Detecting runs of homozygosity from low-coverage ancient DNA

Harald Ringbauer[1,2,3,†], John Novembre[3,4,*] and Matthias Steinrücken[3,4,*]

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA.
[2]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA.
[3]Department of Human Genetics, University of Chicago, Chicago, IL, USA.
[4]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA.
[†]Corresponding author.
*These authors contributed equally to this work.

April 2020

Abstract

We present a novel method to detect runs of homozygosity (ROH) from low-coverage genotype data typical for ancient human DNA. ROH are the genetic signature of matings between related parents, and as such, the frequency and length distribution of these blocks can give insight into recent population history and mating patterns. Existing methods identify ROH by scanning for regions that lack heterozygote genotypes, but this strategy frequently fails for ancient individuals: The vast majority of ancient DNA data has low read depth ($<3\times$), which makes reliable diploid genotype calling infeasible. To overcome this limitation, we make use of linkage disequilibrium information from a panel of modern reference haplotypes using a Hidden Markov Model. Our method scans for long stretches where the read data are consistent with only a single haplotype. When tested on simulated and down-sampled pseudo-haploid data from a targeted set of 1.24 million single nucleotide polymorphisms ("1240k SNPs") widely used in ancient DNA, our implementation robustly works for coverage down to $0.5\times$ and can tolerate error rates up to 3%, with high power and low false positive rate for blocks longer than 4 centiMorgans. Therefore, the method can screen a substantial fraction of human genome-wide ancient DNA data for parental relatedness, which will yield new evidence for questions regarding past demography and social organization.

Introduction

Little is known about the past prevalence of human consanguinity, in particular from pre-historical times for which no written record on mating patterns exists. A promising way forward is provided by ancient DNA (aDNA) data, which has enabled researchers to analyze genetic data from human remains of deceased individuals. Throughout the last decade, generation of such data has accelerated (Skoglund and Mathieson, 2018). Most aDNA data is from whole genome sequencing or capture techniques that first enrich human variation for single nucleotide polymorphisms (1240k capture technology, Fu et al., 2015). A major challenge is the generally very low coverage. Typical aDNA studies achieve average coverage per site only around or less than $\sim 1\times$. Moreover, contamination and aDNA degradation can introduce sequencing errors at a rate higher than for present-day individuals (Furtwängler et al., 2018).

Existing approaches can robustly identify ROH for present-day data (e.g. Purcell et al., 2007; Narasimhan et al., 2016). As these methods are designed for high coverage data, such methods have been extended to high coverage ancient individuals where read depth is high (Sikora et al., 2017; Racimo et al., 2020). But these only constitute a very small fraction of the existing aDNA record. Recently, a method to jointly infer heterozygosity and ROH has been introduced (Renaud et al., 2019). This method has been reported to work down to ca. $5\times$ average coverage, which again precludes its use on the vast majority of published aDNA data.

To improve the analysis of ROH in ancient data, we developed a method to analyze low coverage aDNA data. Specifically, we developed a Hidden Markov Model (HMM) that is an extension of the widely used Li & Stephens haplotype copying HMM model (e.g. Li and Stephens, 2003; Hellenthal et al., 2014). Briefly, we use the copying states of the original model to represent ROH segments, and we introduce one additional state, designed to represent when the observed data is not from an ROH segment (for a graphical summary, see Fig. 1A,B). A rationale for this model is that inside an ROH segment, the observed read data are observations from a single haplotype that is carried on both the maternal and paternal chromosomes of the target individual, and that single haplotype can be modelled well by the Li & Stephens model. Outside an ROH segment, the read data arise from two distinct haplotypes, and thus the Li & Stephens model will provide a poor fit relative to a state with Hardy-Weinberg emissions. Testing various scenarios with observed reads down-sampled at standard 1240k SNP capture sites (Fu et al., 2015) demonstrates that our novel method can robustly infer ROH longer than 4 cM using pseudo-haploid data down to ca. $0.5\times$ coverage.

Methods

1.1 The hidden Markov model

We first describe how to model diploid genotype data y from a focal individual and a reference panel of n phased haplotypes $x_1, \ldots, x_n$ at a set of L loci, assuming biallelic markers. Thus, $y \in \{0, 1, 2\}^L$ and $x_i \in \{0, 1\}^L$. In section 1.1.3, we will describe how types

