# Estimating genetic kin relationships in prehistoric populations

**Jose Manuel Monroy Kuhn**[1,2], **Mattias Jakobsson**[1,*], **and Torsten Günther**[1,*]

[1]Uppsala University, Department of Organismal Biology and SciLifeLab, Uppsala, Sweden
[2]University of Freiburg, Institute of Biology I (Zoology), Freiburg, Germany
[*]Corresponding authors: mattias.jakobsson@ebc.uu.se, torsten.guenther@ebc.uu.se

## ABSTRACT

Archaeogenomic research has proven to be a valuable tool to trace migrations of historic and prehistoric individuals and groups, whereas relationships within a group or burial site have been more challenging to investigate. Knowing the genetic kinship of historic and prehistoric individuals would give important insights into social structures of ancient and historic cultures. Most archaeogenetic research concerning kinship has been restricted to uniparental markers, while studies using genome-wide information were mainly focused on comparisons between populations. Applications which infer the degree of relationship based on modern-day DNA information typically require diploid SNP data. Low concentration of endogenous DNA, fragmentation and other post-mortem damage to ancient DNA (aDNA) makes the application of such tools unfeasible for most archaeological samples. To infer family relationships for degraded samples, we developed the software READ (Relationship Estimation from Ancient DNA). We show that our heuristic approach can successfully infer up to second degree of relationship with as little as 0.1x shotgun coverage per genome for pairs of individuals. We uncover previously unknown relationships by applying READ to published aDNA datasets from different cultures. In particular we find a group of five closely related males from the same Corded Ware culture site in Germany suggesting patrilocality, which highlights the possibility to uncover social structures of ancient populations by applying READ to genome-wide aDNA data.

## Introduction

An individual's genome is a mosaic of different segments inherited from our various direct ancestors. These segments, shared between individuals, can be referred to as identical by descent (IBD). Knowledge about IBD segments has been used for haplotype phasing[1,2], heritability estimation[3,4], population history[5], inference of natural selection[6] and to estimate the degree of biological relationship among individuals[7]. A number of methods have been developed to estimate the degree of biological relationship by inferring IBD from SNP genotype or whole genome sequencing data. The methods for estimating relationship levels implemented in PLINK[8], SNPduo[9], ERSA[10,11], KING[12], REAP[13] and GRAB[14] greatly benefit from genome wide diploid data, information about phase, recombination maps and population allele frequency, and are sometimes able to successfully infer relationships up to 11th degree[11].

Knowing whether a pair of individuals is directly related or not, and estimating the degree of relationship is of interest in various fields: Genome-wide association studies and population genetic analyses often try to exclude related individuals since they do not represent statistically independent samples; in forensics, archaeology and genealogy, individuals and their relatives can be identified based on DNA extracted from human remains[15,16]; Breeders and conservation biologists are interested in the relatedness of mating individuals[17,18]. Current methods present significant limitations for the analysis of degraded samples. Especially in the fields of forensics and archaeology, where specimens are subject to taphonomic processes and postmortem damage resulting in incomplete data due to low concentrations and fragmentation of endogenous DNA in the sample[19]. In archaeology, the analysis of IBD has the potential to provide an independent means to test kinship behavior, biological and/or cultural, on the basis of social organization, socioeconomic dynamics, gender relationships, agency and identity[20], but current methods would be restricted to exceptionally preserved samples. In forensic science and practice, the dominant approach has been to type several short tandem repeat (STR) markers, which in most cases provide sufficient information for relatedness assessment, but the STRs might be hard to type in degraded samples[21]. In addition to nuclear STRs, mitochondrial and Y-chromosome haplogroups have been widely used to infer family relationships (e.g.[15,16,22,23]), although they can formally only exclude certain direct relationships since most mitochondrial and Y-chromosome haplogroups are relatively common among unrelated individuals. These uniparental markers can be typed from degraded samples, and can be used to exclude maternal or paternal relationships but not to infer the actual degree of relationship. Genome-wide data, however, can be obtained from degraded samples at a higher success rate than STRs and it can be used to confidently identify individuals[24].

SNP data can be achieved from genotyping experiments (e.g. SNP arrays or RAD sequencing), targeted capture[25] and whole-genome shotgun sequencing (e.g.[26,27]). The field of ancient DNA has developed rapidly over the last few years which allowed sequencing the genomes of extinct hominins[28,29], as well as studying population history in Europe[25–27,30–37] and the peopling of the Americas[36,38,39]. However, both whole-genome shotgun sequencing (e.g.[27,31,32]) and genome-wide SNP capture (e.g.[25,33]) usually achieve coverages $<$1x per informative site for most individuals which makes diploid genotype calls at all sites virtually impossible. Methods to infer relationships, however, rely on such ideal data to identify IBD blocks which is a major limitation for applying them to ancient DNA data.

