relative from 1KG³⁰

The Table 1 shows that RPsH does not significantly alter the result of the kinship coefficient calculations.

7.3. The effect of overlapping marker fraction on the kinship coefficient calculation

The original PC-Relate algorithm was created to analyse modern fully genotyped diploid samples. However, in the case of ancient data, genome coverage and partial genotyping are one of the factors that have the greatest variability between samples. The calculation of the kinship coefficient is based on the identity-by-descent (IBD) segments shared between two samples, which can only be assessed at marker positions where both samples are genotyped. To investigate the effect of this factor, a metric called the overlap marker fraction was defined. This was calculated metric by dividing the number of markers where both samples are genotyped by the total number of markers in the data set. Using the 100 random pseudo-haploidized fully typed dataset of the previous experiment, markers were randomly depleted between the selected sample dups and true 1KG relatives to a marker overlap fraction between 5 and 100% using different random seeds. The results revealed that the kinship

coefficient is a linear function of the marker overlap fraction, as shown in Figure 4/A. This allows for a simple method to correct the value of the kinship coefficient for low-coverage genomes by dividing the estimated kinship coefficient with the marker overlap fraction between the two samples. According to the simulation, the correction of the estimated kinship coefficient for sparsely genotyped data resulted in a reproducible kinship estimation regardless of the overlap fraction, as shown in Figure 4/B. As expected, a low partially overlapping subset of markers would lead to less complete representation of the reference population and the test individuals, thus regression of IBD/IBS based on the PC-Relate algorithm would have higher SD at low marker counts. Although the marker overlap fraction correction differs more than one magnitude between very low and high marker overlap fraction sample pairs, the analysis shows (³⁰, supplementary table 1) that the correction itself does not introduce overall bias (the mean is statistically the same) and does not significantly multiply the error rate ($4\times$ increase in SD at $20\times$ correction factor). The results suggest that the increase in SD of the method is likely due to the higher uncertainty of PCA.

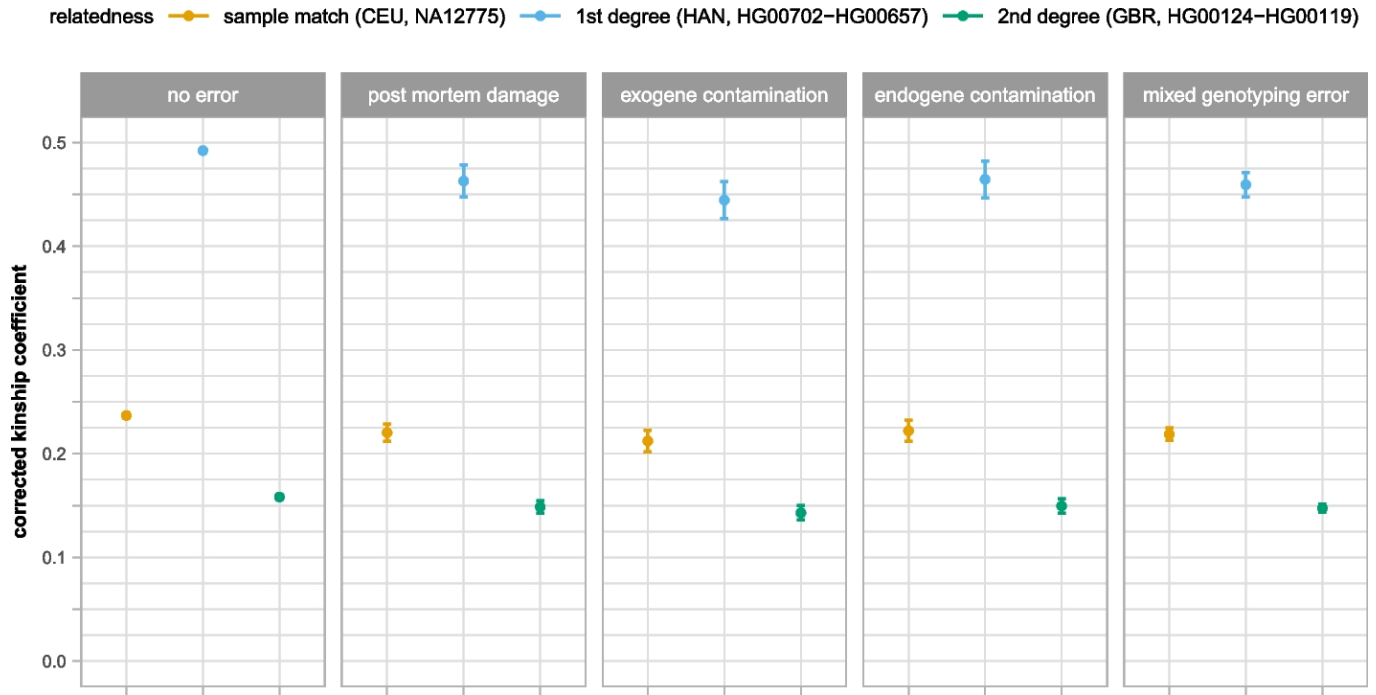


4. Figure The mean and the 95% confidence interval of the A uncorrected and B corrected kinship coefficient

7.4. The effect of genotyping errors on the corrected kinship coefficient

The effect of genotyping errors was tested using the same data sets and sample dups/known first- and second-degree 1KG relatives as in previous experiments. The simulation setup has 4 scenarios: (1) only post mortem damage, (2) only endogenous contamination (using a random Yoruba (YRI) individual as the contaminant), (3) only exogenous contamination, and (4)

equal combination of the first three sources of genotype errors. In each scenario, approximately twice as many genotyping errors per individual were introduced compared to a typical aDNA dataset. The mean and standard deviation of the corrected kinship coefficients were calculated (³⁰ Additional file 3: Table S2) between the selected samples. The mean and 95% confidence interval of the corrected kinship coefficients of the selected sample dup (CEU; NA12775.SG), true first-degree relationship (HAN; HG00702-HG00657), and true second-degree relationship (GBR; HG00119.SG-HG00124.SG) were visualized in Figure 5. Generally, genotyping errors decreased the mean and increased the standard deviation of the estimated kinship coefficient. The decrease in the corrected coefficients were proportional to the expected kinship coefficients. The greatest effect (approximately 9.6% lower kinship coefficient) was observed in the case of exogenous contamination. It is speculated that this effect was likely due to the fact that, in this scenario (although the genotyping error affects a different subset of markers in different samples), the markers' states were uniformly set to the homozygote major state, leading to higher bias than that of random flips of the minor/major state in different samples in the other scenarios. In our simulations, even with relatively high error rates (compared to experimental aDNA error rates), the largest effect was still significantly smaller than the 50% difference of expected kinship coefficients between different degrees of relations, thus the estimated degree of relatedness for the analysed relatives remained the same.



5. Figure Corrected kinship coefficients with different genotyping errors

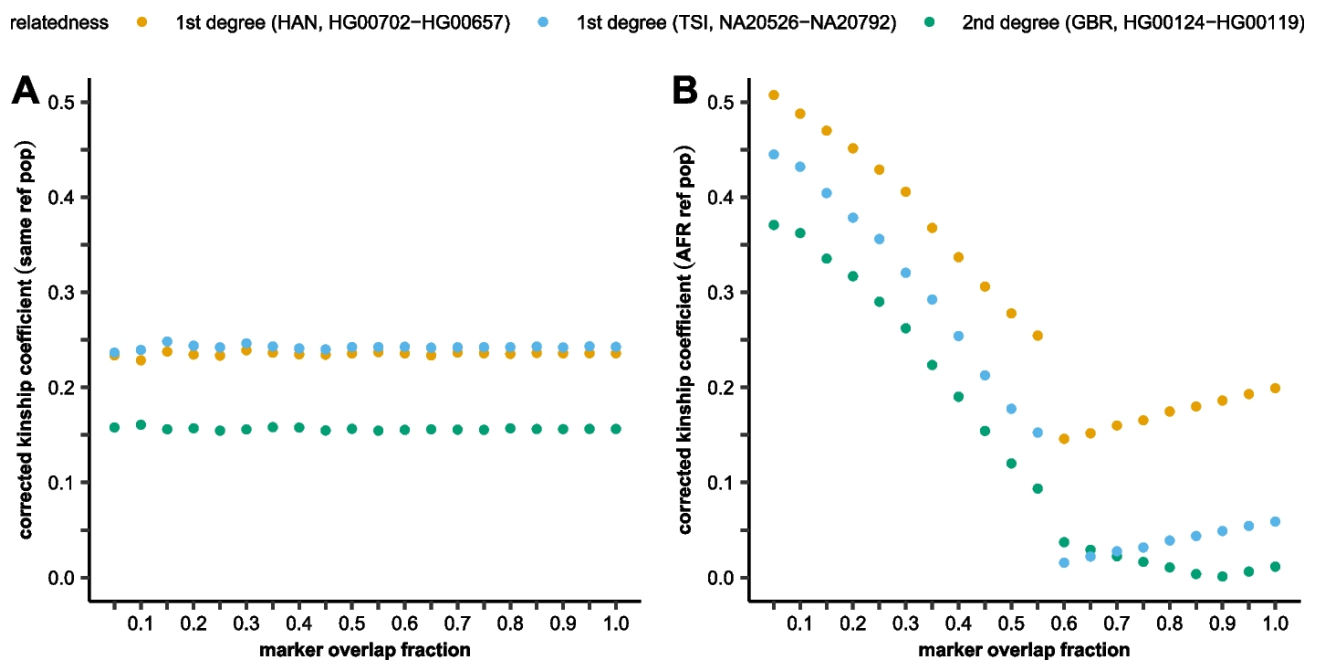
7.5. The effect of reference population selection on kinship analysis

In this analysis, the focus is on investigating scenarios where a proper reference population is unknown or unavailable, a situation frequently encountered with ancient samples. Using the same public 1KG phase 3 data set and the three known relatives HG00702-HG00657 parent-child of the Han population, NA20526-NA20792 siblings of the TSI population, and HG00124-HG00119 second-order relatives of the GBR population, the effect of the reference population on the calculated kinship coefficients were investigated. Three different scenarios were tested: (1) the reference population was the same as that to which the selected individual belonged to; (2) the reference population was from a different superpopulation (AFR); (3) the reference population was from the same superpopulation as the selected individual (JPT for Han, IBS for TSI, FIN for GBR). To study the effect of overlapping marker fractions in these more complex cases, the selected sample pairs were also marker-depleted in the range of 100 to 5% overlap fractions.

The results revealed that a reference population with significantly different genetic backgrounds, such as African for European samples, strongly corrupts the results. In this experimental setup, there is a deficiency in the availability of a sufficient number of unrelated references; hence, IBS fractions are likely not represented and cannot be properly regressed

out. Furthermore, when only a couple non-AFR individuals are included in the analysis depending on the marker overlap fraction, the EUR/EAS-specific markers are mostly excluded as the PCAngsd implementation uses a default 0.05 MAF marker pruning. Thus, at low marker overlap, mainly the AFR-specific markers are kept, while at higher marker overlap slightly more EUR/EAS-specific markers are also included in the analysis. Based on the used marker set, the optimal number of eigenvectors and the underlying PCA based regression of IBS components are expected to be different. The observed nonlinear kinship coefficient estimates in Figure 6 could likely be caused by these differences.

Using a superpopulation with similar genetic background to the sample gives very similar results as if its own reference population were used (³⁰ Additional file 1: Figure S2).

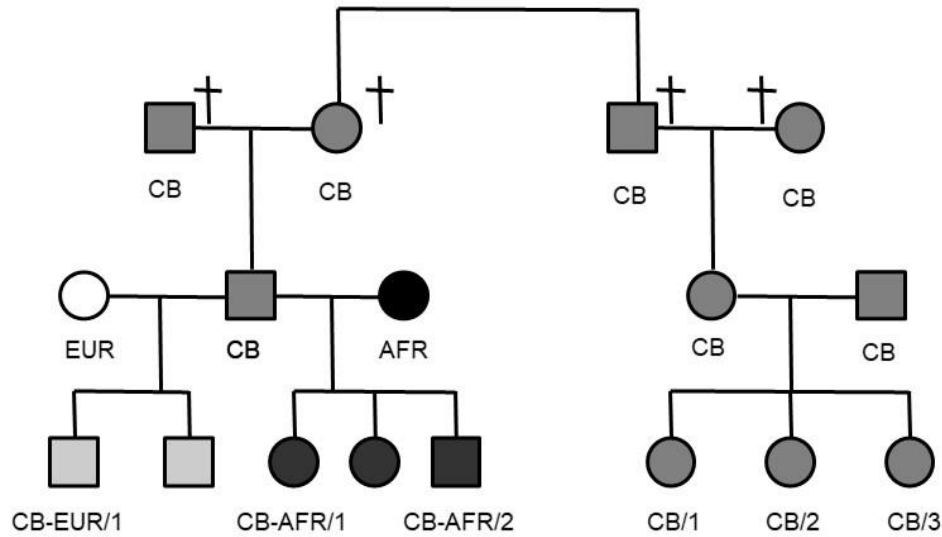


6. Figure A. Corrected kinship coefficients with the same reference population like the samples. B. Corrected kinship coefficients with African reference populations.

7.6. Effect of reference population selection on kinship analysis in a complex admixed family with multiple ethnic relations

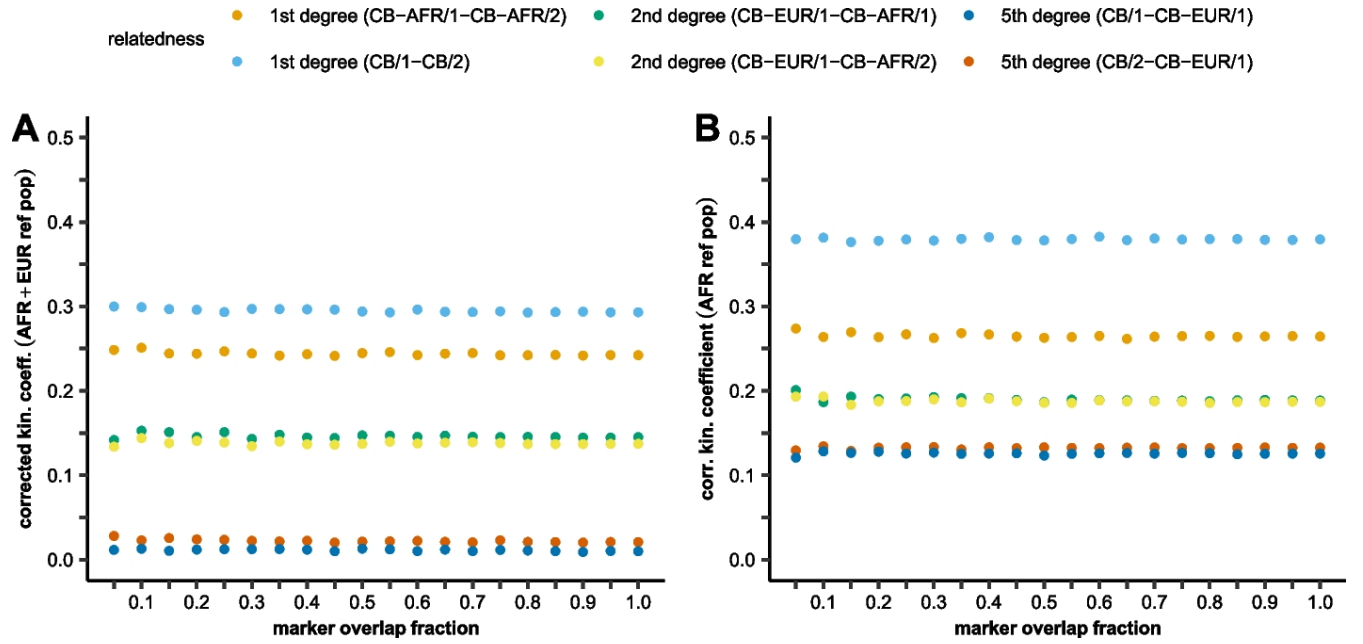
To assess the choice of reference population in the kinship analysis of admixed individuals (often the case in ancient populations), a complex admixed Cabo Verdean-Hungarian family with known pedigree was analysed. In this family, multiple instances of historical and recent admixtures have occurred, leading to individuals with varying ratios of admixture

components. WGS data were available for siblings (1st order), differently admixed half-sibs (2nd order), and fifth order relatives as shown in Figure 7.



7. Figure Pedigree of a complex admixed modern family with Cabo-Verdean (CB, ~50%-50% old EUR/AFR admix), AFR (100% African), EUR (100% Hungarian) family members and recently admixed offsprings. WGS data was only available from third generation individuals denoted with numbers in their ID.

Two scenarios were tested: the reference population was (1) only African (AFR) representing the majority of the admix sources in the tested samples; (2) African and European (EUR) populations were included. Additionally, marker depletion was performed to investigate the effect of coverage in this complex scenario. The results in Figure 8 demonstrate that to obtain realistic coefficient values in the case of a complex admixture, a combined set of reference populations is required that represents the population structure of all ancestors. Using just the majority source as a reference significantly distorts the result. To test whether random haploidization alters the calculation of the kinship coefficient compared to the better phased diploid data in this complex admixed case, this analysis was also performed from the original diploid dataset (³⁰ Additional file 4: Table S3) with the AFR + EUR reference population. It was confirmed again that even in such a complex admixed family, the differences due to RPsH were negligible.