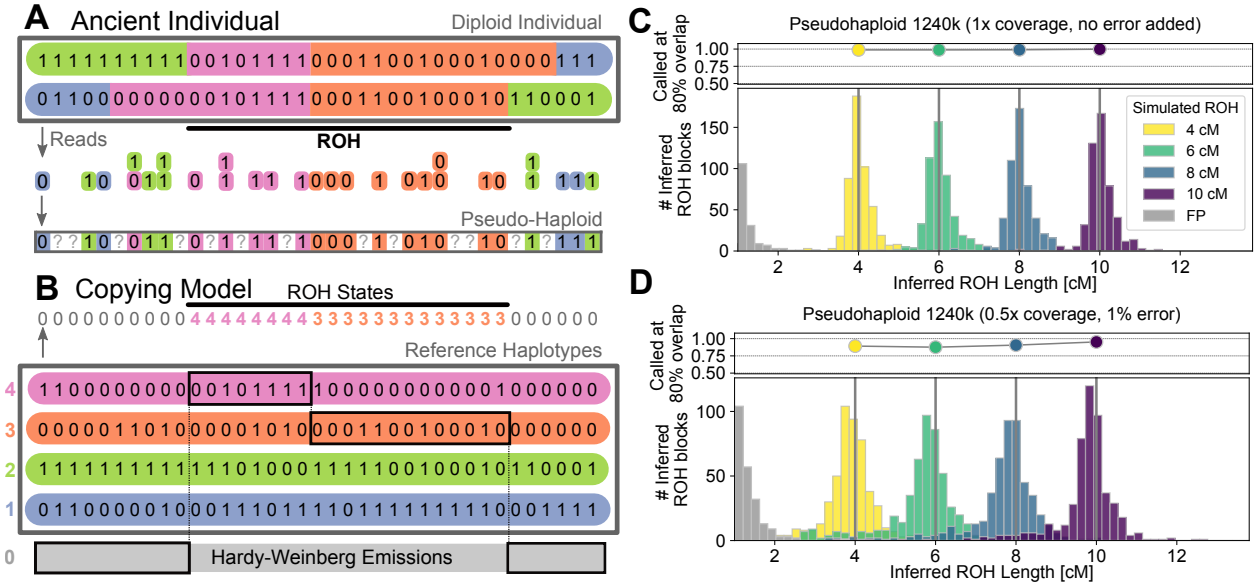


Figure 1: Detecting runs of homozygosity using a reference panel. Panel A: Illustration of genotype data from a diploid individual. Sequencing reads mapping to a biallelic SNP produces counts of reads for each allele, from which in turn pseudo-haplotype genotypes, i.e. single reads per site, are sampled (at random). Panel B: Schematic of Method. A target individuals genotype data is modelled as being copied from a reference panel (colored) and one additional non-ROH state, where copying probabilities are given by Hardy-Weinberg proportions. Panel C: We applied our method to simulated data with known ROH copied in (see Methods for details). We copied in ROH of either 4,6,8 and 10 cM length (5 of every length class into each of 100 simulated chromosomes, 1.4), and depict histograms of inferred ROH lengths (in color) as well as false positives (in gray). Panel D: Same as panel C, but a simulation with erroneous and missing data more typical for ancient DNA.

of data y relevant to applications using low-coverage sequencing data, like ancient DNA, can be modeled by treating the unobserved diploid genotypes as latent variables and using appropriate emission probabilities. Throughout, we measure the distance between loci along haplotypes in genetic map units (i.e. Morgans) $\mathbf{r} = r_1, \ldots, r_{L-1}$, where r_l denotes the distance between locus $l+1$ and l. We assume that a genetic map is available, which is the typical case for humans and model organisms. If no genetic map is available, the map distances can be approximated using the average recombination rate, but we note that here we only tested scenarios where a map is available.

1.1.1 State Space

The Hidden Markov model (HMM) can assume any of $n+1$ hidden states $0, \ldots, n$ at every marker l, where n is the number of haplotypes in the reference panel. As we outline below, the 0-th state represents that the focal individual is not in a ROH at the respective marker, and has emission probabilities according to Hardy-Weinberg proportions, while the states $1, \ldots, n$ are the classical copying states. In each of these copying states (denoted

here as the ROH states), we model the copying as in the original Li & Stephens model, with one important modification: We assume that the genotype of the focal individual y is homozygous for the allele of the reference haplotype that it copies from. The emission probabilities are specific to the exact kind of data that is analyzed, and can include various types of error models, which we discuss in more detail in Section 1.1.3.

In the Hardy-Weinberg state 0, the probabilities of observing a diploid genotype reflect the probabilities of an underlying genotype in Hardy-Weinberg equilibrium, with probabilities of the alleles according to the underlying allele frequency in the reference panel at this locus. We note this state is identical to the non-ROH state used in a previously developed HMM to call ROH (Narasimhan et al., 2016).

1.1.2 Infinitesimal Transition Rates

To define a hidden Markov model, one needs to specify the transition probabilities between the hidden states for each pair of successive loci l and $l+1$. In our model, we do so by using an infinitesimal rate matrix Q of dimension $(n+1)\times(n+1)$, from which the transition probability matrix $A_{l\to l+1}$ can be obtained via exponentiation: $A_{l\to l+1} = \exp(Q \cdot r_l)$, where r_l is the genetic distance between the respective loci.

Following Li & Stephens, the copying states $i = 1,\ldots,n$ are symmetric in our model. We can thus specify the infinitesimal rate matrix by three parameters: A single rate for the transition from the non-copying into a copying state Q_{0j} for all $j > 0$, a single rate for leaving a copying state Q_{j0} for all $j > 0$ and a third rate for transitioning from one copying state to another $\phi_{\text{ROH}} = Q_{jk}$ for all $j,k > 0, j \neq k$. The diagonal entries of the rate matrix Q are determined by the rate matrix condition $Q_{ii} = -\sum_{j\neq i} Q_{ij}$.

We point out that in the limit of infinite jumping rates within ROH ($\phi_{\text{ROH}} \to \infty$), our model converges to the full model of Narasimhan et al. (2016), as the probabilities of being in one of the allelic states (the sum of probabilities of copying from all reference haplotypes that have this allelic state) will then reflect its frequency, as in this limit jumps occur between any two consecutive markers.

1.1.3 Emission Probabilities

In our model, the emission probabilities that specify the probability of observing the data at locus l given some hidden state i, $e_i(y_l)$ depend on the type of data. We implemented three emission models: diploid genotype data, pseudo-haploid genotype data and read count data, with all three of them incorporating a model for genotype error. Throughout, we always disregard markers with missing data by removing them both from the reference panel as well as the target and adjusting the transition rates accordingly.

We implemented the emission model for diploid genotypes as follows. In the non-ROH state (i=0), the Hardy-Weinberg emission probabilities for the genotypes are $(1-p_l)^2$, $2p_l(1-p_l)$, and p_l^2, for observing homozygosity for the ancestral allele, heterozygosity, and homozygosity for the derived allele, respectively, where p_l is the frequency of the derived allele in the reference panel at locus l. For the ROH-states ($i = 1,\ldots,n$), the genotype probabilities are 1 to be homozygous for the allelic type of the source haplotype in the reference panel, and 0 for the two other possible diploid genotypes. We extend

these genotype probabilities to model possibly erroneous genotypes by assuming that with probability ϵ a genotype is flipped to one of the two other genotypes at random. This simplified error model has the advantage of having only a single parameter while broadly modeling a wide range of possible errors, including genotyping error in the reference as well as in the target, or new mutations that are private to the target individual. We note that for ancient DNA data, where genotyping error rates (including errors due to contamination) are typically on the order of $10^{-2} - 10^{-3}$ (Racimo et al., 2016), the genotyping error rate will be the main driver of ϵ, as for modern human populations the reference panel is almost always separated no more than 10^5 generations from the target. The per base-pair mutation rate is on the order of 10^{-8} per generation, which results in an upper bound for the substitution rate of order 10^{-3}.

The second emission model we implemented is for pseudo-haploid genotype data, a widely available data type for human ancient DNA. For the copying states ($i = 1, \ldots, n$), the allele on haplotype i is emitted with probability $1 - \epsilon$, and the alternative allele is emitted with probability ϵ. For the non-ROH state (i=0), the emission probabilities model sampling one read from an underlying genotype in Hardy-Weinberg equilibrium under the allele frequencies in the reference panel: A derived pseudo-haploid marker is observed with probability p_l, and an ancestral marker with probability $1 - p_l$. To account for possible errors, with probability ϵ the observed read actually reflects the opposite allelic state. As in the case of diploid genotypes, this error rate ϵ models both the disagreement rate due to new mutations occurring on the genealogical lineage between the reference haplotype and the target, as well as the rate of genotyping errors.