However, even low coverage data contain information about the degree of relationship. To utilize this information, we developed READ (Relationship Estimation from Ancient DNA), a heuristic method to infer family relationships up to second degree from samples with extremely low coverage. The method is tested on publicly available data with known relationship which we sub-sample to resemble the properties of degraded samples. We also apply our pipeline to a number of ancient samples from the literature and confidently classify individual pairs as being related.

# Results

## Method Outline

We divide the genome into non-overlapping windows of 1 Mbps each and for each pair of individuals calculate the proportion of non-matching alleles inside each window $P0$. The genome-wide distribution of $P0$ is then normalized using the average $P0$ of an unrelated pair of individuals which accounts for effects of SNP ascertainment and population diversity. Depending on the normalized proportion of shared alleles, each pair of individuals is classified as unrelated, second-degree (i.e. nephew/niece-uncle/aunt, grandparent-grandchild or half-siblings), first-degree (parent-offspring or siblings) or identical individuals/identical twins (Figure 1).

## Simulations based on modern data with known relationship

READ's performance was tested on 1,326 individuals of 15 different populations from the phase 3 data of the 1000 genomes project[40]. A total of 86,336 pairwise comparisons were tested. READ showed an overall good performance with false negative and false positive rates below four percent for as little as 1,000 overlapping SNPs (Figure 2A). The proportion of related individuals that were classified as related but not to the correct degree increased with less data. Separating the error rates between first and second degree relatives shows that most of this increase is due to first degree relatives classified as second degree relatives when the number of SNPs is low (Figure 2B). False positive rates are low for both degrees of relationship and false negative rate is below one percent for first degree relatives (Figure 2B and C). The rate of false negatives is considerably high for second degree relatives and it increases up to 39% for low numbers of SNPs (Figure 2C).

## Relationships among prehistoric Eurasians

To investigate READ's performance on empirical aDNA data, we analyzed a large published genotype data set of 230 ancient Eurasians from the Mesolithic, Neolithic and Bronze Age periods[33]. In accordance with the original publications[25,27,33], READ inferred RISE507 and RISE508 to be the same individual and all nine known relationships were correctly identified as first degree relatives (Table 1). In addition to those, READ identified one additional pair of first degree relatives as well as six new second degree relationships. All relatives are from the same location and their radiocarbon dates (if available) are overlapping.

Combining the information obtained from radiocarbon dating, READ as well as uniparental haplotypes can help to narrow down the possible form of relationship. For instance, I0111 (female) and I1530 (male) are inferred to be first degree relatives, which means they are either full-siblings, mother/son or father/daughter. The shared mitochondrial haplogroup (H3ao) makes father/daughter less likely, while the slightly older radiocarbon date for I0111 (2475-2204 calBCE versus 2345-2198 calBCE) rather suggests mother/son than siblings.

READ identified an unknown pair of first degree relationship between two Srubnaya individuals (I0360 and I0354). Notably, Mathieson et al (2015)[33] have excluded I0354 since she was an outlier compared to other Srubnaya individuals. The shared mitochondrial haplogroup (U5a1) and the slightly older age of I0354 make her the putative mother of I0360. The classification of I0360 and I0354 as first degree relatives could be a false positive, but it is very likely that they are at least second degree relatives as the fraction of unrelated individuals wrongly classified as first degrees is extremely low (Figure 2B). Furthermore, a highly distinct genetic background of one of the individuals should rather cause false negatives and not false positives which increases the likelihood that the two individuals are in fact related. I0354 could have been a recent migrant to the region who produced offspring (I0360) with a local male, which would explain both the relationship between I0354 and I0360 and the genomic dissimilarity between I0354 and other Srubnaya individuals.

Particularly interesting is a group of five related males from the Corded Ware site in Esperstedt, Germany (Table 1, Figure 3). Mathieson et al (2015)[33] described two first degree relationships between I1540 and I1541 as well as between I1541 and I1538. Notably, READ missed the second degree relationship between I1540 and I1538, which is likely to be a false negative as the false negative rate for second degree relatives is known to be high (Figure 2C) and the value for that pair (0.91) is only 1.2

standard errors above the threshold for second degree relatives (0.9). Identical radiocarbon dates do not help to indicate a chronological order, but based on their Y chromosomes (all R1a), one can assume that they represent a paternal line of ancestry. I1540 is classified as R1a1 in Mathieson et al (2015)[33], but the specific marker in the Y-chromosome this call is based on (L120) is missing in individuals I1538 and I1541, so they could all carry the same haplotype. In addition to these three individuals, I1534 is a second degree relative of I1538 and I1541, but he was carrier of a different Y-haplogroup (R1b1a2), so a direct paternal relationship can be excluded. I0104 who is a second degree relative to I1541 might also carry the same Y chromosome as I1538, I1540 and I1541, but that cannot be determined due to low coverage in those individuals. In total, 13 Corded Ware individuals from Esperstedt were genotyped, nine of them were males. It is notable that all five related Esperstedt individuals discussed here were males and only one pair of related Corded Ware individuals from Esperstedt involved a female (I1539 and I1532; Table 1).