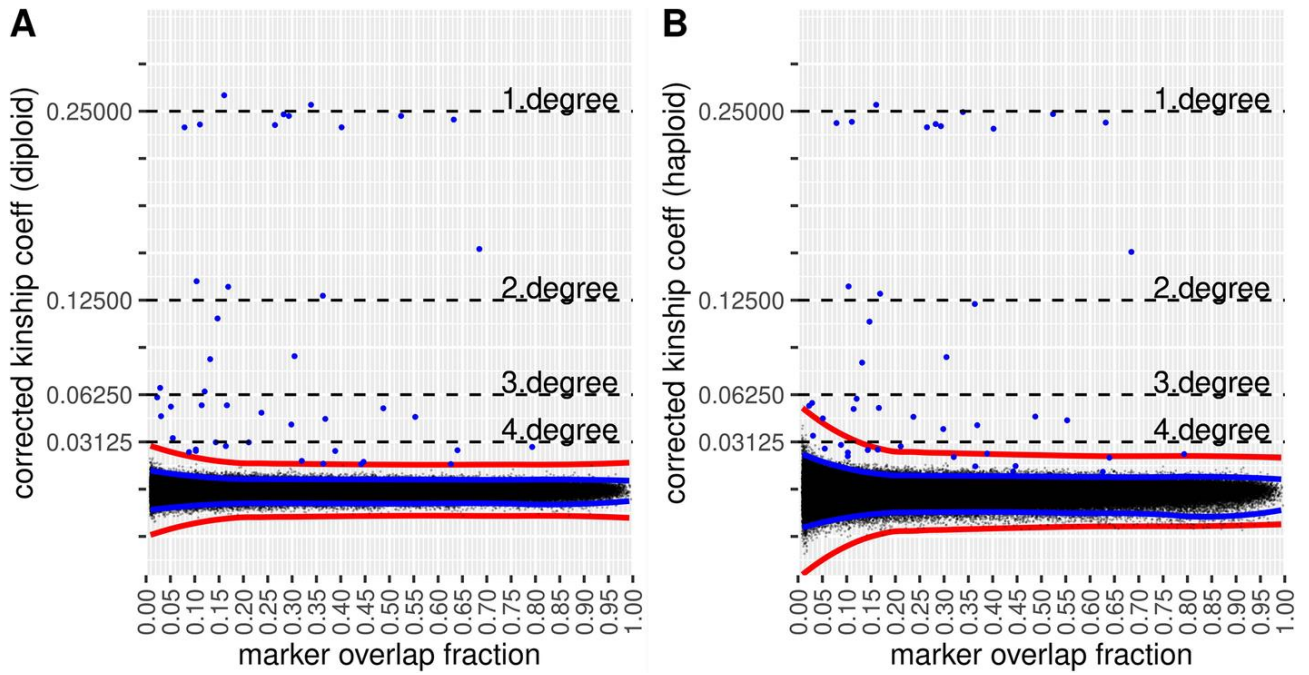
8. Figure A. Corrected kinship coefficients with African and European reference populations. B. Corrected kinship coefficients with only African populations.

7.7. Statistical validation, assessment of technical errors

The EUR and EAS individuals were selected from the 1KG phase 3 dataset ($n=1020$) and estimated the kinship coefficients between these individuals. Although there are a few true relatives in the selected individuals, the overwhelming majority of pairwise relations are expected to be unrelated, thus representing the variance in technical error of the entire analysis. To test the effect of the overlapping genotyping fraction on the mean and standard error of the corrected kinship coefficient, the set of markers in these individuals was randomly depleted between 100,000 markers and the fully typed marker count ($\sim 1.2M$), amounting to 100% of the marker count of the original data set. Using RPsH, a pseudo-haploid dataset was also created for comparison. The pairwise kinship coefficient matrix and the estimated kinship coefficients were created by the overlap fraction of the markers. Using the pairwise matrix of 1020 individuals, the 519,690 kinship coefficients were plotted between all combinations of individuals for the diploid and haploid dataset (Figure 9).

The variance of the corrected kinship coefficient depends on the marker overlap fraction between the test individuals (Figure 9). Since the marker overlap fractions between any two ancient samples are different, applying a predefined kinship coefficient threshold to identify relatives would lead to decreased sensitivity or specificity depending on the marker overlap fraction. In other words, the statistical power to differentiate relatives from unrelated depends

on the marker overlap fraction, and the same threshold should not be applied. However, based on the experimental variance of the corrected kinship coefficient observed in the analysed dataset, the Z score (N standard deviation from the mean) could be used as a criteria to differentiate relatives from unrelated with the same sensitivity and specificity independent of the marker overlap fraction. Given the inability to eliminate errors that arise from missing genome components in the reference populations during this experiment, a conservative N=6 sigma threshold was employed to detect biological differences. As expected, the haploid dataset resulted in higher variance due to random information loss especially at very low (<5%) marker overlap fractions. Although the marker overlap fraction correction differs two magnitudes (0.007–0.996) between very low and high marker overlap fraction sample pairs, the analysis shows that similar to the corrected kinship coefficient between relatives (³⁰ Additional file 2: Table S1) the correction itself does not significantly multiply the error rate of unrelated samples. Most technical errors are expected to be the result of using very sparse data to regress IBS to identify IBD fragments using PC-Relate. Despite the higher variance, the estimated kinship coefficients show a very high correlation between the fully typed (~1.2M high-quality diploid markers) diploid and the corrected coefficients of the partially typed diploid and haploid datasets (R=0.9998 and R=0.9993 respectively) compared to the correlation with the uncorrected estimates (R=0.749 and R=0.751; ³⁰ Additional file 5: Table S4). The results indicate that the applied pseudo-haploidization and correction in the marker-depleted experimental data do not introduce overall bias. The method successfully identified all known first- and second-degree relatives within the analysed subset of 1KG EUR/EAS individuals, and it also revealed a few additional distant third or fourth relatives (³⁰ Additional file 5: Table S4). The analysis shows that 4th-degree relatives are expected to surpass the 6 sigma threshold in diploid data, except in cases of very low overlap marker fractions (<2%, approximately 17,000 overlap markers). In the case of haploid data, the establishment of 3rd degree relatedness is possible even from low marker overlap fractions, and the establishment of fourth degree relatedness is possible when the sample pair has >10% marker overlap fraction (equal to roughly ~85,000 markers). However, in case of 4th degree relatives depending on the overlapping genotyping fraction, the estimated confidence interval of corrected kinship coefficient, and true biological variation, it is expected to have more false positive/negative and uncertain kinship estimations at low marker overlap fractions.



9. Figure Statistical validation of corrected kinship estimations. A. Diploid data. B. Haploid data

7.8. Kinship analysis of ancient samples with known relations using kinship coefficient correction

To demonstrate the suitability of the methodology for ancient data, low-coverage ancient sequences with known family relations were analysed. In the first example, the analysis focusses on a known father-son (first-degree) relation involving two Medieval samples. Both remains provided two types of biological samples: bone powder extracted from the teeth and from the pars petrosa. The father's samples included three parallel DNA isolates and NGS libraries, with two prepared from pars petrosa and an additional one from teeth. For the offspring, one DNA isolate and an NGS library were prepared from both types of biological samples. In total, there were 3+2 NGS sequences with significantly different genome coverages (ranging from $0.87\times$ to $11.9\times$) from these two ancient individuals.

Sample 1	Sample 2	Marker overlap fraction	Relation	Expected kinship coeff.	Uncorrected kinship coeff.	Corrected kinship coeff.
Father high	Child high	0.933894	First degree	0.25	0.22883	0.24502
Father high	Child low	0.377769	First degree	0.25	0.09179	0.24297
Father medium	Child high	0.730011	First degree	0.25	0.17604	0.24114
Father medium	Child low	0.296127	First degree	0.25	0.07150	0.24144
Father low	Child high	0.511753	First degree	0.25	0.12239	0.23915
Father low	Child low	0.207545	First degree	0.25	0.05014	0.24160
Father high	Father medium	0.778247	Sample matching	0.5	0.38284	0.49193
Father high	Father low	0.545866	Sample matching	0.5	0.27020	0.49499
Father medium	Father low	0.427389	Sample matching	0.5	0.20863	0.48814
Child high	Child low	0.354511	Sample matching	0.5	0.17286	0.48761
Sample	Sample type	Coverage	Typed marker count (1240k)			
Father high	Teeth	11.997	1,148,973			
Father mid	Petrosa	3.057	896,623			
Father low	Petrosa	1.546	630,405			
Child high	Petrosa	5.510	1,075,845			
Child low	Petrosa	0.879	435,005			

2. Table Result of the two, known related medieval person

Consequently, the robustness of the correction method was assessed in the six combinations of these data sets. Table 2 shows the uncorrected and corrected kinship coefficients calculated from these data. In the second example, a reanalysis was performed on a published group of five related males from Corded Ware Culture (2500–2050 BCE) with first-, second-, third-, and fourth-degree kinship relationships³⁸. Table 3 presents family relations with uncorrected and corrected kinship coefficients calculated by the methodology from ancient public 1240k data. These individuals were initially analysed in the original READ manuscript²⁴, where relations were identified to the second degree, except for the one between I1538 and I1540, and all third and fourth relations were inferred solely from family relations.

Sample 1	Sample 2	Marker overlap fraction	Relation	Expected kinship coeff.	Uncorrected kinship coeff.	Corrected kinship coeff.
I1538	I1541	0.039828	First degree	0.25	0.01036	0.26006
I1540	I1541	0.079863	First degree	0.25	0.01959	0.24534
I1534	I1541	0.047882	Second degree	0.125	0.00715	0.14938
I1538	I1534	0.025735	Second degree	0.125	0.00335	0.13003
I1538	I1540	0.042478	Second degree	0.125	0.00456	0.10730
I1541	I0104	0.216185	Second degree	0.125	0.03394	0.15700
I1534	I1540	0.049813	Third degree	0.0625	0.00405	0.08123
I1538	I0104	0.107634	Third degree	0.0625	0.00846	0.07864
I1540	I0104	0.22405	Third degree	0.0625	0.01812	0.08088
I1534	I0104	0.129456	Fourth degree	0.03125	0.00645	0.04982
Sample ID	Coverage	Typed marker count (1240k)				
I0104	4.184	962767				
I1534	0.158	164095				
I1538	0.126	135269				
I1540	0.298	285866				
I1541	0.294	276299				

3. Table Results of the Corded Ware Culture family relations

7.9. Kinship analysis of ancient samples from the AADR 1240K dataset

Kinship analysis was conducted on 2136 ancient Eurasian individuals from the AADR 1240K dataset, each having more than 100K genotyped markers. Without manual curation or the use of an additional matching reference population, potential relatives above a corrected kinship coefficient of ~ 0.046875 (indicative of 3rd–4th degree kinship) were filtered. The analysis revealed 410 related individuals organized into 184 kin groups (³⁰ Additional file 6: Table S5). Sample duplicates (N=26) and joint datasets (N=30) from the same sample were identified. Notably, sample duplicates with different master IDs were found in the AADR dataset published in different manuscripts (I1526-NEO232; I7782-NEO298; I8295-NEO230; I8296-NEO231), where all four sample pairs originated from the same geological site, belonged to the same population, and exhibited identical or nearly identical haploid typing, with

differences likely attributed to missing branch-defining markers due to coverage variations. Additionally, a probable sample mix was identified³⁹, involving two individuals (MJ-15 Ukraine_IA_Western- Scythian.SG and MJ-35 Ukraine_Cimmerians_o2.SG) with a corrected kinship coefficient of 0.5, equivalent to a sample match (or monozygotic twin), despite having different population assignments. All of these individuals had same sex and identical/nearly identical mitochondrial and Y haplogroups as well. Furthermore, all 111 previously identified kinship relations from the AADR dataset were identified. Three relatives (I8502, I8524; MK5001, MK5004; KBD001, KBD002) previously labeled as uncertain (1st or 2nd degree) in the AADR dataset were reclassified as 2nd-degree relatives by our analysis. Additionally, three kin pairs initially indicated as 1st-degree relatives were reclassified as 2nd-degree relatives (RISE1163, RISE1169; RISE1168, RISE1173; RISE1168, RISE1169). Beyond the published data, our approach within the 184 kin groups identified 6 new 1st-degree, 108 2nd-degree, 144 3rd-degree, and 40 4th-degree relations, involving a total of 279 new relatives (³⁰ Additional file 6: Table S5).

In instances where an appropriate reference population was absent in the dataset, hindering the establishment of suitable kinship relations, the correction resulted in invalid distant 3rd–4th-degree kinship relations highlighted in red in ³⁰ Additional file 6: Table S5. To assess the sensitivity of our analysis, READ ²⁴ was utilized for validation. As READ relies on the proper reference population and utilizes a global threshold to differentiate between unrelated and potential kin, the joint set of 2136 individuals could not be analysed together. The top 10 populations with the highest number of individuals were selected, and READ analysis was performed separately. READ identified relatives up to the 2nd degree, with no additional relatives identified compared to our methodology in the selected populations. The degree of kinship matched for each identified relative between the two methods. Notably, our method identified one additional 2nd-degree relation (AITI_95_d and AITI_98) that was missed by READ. In this case, the samples had very low genome coverage (0.145× and 0.402×) and only ~0.05% marker overlap fraction. Although the corrected kinship coefficient was significantly above the 3rd degree (0.0625), it was less (0.1004) than the expected 0.125 corresponding to 2nd degree, suggesting that these relatives share less than expected genome portions due to true biological variation (³⁰ Additional file 7: Table S6). A similar scenario was observed in the case of the missed 2nd-degree relation between the I1538 and I1540 CWC individuals (Table 2), indicating that READ is less sensitive when the marker overlap is low, and the shared genome fraction significantly differs from the statistically expected mean.

7.10. Genetic validation of St.Ladiuslaus

DNA extraction from the herma of St.Ladislaus was performed, specifically from the *pars petrosa* and tooth root, followed by analysis. Due to the careful sampling to not to damage the relic, the DNA extraction of the tooth was low on endogenous content, but subsequently, the tooth root, exhibiting a very high endogenous content (72.2%), underwent whole genome sequencing.⁴⁰ The whole genome sequencing resulted really high coverage, 16,4 fold in average.

Within the Matthias Church in Budapest, a total of three Árpád dynasty remains are housed: King III. Béla (HU3B), his wife Anna of Antioch (HUAA), and an unidentified relative (HU52) whose identification remains uncertain to this day. These remains sequenced in an earlier publications, and were used for kinship analysis for the sequence of St.Ladislaus.⁴¹

The samples were analysed with medieval Carpathian Basin reference from Maróti et al 2022 publication.⁴² The result of the analysis summarized in Table 4.

ID1	ID2	Corrected kinship coefficient
HU3B	HU52	0,123
HU3B	HUAA	-0,011
HU3B	SZTLF	0,013
HUAA	SZTLF	-0,002
HU52	HUAA	0,102
HU52	SZTLF	0,012

4. Table Result of corrected kinship analysis of St. Ladislaus and other members of House of Árpád. Expected kinship coefficient for 2nd degree relatives are 0.125 and for 5th degree relatives are 0,015

As shown in Table 4 the relationship between the HU52, the unknown royal family member, and both Anna of Antioch and King III. Béla have second degree relatives. Between the sample from the relic and King III. Béla there is a 5th degree relationship, confirming, considering pedigree of the House of Árpád, that the skull preserved in the relic indeed belongs to St. Ladislaus.

8. Discussion

Identification of relatives from the genomic data of ancestors is of great interest as it allows the study of family relationships, but it is also a precondition for most population genetic analyses to exclude close relatives from datasets (e.g., ADMIXTURE, PCA). To date, the best analysis tools were able to indicate mainly first- and second-degree relatedness from very low-coverage ancient samples^{24,25,43,44}. Based on simulated data, lcMLkin can accurately infer kinship up to the 3rd degree from $2\times$ genome coverage when the F_{ST} is low between the reference population and analysed data²⁵. However, the majority of aDNA data is below $2\times$ genome coverage. In these data, most markers are represented by one read/genotype only. It is untested whether it is possible to infer comparable diploid genotype likelihoods suitable for lcMLkin from very low-coverage data. The recent heuristic method READ (Relationship Estimation from Ancient DNA) infers relatedness up to 2nd degree from as low as $0.1\times$ coverage sequence data²⁴. In the most comprehensive AADR ancient genome data set²⁶, the majority of the indicated kinship relations are 1st degree and the handful of indicated 2nd-degree relations in all cases are uncertain. These samples are labelled with 1d.or.2d.rel tag. Diploid variant calling and genotype likelihood-based methods with the extra information of rare alleles allow better phasing and identification of IBD fragments leading to improved kinship coefficient estimations from deeply genotyped WGS data. Accordingly, some methods attempt to infer genotype likelihoods or diploid genotype calls from low-mid genome coverage ($2\text{--}4\times$) data^{25,45,46}. KING, a method that was developed to be used for fast and robust kinship coefficient estimation from low amounts of fully typed diploid markers ($5\text{--}150\text{k}$), can infer up to 3rd-degree relations from approximately 150k markers or 1st–2nd-degree relation from even as low as 5k diploid markers⁴⁵. Even though these tools are used to analyse low marker count ancient samples, the assumption implicit in these methods that the data is sufficiently high-quality diploid is often false in case of low marker count extremely low-coverage ancient samples. Therefore, when comparing samples of different genome coverage, the inferred genotype likelihoods or diploid variants from low/variable genome coverage samples could lead to major bias. To overcome these difficulties and mitigate the main genotyping biases in case of low coverage ancient samples, we used a combination of strategies to account for the effects caused by PMD and varying low genome coverage. We used random allele sampling that is the gold standard methodology when performing PCA and other population genetic analyses on ancient samples, as it leads to statistically equal genotype likelihoods of genotyped markers regardless of the genome coverage. To avoid

excessive, variable amounts of false positive variants due to the variable rate of PMD, exogenous DNA contamination, and technical errors (alignment artifacts), we restricted our analysis to the already known biallelic, high-frequency, and population-informative SNPs of the 1240K AADR dataset. This strategy perfectly aligned with our choice of kinship analysis method since the PC-Relate algorithm uses PCA to differentiate between IBD/IBS fragments.

We have demonstrated that random pseudo-haploidization of data in our analysis pipeline does not affect the result of kinship analysis. This is also confirmed by the PCA analysis, showing that the same modern individual from diploid or different pseudo-haploidized data had nearly identical PCA components.

Overlapping marker fraction, according to our study, is the major factor influencing the calculated kinship coefficient of partially genotyped samples in our analysis pipeline. Our simulations revealed that the overlapping marker fraction and the calculated kinship coefficient had a strong linear correlation. Although the PC-Relate algorithm does not require the specification of the underlying population structure of the analysed relatives, we have shown that a proper reference set is required for the analysis. As expected, the samples' own reference population resulted in proper kinship coefficients, but using reference from a different super-population corrupted the results. On the other hand, using the samples' super-population as reference resulted in comparable although slightly higher kinship coefficients compared to the proper reference population proving the robustness of the PC-Relate algorithm. This reference bias is not amplified by the applied correction for marker overlap; however, it could lead to the false identification of distant relatives. We also tested the effect of reference population choice in a complex Creole/ European admixed Cabo Verdean-Hungarian family with known 1st- to 5th-degree family relations. We have shown that the best result is achieved when all super-populations of the sources are included in the reference population set. Comparing the analyses of pseudo-haploid and diploid data for this complex admixed family confirmed the robustness of our approach, as we got nearly identical results.