The third emission model we implemented is for read count data, where the data for a specific locus consists of n reads, with k of them mapping to the derived allele and $n - k$ to the reference allele. Given the underlying genotype, modelled probabilistically as in the diploid genotype case described above, we add a second layer that describes the sampling of the n reads. We use a binomial model, where the probability of observing k out n marker to be derived is binomial with probability $p = 0$, $p = 0.5$, and $p = 1$ given the heterozygous ancestral, homozygous, and heterozygous derived genotype, respectively. We add two levels of error: One at the read level, where each read is flipped to the opposite allele with probability ϵ, which can be absorbed into the binomial probabilities. We add an additional level of error at the genotype level, corresponding to the error model of erroneous diploid genotypes described above, where a diploid genotype is flipped to one of the other two possibilities with probability ϵ_{ref}. The intuition for the genotype level of errors is to account for rare errors in the reference panel that would induce mismatches between the target individual's genotype.

We note that extensions for more complex error models that include position and context-specific effects or leverage base quality scores from the sequencing, and models for other kind of data could be incorporated by adjusting the emission probabilities linking the unobserved diploid genotypes to the data. Importantly, such extensions can be naturally modelled using a genotype likelihood framework that describes the likelihood of the observed data under each of the three possible latent diploid genotype states.

1.2 Posterior Decoding

We use standard Hidden Markov model algorithms to calculate the posterior probability $P(\pi_l = i|y)$ of the hidden state i at locus l observing the data $y_1, \dots y_L$ (Durbin et al., 1998). Specifically, we compute the forward probabilities,

$$f_i(l) := P(y_1, \dots, y_l, \pi_l = i) = e_i(y_l) \sum_k f_k(l-1) A_{ki}, \tag{1}$$

as well as the backward probabilities,

$$b_k(l) := P(y_{l+1}, \dots, y_L | \pi_l = k) = \sum_i A_{ki} e_i(x_{l+1}) b_i(l+1), \tag{2}$$

using dynamic programming, where A denotes the transition matrix $A_{l-1 \to l}$. Together, these are combined to obtain the posterior:

$$P(\pi_l = i|y) = \frac{f_i(l) b_i(l)}{P(y)}, \tag{3}$$

where $P(y)$ denotes the full probability of the data, which can be computed as $P(y) = \sum_k f_k(L)$.

To complete the posterior decoding and thereby call ROH segments, we use posterior thresholding. We return consecutive regions where the posterior probability of the non-ROH state remains below the threshold $1 - T$, or equivalently the sum of the posteriors of the copy states is above T. In section 1.5 we describe the procedure for how we set the default value of T for our implementation of the method.

1.3 Computational Speedup

The run-time (and memory requirement) of the algorithm for the posterior decoding of the HMM scales linearly with the number of loci L that are analyzed. In the naive implementation, the scaling with the number of hidden states K (the number of reference haplotypes plus one here) is quadratic, since the full transition matrix has to be computed and each entry employed in Equation (1) and (2). Thus, the run-time of the naive implementation is $\mathcal{O}(LK^2)$.

However, as is standard for these models, we can reduce this run-time to linear in the number of hidden states, to $\mathcal{O}(LK)$, by using the symmetry of the copying states: For hidden state $i > 0$, the sum in Equation (1) can be split up into three parts (we suppress dependencies on $l-1$ here):

$$\sum_k f_k A_{ki} = \underbrace{f_0 A_{0i}}_{I} + \underbrace{\sum_{k>0} f_k A_{12}}_{II} + \underbrace{f_i (A_{ii} - A_{12})}_{III}, \tag{4}$$

where we used that $A_{ki} = A_{12}$ for all $k, i > 0$, which follows from the symmetry of the transition rate matrix Q. Similarly, for k=0 we get:

$$\sum_k f_k A_{k0} = \underbrace{f_0 A_{00}}_{I} + \underbrace{\sum_{k>0} f_k A_{10}}_{II}, \tag{5}$$

because $A_{k0} = A_{10}$ for all $k > 0$.

The quadratic dependence of the run-time on the number of states is caused by the sum in II in Equation (4), and similarly in Equation (5). However, when updating the forward probabilities $f_i(l)$ for all states i, we only need to pre-compute $\sum_{k>0} f_k$ once for every locus. Doing so achieves the reduction to linear run-time. The backward algorithm can be modified analogously, with first splitting the sum in Equation (2) and then pre-computing $\sum_{i>0} A_{ki} e_i b_i$ only once when updating $b_k(l)$ for all states k.

1.3.1 Efficient computation of the transition matrices

In the naive implementation of our algorithm, the infinitesimal rate matrix Q has to be exponentiated at every locus l, which would be computationally costly (depending on the implementation scaling quadratic or worse with number of states). However, due to the speed-up described in Section 1.3, we only require a small subset of the entries of the full transition matrix, namely A_{00}, A_{11}, A_{12}, A_{01} and A_{10}. We note that a truly symmetric model (such as the original Li & Stephens copying model) could be reduced even further into a single transition rate (the probability of staying in a copy state, Price et al., 2009). However, due to the additional non-ROH state here, one has to keep track of at least three rates, and these can be efficiently pre-compute as follows.

Using the symmetry of the copying states $1, \ldots, n$, we can collapse these states into state 1 and a single surrogate state for $2, \ldots, n$. We then only need to consider the states 0,1, and the surrogate state, thus arriving at a 3×3 transition rate matrix $\tilde{Q}$, where $\tilde{Q}_{ij} = Q_{ij}$ for $i \leqslant 2$, $j < 2$ and $\tilde{Q}_{i2} = \sum_{j>1} Q_{ij} = (n-1)Q_{i2}$ for $i < 2$. Importantly, by exponentiation of $\tilde{Q}$ the three relevant entries of A can be recovered by first computing $\tilde{A} = \exp(\tilde{Q})$ and then using $A_{ij} = \tilde{A}_{ij}$ for $i, j < 2$ and $A_{12} = \tilde{A}_{12}/(n-1)$.

To efficiently incorporate variable recombination distances between loci, we first diagonalize the common collapsed rate matrix: $\tilde{Q} = P^{-1}\tilde{D}P$. For each locus l, we can then exponentiate using $\exp(\tilde{A} \cdot r) = P^{-1} \exp(\tilde{D} \cdot r_l)P$, which only requires exponentiation of a diagonal matrix, and recover the corresponding entries of $\tilde{A}$ and consequently A required for calculating the full posterior. In section 1.5 we describe the procedure for how we set the default rates of Q for our implementation.