## Discussion

Several methods to estimate the degree of relationship between pairs of individuals have been developed. For ideal data (i.e. genome-wide diploid data without errors), they successfully infer relationships up to 11th degree[11]. Since such data cannot be obtained from degraded samples, a loss in precision was expected. Estimation of second degree relationships (i.e. niece/nephew-aunt/uncle, grandparent-grandchild, half-siblings) is sufficient to identify individuals belonging to a core family which were buried together. We can show that obtaining as little as 2,500 overlapping common SNPs is enough to classify up to second degree relationships from effectively haploid data. The biggest limitations when using such low numbers of SNPs is the high rate of false negatives for second degree relatives. Therefore, READ can be considered as conservative as false positives are avoided at the cost of a increased false negative rate. This error rate decreases substantially with more data and missing some second degree relationships seems preferable over wrongly inferring relationships for unrelated individuals. It is very unlikely that first degree relatives are classified as unrelated but some second degree relatives might be wrongly classified as unrelated. Shared uniparental haplotypes or a test result close to the threshold (e.g. less than two standard errors difference) could raise such suspicions and might motivate additional sequencing of the samples in question. The number of SNPs required for a positive classification as first degree can be obtained by shotgun sequencing all individuals to a genome coverage of 10% (or 0.1x), which is in reach for most archaeological samples displaying some authentic DNA. More data would be beneficial to avoid false negatives in the case of second degree relatives. Recently developed methods for modern DNA which use genotype-likelihoods to handle the uncertainty of low to medium coverage data require 2x genome coverage to estimate third degree relationships[41,42].

An important part of the READ pipeline is the normalization step. This step makes the classification independent of within population diversity, SNP ascertainment and marker density. This property, however, requires at least one additional and unrelated individual from the same population. In practice, such a supposedly unrelated individual might be sequenced as part of the same study or a pair of individuals from a surrogate population with similar expected diversity as groups from similar cultural and geographical backgrounds show very similar normalization scores (Figure 4). The assignment of all individuals to a population can be checked with established methods as principal component analysis (PCA) or outgroup $f_3$ statistics[39]. Furthermore, obtaining just one unrelated individual (or a pair of unrelated individuals from a surrogate population) seems to be more feasible than obtaining data for a whole population as required by other methods[41]. A certain limitation for all kinship estimation methods is if the sampled population itself cannot be considered homogeneous, for example due to varying degrees of admixture. Only quite recent developments in inferring relationships can efficiently deal with such cases for modern data[43].

We successfully applied READ to data obtained from ancient individuals. READ confidently found all known relationships in the dataset. Furthermore, it identified a number of previously unknown relationships, mainly of second degree. The combination of genomic data, uniparental markers and radiocarbon dating allowed to conclude how two individuals are related to each other. Additional information such as osteological data on the age of the samples or stratigraphic information as burial location or depth could further help to assess the direction of a kinship. Of particular interest was a group of five males from Esperstedt in Germany who were associated with the Corded Ware culture - a culture which arose after large scale sex-biased migrations[25,27,44]. The close relationship of this group of only male individuals from the same location suggest patrilocality and female exogamy, a pattern which has also been concluded from Strontium isotopes at another Corded Ware site just 30 kilometers from Esperstedt[15] and suggested for the culture in general[45]. This represents just one example of how the genetic analysis of relationships can be used to uncover and understand social structures in ancient populations. More data from additional sites, cultures and species other than humans will offer various opportunities for the analysis of relationships based on genome-wide data.

# Materials and Methods

## Approach to detect related individuals

Our approach is based on the methodology used by GRAB[14] which was designed for unphased and diploid genotype or sequencing data. This approach divides the genome into non-overlapping windows of 1 Mbps each and compares for a pair of individuals the alleles inside each window. Each SNP is classified into three different categories: IBS2 when the two alleles are shared, IBS1 when only one allele is shared and IBS0 when no allele is shared. The program calculates the fractions for each category ($P2$, $P1$ and $P0$) per window and, based on certain thresholds, uses them for relationship estimation. GRAB can estimate relationships from 1st to 5th degree.