In the statistical evaluation using the down sampled modern diploid/pseudo-haploid data, we simulated marker counts similar to aDNA data. We applied the same minimum 100,000 genotyped markers per individual threshold that was used in the analysis of 2136 selected ancient individuals from the AADR dataset. This equals roughly $0.08\times$ genome coverage considering the $\sim 1.15\text{M}$ autosomal markers of the 1240K marker set. Thus, the simulated data had similar marker counts and distribution as the analysed AADR dataset. Accordingly, pairwise marker overlap was $<5\%$ ($<57,000$ markers) between 3.72 and 3.46% of the analysed

sample pairs (ancient and modern respectively). Our analysis shows that when proper reference population is available, the applied method is suitable to identify relations up to the 4th degree from low to high coverage mixed samples.

We also confirmed the robustness of our methodology on real ancient data with known family relations. Our analysis showed that in the case of a medieval Hungarian family, a general modern European reference super-population gave appropriate results. Despite the fact that the uncorrected kinship coefficients varied highly due to the different genome coverages, our methodology resulted in reproducible corrected kinship coefficients consistent with the known family relation in each case. In the second example, we reanalysed published kinship relations from Corded Ware Culture samples ²⁴. Compared to the READ software which could indicate relations up to the second degree of kinship and even missed one second-degree relation, our approach could properly identify all relations up to 4th degree from this large ancient family with very low/variable genome coverages ($0.12\times$ – $4.18\times$), underlining the efficiency and usefulness of our approach.

Our results exposed both the advantages and the limitations of our method. Although RPsH combined with the choice of the 1240K marker set in our study allowed us to overcome genotyping bias of low-coverage ancient samples, it clearly restricts the analysis to populations that are properly represented by these markers. In the PC-Relate algorithm, PCA is used to regress out the population-specific IBS components. Using linear regression to fit individuals to the model, all the remaining non-regressed PC components are calculated as IBD. Thus, insufficient amount of reference individuals, improper or missing population components in the reference, or marker sets that are lacking informative markers of the tests lead to underestimation of IBS and inflated kinship coefficient estimation. The greater the difference between the structure of related individuals and the reference populations, the greater fraction of IBS is accounted incorrectly as IBD which can seriously bias small kinship coefficients representing very distant kinship relations. Accordingly, the current 1240K marker set is less suitable for the analysis of extremely old samples, and for small isolated populations, because these supposedly have less informative markers in this marker set, and also have insufficient reference populations in the current genome databases. Furthermore, while PC-Relate kinship coefficient estimator is known to be appropriate even in inbred populations ³⁷, we have to caution that in case of inbred or small drifting populations extra care has to be taken to confirm that the test individuals are analysed with their own reference population. When no prior knowledge exists on the reference population, F_{ST} or

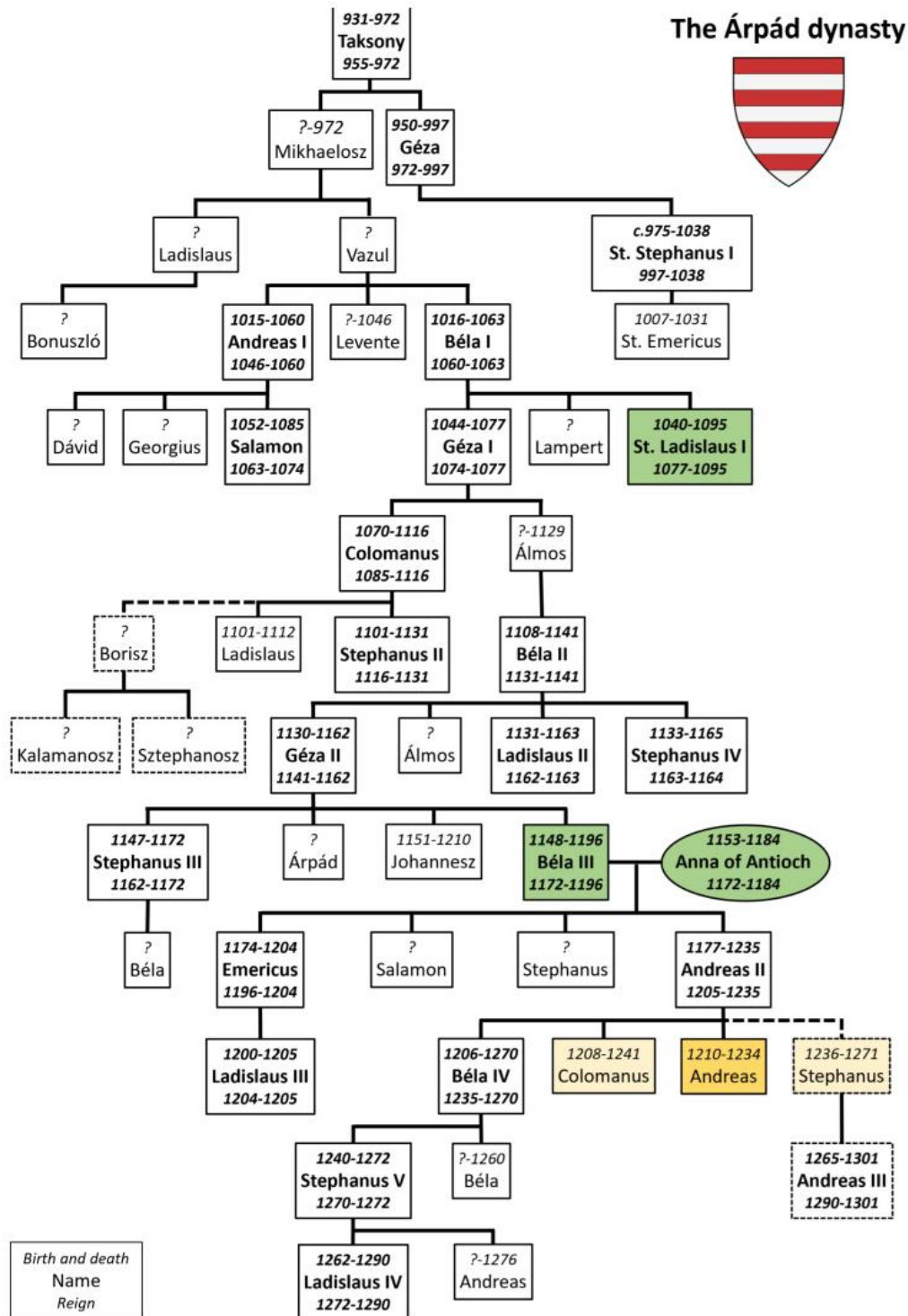
FastNGSAdmix ⁴⁷ analysis could be used as an objective method to select individuals best matching our test individual's genome structure as a reference population.

Genotyping error simulations show that approximately double error rate compared to typical experimental aDNA data leads to 5.4–10.4% proportionally lower corrected kinship coefficient than the expected kinship coefficient in our workflow. The mean corrected kinship coefficient of the validated sample dups and 1st relatives of the experimental 2136 AADR individual was 0.48 and 0.24 respectively (approximately ~4% lower from the expected). This is in accordance with the mean X contamination rate (1.28%) of ancient individuals of the AADR V42.2 dataset suggesting that our kinship-estimation method can be safely used on typical aDNA data. Nevertheless, analysis of highly contaminated (CRITICAL/FAIL) samples containing higher rate of genotyping errors (>5%) could lead to underestimation of corrected kinship coefficient and as a result to underestimation of the degree of relation especially in case when the relatives share less than the expected IBD fragments due to true biological variation. The unsupervised analysis of 2136 ancient individuals of the 1240K AADR dataset demonstrated that our method could identify real 1st–4th degree of relatedness from very low-coverage ancient damaged samples and fails only when the proper reference population is not present in the dataset. Comparison with READ showed that our method has better sensitivity, offers improved performance, and scales better on multi-core machines. On the other hand, our results show that the 1240K marker set was sufficient to properly analyse 4000-year -old ancient Corded Ware Culture individuals with a modern Eurasian reference population, suggesting that the majority of the high-frequency EUR informative markers were already present at this age.

According to our results, the used method had slight downward (2–4%) bias in the analysed 1KG dups and first-degree relatives and also in the validated first-degree ancient samples. However, this downward bias is also present in the kinship coefficient estimation of fully typed diploid 1KG relatives suggesting that the original PC-Relate algorithm and not the applied correction or pseudo-haploidization is accountable for this bias. This is also supported by our simulations on the corrected kinship coefficient calculated from marker depleted pseudo-haploid data and the original fully typed diploid data and the very high correlation between the kinship coefficient calculation of marker depleted pseudo-haploidized and the original fully typed 1KG data. On the other hand, the mean of the corrected kinship coefficient of the indicated 2nd-degree relatives (n=119) of experimental AADR data is 0.1274 that is a slightly over the expected value (~2% relative difference) suggesting that the

bias could originate from more than a single factor. Our analysis revealed new possibilities to improve kinship analysis from low coverage ancient data. Diploid typing with pre-capture enrichment could result in higher sensitivity even at lower marker overlap fraction. However, this is only feasible when a sufficient number of individuals are available from the matching reference populations. According to our analysis, ~50–100 unrelated individuals are sufficient as a reference in case of modern samples. We speculate that in case of populations with less complex genome structures (like pre-iron age populations), a smaller number of unrelated individuals could likely represent the population structure properly. This is also demonstrated in the case of the validated relations of the analysed CWC individuals where the analysis resulted comparable kinship coefficient estimates using modern EUR individuals as a reference or the 25 Czech, Latvian, Estonian, and German CWC individuals of the AADR dataset. Alternatively, using larger marker sets would increase the number of overlapping markers between individuals resulting in higher sensitivity from the already available low-coverage WGS data. The increasing number of aDNA studies should identify proper reference populations and suitable high frequency marker sets for cases that are difficult to analyse at present.

St.Ladislaus is one of our most renowned kings, who combined knightly virtues with the development of state governance, a path initiated by his predecessor, St.Stephan. Following his death in 1095, King III. Béla, upon a request in 1196 was canonized. Consequently, the remains of the king interred in Várad (present-day Nagyvárad, Romania) were exhumed and utilized as relics. Among these relics, the most notable is the herma of the skull, preserved in the cathedral of Győr after enduring numerous trials. Due to the long history of the herma, historians and archaeologists harboured doubts about the authenticity of the relic, prompting the need for genetic examinations. The closest known ruler to have descended from the Árpád dynasty was III. Béla, who was buried in the Basilica in Székesfehérvár, later in the Matthias Church in Budapest. His Y-chromosome sequence were published in 2020 ^{40,41}, and whole genome sequence in 2022. Given the significant genetic distance, traditional methods for detection were not feasible. However, the method described above opened the doors to potential genetic validation. Figure 10 contains an extract of the family tree of Árpád Dynasty where samples from the analysis showed in green and the suggested identity for HU52 sample shown in orange. Our analysis shown that King Bela III, whose remains were reburied in Mathias Church in Budapest was in 5th degree relation from the sample from the relic, with this, we can validate the originality of the relic as St.Ladislaus.



10. Figure The extract of the family tree of the Árpád dynasty starts with Taksony, the last common male ancestor of the kings of the family. It exclusively contains the male members of the house known from historical data, and Anna of Antioch, the wife of Béla III. **Bald** – ruler; dashed line and frame – uncertain family member; green – identified persons with WGS data; orange – suggested identity for HU52.⁴⁰

9. Conclusion

During our work, we examined the PC-Relate algorithm, along with its potential for development. We investigated the limitations and effectiveness of the algorithm under various parameters.

One such parameter was the effect of random pseudo-haploidization on the Principal Component Analysis-based method and the determination of kinship coefficients, which we demonstrated to have no significant impact on either. Subsequently, we examined the effect of the number of overlapping markers on the kinship coefficient value through random down sampling, which resulted in a linear relationship. Using this, we developed a correction method in which the number of overlapping markers is divided by the total number of markers, and this value is then divided into the obtained kinship coefficient. We showed that with this correction, accurate kinship estimates can be obtained even with a low number of overlapping markers.

Using our own algorithm, we examined the corrected kinship coefficient under various genotyping errors. A total of five experimental setups were investigated: genotyping error-free examination; endogenous contamination; exogenous contamination; postmortem damage; and all genotyping errors combined. The largest difference was observed with exogenous contamination, which reduced the coefficient value by nearly 10%.

In the last set of simulation studies, we were interested in the effect of the reference population on the study results. In this case, we examined three known relatives from the 1KG database and a highly admixed Hungarian-Greenlandic family. In the former case, the result was that if we do not know the exact population of a given sample pair, we can achieve satisfactory results by using only the appropriate super-population. In the case of the highly admixed family, the use of both superpopulations was necessary.

The method's validation was conducted on two datasets. One was published using another method, READ, on a Corded Ware Culture family containing 5 male members. We were able to detect all kinship relationships, including those identified by the other algorithm, as well as second-degree and several third- and fourth-degree kinship relationships. In another case, we compared multiple sequences of a known father-son relationship medieval family with different coverages, successfully identifying the kinship relationship despite any coverage differences.

We examined kinship relationships in version 44.2 of the Allen Ancient DNA Repository, after filtering for age and geographic location, as mentioned in the Materials and Methods section. We successfully detected several new kinship relationships and also sample contamination.

Finally, using the method, we successfully confirmed the authenticity of the skull relic of St. Ladislaus preserved in Győr, what is the first catholic Saint that was confirmed genetically.

In summary, our proposed methodology is capable of reliably identifying the relatedness up to the 4th degree from low-coverage genome data, redefining the limits of kinship analysis from low-coverage ancient or badly degraded forensic WGS data.

10. Acknowledgements

I am grateful to my supervisor, Dr. Zoltán Maróti, for his unwavering support over the years and for introducing me to the world of bioinformatics and research. The critical and scientific thinking he thought me were invaluable lessons for my further career.

I would also like to express my thanks to Dr. Tibor Kalmár, the head of the Genetic Diagnostic Laboratory at the Department of Pediatrics and Pediatric Health Center, for providing me with the opportunity to conduct research in the lab. I had the chance to learn the ins and outs of next-generation sequencing and received numerous invaluable human and scientific pieces of advice over the years.

I extend my gratitude to all the staff at the Genetic Diagnostic Laboratory for creating an excellent working atmosphere for my research throughout the years.

I would also thank the Archaeogenetic Research Group (Faculty of Sciences and Informatics, Department of Genetics) lead by Dr. Tibor Török, for accepting me in the group and always give me work to improve myself. I received generous support and help from Dr. Tibor Török, Dr. Endre Neparáczki, Dr. Gergely István “Bobek” Varga, Dr. Alexandra Ginguta, Dr. Balázs Tihanyi, Dr. Olga Spekker, Luca Kis, Bence Kovács, Oszkár Schütz and all our bachelor and master thesis students.

I am deeply thankful for the support of my family, especially my father Dr. Emil Nyerki, my mother Zsuzsanna Ágnes Alabert Dr. Nyerkiné, my brother Marcell András Lipták for all the help and support, through all the years of university.

I am grateful for all my friend to always listen to my speeches about my topic, my new findings even when we came together to relax after a long week of work. Thank you guys, you are the best!

I would dedicate this work for my late friends and mentors who always believed me through my early years of education and academic journey.

References

1. G.Garcia M, Pérez-Creles MD, Osuna E, Legaz I. Impact of the Human Microbiome in Forensic Sciences : a Systematic Review. *Appl Environ Microbiol*. Published online 2022.
2. Tillmar A, Sturk-Andreaggi K, Daniels-Higginbotham J, Thomas JT, Marshall C. The FORCE panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications. *Genes (Basel)*. 2021;12(12). doi:10.3390/genes12121968
3. Liu F, Li Z, Yang W, Xu F. Age-Invariant Adversarial Feature Learning for Kinship Verification. *Mathematics*. 2022;10(3). doi:10.3390/math10030480
4. Morimoto C, Manabe S, Kawaguchi T, et al. Pairwise kinship analysis by the index of chromosome sharing using high-density single nucleotide polymorphisms. *PLoS One*. 2016;11(7):1-17. doi:10.1371/journal.pone.0160287
5. Shyla A, Borovko SR, Tillmar AO, et al. Belarusian experience of the use of FamLinkX for solving complex kinship cases involving X-STR markers. *Forensic Sci Int Genet Suppl Ser*. 2015;5:e539-e541. doi:10.1016/j.fsigss.2015.09.213
6. He G, Adnan A, Al-Qahtani WS, et al. Genetic admixture history and forensic characteristics of Tibeto-Burman-speaking Qiang people explored via the newly developed Y-STR panel and genome-wide SNP data. *Front Ecol Evol*. 2022;10(October):1-19. doi:10.3389/fevo.2022.939659
7. Mittnik A, Massy K, Knipper C, et al. Kinship-based social inequality in Bronze Age Europe. *Science (80-)*. 2019;366(6466):731-734. doi:10.1126/science.aax6219
8. Schreiber M, Stein N, Mascher M. Genomic approaches for studying crop evolution. *Genome Biol*. 2018;19(1):1-15. doi:10.1186/s13059-018-1528-8
9. McHugo GP, Dover MJ, MacHugh DE. Unlocking the origins and biology of domestic animals using ancient DNA and paleogenomics. *BMC Biol*. 2019;17(1):1-20. doi:10.1186/s12915-019-0724-7
10. Higuchi R, Bowman B, Freiburger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. *Nature*. 1984;312(5991):282-284. doi:10.1038/312282a0
11. Pääbo S. Molecular cloning of Ancient Egyptian mummy DNA. *Nature*. 1985;314(18).
12. Linderholm A. 2016_Linderholm-Ancient DNA the next generation – chapter and verse - REVIEW. 2016;(December 2015):150-160.
13. Velsko IM, Fagernäs Z, Tromp M, et al. Exploring archaeogenetic studies of dental calculus to shed light on past human migrations in Oceania. *bioRxiv*. Published online 2023:2023.10.18.563027. <https://www.biorxiv.org/content/10.1101/2023.10.18.563027v1%0Ahttps://www.biorxiv.org/content/10.1101/2023.10.18.563027v1.abstract>
14. Henriksen RA, Zhao L, Korneliussen TS. NGSNGS: next-generation simulator for next-generation sequencing data. *Bioinformatics*. 2023;39(1):10-11. doi:10.1093/bioinformatics/btad041