1.4 Simulating genetic data with ROH

To test the performance of our method, we simulated genetic data with known ROH. We use this data below to carry out experiments where we down-sample to lower coverage and add genotyping errors to 1) help determining robust HMM parameters (Section 1.5) and to 2) test the performance (Section 1.5). First, we describe the method we used to generate these simulated datasets with known ROH.

We used a copying approach inspired by Ralph and Coop (2013) to generate ground-truth ROH block sharing data for testing methods. A synthetic mosaic individual without long ROH >1 cM is first generated by concatenating stretches of diploid genotypes in 0.25 cM tracts from randomly chosen individuals of the reference set. The intuition is that the probability of long ROH blocks (>1 cM) arising inadvertently is very low (as multiple ROH blocks would have to be concatenated), while still mostly retaining local

LD structure typical for diploid human individuals. In our simulations, we used the positions of a widely used set of SNPs developed for human ancient DNA studies (1240k capture technology for 1.24 million SNPs Fu et al., 2015), and we focused on chromosome 3, a human chromosome with a typical density of these sites per map unit (Morgan).

We then copied in five ROH blocks of a given length uniformly at random, enforcing that ROH blocks do not overlap by placing them at random in 5 evenly split up sectors of the chromosome. The copied-in stretch originates from one haplotype of the source population (chosen uniformly), and both alleles of the synthetic individual are set to the allele of the copied-in stretch. The source population for the simulations is then excluded from the reference panel. These synthetic mosaic individuals, with known diploid genotypes, serve as test cases for the method, and various types of data (such as read count or pseudo-haploid) was generated based on them. For various tests, these data were downsampled and error added to it, to simulate data of varying quality (Fig. 1C,D).

1.5 Parameter Choice

The model has several parameters that have to be set when analyzing data. Here we describe how we set the parameters we used throughout our empirical analysis and our simulation experiments. We set the infinitesimal transition rates based on the typical tracts we are interested to find. Our target use case here is to detect ROH blocks that are of length 5 cM that occur once every 100 cM. Accordingly, we chose the infinitesimal rate parameters (per Morgan) as 1 (jump from non-ROH into ROH) and 20 (jump from a ROH state into non-ROH). We fixed the transition rate between ROH states (i.e. the haplotype copying model switch rate) to 300 per Morgan, corresponding to an average copy tract length of ca. 0.3 cM. This value was chosen based on performance of ROH calling in pilot simulations and a likelihood profile of a Li & Stephens model of Tuscany haplotypes from all non-Tuscany Europeans in the 1000 Genomes dataset. We fix this set of parameters throughout our analysis.

1.5.1 Choice of Posterior Threshold

To determine a robust posterior threshold, we ran simulation experiments with data typical for our use case, which is analysis of 1240k pseudo-haploid data with the full 1000 Genomes dataset set as a reference panel. As test cases, we simulated mosaics of chromosome 3 with pseudo-haploid data, i.e. one allele chosen at random, and then downsampled randomly to 50% of all 1240k SNPs covered (and the rest indicated as missing). We then flipped the allele at random with probability 0.01 to the other allele to simulate data with low quality. This is a representative use case for our method: As described below (section 2.1) we apply our method to individuals in real datasets with more than 400,000 SNPs covered, for which estimated error rates are below 5% . We point out that error rates cover both sequencing error and contamination, and that not all contamination results in erroneous reads. The reason for choosing the cutoff based on low quality data is that we want the cutoff to be robust in these cases. We tradeoff maximum specificity for high quality data (where more aggressive cutoff settings would be possible) to allow our method being applicable to a wide range of use cases with default parameters.

Using the TSI (Tuscany, Italy) samples from the 1000 Genomes dataset, we simulated 100 replicates of mosaics of chromosome 3 for two scenarios: 1) with 4 cM ROH blocks copied in (to determine power and bias of inferred ROH length) 2) no blocks copied in as well (to assess false positives). We then ran the method using the 1000 Genomes dataset and only TSI individuals removed as reference panel, tested various posterior cutoffs, and monitored false positive rate, power, length bias, and standard deviation of the longest block overlapping the true ROH blocks, with blocks of length 4 cM as the test case. When analyzing 100 replicates with various posterior cutoffs, we found that a cutoff of 0.998 lead to a good performance in terms of the magnitude of bias for ROH, as well as standard deviation of inferred length of ROH (Table 1). As our overall goal is to call ROH with little bias and also with little variation in length, we chose this value of 0.998 as posterior cutoff in our implementation.

Posterior Cutoff	Rep.	STD 4cM	FP ROH>1cM	FP ROH>2cM	Avg. Bias 4 cM [cM]	Frac. 80% of 4 cM called
0.996	100	0.61	5.39	0.47	0.06	0.958
0.997	100	0.59	4.70	0.35	0.02	0.950
0.998	100	0.57	3.78	0.21	-0.03	0.930
0.999	100	0.60	2.34	0.11	-0.15	0.892

Table 1: Varying the posterior cutoff on various performance metrics. We varied the posterior cutoff used for calling ROH, calculated several summary statistics when calling ROH for mosaic individuals (TSI). For each line, 100 replicates for chromosome 3 with five 4 cM ROH copied or no ROH copied in were simulated to calculate the performance statistics. False positive rates (FP) are calculated as the average number of falsely inferred blocks per replicate chromosome.

For applications on 1240k pseudo-haploid SNPs with at least 400,000 autosomal SNPs covered and using the 1000 Genomes data as the reference panel, this set of parameters can be readily applied, and we provide these parameters as the default settings in our software package that implements the method. For users who wish to apply our method to another set of SNPs, a different reference panel, or non-human data, we strongly recommend to repeat a similar strategy to find a suitable threshold in the respective scenario.

1.5.2 Merging of Gaps between ROH

Motivated by the observation that the vast majority of false positive ROH are shorter than 2 cM (Fig. 1), we only record ROH blocks >2 cM. We observed that long ROH are sometimes broken up by spurious gaps (Fig. 2 and manual inspection of blocks where the length was substantially underestimated), as similarly seen in methods that call long IBD blocks between individuals (Browning and Browning, 2015). Such gaps may arise due to genotyping error, structural variation or very low SNP density. Following a standard procedure of IBD block calling (Ralph and Coop, 2013) and of genomic feature annotation with HMMs (Durbin et al., 1998), we decided to merge gaps, as experiments with lowering the posterior threshold or with decreasing the jump rate introduced a large surplus of additional false positives. To ensure that we do not merge two false positives (the false positive rate >2 cM is non-zero), we additionally require at least one of the merged blocks

to be longer than 4 cM, and the gaps to be less than 0.5 cM in length. Fig. 2 shows that this procedure improves the performance substantially.