We assume that our input data stems from whole genome shotgun sequencing of an ancient sample resulting in low coverage sequencing data. Therefore, we only expect to observe one allele per individual and site which is either shared or not shared between the two individuals. READ does not model aDNA damage, so it is expected that the input is carefully filtered, e.g. by restricting to sites known to be polymorphic, by excluding transition sites or by rescaling base qualities before SNP calling[46]. Analogous to GRAB[14], we partition the genome in non-overlapping windows of 1 Mbps and calculate the proportions of haploid mismatches and matches, $P0$ and $P1$, for each window. Since $P0+P1=1$, we can use $P0$ as a single test statistic. To reduce the effect of SNP ascertainment and population diversity, each individual pair's $P0$ scores are normalized by dividing by the average $P0$ score from an unrelated pair of individuals from the same population ascertained in the same way as for the tested pairs. This normalization sets the expected score for an unrelated pair to 1 and we can define classification cutoffs which are independent of the diversity within the particular data set. We define three thresholds to identify pairwise relatedness as unrelated, second-degree (i.e. nephew/niece-uncle/aunt, grandparent-grandchild or half-siblings), first-degree (parent-offspring or siblings) and identical individuals/identical twins. The general work flow and the decision tree used to classify relationships is shown in Figure 1. There are four possible outcomes when running READ: unrelated (normalized $\text{P0}\geq 0.9$), second degree ($0.9\geq$normalized $\text{P0}\geq 0.8$), first degree ($0.8\geq$normalized $\text{P0}\geq 0.65$) and identical twins/identical individuals (normalized $\text{P0}<0.65$) (Figure 1). The cutoffs were chosen to maximize precision in the pseudo-haploidized 1000 genomes dataset (see below) before randomly subsampling SNPs. The option of classifying two individuals as third degree was not implemented as the few known third degree relationships in the empirical datasets showed values similar to unrelated individuals (data not shown). Furthermore, we calculate the standard error of the mean of the distribution of normalized $P0$ scores and use the distance to the cutoffs in multiples of the standard error (similar to a Z score) as a measurement of confidence.

Relationship Estimation from Ancient DNA (READ) was implemented in Python 2.7[47] and GNU R[48]. The input format is TPED/TFAM[8] and READ is publicly available from `https://bitbucket.org/tguenther/read`

## Modern data with reported degrees of relationships

Autosomal Illumina Omni2.5M chip genotype calls from 1326 individuals from 15 different populations were obtained from the 1000 genomes project (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/`)[40]. We used vcftools version 0.1.11[49] to extract autosomal biallelic SNPs with a minor allele frequency of at least 10% (1,156,468 SNPs in total) and to convert the data to TPED/TFAM files. The data set contains pairs of individuals that were reported as related, 851 of them as first degree relationships and 74 as second degree. We randomly sub-sampled 1000, 2500, 5000 and 50000 SNPs and also randomly picked one allele per site in order to mimic extremely low coverage sequencing of ancient samples. READ was then applied to these reduced data sets and the median of all average $P0$s per population was used to normalize scores assuming that this would represent an unrelated pair. Individual pairs with known relationship, their degree of relatedness as well as the relatedness inferred by READ for different data subsets are shown in Supplementary Files **XXX**. Related individuals classified by READ as unrelated were considered as false negatives, unrelated individuals classified as related were considered as false positives and related individuals classified as related but not on the proper degree were considered as incorrect related. The false negative rate was obtained by dividing the number of false negatives by the total number of true related pairs, the false positive rate by dividing the number of false positives by the total number of unrelated pairs and the incorrect related rate by dividing the number of incorrectly classified related pairs by the total number of true related pairs.

## Ancient data

In addition to the modern data, published ancient data was obtained from the study of Mathieson et al. (2015)[33]. The data set consisted of 230 ancient Europeans from a number of publications[25,27,30,31,50,51] as well as new individuals from various time periods during the last 8,500 years. The data set consisted of haploid data for up to 1,209,114 SNPs per individual. We extracted only autosomal data for all individuals and applied READ to each cultural or geographical group (as defined in the original data set of Mathieson et al (2015)[33]) with more than four individuals separately, using the median of all average $P0$s per group for normalization assuming that this would represent an unrelated pair. Mathieson et al (2015)[33] report nine pairs of related individuals and they infer all of them to be first degree relatives.

## References

**1.** Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068–1075 (2008). URL http://dx.doi.org/10.1038/ng.216.

**2.** Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. & Durbin, R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol* **35**, 853–860 (2011). URL http://dx.doi.org/10.1002/gepi.20635.

**3.** Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193–1198 (2012). URL http://dx.doi.org/10.1073/pnas.1119675109.

**4.** Browning, S. R. & Browning, B. L. Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* **46**, 617–633 (2012). URL http://dx.doi.org/10.1146/annurev-genet-110711-155534.

**5.** Ralph, P. & Coop, G. The geography of recent genetic ancestry across europe. *PLoS Biol* **11**, e1001555 (2013). URL http://dx.doi.org/10.1371/journal.pbio.1001555.

**6.** Albrechtsen, A., Moltke, I. & Nielsen, R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**, 295–308 (2010). URL http://dx.doi.org/10.1534/genetics.110.113977.