15. Latorre SM, Lang PLM, Burbano HA, Gutaker RM. Isolation, Library Preparation, and Bioinformatic Analysis of Historical and Ancient Plant DNA. *Curr Protoc Plant Biol.* 2020;5(4):1-28. doi:10.1002/cppb.20121
16. Liagre EBK, Hoogland MLP, Schrader SA. It runs in the family: Kinship analysis using foot anomalies in the cemetery of Middenbeemster (Netherlands, 17th to 19th century). *Int J Osteoarchaeol.* 2022;32(4):769-782. doi:10.1002/oa.3100
17. Udall AM, de Groot JIM, De Jong SB, Shankar A. How I See Me—A Meta-Analysis Investigating the Association Between Identities and Pro-environmental Behaviour. *Front Psychol.* 2021;12(March). doi:10.3389/fpsyg.2021.582421
18. Hajduk GK, Cockburn A, Margraf N, Osmond HL, Walling CA, Kruuk LEB. Inbreeding, inbreeding depression, and infidelity in a cooperatively breeding bird*. *Evolution (N Y).* 2018;72(7):1500-1514. doi:10.1111/evo.13496
19. Snedecor J, Fennell T, Stadick S, et al. Fast and accurate kinship estimation using sparse SNPs in relatively large database searches. *Forensic Sci Int Genet.* 2022;61(August):102769. doi:10.1016/j.fsigen.2022.102769
20. Sticca EL, Belbin GM, Gignoux CR. Current Developments in Detection of Identity-by-Descent Methods and Applications. *Front Genet.* 2021;12(September):1-6. doi:10.3389/fgene.2021.722602
21. Henden L, Grosz BR, Ellis M, Nicholson GA, Kennerson M, Williams KL. Identity-by-descent analysis of CMTX3 links three families through a common founder. *J Hum Genet.* 2023;68(1):47-49. doi:10.1038/s10038-022-01078-1
22. Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating relatedness between malaria parasites. *Genetics.* 2019;212(4):1337-1351. doi:10.1534/genetics.119.302120
23. Glodzik D, Navarro P, Vitart V, et al. Inference of identity by descent in population isolates and optimal sequencing studies. *Eur J Hum Genet.* 2013;21(10):1140-1145. doi:10.1038/ejhg.2012.307
24. Kuhn JMM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. *PLoS One.* 2018;13(4):1-21. doi:10.1371/journal.pone.0195491
25. Lipatov M, Sanjeev K, Patro R, Veeramah KR. Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. Published online 2015:1-20. doi:10.1101/023374
26. Allen Ancient DNA Resource. No Title. <https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data>
27. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74. doi:10.1038/nature15393
28. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575. doi:10.1086/519795
29. Nyerki E, Kalmár T, Schütz O, Lima RM, Neparáczi E, Török T KT. correctKin Plink dataset.
30. Nyerki E, Kalmár T, Schütz O, et al. correctKin: an optimized method to infer

- relatedness up to the 4th degree from low-coverage ancient human genomes. *Genome Biol.* 2023;24(1):1-21. doi:10.1186/s13059-023-02882-4
31. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-909. doi:10.1038/ng1847
 32. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2(12):2074-2093. doi:10.1371/journal.pgen.0020190
 33. Team Rs. RStudio: Integrated Development Environment for R. Published online 2019. <http://www.rstudio.com/>
 34. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016. <https://ggplot2.tidyverse.org>
 35. Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics.* 2018;210(2):719-731. doi:10.1534/genetics.118.301336
 36. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics.* 2014;15(1):1-13. doi:10.1186/s12859-014-0356-4
 37. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet.* 2016;98(1):127-148. doi:10.1016/j.ajhg.2015.11.022
 38. Mathieson I, Lazaridis I, Rohland N, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature.* 2015;528(7583):499-503. doi:10.1038/nature16152
 39. Järve M, Saag L, Scheib CL, et al. Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance. *Curr Biol.* 2019;29(14):2430-2441.e10. doi:10.1016/j.cub.2019.06.019
 40. Varga GIB, Kristóf LA, Maár K, et al. The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy. *J Genet Genomics.* 2023;50(1):58-61. doi:10.1016/j.jgg.2022.06.008
 41. Nagy PL, Olasz J, Neparácski E, et al. Determination of the phylogenetic origins of the Árpád Dynasty based on Y chromosome sequencing of Béla the Third. *Eur J Hum Genet.* 2021;29(1):164-172. doi:10.1038/s41431-020-0683-z
 42. Maróti Z, Neparácski E, Schütz O, et al. The genetic origin of Huns, Avars, and conquering Hungarians. *Curr Biol.* 2022;32(13):2858-2870.e7. doi:10.1016/j.cub.2022.04.093
 43. Kennett DJ, Plog S, George RJ, et al. Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat Commun.* 2017;8. doi:10.1038/ncomms14115
 44. Ringbauer H, Steinrücken M, Fehren-Schmitz L, Reich D. Increased rate of close-kin unions in the central Andes in the half millennium before European contact. *Curr Biol.* 2020;30(17):R980-R981. doi:10.1016/j.cub.2020.07.072
 45. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26(22):2867-2873. doi:10.1093/bioinformatics/btq559

46. Severson AL, Korneliussen TS, Moltke I. LocalNgsRelate: a software tool for inferring IBD sharing along the genome between pairs of individuals from low-depth NGS data. *Bioinformatics*. 2022;38(4):1159-1161. doi:10.1093/bioinformatics/btab732
47. Jørsboe E, Hanghøj K, Albrechtsen A. FastNGSadmix: Admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics*. 2017;33(19):3148-3150. doi:10.1093/bioinformatics/btx474

Annex

Co-author certification

3. számú melléklet: Társ szerzői lemondó nyilatkozat

Co-author certification

I, myself as a corresponding author of the following publication(s) declare that the authors have no conflict of interest, and Nyerki Emil Ph.D. candidate had significant contribution to the jointly published research(es). The results discussed in her thesis were not used and not intended to be used in any other qualification process for obtaining a PhD degree.

Szeged, 2024.03.08



Gergely István Varga

first author



Endre Neparáczki

last author

The publication(s) relevant to the applicant's thesis:

Gergely I.B. Varga, Lilla Alida Kristóf, Kitti Maár, Luca Kis, Oszkár Schütz, Orsolya Váradi, Bence Kovács, Alexandra Gînguță, Balázs Tihanyi, Péter L. Nagy, Zoltán Maróti, **Emil Nyerki**, Tibor Török, Endre Neparáczki. *The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy*, Journal of Genetics and Genomics, Volume 50, Issue 1, (2023)

I.


Emil Nyerki, Tibor Kalmár, Oszkár Schütz, M. Lima Rui, Endre Neparácski, Tibor Török, Zoltán Maróti. *Correctkin: An Optimized Method to Infer Relatedness up to the 4th Degree from Low-Coverage Ancient Human Genomes*. Genome Biology 24, no. 1 (2023). <https://doi.org/10.1186/s13059-023-02882-4>. **IF:12.3 D1**

METHOD

Open Access



correctKin: an optimized method to infer relatedness up to the 4th degree from low-coverage ancient human genomes

Emil Nyerki^{1,2†}, Tibor Kalmár^{1†}, Oszkár Schütz³, Rui M. Lima⁴, Endre Neparáczki^{2,3}, Tibor Török^{2,3} and Zoltán Maróti^{1,2*} 

[†]Emil Nyerki and Tibor Kalmár contributed equally to this work.

*Correspondence: maroti.zoltan@med.u-szeged.hu

¹ Department of Pediatrics, University of Szeged Albert Szent-Györgyi Medical Center Faculty of Medicine, Szeged, Hungary

² Department of Archaeogenetics, Institute of Hungarian Research, Budapest, Hungary

³ Department of Genetics, University of Szeged, Szeged, Hungary

⁴ Institute of Plant Biology, Biological Research Centre, Szeged, Hungary

Abstract

Kinship analysis from very low-coverage ancient sequences has been possible up to the second degree with large uncertainties. We propose a new, accurate, and fast method, correctKin, to estimate the kinship coefficient and the confidence interval using low-coverage ancient data. We perform simulations and also validate correctKin on experimental modern and ancient data with widely different genome coverages ($0.12\times-11.9\times$) using samples with known family relations and known/unknown population structure. Based on our results, correctKin allows for the reliable identification of relatedness up to the 4th degree from variable/low-coverage ancient or badly degraded forensic whole genome sequencing data.

Keywords: Kinship, Genomics, Low coverage, Ancient DNA, Forensic

Background

Kinship analysis is a method for determination of the familial relationship between individuals from genome data. The kinship coefficient is defined as the probability that two homologous alleles drawn from each of two individuals are the result of identity by descent (IBD). This is a classic measurement of relatedness [1, 2]. Several algorithms have been developed to perform kinship analysis [3] including GERMLINE [4], fastIBD [5], GRAB [6], and ANGSD [7]. These are based on different strategies and metrics of IBD segments for calculating relatedness from microarray or WGS data. Distinguishing IBD which represents familial relatedness from identity-by-state (IBS) that represents population relatedness is difficult as both result in genetic similarity based on shared alleles. Despite the biological variation in IBD sharing due to the outcome of the stochastic nature of recombination and segregation during meiosis in gametogenesis, it is possible to infer kinship up to the 5–6th degree of relatedness from microarray or deeply typed WGS data. Achieving such a high level of certainty also requires an appropriate



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

set of reference data (except for methods like KING [8] or IBIS [9]), and clever algorithms that account for biological variations resulting from familial (IBD) and population relatedness (IBS).

Recently huge genomic datasets have been generated from ancient samples in order to uncover the genetic relations of ancient and modern populations. From this data, it is also of high interest to study the family organization of ancient populations. However, analyzing ancient DNA (aDNA) poses additional difficulties due to the widely different but generally low genome coverage and postmortem damage (PMD) observed in these samples. The aDNA databases usually contain sequence data between 0.05 and 3× average genome coverage [10–12], since the sequencing of ancient samples with low endogenous DNA content is still challenging and costly. Differences in coverage and only partial overlap of genetic markers between samples can lead to significant bias when comparing the frequencies and genotype likelihoods of genetic variants, leading to uncertainties of the inferred genotype probabilities. An additional problem in the analysis of ancient data is that in most cases there is limited or no information on the appropriate reference population data to distinguish IBD from IBS.

In the present study, we wanted to address the difficulties of low-coverage aDNA data and dissect the main factors that affect kinship calculations. To overcome typing bias, random sampling of one allele per site (pseudo-haploid calling) was used successfully in aDNA studies [13–22]. In order to compare diploid and pseudo-haploid datasets, heterozygous alleles of diploid data need to be random pseudo-haploidized (RpsH) by randomly assigning heterozygote alleles as either homozygote reference (REF) or homozygote alternative (ALT). Although rare alleles can offer significant improvement in some kinship calculation methods when analyzing high-quality WGS data, genotype calling from the whole human genome could lead to excessive, variable amounts of false positive calls from low-coverage, degraded aDNA datasets. To minimize this bias, we restricted our analysis to the already known biallelic, high-frequency, and population-informative SNPs of the V42.2 1240K Allen Ancient DNA Resource (AADR) dataset [23]. To address the issue of unknown reference populations, we used the PC-Relate algorithm [24]. In the presence of unspecified population structure, this algorithm proposes a principal component-based, model-free approach for estimating kinship coefficients and IBD sharing probabilities. We applied a combination of techniques to mitigate genotyping uncertainties and tested their effects and limitations on kinship analysis of low-coverage ancient sequences. We used simulation to downsample fully genotyped real NGS data to examine the effect of partial marker overlap between samples and we also explored the effect of reference population choice on the kinship coefficient calculation. Based on our results, we developed a new computational approach which can reliably calculate corrected kinship coefficient from poorly genotyped data.

Here we offer guidelines and a list of the necessary bioinformatics tools required to calculate the corrected kinship coefficient. These guidelines overcome the technical limitations of generally low genome coverage, postmortem damage, genotyping uncertainties, and the partial overlapping of genetic markers between samples. As a proof of concept, we validated our proposed methodology on both experimental modern and ancient data with widely different genome coverages, using samples with known family relations and known or unknown population structure.

Results

To use marker counts similar to aDNA data, all modern dataset were downsampled to the autosomal marker positions of the 1240K SNP set of the AADR dataset [23] in all of our simulations.

The effect of random pseudo-haploidization (RPsH) on PCA calculations

Since the selected kinship methodology, PC-Relate, applies principal component analysis (PCA) to identify population structure, we first tested the effect of RPsH on PCA. We selected the British (GBR), Toscani (TSI), Iberian (IBS), and Finnish (FIN) populations from the 1000 Genome Project Phase 3 (1KG phase 3) dataset (404 samples) and randomized the diploid dataset with three different seeds. We performed smartpca analysis on the original diploid and the three random pseudo-haploidized dataset. According to our results, RPsH does not alter the PCA calculations significantly (Additional file 1: Figure S1).

The effect of random pseudo-haploidization (RPsH) on kinship coefficient calculation

We assessed the effect of RPsH on kinship calculation by selecting 509 individuals from five populations with different population structure from the 1KG phase 3 dataset [25]. The five populations were as follows: FIN, GBR, TSI, Han (HAN), and Utah residents with Northern and Western European ancestry (CEU). In our experiments, we generated 100 different pseudo-haploid datasets from the original diploid data using different random seeds.

To study exclusively the effect of RPsH on kinship coefficient calculation, we included sample duplicates with different random pseudo-haploidization. This setup does not exclude differences between the genome structure of the test sample and the reference population, thus we selected random individuals from the GBR (HG00244.SG), FIN (HG00356.SG), CEU (NA12763.SG, NA12775.SG), and TSI (NA20798.SG) populations. This allowed us to overcome the interference of other effects, such as skewed recombination/segregation, differences in sequence alignment, genotyping, genome composition, or the population structure between the relatives. This idealized experimental setup is the equivalent of the monozygotic twin kinship relation, in forensics referred to as sample matching, while the maximal expected kinship coefficient (0.5) allows for the most sensitive analysis.

To study the effect of RPsH on true first/second-order relatives, we also selected samples with known family relations from the 1KG phase 3 dataset. (a) HG00702-HG00657 a parent-child relation from a Han population, where exactly 50% of genome is shared between the two samples, (b) NA20526-NA20792 siblings from a TSI population, where 50% of the genome comes from the same parents; however, a different subset of the markers are found in the sibs due to segregation and recombination, and (c) HG00124-HG00119 second-order relatives from a GBR population, where statistically 25% of genomes are shared. We calculated the kinship coefficient for each of the selected relatives (kin1/kin2) using their own reference population for the 100 different randomization and calculated the mean and the standard deviation of the estimated kinship coefficients. Knowing the expected kinship coefficients (0.5 for sample match, 0.25

for 1st-, 0.125 for 2nd-degree relations), we were able to validate that RPsH does not significantly alter the calculated kinship coefficient in these settings (Additional file 2: Table S1).

The effect of overlapping marker fraction on the kinship coefficient calculation

The original PC-Relate algorithm was created to analyze modern fully genotyped diploid samples. However, in the case of ancient data, genome coverage and partial genotyping is one of the factors that has the greatest variability between samples. Kinship coefficient calculation is based on the IBD segments shared between two samples which can only be assessed at marker positions where both samples are genotyped. To investigate the effect of this factor, we defined a metric called overlapping marker fraction. We calculate this metric by dividing the number of markers where both samples are genotyped with the total number of markers in the dataset (1240K).

Using the 100 random pseudo-haploidized fully typed dataset of the previous experiment, we randomly depleted the markers between the selected sample dups and true 1KG relatives to a marker overlap fraction between 5 and 100% using different random seeds. We calculated the mean and SD of the estimated kinship coefficients between the selected sample pairs of the different randomizations for each overlap fraction (Additional file 2: Table S1). We visualized the mean and SD of the uncorrected kinship coefficients of a sample dup (CEU; NA12775.SG), a known first-degree relation (HAN, HG00702-HG00657) and a true second-degree relation (GBR, HG00119.SG-HG00124.SG) in Fig. 1A.

The results revealed that the kinship coefficient is a linear function of the marker overlap fraction. This allows a simple method for correcting the kinship coefficient value for low-coverage genomes by dividing the estimated kinship coefficient with the marker overlap fraction between the two samples. According to our simulation, the correction of estimated kinship coefficient for sparsely genotyped data resulted in reproducible kinship estimation regardless of the overlap fraction (Fig. 1B). As expected, low partially overlapping subset of markers would lead to less complete representation of the

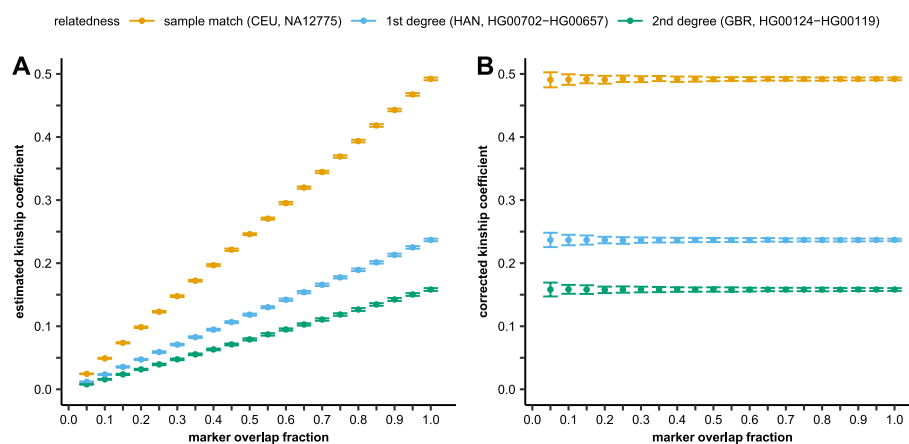


Fig. 1 The mean and the 95% confidence interval of the **A** uncorrected and **B** corrected kinship coefficient between selected 1KG individuals (sample dup, known 1st and 2nd degree) at different marker overlap fractions

reference population and the test individuals thus regression of IBD/IBS based on the PC-Relate algorithm would have higher SD at low marker counts. Although the marker overlap fraction correction differs more than one magnitude between very low and high marker overlap fraction sample pairs, the analysis shows (Additional file 2: Table S1) that the correction itself does not introduce overall bias (the mean is statistically the same) and does not significantly multiply the error rate ($4\times$ increase in SD at $20\times$ correction factor). Our results suggest that the increased SD of the method are likely due to the higher uncertainty of PCA.

The effect of genotyping errors on the corrected kinship coefficient

We tested the effect of genotyping errors using the same dataset and sample dups/known first- and second-degree 1KG relatives as in the previous experiments. We simulated 4 scenarios: (1) only post mortem damage, (2) only endogenous contamination (using a random YRI individual as the contaminant), (3) only exogenous contamination, and (4) equal combination of the first three sources of genotype errors. In each scenario, we had approximately twice as many genotyping errors per individual introduced as in a typical aDNA dataset (see “Methods”). We calculated the mean and SD of the corrected kinship coefficients (Additional file 3: Table S2) between the selected samples. We visualized the mean and 95% confidence interval of the corrected kinship coefficients of the selected sample dup (CEU; NA12775.SG), true first-degree relation (HAN; HG00702-HG00657), and true second-degree relation (GBR; HG00119.SG-HG00124.SG) in Fig. 2.

