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

List of publications

I. **Emil Nyerki**, Tibor Kalmár, Oszkár Schütz, M. Lima Rui, Endre Neparáczki, 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 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) **IF:5,9 Q1**

Introduction

Kinship analysis is a crucial tool in both archaeogenetics and forensic sciences for determining relationships between individuals based on genetic information.

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 [1]. 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 [2].

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)).

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

Materials and Methods

Used Data

In all analyses conducted, the genome coordinates of the 1240K SNP set from Allen Ancient DNA Resource (AADR) [3] were used. For marker overlap simulations, two distinct fully-typed modern datasets were employed: the 1000 Genomes Project Phase 3 data [4], 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.

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.

The public AADR V42.4 1240K dataset [3] 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 ([5] 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 ([5] Additional file 7: Figure S5). Following filtration, the dataset comprised 2136 ancient individuals ([5] Additional file 6: Table S5).

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.

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. 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: Iberian(IBS),Great Britain (GBR), Finnish (FIN), Toscani (TSI), Utah residents with Northern and Western European Ancestry (CEU), Dai Chinese (CDX), Han Chinese in Bejing (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.

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
- Exogenous (non-human DNA) contamination
- Endogenous (human DNA) contamination

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.

References

1. 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

2. 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

3. Allen Ancient DNA Resource. No Title. https://reich.hms.harvard.edu/allen-ancient-dna-

resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data

4. 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

5. 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