**7.** Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**, 771–780 (2006). URL http://dx.doi.org/10.1038/nrg1960.

**8.** Purcell, S. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007). URL http://dx.doi.org/10.1086/519795.

**9.** Roberson, E. D. O. & Pevsner, J. Visualization of shared genomic regions and meiotic recombination in high-density snp data. *PLoS One* **4**, e6711 (2009). URL http://dx.doi.org/10.1371/journal.pone.0006711.

**10.** Huff, C. D. *et al.* Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome research* **21**, 768–774 (2011).

**11.** Li, H. *et al.* Relationship estimation from whole-genome sequence data. *PLoS Genet* **10**, e1004144 (2014). URL http://dx.doi.org/10.1371/journal.pgen.1004144.

**12.** Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

**13.** Thornton, T. *et al.* Estimating kinship in admixed populations. *The American Journal of Human Genetics* **91**, 122–138 (2012).

**14.** Li, H., Glusman, G., Huff, C., Caballero, J. & Roach, J. C. Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS One* **9**, e85437 (2014). URL http://dx.doi.org/10.1371/journal.pone.0085437.

**15.** Haak, W. *et al.* Ancient dna, strontium isotopes, and osteological analyses shed light on social and kinship organization of the later stone age. *Proc Natl Acad Sci U S A* **105**, 18226–18231 (2008). URL http://dx.doi.org/10.1073/pnas.0807592105.

**16.** King, T. E. *et al.* Identification of the remains of king richard iii. *Nature communications* **5** (2014).

**17.** Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M. & Bijma, P. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**, 483–496 (2006). URL http://dx.doi.org/10.1534/genetics.105.049940.

**18.** Habier, D., Fernando, R. & Dekkers, J. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397 (2007).

**19.** Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and dna fragmentation in ancient dna. *PLoS One* **7**, e34131 (2012). URL http://dx.doi.org/10.1371/journal.pone.0034131.

**20.** Ensor, B. E. *The archaeology of kinship: Advancing interpretation and contributions to theory* (University of Arizona Press, 2013).

**21.** Canturk, K. M. *et al.* Current status of the use of single-nucleotide polymorphisms in forensic practices. *Genet Test Mol Biomarkers* **18**, 455–460 (2014). URL http://dx.doi.org/10.1089/gtmb.2013.0466.

**22.** Deguilloux, M. *et al.* Ancient dna and kinship analysis of human remains deposited in merovingian necropolis sarcophagi (jau dignac et loirac, france, 7th–8th century ad). *Journal of Archaeological Science* **41**, 399–405 (2014).

**23.** Cui, Y. *et al.* Identification of kinship and occupant status in mongolian noble burials of the yuan dynasty through a multidisciplinary approach. *Philos Trans R Soc Lond B Biol Sci* **370**, 20130378 (2015). URL http://dx.doi.org/10.1098/rstb.2013.0378.

**24.** Hughes-Stamm, S. R., Ashton, K. J. & van Daal, A. Assessment of dna degradation and the genotyping success of highly degraded samples. *International journal of legal medicine* **125**, 341–348 (2011).

**25.** Haak, W. *et al.* Massive migration from the steppe was a source for indo-european languages in europe. *Nature* **522**, 207–211 (2015). URL http://dx.doi.org/10.1038/nature14317.

**26.** Skoglund, P. *et al.* Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science* **336**, 466–469 (2012). URL http://dx.doi.org/10.1126/science.1216304.

**27.** Allentoft, M. E. *et al.* Population genomics of bronze age eurasia. *Nature* **522**, 167–172 (2015). URL http://dx.doi.org/10.1038/nature14507.

**28.** Meyer, M. *et al.* A high-coverage genome sequence from an archaic denisovan individual. *Science* **338**, 222–226 (2012). URL http://dx.doi.org/10.1126/science.1224344.

**29.** Prüfer, K. *et al.* The complete genome sequence of a neanderthal from the altai mountains. *Nature* **505**, 43–49 (2014). URL http://dx.doi.org/10.1038/nature12886.

**30.** Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature* **513**, 409–413 (2014). URL http://dx.doi.org/10.1038/nature13673.

**31.** Skoglund, P. *et al.* Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers. *Science* **344**, 747–750 (2014). URL http://dx.doi.org/10.1126/science.1253448.

**32.** Günther, T. *et al.* Ancient genomes link early farmers from atapuerca in spain to modern-day basques. *Proc Natl Acad Sci U S A* **112**, 11917–11922 (2015). URL http://dx.doi.org/10.1073/pnas.1509851112.

**33.** Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient eurasians. *Nature* **528**, 499–503 (2015). URL http://dx.doi.org/10.1038/nature16152.

**34.** Cassidy, L. M. *et al.* Neolithic and bronze age migration to ireland and establishment of the insular atlantic genome. *Proceedings of the National Academy of Sciences* **113**, 368–373 (2016).

**35.** Hofmanová, Z. *et al.* Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences* 201523951 (2016).

**36.** Slatkin, M. & Racimo, F. Ancient dna and human history. *Proceedings of the National Academy of Sciences* **113**, 6380–6387 (2016).