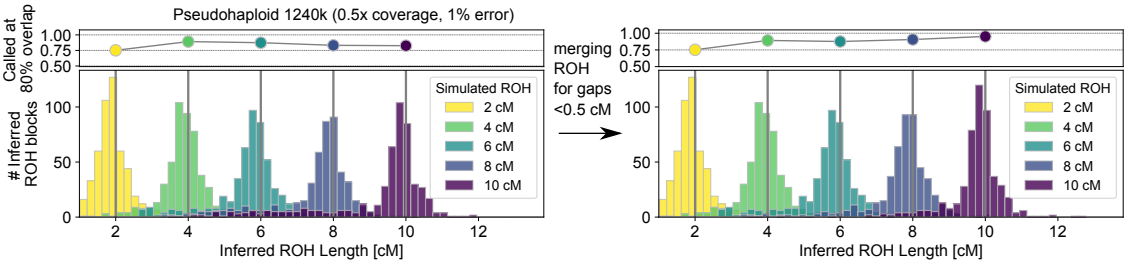


Figure 2: Improving power for long ROH blocks by merging gaps between ROH stretches We depict the effect of merging ROH gaps for the "worst case" simulation scenario where we expect our method to have the least power to detect uninterrupted segments of ROH. Merging gaps <0.5 cM for between blocks where the longer block >4 cM markedly improves performance for long ROH blocks (>8 cM), without changing the distribution of shorter ROH blocks (4 cM).

Results

2.1 Performance on simulated data

To test its performance, we applied our implementation of the method with default parameters chosen as described in Section 1.5 to mosaic individuals with copied in ROH blocks as detailed in Section 1.4. The majority of published ancient DNA is released as pseudo-haploid data, i.e. with one read randomly picked per site. When applying the method to such pseudo-haploid data sampled from the test individuals and additionally down-sampled to varying degree, our analysis revealed that it has high power (>95%) to detect ROH blocks >4 cM while having simultaneously a low false positive rate (Fig. 3A) down to at least 0.5× covered 1240k sites. Moreover, we find that, when first applying random genotype errors, the method can tolerate genotype error rates up to 5% (Fig. 3B).

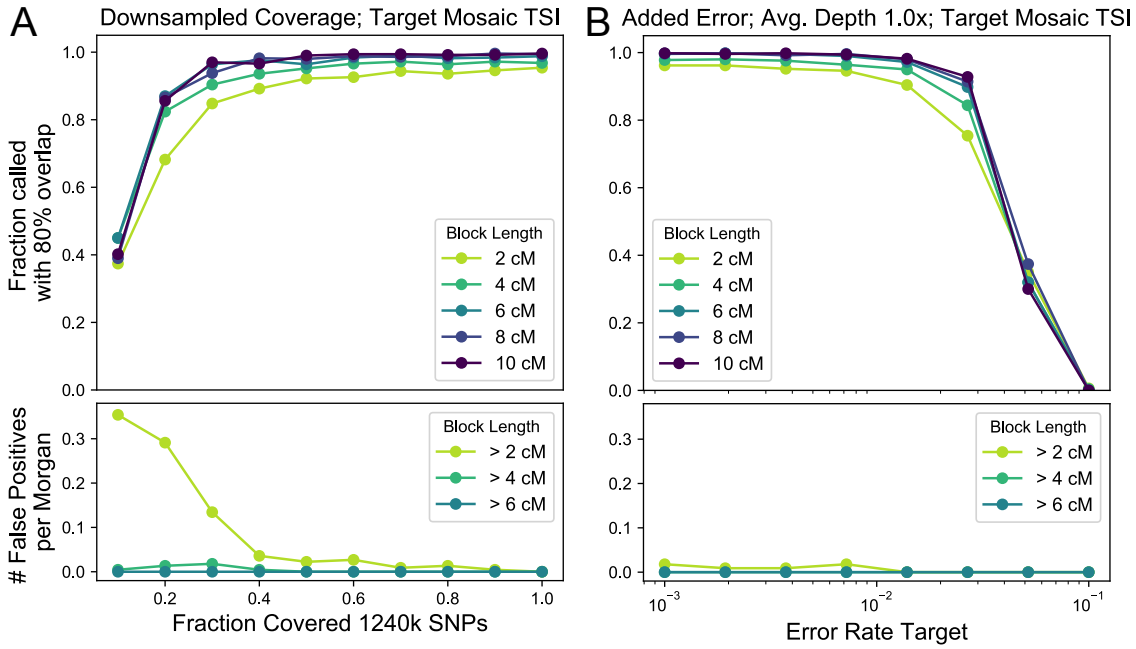


Figure 3: Performance of the method to detect ROH within mosaic individuals We analyzed 100 individual chromosomes 3 which have been copied together as mosaics from 0.25 cM stretches from TSI individuals (Tuscany) of the 1000 genomes dataset (Section 1.4) on the 1240k sites. For each site, we then sampled one read from the diploid genotype at random, creating pseudo-haploid data. We further down-sampled to varying degrees (0.1-1.0×, Panel A), or introduced random genotype genotype errors at different rates (0.001-0.1) and applied the method with a copying error rate set to 1% (Panel B), using the 1000 genome data with the TSI haplotypes removed as reference panel (4794 haplotypes).

2.2 Reference panels with varying genetic distance

To test the impact of different coalescence time distributions to the reference panel, we tested the method on simulated mosaic individuals from various global populations when using a reference panel consisting of European haplotypes. We note that under a simple model of a clean population split, the divergence time between the target and the reference population introduces a minimum boundary for coalescence times of the reference haplotype with the reference panel, similar to a temporal separation of an ancient target from the reference panel.

We tested how well the method works when using a European reference panel (with TSI removed, 792 out of 1,006 haplotypes remaining) for mosaic individuals generated from several target populations of the 1000 Genomes dataset. We tested four target populations, chosen to cover a wide range of population genetic distances. We tested with pseudo-haploid data on 1240k SNPs, picking one allele at random at each 1240k site (Tab. 2 and Fig. 4).