In general, genotyping errors lower the mean and increase the SD of estimated kinship coefficient. The decrease in the corrected coefficients was proportional to the expected kinship coefficients. The largest effect ($\sim 9.6\%$ lower kinship coefficient) was seen in case of the exogenous contamination. We speculate that it was likely due to the fact that in this scenario (although the genotyping error affects different subset of markers in different samples) the states of the markers were uniformly set to the homozygote major state



Fig. 2 Effect of different aDNA-related genotype errors on the corrected kinship coefficient. In the last case (mixed error), we introduced an equal amount of post mortem damage, exogenous and endogenous contamination in the simulated data. The points represent the mean, and the error bars represent the 95% confidence interval of the corrected kinship coefficient between selected 1KG individuals (sample dup, known first and second degree)

leading to higher bias than that of random flip of minor/major state in different samples in the other scenarios. In our simulations even at the applied relatively high error rates (compared to experimental aDNA error rate), the largest effect was still significantly smaller than the 50% difference of expected kinship coefficients between different degrees of relations thus the estimated degree of relatedness for the analyzed relatives remained the same. However, we have to note that skewed IBD sharing and high genotyping error in the test individuals could lead to false (one-degree higher) classification of the analyzed relation.

The effect of reference population selection on kinship analysis

In this analysis, we wanted to investigate the scenario in which proper reference population is unknown or unavailable which is often the case for ancient samples. Using the same public 1KG phase 3 dataset and the three known relatives HG00702-HG00657 parent-child of Han population, NA20526-NA20792 siblings of TSI population, and HG00124-HG00119 second-order relatives of GBR population, we investigated the effect of the reference population on the calculated kinship coefficients. We tested three different scenarios: (1) the reference population was the same as that to which the selected individual belonged to; (2) the reference population was from a different super-population (AFR); (3) the reference population was from the same super-population as the selected individual (JPT for Han, IBS for TSI, FIN for GBR) (Fig. 3, Additional file 1: Figure S2). To study the effect of overlapping marker fractions in these more complex cases, the selected sample pairs were also marker depleted in the range of 100 to 5% overlap fractions.

The results revealed that a reference population with significantly different genetic background, like African for European samples, strongly corrupts the results. In this experimental setup, we lack a proper number of unrelated references; hence, IBS fractions are likely not represented and cannot be properly regressed out. Furthermore, when only a couple non-AFR individuals are included in the analysis depending on

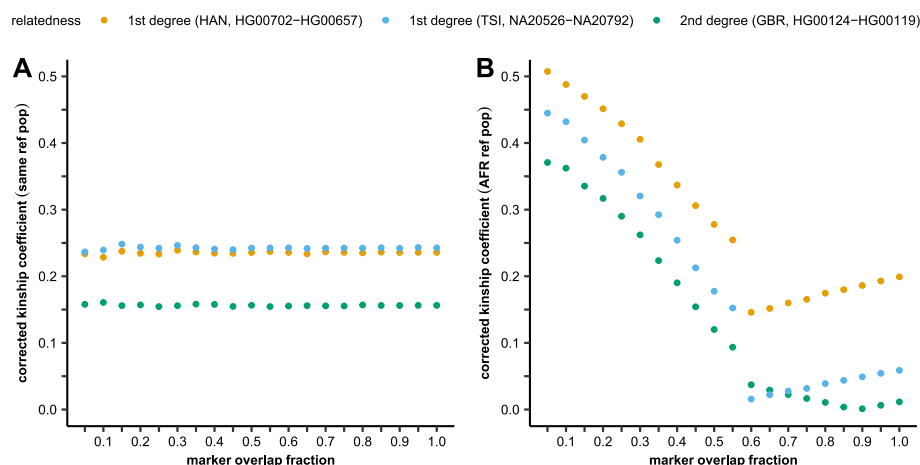


Fig. 3 The effect of reference population choice and marker overlap fraction on the corrected kinship coefficients between known 1st–2nd-degree 1KG relatives. **A** The samples' own populations used as references. **B** The AFR super-population was used for each sample

the marker overlap fraction, the EUR/EAS-specific markers are mostly excluded as the PCAngsd implementation uses a default 0.05 MAF marker pruning. Thus, at low marker overlap, mainly the AFR-specific markers are kept while at higher marker overlap slightly more EUR/EAS-specific markers are also included in the analysis. Based on the used marker set, the optimal number of eigenvectors and the underlying PCA-based regression of IBS components are expected to be different. We speculate that these differences could be the likely cause of the observed non-linear kinship coefficient estimates in Fig. 3B.

Using a super-population with similar genetic background to the sample gives very similar results as if its own reference population were used (Additional file 1: Figure S2).

Effect of reference population selection on kinship analysis in a complex admixed family with multiple ethnic relations

To assess the choice of reference population in the kinship analysis of admixed individuals (often the case in ancient populations), we analyzed a complex admixed Cabo Verdean-Hungarian family with known pedigree. In this family, we had multiple old as well as recent admixes resulting in various admix component ratio individuals. WGS data was available for siblings (1st order), differently admixed half-sibs (2nd order), and 5th-order relatives as shown in Additional file 7: Figure S3.

We tested two scenarios: the reference population was (1) only African (AFR) representing the majority of the admix sources in the tested samples; (2) both African and European (EUR) populations were included. Additionally, we performed marker depletion to investigate the effect of coverage in this complex scenario (Fig. 4).

The results in Fig. 4 demonstrate that in order to obtain realistic coefficient values in case of a complex admixture, a combined set of reference populations is required representing the population structure of all ancestors. Using just the majority source as reference significantly distorts the result.

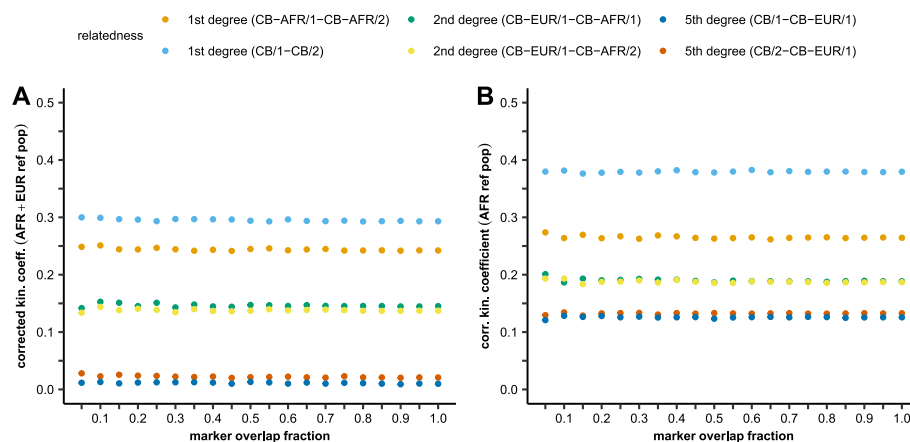


Fig. 4 The effect of reference population choice and marker overlap fraction on the calculated kinship coefficient in a complex admixed modern family with 1st- (sibs), 2nd- (half-sibs), and 5th-degree relations, where individuals originated from populations with largely different genetic structure. Markers were depleted between the relatives to 5–100% overlap fractions. **A** The reference population was a combined set of AFR+EUR populations. **B** The reference populations were AFR populations

To test whether random haploidization alters the kinship coefficient calculation compared to the better phased diploid data in this complex admixed case, we also performed this analysis from the original diploid dataset (Additional file 4: Table S3) with the AFR + EUR reference population. It was confirmed again that even in such complex admixed family, the differences due to RPsH were negligible.

Statistical validation, assessment of technical errors

We selected EUR and EAS individuals from 1KG phase 3 dataset ($n=1020$) and estimated the kinship coefficients between these individuals. Although there are a few true relatives in the selected individuals, the overwhelming majority of pairwise relations are expected to be unrelated, thus representing the variance of technical error of the whole analysis. To test the effect of overlapping genotyping fraction on the mean and standard error of the corrected kinship coefficient, we randomly depleted the marker set in these individuals between 100,000 markers and the fully typed marker count ($\sim 1.2\text{M}$), amounting to 10–100% of the marker count of the original dataset. Using RPsH, we also created a pseudo-haploid dataset for comparison. We calculated the pairwise kinship coefficient matrix and corrected the estimated kinship coefficients by the marker overlap fraction. Using the pairwise matrix of 1020 individuals, we plotted the 519,690 kinship coefficients between all combinations of individuals for the diploid and haploid dataset (Fig. 5).

The variance of the corrected kinship coefficient depends on the marker overlap fraction between the test individuals (Fig. 5). Since the marker overlap fractions between any two ancient samples are different applying a pre-defined kinship coefficient threshold to identify relatives would lead to decreased sensitivity or specificity depending on the marker overlap fraction. In other words, the statistical power to differentiate relatives from unrelated depends on the marker overlap fraction and the same threshold should

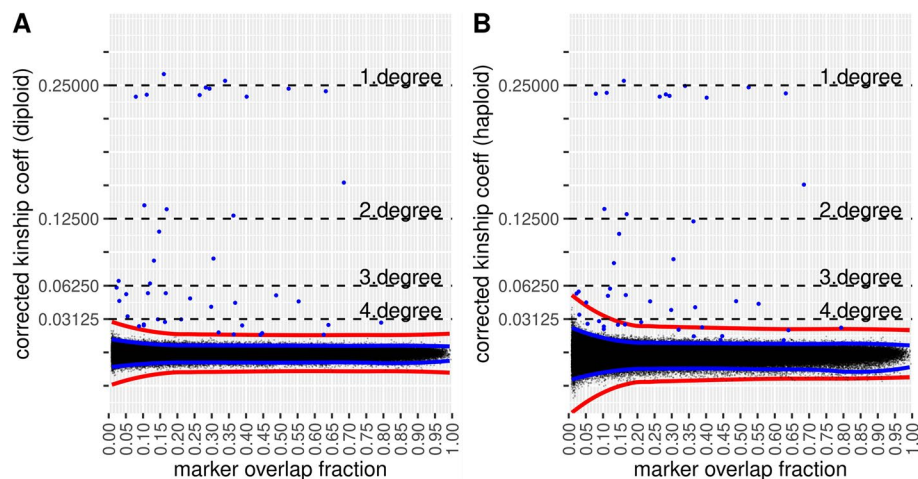


Fig. 5 Corrected kinship coefficients calculated between 1020 EUR and EAS individuals from **A** diploid and **B** haploid dataset. In both cases, individuals were marker depleted between 100,000 and 1.2M markers to simulate partially typed data. The red line represents 6 sigma threshold from the mean. Individual kinship coefficients below the threshold are displayed as small black dots, while individuals above the threshold are displayed as larger blue dots. In case of haploid data, we marked all kinship coefficients that were above the 6 sigma threshold in the diploid dataset for better comparison. The blue lines show the 99% confidence interval of estimated kinship coefficients between unrelated individuals

not be applied. However, based on the experimental variance of the corrected kinship coefficient observed in the analyzed dataset, the Z score (N standard deviation from the mean) could be used as a criteria to differentiate relatives from unrelated with the same sensitivity and specificity independent of the marker overlap fraction. Since in this experiment we could not exclude errors due to missing genome components from the reference populations, we used a conservative $N=6$ sigma threshold to identify biological differences. As expected, the haploid dataset resulted in higher variance due to random information loss especially at very low ($<5\%$) marker overlap fractions. Although the marker overlap fraction correction differs two magnitudes (0.007–0.996) between very low and high marker overlap fraction sample pairs, the analysis shows that similar to the corrected kinship coefficient between relatives (Additional file 2: Table S1) the correction itself does not significantly multiply the error rate of unrelated samples. Most of the technical errors are expected to be the result of using very sparse data to regress out IBS to identify IBD fragments by PC-Relate. In spite of the higher variance, the estimated kinship coefficients show very high correlation between the fully typed ($\sim 1.2\text{M}$ high-quality diploid markers) diploid and the corrected coefficients of the partially typed diploid and haploid datasets ($R=0.9998$ and $R=0.9993$ respectively) compared to the correlation with the uncorrected estimates ($R=0.749$ and $R=0.751$; Additional file 5: Table S4). Our result suggests that the applied pseudo-haploidization and correction in the marker depleted experimental data does not introduce overall bias. Our method identified all known 1st- and 2nd-degree relatives that were included in the analyzed subset of 1KG EUR/EAS individuals and indicated a couple of additional distant 3rd or 4th relatives (Additional file 5: Table S4). Our analysis shows that 4th-degree relatives are expected to be above the 6 sigma threshold in the diploid data except at very low overlapping marker fractions ($<2\% \sim 17,000$ overlapping markers). In case of haploid data, establishment of 3rd-degree relatedness is possible even from low marker overlap fractions, and establishment of fourth-degree relatedness is possible when the sample pair has $>10\%$ marker overlap fraction (equal to roughly $\sim 85,000$ markers). However, in case of 4th-degree relatives depending on the overlapping genotyping fraction, the estimated confidence interval of corrected kinship coefficient, and true biological variation, it is expected to have more false positive/negative and uncertain kinship estimations at low marker overlap fractions.

Kinship analysis of ancient samples with known relations using kinship coefficient correction

To show that our methodology is also suitable for ancient data, we analyzed low-coverage ancient sequences with known family relations. In the first example, we present the analysis of a known father-son (first degree) relation of two Medieval samples. From both remains, we had two types of biological samples, bone powder taken from the teeth and from the *pars petrosa*. From the father, we had three parallel DNA isolates and NGS libraries, two prepared from *pars petrosa* and an additional one from teeth. From the offspring, we had one DNA isolate and NGS library prepared from both types of biological samples. Altogether, we had 3+2 NGS sequences with largely different genome coverages ($0.87\times\text{--}11.9\times$) from these two ancient individuals. Accordingly, we assessed the robustness of our correction method on the 6

Table 1 Correction of kinship coefficient for different genome coverage of ancient samples with known first-degree relation. High, medium, and low refer to coverage levels

Sample 1	Sample 2	Marker overlap fraction	Relation	Expected kinship coeff	Uncorrected kinship coeff	Corrected kinship coeff
Father high	Child high	0.933894	First degree	0.25	0.22883	0.24502
Father high	Child low	0.377769	First degree	0.25	0.09179	0.24297
Father medium	Child high	0.730011	First degree	0.25	0.17604	0.24114
Father medium	Child low	0.296127	First degree	0.25	0.07150	0.24144
Father low	Child high	0.511753	First degree	0.25	0.12239	0.23915
Father low	Child low	0.207545	First degree	0.25	0.05014	0.24160
Father high	Father medium	0.778247	Sample matching	0.5	0.38284	0.49193
Father high	Father low	0.545866	Sample matching	0.5	0.27020	0.49499
Father medium	Father low	0.427389	Sample matching	0.5	0.20863	0.48814
Child high	Child low	0.354511	Sample matching	0.5	0.17286	0.48761
Sample	Sample type	Coverage	Typed marker count (1240k)			
Father high	Teeth	11.997	1,148,973			
Father mid	Petrosa	3.057	896,623			
Father low	Petrosa	1.546	630,405			
Child high	Petrosa	5.510	1,075,845			
Child low	Petrosa	0.879	435,005			

Table 2 Kinship analysis of a large ancient Corded Ware family with multiple 1st- to 4th-degree of relations

Sample 1	Sample 2	Marker overlap fraction	Relation	Expected kinship coeff	Uncorrected kinship coeff	Corrected kinship coeff
I1538	I1541	0.039828	First degree	0.25	0.01036	0.26006
I1540	I1541	0.079863	First degree	0.25	0.01959	0.24534
I1534	I1541	0.047882	Second degree	0.125	0.00715	0.14938
I1538	I1534	0.025735	Second degree	0.125	0.00335	0.13003
I1538	I1540	0.042478	Second degree	0.125	0.00456	0.10730
I1541	I0104	0.216185	Second degree	0.125	0.03394	0.15700
I1534	I1540	0.049813	Third degree	0.0625	0.00405	0.08123
I1538	I0104	0.107634	Third degree	0.0625	0.00846	0.07864
I1540	I0104	0.22405	Third degree	0.0625	0.01812	0.08088
I1534	I0104	0.129456	Fourth degree	0.03125	0.00645	0.04982
Sample ID	Coverage	Typed marker count (1240k)				
I0104	4.184	962767				
I1534	0.158	164095				
I1538	0.126	135269				
I1540	0.298	285866				
I1541	0.294	276299				

combinations of these datasets. We present the uncorrected and corrected kinship coefficients calculated from this data in Table 1.

In the second example, we reanalyzed a published group of five related males from the Corded Ware Culture (2500–2050 BCE) with first-, second-, third-, and

fourth-degree kinship relations [19]. In Table 2, we present the family relations with uncorrected and corrected kinship coefficients calculated by our methodology from public ancient 1240k data.

These individuals were analyzed in the original READ manuscript [26] where relations were identified up to the 2nd degree except the one between I1538 and I1540, and all of the 3rd and 4th relations were inferred from only the family relations.

Kinship analysis of ancient samples from the AADR 1240K dataset

We performed kinship analysis on 2136 ancient Eurasian individuals from the AADR 1240K dataset that had more than 100K genotyped markers. Since no manual curation or additional matching reference population was used, we filtered potential relatives above 0.046875 corrected kinship coefficient (~3rd–4th degree of kinship). Our analysis identified 410 related individuals in 184 kin groups (Additional file 6: Table S5). All sample duplicates ($N=26$), and joint datasets ($N=30$) of the same sample were identified. Curiously, we identified sample duplicates with different master IDs in the AADR dataset published in different manuscripts (I1526-NEO232; I7782-NEO298; I8295-NEO230; I8296-NEO231) where all four sample pairs were from the same geological site and belonged to the same population and haploid typing was identical or nearly identical as some branch defining markers were likely missing due to coverage differences. We also identified a likely sample mix [27] where two individuals (MJ-15 Ukraine_IA_Western-Scythian.SG and MJ-35 Ukraine_Cimmerians_o2.SG) had 0.5 corrected kinship coefficient equivalent with sample match (or monozygotic twin) but had different population assignment. All of these individuals had same sex and identical/nearly identical mitochondrial and Y haplogroups as well. Furthermore, we identified all of the 111 previously identified kinship relations from the AADR dataset. Three uncertain (1st or 2nd) relatives indicated in the AADR dataset (I8502, I8524; MK5001, MK5004 and KBD001, KBD002) could be classified as 2nd-degree relatives by our analysis. We reclassified three kin pairs indicated as 1st-degree relatives as 2nd-degree relatives (RISE1163, RISE1169; RISE1168, RISE1173 and RISE1168, RISE1169). In addition to the published data, within the 184 kin groups our approach indicated 6 new 1st-degree, 108 2nd-degree, 144 3rd-degree, and 40 4th-degree relations between a total of 279 new relatives (Additional file 6: Table S5). In a few cases, when an appropriate reference population was not present in the dataset, it is not possible to establish appropriate kinship relations as the IBS of minor genetic components cannot be regressed out. Consequently, in those cases, the correction resulted in invalid distant 3rd–4th-degree kinship relations highlighted with red in Additional file 6: Table S5.