**37.** Günther, T. & Jakobsson, M. Genes mirror migrations and cultures in prehistoric europe—a population genomic perspective. *Current Opinion in Genetics & Development* **41**, 115–123 (2016).

**38.** Rasmussen, M. *et al.* The genome of a late pleistocene human from a clovis burial site in western montana. *Nature* **506**, 225–229 (2014). URL http://dx.doi.org/10.1038/nature13025.

**39.** Raghavan, M. *et al.* Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature* **505**, 87–91 (2014). URL http://dx.doi.org/10.1038/nature12736.

**40.** Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL http://dx.doi.org/10.1038/nature15393.

**41.** Lipatov, M., Sanjeev, K., Patro, R. & Veeramah, K. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *bioRxiv* 023374 (2015).

**42.** Korneliussen, T. S. & Moltke, I. Ngsrelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* **31**, 4009–4011 (2015). URL http://dx.doi.org/10.1093/bioinformatics/btv509.

**43.** Moltke, I. & Albrechtsen, A. Relateadmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics* **30**, 1027–1028 (2014). URL http://dx.doi.org/10.1093/bioinformatics/btt652.

**44.** Goldberg, A., Günther, T., Rosenberg, N. A. & Jakobsson, M. Familial migration of the neolithic contrasts massive male migration during bronze age in europe inferred from ancient x chromosomes. *bioRxiv* 078360 (2016).

**45.** Sjögren, K.-G., Price, T. D. & Kristiansen, K. Diet and mobility in the corded ware of central europe. *PLoS One* **11**, e0155083 (2016). URL http://dx.doi.org/10.1371/journal.pone.0155083.

**46.** Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapdamage2. 0: fast approximate bayesian estimates of ancient dna damage parameters. *Bioinformatics* btt193 (2013).

**47.** Van Rossum, G. *et al.* Python programming language. In *USENIX Annual Technical Conference*, vol. 41 (2007).

**48.** R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2016). URL https://www.R-project.org/.

**49.** Danecek, P. *et al.* The variant call format and vcftools. *Bioinformatics* **27**, 2156–2158 (2011). URL http://dx.doi.org/10.1093/bioinformatics/btr330.

**50.** Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of european prehistory. *Nat Commun* **5**, 5257 (2014). URL http://dx.doi.org/10.1038/ncomms6257.

**51.** Keller, A. *et al.* New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* **3**, 698 (2012). URL http://dx.doi.org/10.1038/ncomms1701.

## Acknowledgements

We thank Federico Sanchez-Quinto, Jan Storå and Rita Peyroteo Stjerna for comments on the manuscript as well as Gülşah Merve Dal Kılınç and Mehmet Somel for discussions. JMMK received scholarships from the Erasmus Mundus Master Programme in Evolutionary Biology (MEME) and the Consejo Nacional de Ciencia y Tecnología (CONACYT). MJ was supported by an ERC starting grant (no. 311413) and a Swedish Research Council grant. TG was supported by a Wenner-Gren-Foundations fellowship. Computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Projects b2013203 and b2015094.

## Author contributions statement

TG and MJ conceived the study. JMMK and TG designed READ. JMMK implemented READ and conducted simulations. TG analyzed aDNA data. All authors contributed to writing the manuscript.