With divergence occurring tens of thousands of years ago, such as target for CHB (Han Chinese) with European reference haplotypes, 95.0% of copied-in blocks are identified with at least 80% overlap with the true ROH block. However, this behavior does not continue across all pairs of populations, we observe little power to infer ROH in mosaic individuals constructed from YRI haplotypes when using European haplotypes as reference. In this case, while some ROH blocks are still identified, only less than 10% of copied in ROH blocks are inferred with at least with 80% overlap.

Target	Panel	Power at 80% overlap [4cM]	Bias in Length [4cM]	Standard Deviation Length [4cM]
TSI	**EUR***	0.986	0.151	0.46
CHB	**EUR***	0.950	0.138	0.54
CLM	**EUR***	0.882	-0.10	0.69
YRI	**EUR***	0.096	-2.01	0.90

Table 2: Effect of varying distance from reference panel to target. We tested the performance with mosaic individuals from Tuscany, Italy (TSI); Han Chinese from Beijing (CHB); Colombians from Medellin (CLM) and Yoruba from Ibadan (YRI), and tested the power to call ROH blocks of length 4 cM. we define a successful inference when at least 80% of the original ROH block are inferred to be within a single inferred ROH. EUR*: European reference haplotypes with TSI (Tuscany) removed.

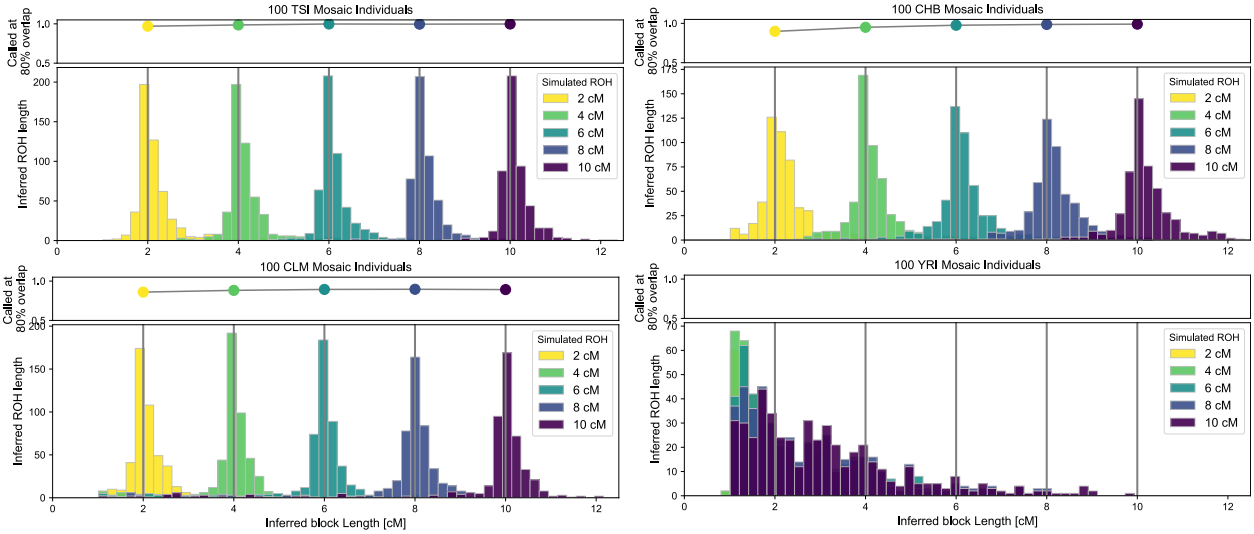


Figure 4: Effect of varying distance from reference panel to target. We tested the performance with target individuals that were simulated as mosaics of haplotypes from Tuscany, Italy (TSI); Han Chinese from Beijing (CHB); Colombians from Medellin (CLM) and Yoruba from Ibadan (YRI), and using European reference haplotypes (without TSI haplotypes).

2.3 Performance on down-sampled Ust Ishim man

High-coverage ancient DNA data provides a useful test case to assess ROH inference. Here we analyzed a Western Siberian individual radio carbon dated to about 45,000 years before present, called "Ust Ishim man". His complete genome has been sequenced to remarkable depth (ca. 40×) from a femur bone (Fu et al., 2014), allowing for robust diploid genotype calls.

Importantly, high-coverage data allows one to call ROH with high reliability by simply identifying stretches that lack sites where many reads indicate heterozygosity (Fig. 5). Moreover, as "Ust Ishim man" is the oldest anatomically modern human sequenced to high coverage to date, it provides us an opportunity to examine an extreme case in terms of how much temporal distance from the reference panel our method can tolerate.

We analyzed read count data for the 1240k SNPs from Ust Ishim man (40× read depth on the target) - using the post-processed publicly available data from Marcus et al. (2020). We then down-sampled these reads to lower coverage (0.2-40×) at random. Furthermore, we created pseudo-haploid data for all SNPs covered (1,115,315 of the 1240k variants were covered) by choosing one read at random per site, and then created artificial data down-sampled to subsets (0.3-1.0× smaller) of the 1240k sites. We analyzed both read-count data and pseudo-haploid data and summed up all ROH blocks longer than a given threshold.

Our results show that we can consistently infer ROH blocks >4 cM when down-sampling to low coverage (0.5× mean depth) of the 1240k markers (Fig. 5). Importantly, even for pseudo-haploid data, which effectively only uses LD information as signal, inference seems to work reliably with as low as 0.3× coverage, with little observable bias for blocks >8 cM and a small false positive rate for blocks >4 cM (Fig. 5). We hypothesize that this is at least in part caused by the extension of a large number of shorter ROH that then get pushed beyond the 4 cM detection threshold.