To test the sensitivity of our analysis, we used READ [26] to validate our findings. As READ depends on the proper reference population and uses a global threshold to distinguish between unrelated and potential kins, the join set of 2136 individuals cannot be analyzed together. We selected the top 10 populations with the highest number of individuals and performed the READ analysis separately. READ identified relatives up to the 2nd degree. In the selected populations, READ identified no additional relatives compared to our methodology. For each identified relative, the degree of kinship was matching between the two methods. In this comparison, our method indicated one additional 2nd-degree relation (AITI_95_d and AITI_98) that was missed by READ. In this case, the

samples have very low genome coverage ($0.145\times$ and $0.402\times$) and only $\sim 0.05\%$ marker overlap fraction. While the corrected kinship coefficient was significantly above the 3rd-degree (0.0625) relation, it was less (0.1004) than the expected 0.125 corresponding to 2nd degree suggesting that these relatives share less than expected genome portions due to true biological variation (Additional file 7: Table S6). We had very similar scenario in case of the missed 2nd-degree relation between the I1538 and I1540 CWC individuals (Table 2) suggesting that READ is less sensitive when the marker overlap is low and the shared genome fraction differs significantly from the statistically expected mean.

Discussion

Identification of relatives from the genomic data of ancestors is of great interest as it allows the study of family relationships, but it is also a precondition for most population genetic analyses to exclude close relatives from datasets (e.g., ADMIXTURE, PCA). To date, the best analysis tools were able to indicate mainly first- and second-degree relatedness from very low-coverage ancient samples [26, 28–30]. Based on simulated data, lcMLkin can accurately infer kinship up to the 3rd degree from $2\times$ genome coverage when the F_{ST} is low between the reference population and analyzed data [28]. However, the majority of aDNA data is below $2\times$ genome coverage. In these data, most markers are represented by one read/genotype only. It is untested whether it is possible to infer comparable diploid genotype likelihoods suitable for lcMLkin from very low-coverage data. The recent heuristic method READ (Relationship Estimation from Ancient DNA) infers relatedness up to 2nd degree from as low as $0.1\times$ coverage sequence data [26]. In the most comprehensive AADR ancient genome data set [23], the majority of the indicated kinship relations are 1st degree and the handful of indicated 2nd-degree relations in all cases are uncertain. These samples are labeled with 1d.or.2d.rel tag.

Diploid variant calling and genotype likelihood-based methods with the extra information of rare alleles allow better phasing and identification of IBD fragments leading to improved kinship coefficient estimations from deeply genotyped WGS data. Accordingly, some methods attempt to infer genotype likelihoods or diploid genotype calls from low-mid genome coverage ($2\text{--}4\times$) data [8, 28, 31]. KING, a method that was developed to be used for fast and robust kinship coefficient estimation from low amounts of fully typed diploid markers ($5\text{--}150\text{k}$), can infer up to 3rd-degree relations from approximately 150k markers or 1st–2nd-degree relation from even as low as 5k diploid markers [8]. Even though these tools are used to analyze low marker count ancient samples, the assumption implicit in these methods that the data is sufficiently high-quality diploid is often false in case of low marker count extremely low-coverage ancient samples. Accordingly, when comparing samples of different genome coverage, the inferred genotype likelihoods or diploid variants from low/variable genome coverage samples could lead to major bias.

To overcome these difficulties and mitigate the main genotyping biases in case of low-coverage ancient samples, we used a combination of strategies to account for the effects caused by PMD and varying low genome coverage. We used random allele sampling that is the gold standard methodology when performing PCA and other population genetic analyses on ancient samples, as it leads to statistically equal genotype likelihoods of genotyped markers regardless of the genome coverage. To avoid excessive, variable amounts

of false positive variants due to the variable rate of PMD, exogenous DNA contamination, and technical errors (alignment artifacts), we restricted our analysis to the already known biallelic, high-frequency, and population-informative SNPs of the 1240K AADR dataset. This strategy perfectly aligned with our choice of kinship analysis method since the PC-Relate algorithm uses PCA to differentiate between IBD/IBS fragments.

We have demonstrated that random pseudo-haploidization of data in our analysis pipeline does not affect the result of kinship analysis (Additional file 2: Table S1). This is also confirmed by the PCA analysis, showing that the same modern individual from diploid or different pseudo-haploidized data had nearly identical PCA components (Additional file 7: Figure S1).

Overlapping marker fraction, according to our study, is the major factor influencing the calculated kinship coefficient of partially genotyped samples in our analysis pipeline. Our simulations revealed that the overlapping marker fraction and the calculated kinship coefficient had a strong linear correlation (Fig. 1, Additional file 2: Table S1). Although the PC-Relate algorithm does not require the specification of the underlying population structure of the analyzed relatives, we have shown that a proper reference set is required for the analysis. As expected, the samples' own reference population resulted in proper kinship coefficients, but using reference from a different super-population corrupted the results (Fig. 3). On the other hand, using the samples' super-population as reference resulted in comparable although slightly higher kinship coefficients compared to the proper reference population (Additional file 7: Figure S2) proving the robustness of the PC-Relate algorithm. This reference bias is not amplified by the applied correction for marker overlap (Fig. 4); however, it could lead to the false identification of distant relatives. We also tested the effect of reference population choice in a complex Creole/European admixed Cabo Verdean-Hungarian family with known 1st- to 5th-degree family relations. We have shown that the best result is achieved when all super-populations of the sources are included in the reference population set (Fig. 4). Comparing the analyses of pseudo-haploid and diploid data for this complex admixed family confirmed the robustness of our approach, as we got nearly identical results (Additional file 4: Table S3).

In the statistical evaluation using the downsampled modern diploid/pseudo-haploid data, we simulated marker counts similar to aDNA data. We applied the same minimum 100,000 genotyped markers per individual threshold that was used in the analysis of 2136 selected ancient individuals from the AADR dataset. This equals roughly $0.08 \times$ genome coverage considering the ~ 1.15 M autosomal markers of the 1240K marker set. Thus, the simulated data had similar marker counts and distribution as the analyzed AADR dataset. Accordingly, pairwise marker overlap was $<5\%$ ($<57,000$ markers) between 3.72 and 3.46% of the analyzed sample pairs (ancient and modern respectively). Our analysis shows that when proper reference population is available, the applied method is suitable to identify relations up to the 4th degree from low to high coverage mixed samples (Fig. 5).

We also confirmed the robustness of our methodology on real ancient data with known family relations. Our analysis showed that in the case of a medieval Hungarian family, a general modern European reference super-population gave appropriate results. Despite the fact that the uncorrected kinship coefficients varied highly due to the different genome coverages, our methodology resulted in reproducible corrected kinship

coefficients consistent with the known family relation in each case (Table 1). In the second example, we reanalyzed published kinship relations from Corded Ware Culture samples [26]. Compared to the READ software which could indicate relations up to the second degree of kinship and even missed one second-degree relation, our approach could properly identify all relations up to 4th degree from this large ancient family with very low/variable genome coverages ($0.12\times-4.18\times$), underlining the efficiency and usefulness of our approach (Table 2).

Our results exposed both the advantages and the limitations of our method. Although RPsH combined with the choice of the 1240K marker set in our study allowed us to overcome genotyping bias of low-coverage ancient samples, it clearly restricts the analysis to populations that are properly represented by these markers. In the PC-Relate algorithm, PCA is used to regress out the population-specific IBS components. Using linear regression to fit individuals to the model, all the remaining non-regressed PC components are calculated as IBD. Thus, insufficient amount of reference individuals, improper or missing population components in the reference, or marker sets that are lacking informative markers of the tests lead to underestimation of IBS and inflated kinship coefficient estimation. The greater the difference between the structure of related individuals and the reference populations, the greater fraction of IBS is accounted incorrectly as IBD which can seriously bias small kinship coefficients representing very distant kinship relations. Accordingly, the current 1240K marker set is less suitable for the analysis of extremely old samples, and for small isolated populations, because these supposedly have less informative markers in this marker set, and also have insufficient reference populations in the current genome databases. Furthermore, while PC-Relate kinship coefficient estimator is known to be appropriate even in inbred populations [24], we have to caution that in case of inbred or small drifting populations extra care has to be taken to confirm that the test individuals are analyzed with their own reference population. When no prior knowledge exists on the reference population, F_{ST} or FastNGSAdmix [32] analysis could be used as an objective method to select individuals best matching our test individual's genome structure as a reference population.

Genotyping error simulations show that approximately double error rate compared to typical experimental aDNA data leads to 5.4–10.4% proportionally lower corrected kinship coefficient than the expected kinship coefficient in our workflow (Fig. 2, Additional file 3: Table S2). The mean corrected kinship coefficient of the validated sample dups and 1st relatives of the experimental 2136 AADR individual was 0.48 and 0.24 respectively (approximately ~4% lower from the expected). This is in accordance with the mean X contamination rate (1.28%) of ancient individuals of the AADR V42.2 dataset suggesting that our kinship-estimation method can be safely used on typical aDNA data. Nevertheless, analysis of highly contaminated (CRITICAL/FAIL) samples containing higher rate of genotyping errors (>5%) could lead to underestimation of corrected kinship coefficient and as a result to underestimation of the degree of relation especially in case when the relatives share less than the expected IBD fragments due to true biological variation.

The unsupervised analysis of 2136 ancient individuals of the 1240K AADR dataset (Additional file 6: Table S5) demonstrated that our method can identify real 1st–4th degree of relatedness from very low-coverage ancient damaged samples and fails only when the proper reference population is not present in the dataset. Comparison

with READ showed that our method has better sensitivity, offers improved performance, and scales better on multi-core machines (Additional file 7: Table S6). On the other hand, our results show that the 1240K marker set was sufficient to properly analyze 4000-year-old ancient Corded Ware Culture individuals with a modern Eurasian reference population, suggesting that the majority of the high-frequency EUR informative markers were already present at this age.

According to our results, the used method had slight downward (2–4%) bias in the analyzed 1KG dups and first-degree relatives and also in the validated first-degree ancient samples. However, this downward bias is also present in the kinship coefficient estimation of fully typed diploid 1KG relatives suggesting that the original PC-Relate algorithm and not the applied correction or pseudo-haploidization is accountable for this bias. This is also supported by our simulations on the corrected kinship coefficient calculated from marker depleted pseudo-haploid data and the original fully typed diploid data (Additional file 2: Table S1, Fig. 1B) and the very high correlation between the kinship coefficient calculation of marker depleted pseudo-haploidized and the original fully typed 1KG data (Additional file 5: Table S4). On the other hand, the mean of the corrected kinship coefficient of the indicated 2nd-degree relatives ($n=119$) of experimental AADR data is 0.1274 (Additional file 6: Table S5) that is a slightly over the expected value (~2% relative difference) suggesting that the bias could originate from more than a single factor.

Our analysis revealed new possibilities to improve kinship analysis from low-coverage ancient data. Diploid typing with pre-capture enrichment could result in higher sensitivity even at lower marker overlap fraction as seen in Fig. 5. However, this is only feasible when a sufficient number of individuals are available from the matching reference populations. According to our analysis, ~50–100 unrelated individuals are sufficient as a reference in case of modern samples. We speculate that in case of populations with less complex genome structures (like pre-iron age populations), a smaller number of unrelated individuals could likely represent the population structure properly. This is also demonstrated in the case of the validated relations of the analyzed CWC individuals where the analysis resulted comparable kinship coefficient estimates using modern EUR individuals as a reference or the 25 Czech, Latvian, Estonian, and German CWC individuals of the AADR dataset (Table 2, Additional file 6: Table S5). Alternatively, using larger marker sets would increase the number of overlapping markers between individuals resulting in higher sensitivity from the already available low-coverage WGS data. The increasing number of aDNA studies should identify proper reference populations and suitable high-frequency marker sets for cases that are difficult to analyze at present.

To facilitate the evaluation and use of our approach, we provide a practical workflow (Fig. 6, Additional file 8: Note S1) for kinship analysis of low-coverage genome data.

Our workflow is based on publicly available free software ANGSD, PLINK, and PCAngsd. Additionally, we also provide the correctKin tool [<https://github.com/zmaroti/correctKin>] to import and shape data, calculate the pairwise overlapping marker fraction, and filter relatives based on the empirical error model.

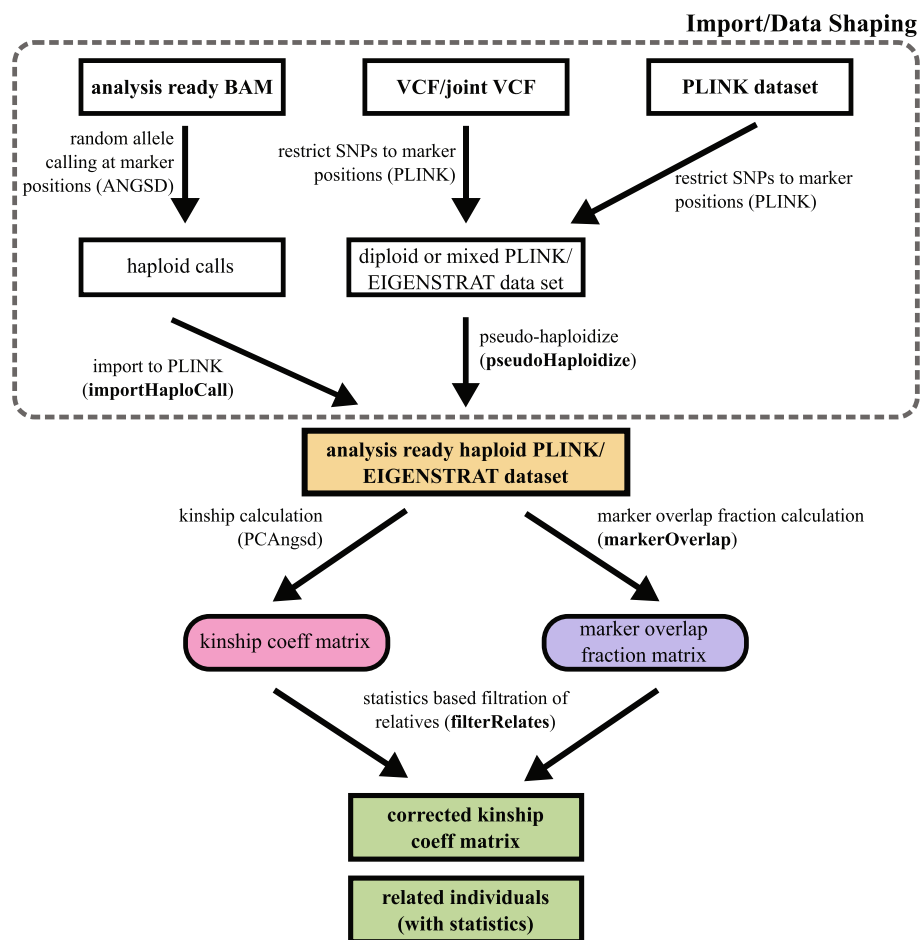


Fig. 6 Step-by-step workflow to analyze kinship relation of low-coverage ancient/modern samples from various data sources. Additional tools presented in this manuscript are denoted with bold letters

Conclusions

In summary, our proposed methodology is capable of reliably identifying the relatedness up to the 4th degree from low-coverage genome data, redefining the limits of kinship analysis from low-coverage ancient or badly degraded forensic WGS data.

Methods

Used software and datasets

The software requirements and the detailed instructions to perform the analysis workflow from various data sources are described in Additional file 8: Note S1.

In all of our analysis, we used the genome coordinates of 1240K SNP set from Allen Ancient DNA Resource (AADR) [23]. For marker overlap simulations, we used two different full-typed modern datasets: the 1000 Genomes Project Phase 3 data [25], and a large admixed Cabo Verdean-Hungarian family of known pedigree with first- (siblings), second- (half siblings), and fifth-degree relatives from our anonymized clinical biobank. The variants of the joint VCF (Variant Call Format) files were filtered for the 1240K SNP coordinates and imported into plink 1.9 binary format [33, 34].

To test the effect of genome coverage on the estimated kinship coefficients from real ancient data, we used our unpublished 1240K genotype data of a known medieval parent offspring. The dataset contains low-coverage partially typed pseudo-haploid genotype data from 3+2 separate library preparations from different biological samples of the analyzed individuals. We deposited the unpublished medieval datasets in PLINK 1240K binary format presented in this manuscript at Zenodo [34].

The public AADR V42.4 1240K dataset [23] was used to validate our methodology on a wide variety of ancient individuals. We included only ancient samples with more than 100K genotyped markers ($N=2810$). We excluded samples older than 8000BC ($N=216$) as older samples were very few and were lacking proper number of samples as a reference population (Additional file 7: Figure S4). To avoid analyzing individuals with very few and/or inappropriate reference populations, we restricted the analyzed samples by their geo location (in between the Longitude -12 – 120 and Latitude 28 – 65) excluding 458 individuals (Additional file 7: Figure S5). After filtration, the resulting dataset contained 2136 ancient individuals (Additional file 6: Table S5).

New bioinformatics tools

To aid easy importing, manipulating, and analyzing the genotype data in our proposed workflow, we created the essential tools:

- *importHaploCall* to import pseudo-haploid genotype calls from the ANGSD
- *pseudoHaplo* to perform RPsH using a diploid dataset
- *markerOverlap* to calculate the pairwise marker overlap fraction matrix
- *filterRelates* to correct kinship coefficient, and filter relatives based on error model and/or hard kinship coefficient threshold

To study the effect of partially genotyped markers in a controlled fashion and comparing results with the analysis of the fully genotyped modern samples we used

- *depleteMarkers* to simulate the desired marker overlap fraction between selected samples
- *depleteIndivs* to simulate a random partially genotyped sample cohort

The tools work with the main genotype data formats (PLINK, EIGENSTRAT, PACKEDANCESTRYMAP). We documented the usage and options of the new tools with command examples in Additional file 8: Note S1. Tools are available in zenodo and the GitHub repository (<https://github.com/zmaroti/correctKin>) [35, 36].