## Additional information

**Competing financial interests:** The authors declare no competing financial interests.

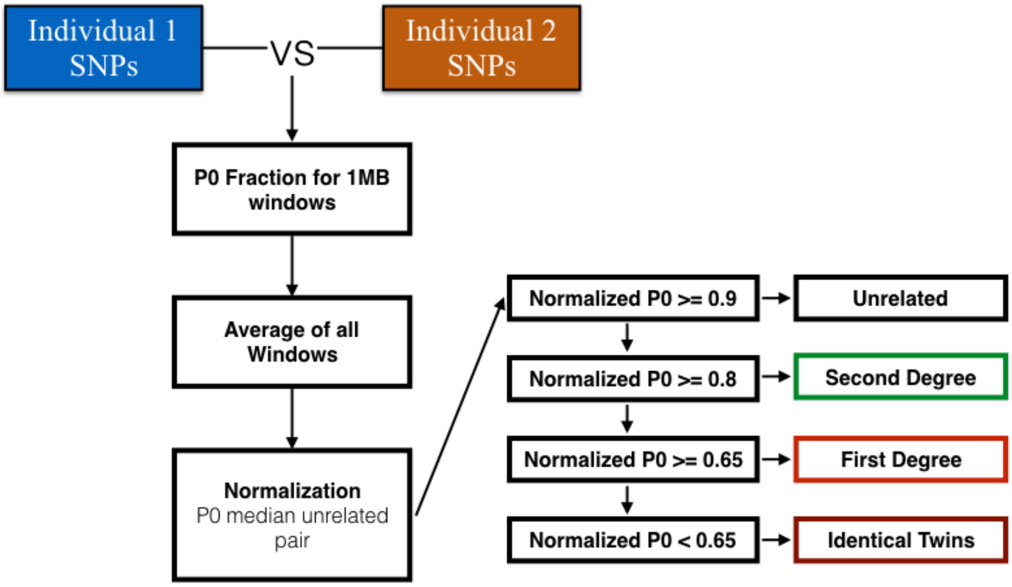

**Figure 1.** Outline of the general READ workflow to estimate the degree of relationship between two individuals.

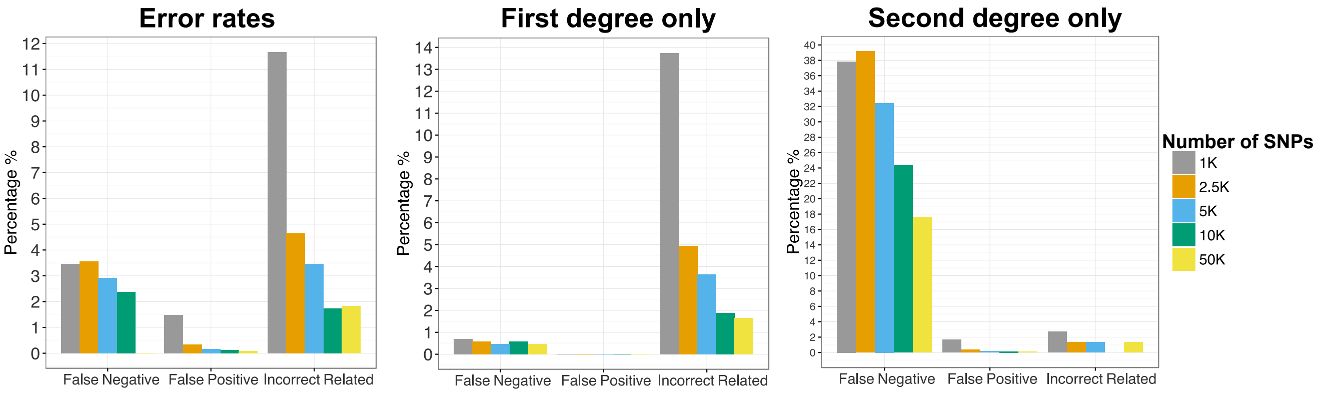

**Figure 2.** Error rates for different numbers of SNPs. The analysis is based on pairs of individuals with known degree of relationship. (A) All pairs of individuals, (B) only first degree relatives and (C) only second degree relatives. Pairs known to be related but classified with the wrong degree are shown as ”Incorrect Related”.

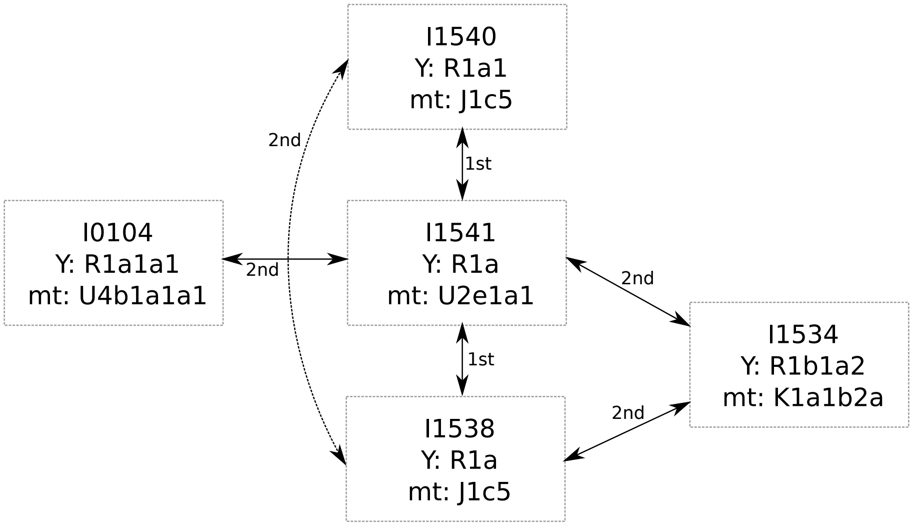

**Figure 3.** Kin-relationship among males at the Corded Ware site in Esperstedt, Germany. The dashed line between I1540 and I1538 shows a second degree relationship missed by READ