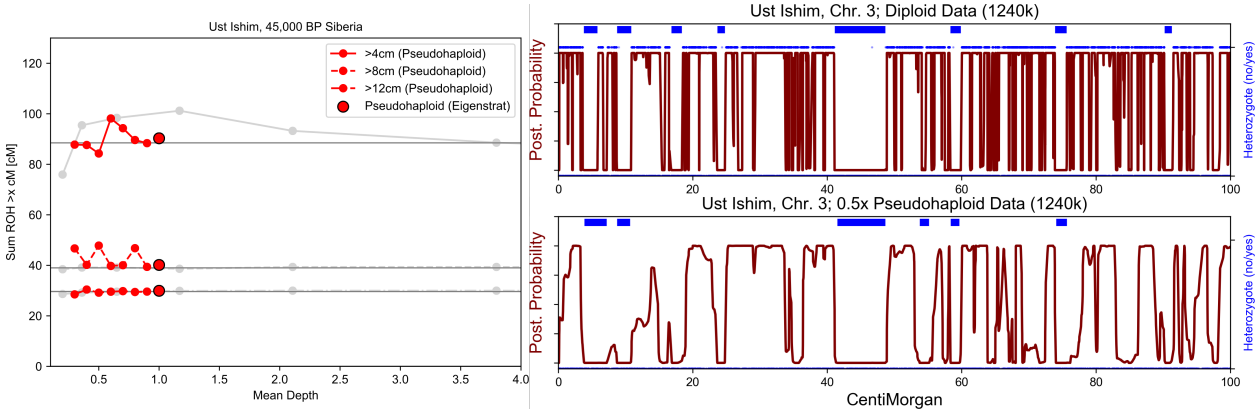


Figure 5: Properties of inferred ROH when downsampling from high coverage data on the Ust Ishim man. Left: Sum of inferred ROH >4, >8, >12 cM when down-sampling to various degrees, both for pseudohaploid data (red), as well as read count data (gray). The gray horizontal line depicts the value when using diploid genotype calls, which we assume as ground truth. Right: Posterior and inferred ROH for a region of Chromosome 3, when using diploid genotype data (top) and pseudo-haploid data down-sampled to 0.5× (bottom). We depict inferred ROH greater than 1 cM before the gap merging as blue lines above the posterior. For the diploid genotype data, we indicate heterozygous genotypes (blue dots above the posterior trace) and homozygous genotypes (blue dots below the posterior line trace. Long gaps of heterozygosity align well with the inferred ROH segments (blue lines).

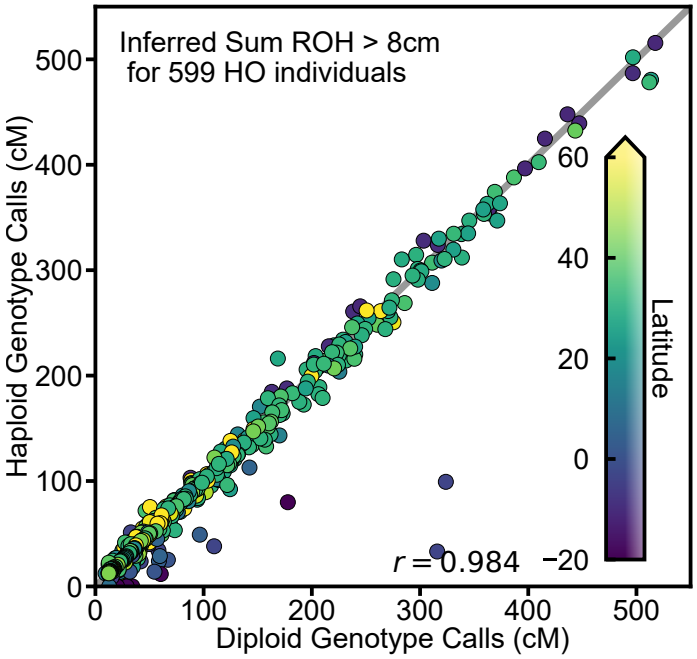
2.4 Performance on present-day populations

We applied our method to the Human Origins dataset of 1,941 present-day humans originating from 162 global populations genotyped at autosomal SNPs (Lazaridis et al., 2014). These SNPs constitute a subset of the 1240k enrichment targets (≈0.6 of ≈1.24 million SNPs). Because this dataset provides diploid genotype calls, we ran our method with the diploid mode and called ROH >4 cM in all 1,941 individuals, using 5,008 global haplotypes from the 1000 Genomes reference panel. We manually checked several called ROH, and confirmed that ROH calls correctly identify regions with almost no heterozygous markers.

To test the pseudo-haploid mode of our method on a global panel of variation, we used all HO individuals with at least one ROH longer than 12 cM identified (599 individuals) as a test set. In addition to the high quality diploid ROH calls, we ran the pseudo-haploid mode on these individuals, choosing one allele at random for each diploid genotype call (ca. 550,000 SNPs per individual, ranging from individuals with 537,000 to 556,000 called gentoypes). Our tests confirmed that the ROH calls from the haploid and the diploid data closely agree for the majority of individuals (Fig. S6), with a correlation between datasets of r = 0.984 when comparing ROH >8 cM (Fig. 6). A notable exception are certain Sub Saharan populations, in particular South and East African hunter gatherers, for which a substantial fraction of long ROH are not correctly identified in the haploid data (Tab. 6).

When investigating these African Hunter gatherers, we noticed that the typical pattern in the inference from pseudo-haploid data is many gaps dispersed throughout ROH identified in the diploid data (e.g. Fig. 7). This pattern mirrors the one we observed when analyzing mosaic targets created from Yoruba haplotypes using an European only reference panel (Section 1.4), pointing toward some haplotype segments not captured well by the reference panel. Indeed, it has been observed previously that hunter gatherer populations in Sub Saharan Africa possess deeply diverged ancestry (Schlebusch et al., 2012), which together with the fact that the African reference haplotypes from the the 1000 Genomes data only include a single population from Central, Southern and Eastern Africa (i.e. the Luhya), yields a plausible explanation for the limited power of our method to call ROH.

After removing Sub Saharan African populations from Central, South and Eastern Africa, the correlation increases to r= 0.997, and the average difference between the sum of ROH > 8 cM inferred from pseudo-haploid and diploid genotype data is -0.53 cM (the mean of the sum of ROH inferred from diploid data is 98.03 cM). Upon inspecting specific length categories, ROH calls from all length classes are highly correlated, ranging from r = 0.925 for ROH 4-8 cM to r = 0.988 for ROH longer than 20 cM (Fig. 8). No population other the Sub Saharan African exhibits a substantial bias, which provides evidence that the reference panel is suitable for all other groups in the HO panel.



HO Population	Failed	Total
Ju_hoan_North	4	4
Hadza	3	3
Mbuti	3	4
Khomani	3	3
Biaka	2	3
Ethiopian_Jew	1	2
Somali	1	6

Figure 6 & Table 3: Comparison of diploid and pseudo-haploid ROH calls for HO individuals. Left: Comparison of ROH calls >8 cM for pseudo-haploid and diploid data for each HO individual with at least one ROH >12 cM (599 individuals). The scatter plot compares the total sum of all ROH blocks >8 cM. Right: Table summarizing individuals where more than 50% of sum ROH >8 cM are not called with pseudo-haploid data. These individuals correspond to the individuals that deviate substantially downwards from the diagonal line.