Random pseudo-haploidization and pairwise overlapping marker fraction calculation

We defined the overlapping marker fraction between two samples as the number of markers typed in both samples divided by the number of all markers in the dataset.

Using our “*pseudoHaplo*” tool, we created 100 randomly pseudo-haploidized datasets from the fully typed modern diploid dataset using different random seeds. In all of the presented examples, we used our own tool “*markerOverlap*” to calculate the pairwise

overlapping marker fraction matrix of samples used for the kinship coefficient correction [35, 36].

Principal component analysis

We selected the GBR, TSI, IBS, and FIN populations from 1KG dataset (404 samples) and randomized the diploid dataset with three different seeds. We performed smartpca [37, 38] analysis on the original diploid and the three random pseudo-haploidized dataset with the “*inbreed: YES*” option. We used the R (version 4.0.5) [39] and the ggplot2 R package (3.3.5) [40] to visualize the individuals on the PC1 and PC2 axes (Additional file 7: Figure S1).

Simulating the effect of low coverage from fully typed modern datasets

To study the effect of coverage and the resulting lower genotyping percentage on the kinship coefficient calculation in a controlled fashion, we used “*depleteMarkers*” to randomly deplete markers from a fully typed (PLINK, EIGENSTRAT) dataset, resulting in the desired percentage of marker overlap between two samples [35, 36]. Using this tool, we simulated the overlapping marker fraction in the selected samples in the range of 5–100% with step of 5 percentages.

To assess the technical error of low/variable coverage data on the whole workflow, we selected 1020 fully typed diploid Eurasian samples (CEU, IBS, GBR, FIN, TSI, CDX, CHB, CHS, JPT, KHV populations) of the 1KG phase 3 dataset. We applied “*depleteIndivs*” to create a random, partially typed sample cohort with marker count between 100,000 and the full 1,150,639 markers. From the partially typed diploid dataset, we also created a pseudo-haploidized dataset using the “*pseudoHaplo*” tool [35, 36]. We performed kinship analysis with PCAngsd and corrected the estimated kinship coefficients according to the marker overlap fraction of sample pairs on the partially genotyped datasets. We compared the results with the estimated kinship coefficients using the original fully typed diploid dataset.

Simulation of aDNA-related genotyping errors

PLINK and EIGENSTRAT data format were designed for biallelic markers. There are only 4 possible allelic states (homozygote major allele, homozygote minor allele, heterozygote major/minor, and missing), thus any other nucleotide that is different than the minor or major allele cannot be represented and the allelic state of samples with invalid alleles are set to the “missing” state at such marker positions.

Based on the data format restriction, the three typical aDNA-related genotype errors can be simulated in the following ways for pseudo-haploid PLINK dataset:

- Post mortem damage; if the C->T or G->A conversion leads to different nucleotide than the minor or major allele, the state is set to “missing,” otherwise, if the minor and major alleles are C/T, T/C, G/A or A/G, the homozygote minor and major states are flipped.
- Exogenous (non-human DNA) contamination; since the exogenous DNA consist of mainly DNA of microorganisms (usually in ancestral state), it leads to excessive

homozygote major allele, thus random subsets of markers are set to the homozygote major allele state.

- Endogenous (human DNA) contamination; random subsets of markers are set to the state of the same markers genotyped from another sample (theoretically, the largest number of SNPs are expected to be flipped in case the population has the largest F_{ST} from the test individuals or practically if the test is contaminated with sample from a very different population).

From most population genetic analyses, highly contaminated samples are excluded. In the comprehensive AADR ancient dataset, the following criteria is used to mark bad-quality sequences:

- ANGSD X contamination (applicable only for males) 0.02–0.05="QUESTIONABLE", >0.05="QUESTIONABLE_CRITICAL" or "FAIL".
- mtcontam <0.8 is "QUESTIONABLE_CRITICAL", 0.8–0.95 is "QUESTIONABLE", and 0.95–0.98 is recorded but "PASS", gets overridden by ANGSD X contamination.

Accordingly, the 1240K v42.2 AADR dataset ($n=3589$ ancient samples) 157 is marked CRITICAL/FAIL (>5% error rate), while the mean of the X contamination rate of all ancient samples is 1.28%.

We simulated the three different errors separately and also made a mixed case where all three error types were introduced in equal amount leading to the same total error rate. In all cases, we had maximum total genotyping error rate of 5% (the threshold of CRITICAL/FAIL tag of the AADR criteria). Accordingly, each sample had random 0–5% genotyping error, leading to an overall ~2.5% genotype error rate of the whole dataset that is roughly the double of the genotyping error rate of the experimental AADR aDNA dataset. In each simulation, we used 100 different randomizations with different random seed and calculated the mean and SD of the corrected kinship coefficients.

Uncorrected kinship coefficient estimation

Kinship coefficient estimation was performed by the PCAngsd [41] software (version 0.99) from the ANGSD package [7] that implements a fast parallelized kinship calculation from PLINK or EIGENSTRAT format based on the PC-Relate algorithm [24] with the “-inbreed 1 -kinship” parameters.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-02882-4>.

Additional file 1: Figure S1. Comparison of PCA using diploid and pseudo-haploid data. **Figure S2.** Effect of reference population (same super-population) on kinship coefficient. **Figure S3.** Pedigree of complex admixed modern family. **Figure S4.** The date distribution of ancient AADR individuals. **Figure S5.** Geographical distribution of ancient AADR individuals.

Additional file 2: Table S1. Effect of RPsH and marker overlap on kinship coefficient.

Additional file 3: Table S2. Effect of genotyping errors on kinship coefficient.

Additional file 4: Table S3. Comparison of kinship analysis of a complex admixed family using pseudo-haploid and diploid data.

Additional file 5: Table S4. Comparison of kinship coefficient correction using experimental 1KG Phase III data.

Additional file 6: Table S5. correctKin analysis of 2126 ancient AADR individuals.

Additional file 7: Table S6. Comparison of correctKin with READ software.

Additional file 8: Note S1. Installation requirements and a step-by-step method description including detailed usage examples to perform correctKin analysis from various data sources.

Additional file 9. Review history.

Acknowledgements

The authors would like to thank Michael F. Nagy and Peter L. Nagy for their valuable suggestions and their help in proof-reading the final manuscript.

Review history

The review history is available as Additional file 9.

Peer review information

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

Authors' contributions

Conceptualization, methodology, software: Z.M. Formal analysis: E.Ny. and Z.M. Resources E.N., R.M.L., and T.T. Interpretation of results: Z.M., T.K., and E.Ny. Visualization: O.S., E. Ny., and Z.M. Writing original draft: E.Ny., Z.M., and T.K. All authors took part revising the results and contributed to the final manuscript. E.Ny. and T.K. contributed equally to this study. All author(s) read and approved the final manuscript.

Funding

Open access funding provided by University of Szeged. E.Ny. was supported by the ÚNKP-21-3-SZTE-67 New National Excellence Program of the Ministry for Innovation and Technology, from the source of the National Research, Development and Innovation Fund. This research was funded by grants from the National Research, Development and Innovation Office (TUDFO/5157-1/2019-ITM; TKP2020-NKA-23 to E.N.). Rui M. Lima was supported by the NTP-NFTÖ-18 Scholarship. Z. M. was supported by University of Szeged Open Access Fund, award number 5771.

Availability of data and materials

The extra tools described and used in this manuscript were deposited to Zenodo [35]. Our software with potential future updates is also available in the GitHub repository (<https://github.com/zmaroti/correctKin>) [36]. The raw data of the medieval dataset can be accessed at European Nucleotide Archive (<http://www.ebi.ac.uk/ena>) under the accession number PRJEB47418 [42]. We deposited the medieval dataset and the subset of 1KG Phase III individuals presented in this manuscript at in PLINK 1240K binary format [34]. The complex admixed modern family data that supports the finding of this study are not publicly available due to them containing information that could compromise research participant privacy/consent. The anonymized 1240K AADR PLINK subset of this data without phenotype information is available on request from the corresponding author (ZM). Access to the data requires noncommercial, research only usage approved by your ethical committee.

Declarations

Ethics approval and consent to participate

Written informed consent was obtained from the individuals of the family included in this study. Genetic analysis of the included family was approved by the National Scientific and Research Ethics Committee of the Medical Research Council of Hungary (ETT TUKEB, registration number 28676-7/2017/EÜIG). All procedures performed in studies involving human participants were in accordance with the ethical standards of the National Scientific and Research Ethics Committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 20 April 2022 Accepted: 17 February 2023

Published online: 28 February 2023

References

1. Thompson EA. The estimation of pairwise relationships. *Ann Hum Genet.* 1975;39:173–88.
2. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet.* 2015;16:33–44.
3. Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, et al. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics.* 2017;207:75–82.
4. 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;19:318–26.
5. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *Am J Hum Genet.* 2011;88:173–82.
6. Li H, Glusman G, Huff C, Caballero J, Roach JC. Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS One.* 2014;9:1–6.

7. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014;15:1–13.
8. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26:2867–73.
9. Seidman DN, Shenoy SA, Kim M, Babu R, Woods IG, Dyer TD, et al. Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. *Am J Hum Genet*. 2020;106:453–66.
10. Jeong C, Balanovsky O, Lukianova E, Kahbatkzy N, Flegontov P, Zaporozhchenko V, et al. The genetic history of admixture across inner Eurasia. *Nat Ecol Evol*. 2019;3:966–76.
11. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, et al. The formation of human populations in south and Central Asia. *Science*. 2019;365(6457). <https://www.science.org/doi/epdf/10.1126/science.aat7487>.
12. Harney É, Nayak A, Patterson N, Joglekar P, Mushrif-Tripathy V, Mallick S, et al. Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. *Nat Commun*. 2019;10:1–10.
13. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of bronze age Eurasia. *Nature*. 2015;522:167–72.
14. 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;522:207–11.
15. Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, et al. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science*. 2012;336:466–9.
16. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient near east. *Nature*. 2016;536:419–24.
17. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. 2014;513:409–13.
18. Günther T, Valdiosera C, Malmström H, Ureña I, Rodríguez-Varela R, Sverrisdóttir ÓO, et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc Natl Acad Sci U S A*. 2015;112:11917–22.
19. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015;528:499–503.
20. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. 2015;524:216–9.
21. Saag L, Varul L, Scheib CL, Stenderup J, Allentoft ME, Saag L, et al. Extensive farming in Estonia started through a sex-biased migration from the steppe. *Curr Biol*. 2017;27:2185–2193.e6.
22. Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, et al. 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr Biol*. 2017;27:3202–3208.e9.
23. Allen Ancient DNA Resource (V42.4, 02-2020). <https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data>.
24. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. *Am J Hum Genet*. 2016;98:127–48.
25. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
26. Kuhn JMM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. *PLoS One*. 2018;13:1–21.
27. Järve M, Saag L, Scheib CL, Pathak AK, Montinaro F, Pagani L, et al. Shifts in the genetic landscape of the Western Eurasian steppe associated with the beginning and end of the Scythian dominance. *Curr Biol*. 2019;29:2430–2441.e10.
28. Lipatov M, Sanjeev K, Patro R, Veeramah KR. Maximum likelihood estimation of biological relatedness from low coverage sequencing data; 2015. p. 1–20.
29. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, et al. Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat Commun*. 2017;8:14115. <https://doi.org/10.1038/ncomms14115>.
30. Ringbauer H, Steinrücken M, Fehren-Schmitz L, Reich D. Increased rate of close-kin unions in the Central Andes in the half millennium before European contact. *Curr Biol*. 2020;30:R980–1.
31. Severson AL, Korneliussen TS, Moltke I. LocalNgsRelate: a software tool for inferring IBD sharing along the genome between pairs of individuals from low-depth NGS data. *Bioinformatics*. 2022;38:1159–61.
32. Jørsboe E, Hanghøj K, Albrechtsen A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics*. 2017;33:3148–50.
33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
34. Nyerki E, Kalmár T, Schütz O, Lima RM, Neparáczki E, Török T, et al. correctKin PLINK data sets; 2022. <https://doi.org/10.5281/zenodo.7333251>.
35. Maróti Z. correctKin software; 2022. <https://doi.org/10.5281/zenodo.7330922>.
36. Maróti Z. correctKin github repository; 2022. <https://github.com/zmaroti/correctKin/>.
37. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2:2074–93.
38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–9.
39. Team Rs. RStudio: integrated development environment for R; 2019.
40. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
41. Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*. 2018;210:719–31.
42. Neparáczki E, Kis L, Maróti Z, Kovács B, Varga GIB, Makoldi M, et al. The genetic legacy of the Hunyadi descendants. *Heliyon*. 2022;8:e11731.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

II.

Gergely I.B. Varga, Lilla Alida Kristóf, Kitty Maár, Luca Kis, Oszkár Schütz, Orsolya Váradi, Bence Kovács, Alexandra Gînguță, Balázs Tihanyi, Péter L. Nagy, Zoltán Maróti, **Emil Nyerki**, Tibor Török, Endre Neparácski. *The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy* ,Journal of Genetics and Genomics, Volume 50, Issue 1, (2023) **IF:5,9 Q1**



Letter to the editor

The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy



The house of Árpád ruled Hungary for more than four centuries, establishing dynastical relationships with numerous European noble- and ruling houses, and giving many Catholic saints and blessed to the Catholic Church. They were one of the significant dynasties of Medieval Europe (Kristó and Makk, 1996). The knight-king Saint Ladislaus I (c.1040–1095, reign: 1077–1095) is one of the most outstanding kings of this dynasty. In addition to the hand relic (the Holy Right) of King Saint Stephanus I (c. 975–1038, reign: 1000/1001–1038), the skull relic in the Saint Ladislaus' Herma (Fig. 1A–1C) preserved in the Cathedral of Győr, is one of the most important relics for Hungarians (Klaniczai, 2000). By consolidating state power and strengthening Christianity, King Saint Ladislaus completed the work begun by Saint Stephanus I. His charismatic personality, strategic leadership and military talents resulted in the cessation of internal power struggles and foreign military threats (Engel, 2001; Kristóf et al., 2017). He was seen as the embodiment of the knight-king ideal to be emulated all over Europe (Pow, 2018). He was buried in Várad (Oradea, located in modern day Romania) in 1095 and was canonized in 1192 at the request of King Béla III. At this time, his body was exhumed to prepare relics from his skull and other skeletal remains as was customary for the time (Solymosi, 2017). The skull relic had a turbulent history: in 1406 the wooden herm containing it was damaged in a fire, but miraculously the skull has been preserved unharmed. Later it was placed into the current Herma (Fig. 1A) created during the reign of Sigismund of Luxembourg (1387–1437) (Varga, 2017). In the 16th century, the relic had to be rescued from Várad due to the ravage of Transylvania by the protestants. After passing through Prague, Pozsony (now Bratislava, Slovakia) and Veszprém, it reached its current location in the Cathedral of Győr in the first decades of the 17th century. The Herma is still held in high esteem and Saint Ladislaus' memory is cherished with annual processions in the city and throughout the Carpathian Basin (Kristóf, 2017a).

As the above-mentioned events raised doubts concerning the authenticity of the Herma by historians and archaeologists, we set out to compare the Y chromosome sequence of the skull to that of King Béla III, the only Árpád dynasty king with known and identified remains, for whom whole genome data is available (Olasz et al., 2019; Nagy et al., 2021). We extracted DNA from the pars petrosa and the root of a molar tooth from the skull held inside Saint Ladislaus' Herma and built double stranded DNA libraries for Illumina sequencing. Based on low coverage shotgun sequencing results, the library from the petrosa contained minimal amounts of human DNA (0.3%). As we tried to avoid damaging the valuable relic, we drilled very carefully, which may not have reached the most compact

part of the petrous bone. This could be the explanation of such a low endogenous DNA content. At the same time, the library from the tooth root (SZTLF) had 72.2% human reads, so it was suitable for whole genome sequencing. Due to the high quality of the library, we could obtain 16.4-fold mean coverage over the genome, and 7.4-fold coverage of Y-chromosome after collapsing the paired-end reads. This allowed for a high precision phylogenetic analysis of the royal paternal lineage. Sequencing data statistics for Saint Ladislaus and three other reported royal genomes, Béla III (HU3B), his wife Queen Anna (HUAA) and another Árpád dynasty member (HU52) are shown in Table S1 and in Nagy et al. (2021).

The Y chromosome haplogroup of the sample was determined with the Yleaf software, based on the SNP database of ISOGG 2020. Y chromosome sequence of SZTLF was derived for the sub-haplogroup R1a1a1b2a2a1c3~ based on the marker R-Y2632. Analysing onward the sequence manually, it could be assigned to the sub-Hg R-ARP (R1a1a1b2a2a1c3a3b) by virtue of the SNPs ARP1, ARP2, ARP3, ARP4, ARP5, ARP6, ARP7, ARP8 and ARP9 (Nagy et al., 2021). This result supports the originality of Saint Ladislaus' relic, as the Y chromosome of the skull belongs to the exclusive Hg of the Árpád dynasty (for Y chromosome SNP data of SZTLF, see Table S2).

This sub-Hg belongs to the R-Z2125 clade (Nagy et al., 2021), which can be seen in individuals from the Middle-Late Bronze Age on the Caspian Steppe, connected to the Potapovka, Sintashta and Andronovo cultures (Narasimhan et al., 2019). In the Iron Age it was detected in the Turan basin (Narasimhan et al., 2019) and in Scytho-Siberians of the Minusinsk Basin (Narasimhan et al., 2019), later among the Xiongnu (Keyser et al., 2021) and up until the Middle Ages in Mongolia (Jeong et al., 2020) which indicates an eastward and southward spread of the Hg. The phylogenetic analysis of Hg R-Z2123 (a derived subclade of R-Z2125) from modern Y chromosomal data, suggested a Bronze Age BMAC origin of this sub-Hg (Narasimhan et al., 2019; Nagy et al., 2021). The first appearance of R-Z2125 in the Carpathian Basin was detected in 5th-century-CE European Huns (Maróti et al., 2022), and 7th–8th-century-CE Avars (Maróti et al., 2022), but it also arrived with the conquering Hungarians in the 9th–10th century (Neparáczki et al., 2019; Maróti et al., 2022), including Árpád and his family (Nagy et al., 2021). Based on Y-STR markers, a previous study suggested phylogenetic connection between the Árpád house and a Xiongnu elite family (Keyser et al., 2021), supporting the mythical Hun origin of the dynasty.