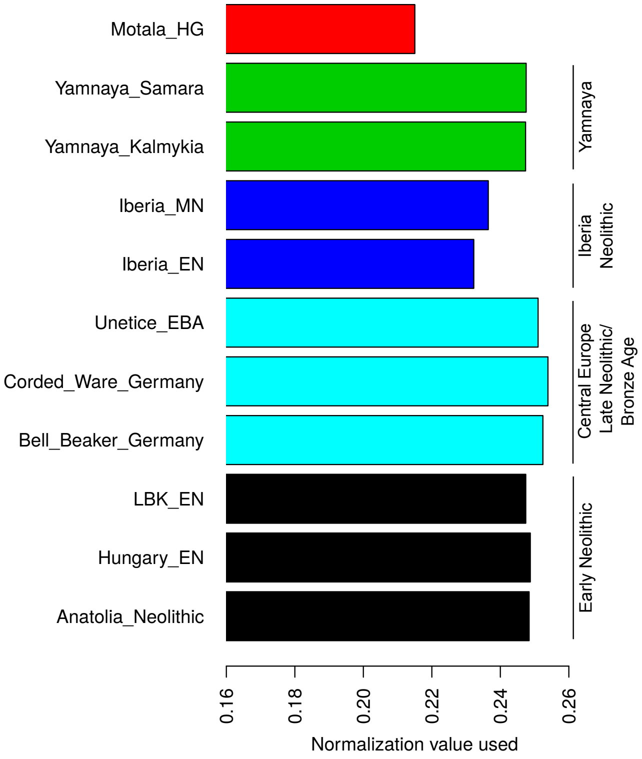

**Figure 4.** Normalization values for a selection of the cultures used in the aDNA analysis. The chronological/geographical context on the right.

**Table 1.** Pairs of relatives among the 230 individuals in the aDNA dataset as inferred by READ.

| Group | Ind1 | MT and Y (Ind1) | C14 date (Ind1) | Ind2 | C14 date (Ind2) | MT and Y (Ind2) | Inferred relationship |
|---|---|---|---|---|---|---|---|
| Neolithic Anatolia | I0736 (female) | N1a1a1a | 6500-6200 BCE | I0854 (female) | 6500-6200 BCE | N1a1a1a | 1st |
| Neolithic Anatolia | I1097 (male) | W1; G2a2b2a | 6500-6200 BCE | I0744 (male) | 6500-6200 BCE | J1c11; G2a2b2a | 2nd* |
| Bell Beaker, Germany | RISE563 (male) | K1c1; R1b1a2a1a2b | NA | RISE564 (male) | NA | H; R1b1a2a1 | 2nd* |
| Bell Beaker, Germany | I0111 (female) | H3ao | 2475-2204 calBCE | I1530 (male) | 2345-2198 calBCE | H3ao; R1 | 1st |
| Corded Ware, Germany | I1538 (male) | J1c5; R1a | 2500-2050 BCE | I1534 (male) | 2500-2050 BCE | K1a1b2a; R1b1a2 | 2nd* |
| Corded Ware, Germany | I1538 (male) | J1c5; R1a | 2500-2050 BCE | I1541 (male) | 2500-2050 BCE | U2e1a1; R1a | 1st |
| Corded Ware, Germany | I1539 (female) | J1c1b1a | 2625-2291 calBCE | I1532 (male) | 2500-2050 BCE | J1c2e; R1a1a | 2nd* |
| Corded Ware, Germany | I1534 (male) | K1a1b2a; R1b1a | 2500-2050 BCE | I1541 (male) | 2500-2050 BCE | U2e1a1; R1a | 2nd* |
| Corded Ware, Germany | I1540 (male) | J1c5; R1a1 | 2500-2050 BCE | I1541 (male) | 2500-2050 BCE | U2e1a1; R1a | 1st |
| Corded Ware, Germany | I1541 (male) | U2e1a1; R1a | 2500-2050 BCE | I0104 (male) | 2559-2296 calBCE | U4b1a1a1; R1a1a1 | 2nd* |
| Chalcolithic Iberia | I1302 (male) | J2b1a3; G2a2b2b | 2880-2630 BCE | I1314 (male) | 2880-2630 BCE | J2a1a1; G2a | 1st |
| Chalcolithic Iberia | I1274 (male) | H;3 I2a2 | 2880-2630 BCE | I1277 (male) | 2830-2820 calBCE | H3; I2a2a | 1st |
| EN Iberia | I0411 (male) | K1a2a; F§ | 5295-5067 calBCE | I0410 (male) | 5295-5066 calBCE | T2c1d or T2c1d2; R1b1 | 1st |
| Srubnaya | I0421 (female) | H3g | 1850-1600 BCE | I0430 (male) | 1850-1600 BCE | H3g; R1a1a1b2a2a | 1st |
| Srubnaya | I0354¶ (female) | U5a1 | 2016-1692 calBCE | I0360 (male) | 1850-1200 BCE | U5a1; R1a1 | 1st* |
| Unetice | I0117 (female) | I3a | 2272-2039 calBCE | I0114 (male) | 2138-1952 calBCE | I3a; I2a2 | 1st |

Radiocarbon dates and uniparental markers as reported by[33]
* newly reported relationship
§ potentially haplogroup R, not enough data
¶ excluded as population outlier in[33]