Nagy et al. (2021) published the Y chromosome sequence of an additional skeletal remain (HU52) proven to be a member of the Árpád dynasty, derived from the Royal Basilica of Székesfehérvár,

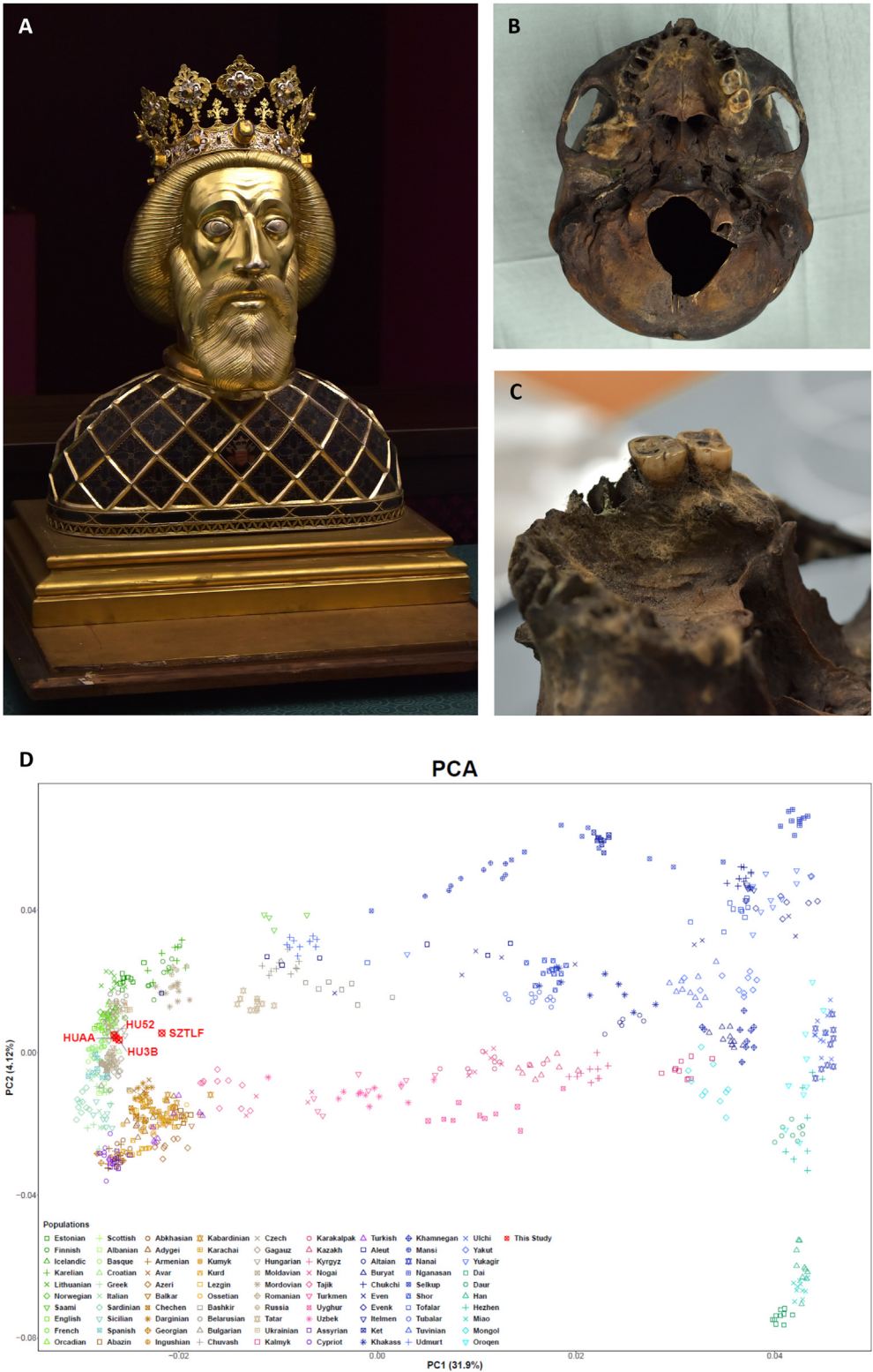


Fig. 1. Saint Ladislaus' Herma and the PCA plot of the royal family members. **A:** The relic made during the reign of Sigismund of Luxemburg (1387–1437). **B** and **C:** The skull of the relic from different viewpoints. **D:** PCA plot of the Árpád dynasty genomes. SZTLF, HU3B, HUA and HU52 genome data (red) were projected on the background generated from modern Eurasian genomes (Table S4). St Ladislaus is shifted eastward from the modern Central European population cluster and other family members suggesting a greater Eastern genomic content.

the burial site of most Hungarian kings, until the Turkish occupation of the city. Since genetic data was available only from a single other Árpád house member, Béla III, the identity of HU52 could not be determined (Olasz et al., 2019; Nagy et al., 2021). Even the identity of the remains HU3B had been questioned by several historians, claiming that the skeleton rather belongs to King Colomanus I than Béla III (e.g., Tóth, 2006). Thus, we examined the kinship relations of the available Árpád dynasty genomes to determine their exact position on the family tree (Fig. S1). The kinship analysis was carried out with an improved approach, which is capable of inferring relatedness up to the 4th degree from low coverage ancient genomes with high certainty (Nyerki et al., 2022). The corrected kinship coefficient indicated 5th level kinship between Saint Ladislaus and HU3B (Table S3), which exactly corresponds to the known family relations of Saint Ladislaus and Béla III. Though the kinship estimation of the novel method becomes uncertain over 4th degree, however, it seems beyond question that the HU3B remains could not belong to Colomanus, who was a second-degree relative of Saint Ladislaus (his nephew) (Fig. S1). Thus, our kinship results indisputably settle the identification of HU3B with Béla III.

The HU52 individual was buried next to Béla III and his wife (Olasz et al., 2019), and although his Y-chromosome unequivocally identified him as a member of the Árpád house (Nagy et al., 2021; Table S2C) his identity remained unknown. Our kinship analysis detected second degree relation between HU52 and Béla III, and further than 4th degree distance from Saint Ladislaus (Table S3). In order to specify the location of HU52 on the family tree, we included in the kinship analysis the genome sequence of Queen Anna, the wife of Béla III. The analysis revealed a two-step distance between the queen and HU52 as well (Table S3). The equal distance of the individual to both the king and the queen confutes the previous suggestion that HU52 could be King Béla II (the grandfather of Béla III) (Olasz et al., 2019). On the contrary it definitively shows that the unidentified man was one of the five grandsons of the royal couple (Fig. S1). It is known that HU52 was buried in the Royal Basilica of Székesfehérvár and his age at death was estimated multiple times by multiple experts to be between 20 and 30 years (Olasz et al., 2019). According to historical data about the burials of the dynasty, the identity of the potential grandsons can be narrowed. King Ladislaus III died as a child, whereas King Béla IV passed away in his 60's and was buried in Esztergom (Kádár, 2012). The age at death of all three known brothers of Béla IV fit in the estimated anthropological age of HU52, however Prince Colomanus was laid to rest in Kloštar Ivanić, Croatia while Stephanus the Posthumous was buried in San Michele in Isola, in Venice, Italy (Kádár, 2012). The third one, Prince Andreas died in Halych, but his final resting place is unknown, thus he is the best candidate to be identified with HU52.

To further characterize the genetic ancestry of the Árpád dynasty members we carried out population genomic analyses. First, we performed Principal Component Analysis (PCA), and projected the studied genomes onto the axes computed from modern Eurasian individuals (Figs. 1D and S2; Table S4). On the PCA plot SZTLF was shifted eastward from the cloud of modern European populations, while the other studied royals were projected near modern Hungarians and Croatians (Fig. S2). Latter results are in agreement with previously reported data from Béla III (Wang et al., 2021), while Saint Ladislaus' PCA position suggests that he retained more from the Eastern genomic heritage of the dynasty than his later relatives.

Next, we carried out qpAdm analysis, to determine the potential genomic ancestries of Saint Ladislaus, as well as of Béla III and HU52. We assembled a source population list including 45 plausible Iron Age and Medieval populations from Central and Northern Europe, as well as from the Eurasian Steppe (for details see Supplementary data; Table S5). We used a Right (reference) population list, which had been optimized for the conquering Hungarian elite (Maróti et al., 2022).

The qpAdm resulted in 168 plausible models for Saint Ladislaus, and after excluding suboptimal models with model competition (Narasimhan et al., 2019; Maróti et al., 2022), we obtained 25 comparable passing models (for further information see Supplementary data). In each model, the major (78%–94%) ancestry showed European origin, while the minor (6%–22%) component was derived from East-Central Asia (Table S5). The best *P*-value models indicated 85% Germany early medieval and 15% conquering Hungarian elite ancestries. This result perfectly reconciles with historical data, as among the maternal ancestors of Saint Ladislaus Polish, Russian, Germanic and Czech persons were recorded, while his paternal lineage goes back to Árpád, leader of the conquering Hungarians.

Model competition gave six passing models for Béla III of which the three with highest *P*-values indicate a major (73%–88%) Lithuania_Late_Antiquity component and a minor (12%–27%) component typical for the local population of the Carpathian Basin (Maróti et al., 2022). The best model for HU52 suggested very similar genome composition (85% Lithuania_Late_Antiquity and 15% local Carpathian basin). Alternative models with lower *P*-values indicated Viking, Sweden Iron Age and Hungarian Langobard affinities. These data also fit historical records, as Polish, Byzantine, Germanic, Russian, French and Serbian maternal ancestors had been documented in the family.

We validated the qpAdm results with *f*₄-statistics: *f*₄ (SZTLF, other royal member; European pop, Mbuti). In accordance with the qpAdm results *f*₄ values indicate that Saint Ladislaus has higher affinity to the conquering Hungarians than the other Royal members have (Table S6).

Our genome analysis results confirm that the Árpád dynasty was not a foreigner family appointed to rule the Hungarians but originated from the same ethnic group as other members of the conquering Hungarian elite, and that then their Central Asian genomes were progressively attenuated during the centuries through marriages with Central European royal families.

This study reports a successful archaeogenomic analysis of a saint's remains. Saint Ladislaus was one of the most significant figures of the Medieval history of Hungary and Europe as well. He was a legendary king and a Catholic saint in one person; thus, the verification of his most important relic has a huge historical and spiritual relevance. Moreover, with the genetic and genomic data of the saint king the personal identity of Béla III's remains, and thanks to this, the identity of Queen Anna remains could be established as well. Thus, the family tree of the Árpád dynasty was fixed on three certain points without any discrepancy, providing a solid base for the personal identification of the Hungarian royal remains in the future. The genomic analyses of the royal family members are in line with the reported conquering Hungarian-Hun origin of the dynasty in harmony with their Y-chromosomal phylogenetic connections.

Data availability

The raw nucleotide sequence data of the samples were deposited to the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>) under the accession number: PRJEB51515.

Conflict of interest

The authors declare no competing financial interests.

Acknowledgment

We are grateful to diocesan András Veres and to commissary Ferenc Reisner from the Diocese of Győr to enable the accession of the bone material. We thank to Miklós Kásler, Gábor Horváth-Lugossy

and László Tamás Vizi for their support and encouragement and Szabolcs Tóth for his administrative work. We are thankful to Michael F. Nagy for revising the manuscript. This research was funded by grants from the National Research, Development and Innovation Office (TUDFO/5157-1/2019-ITM and TKP2020-NKA-23 to E.N.) and NKA 499108/00003 to E.N.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jgg.2022.06.008>.

References

- Engel, P., 2001. The Realm of Saint Stephen. A History of Medieval Hungary. In: Ayton, Andrew (ed.), Pálosfalvi, Tamás (translated). I. B. Tauris Publisher, London-New York, pp. 895–1526.
- Jeong, C., Wang, K., Wilkin, S., Taylor, W.T.T., Miller, B.K., Bemmman, J.H., Stahl, R., Chiovelli, C., Knolle, F., Ulziibayar, S., et al., 2020. A dynamic 6,000-year genetic history of eurasia's eastern Steppe. *Cell* 183, 890–904.
- Kádár, T., 2012. Az Árpád-házi uralkodók és az országlásuk idején hercegi címmel tartományi különhatárat gyakorolt külhoni, fejedelmi származású előkelők, valamint azok családtagjainak elhalálozási és temetkezési adatai 997–1301 között. *Fons (Forráskutatás és Történeti Segédtudományok) XIX* 1, 57–108.
- Keyser, C., Zvenigorosky, V., Gonzalez, A., Fauser, J.-L., Jagorel, F., Gérard, P., Tsagaan, T., Duchesne, S., Crubézy, E., Ludes, B., 2021. Genetic evidence suggests a sense of family, parity and conquest in the Xiongnu Iron Age nomads of Mongolia. *Hum. Genet.* 140, 349–359.
- Klaniczai, G., 2000. Az Uralkodók Szentsége a Középkorban. Balassi Kiadó, Budapest, pp. 114–153.
- Kristóf, G., Makk, F., 1996. Az Árpád-Ház Uralkodói. I. P.C. Könyvek. ISBN 963-7930-97-3.
- Kristóf, L.A., Lukácsi, Z., Patonay, L. (eds.), 2017. Szent Király, Lovagkirály. A Szent László-herma és koponyaereklye vizsgálatai. Győri Hittudományi Főiskola, Győr, pp. 10–13.
- Kristóf, L.A., Lukácsi, Z., Réthelyi, M., Molnár, E., Pálfi, G., Pap, I., Patonay, L., 2017a. Mumifikálási szokások Magyarországon. Királyok, főnemesek, főpapok holttestének konzerválása a középkortól. In: Kristóf, L.A., Lukácsi, Z., Patonay, L. (eds.), Szent Király, Lovagkirály. A Szent László-herma és Koponyaereklye Vizsgálatai. Győri Hittudományi Főiskola, Győr, pp. 203–225.
- Maróti, Z., Neparáczki, E., Schütz, O., Maár, K., Varga, G.I.B., Kovács, B., Kalmár, T., Nyerki, E., Nagy, I., Latinovics, D., et al., 2022. The genetic origin of Huns, Avars, and conquering Hungarians. *Curr. Biol.* 32, 738–741.
- Nagy, P.L., Olasz, J., Neparáczki, E., Rouse, N., Kapuria, K., Cano, S., Chen, H., Di Cristofaro, J., Runfeldt, G., Ekomasova, N., et al., 2021. Determination of the phylogenetic origins of the Árpád Dynasty based on Y chromosome sequencing of Béla the Third. *Eur. J. Hum. Genet.* 29, 164–172.
- Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al., 2019. The formation of human populations in South and Central Asia. *Science* 365, eaat7487.
- Neparáczki, E., Maróti, Z., Kalmár, T., Maár, K., Nagy, I., Latinovics, D., Kustár, A., Pálfi, G., Molnár, E., Marcsik, A., et al., 2019. Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin. *Sci. Rep.* 9, 16569.
- Nyerki, E., Kalmár, T., Schütz, O., Lima, R.M., Neparáczki, E., Török, T., Maróti, Z., 2022. An optimized method to infer relatedness up to the 5th degree from low coverage ancient human genomes. *bioRxiv*. <https://doi.org/10.1101/2022.02.11.480116>.
- Olasz, J., Seidenberg, V., Hummel, S., Szentirmay, Z., Szabados, G., Melegh, B., Kásler, M., 2019. DNA profiling of Hungarian King Béla III and other skeletal remains originating from the Royal Basilica of Székesfehérvár. *Archaeol. Anthropol. Sci.* 11, 1345–1357.
- Pow, S., 2018. Evolving identities: a connection between royal patronage of dynastic saints' cult and Arthurian literature in the twelfth century. In: *Annual of Medieval Studies at CEU. Central European University/Archaeolingua*, pp. 65–74.
- Solymosi, L., 2017. Szent László király sírja, kultusza és szentté avatása. In: Kristóf, Lilla Alida, Lukácsi, Zoltán, Patonay, Lajos (eds.), Szent Király, Lovagkirály. A Szent László-herma és koponyaereklye vizsgálatai, vol. 36. Győri Hittudományi Főiskola, Győr.
- Tóth, E., 2006. III. Béla vagy Kálmán? A székesfehérvári királysír azonosításáról. *Folia Archaeologica* 52, 141–161.
- Varga, P., 2017. A Szent László-herma leírása ötvösművészeti szempontból. In: Kristóf, Lilla Alida, Lukácsi, Zoltán, Patonay, Lajos (eds.), Szent Király, Lovagkirály. A Szent László-herma és koponyaereklye vizsgálatai. Győri Hittudományi Főiskola, Győr, pp. 72–85.
- Wang, C.-C., Posth, C., Furtwängler, A., Sümegi, K., Bánfai, Z., Kásler, M., Krause, J., Melegh, B., 2021. Genome-wide autosomal, mtDNA, and Y chromosome analysis of King Béla III of the Hungarian Arpad dynasty. *Sci. Rep.* 11, 19210.
- Gergely I.B. Varga*
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Lilla Alida Kristóf
Doctoral School of History, University of Szeged, Szeged, Hungary
- Kitti Maár
Department of Genetics, University of Szeged, H-6726 Szeged, Hungary
- Luca Kis
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Department of Biological Anthropology, University of Szeged, H-6726 Szeged, Hungary
- Oszkár Schütz
Department of Genetics, University of Szeged, H-6726 Szeged, Hungary
- Orsolya Váradi
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Department of Biological Anthropology, University of Szeged, H-6726 Szeged, Hungary
- Bence Kovács
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Alexandra Gînguță
Department of Genetics, University of Szeged, H-6726 Szeged, Hungary
- Department of Molecular Biology and Biotechnology, Faculty of Biology and Geology, Babes-Bolyai University, 400006 Cluj-Napoca, Romania
- Balázs Tihanyi
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Department of Biological Anthropology, University of Szeged, H-6726 Szeged, Hungary
- Péter L. Nagy
Praxis Genomics LLC, Atlanta, USA
- Zoltán Maróti, Emil Nyerki
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Department of Pediatrics and Pediatric Health Center, University of Szeged, H-6725 Szeged, Hungary
- Tibor Török, Endre Neparáczki**
Department of Archaeogenetics, Institute of Hungarian Research, H-1014 Budapest, Hungary
- Department of Genetics, University of Szeged, H-6726 Szeged, Hungary

* Corresponding author.

** Corresponding author.

E-mail addresses: varga.gergely@mki.gov.hu (G.I.B. Varga), neparaczki.endre@mki.gov.hu (E. Neparáczki).

7 April 2022

Available online 6 July 2022