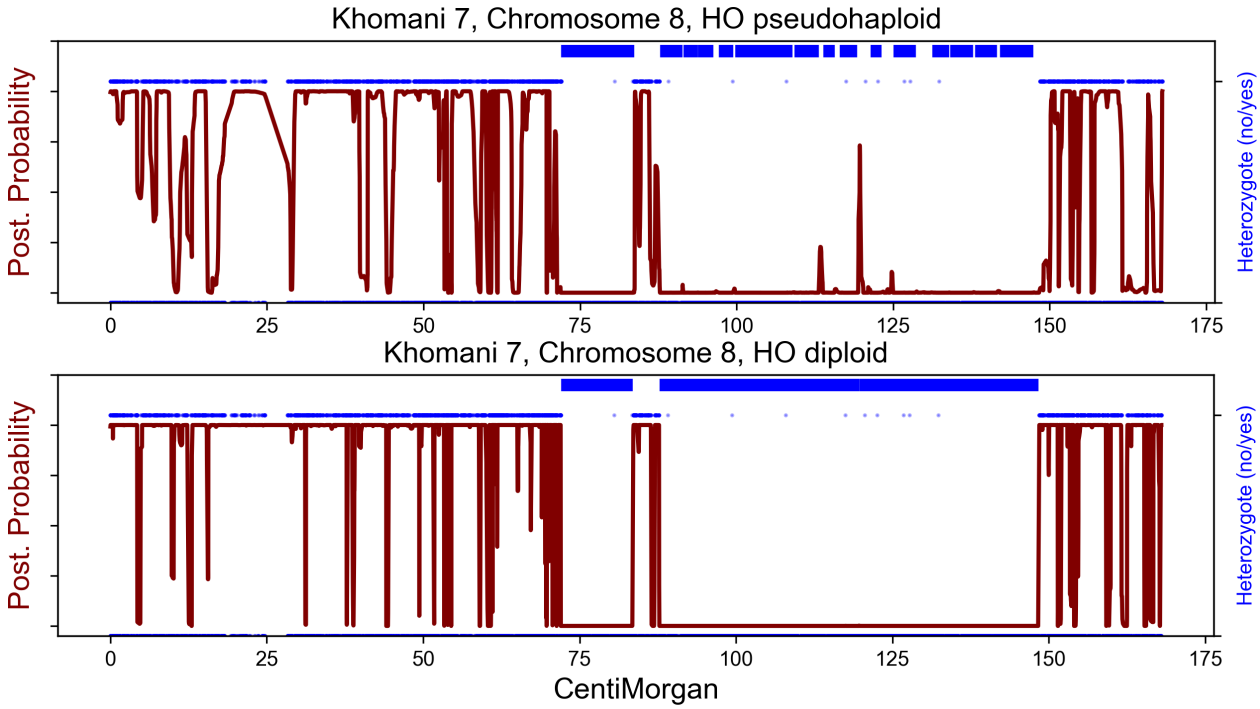


Figure 7: Comparison of diploid and pseudo-haploid ROH calls for a present-day Southern African Hunter gatherer individual. We compare the ROH calls from pseudo-haploid data (top) and diploid genotype data (bottom) from a HO African hunter gatherer in the HO origin dataset (Khomani 7). We show chromosome 8, as this individual has two long ROH on this chromosome that can be identified with high confidence in diploid genotype calls (blue dots above posterior depict heterozygous sites). The diploid mode correctly identifies these regions, whereas the pseudo-haploid mode breaks them up.

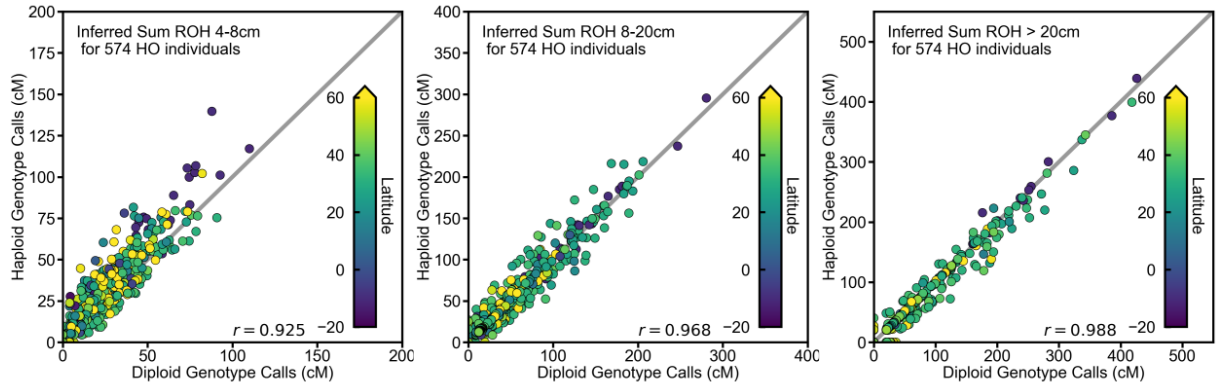


Figure 8: Comparison of diploid and pseudo-haploid ROH calls for HO individuals without Sub Saharan populations. As in Fig. 6 we compare ROH calls for HO populations, with ROH calls from diploid genotype data (x-axis) compared to ROH calls from pseudo-haploid data (y-axis). Here we have removed the Sub Saharan populations from the panel, and show comparison for three length classes: 4-8 cM (left), 8-20 cM (middle) and >20 cM (right).

Conclusion

Our tests show that the new method can robustly identify ROH longer than 4 cM for typical ancient DNA data (including pseudo-haploid genotype calls) down to ca. 0.5× coverage. This application range substantially extends the limit reported by previous methods developed for high quality present-day DNA (Narasimhan et al., 2016; Purcell et al., 2007) and ancient DNA (Renaud et al., 2019). Our tests demonstrated that the new method can tolerate divergence times of up to several tens of thousands of years between the 1000 Genomes reference panel and the target individual, which implies a wide range of applicability to ancient individuals, including all individuals sharing the out-of-Africa bottleneck. We note that we optimized and tested our implementation for genetic variation at more than a million single nucleotide polymorphisms widely used in ancient DNA (1240k capture technology, Fu et al., 2015), and whole genome data can be readily down-sampled to this set of SNPs.

The novel method presented here will allow researchers to screen a substantial fraction of the currently available human genome-wide aDNA data for parental relatedness, which will yield new evidence for questions regarding past demography and social organization (Racimo et al., 2020). In related work, we are carrying out an initial application of the method to analyze ancient DNA from over 1,798 anatomically modern humans from the last 45,000 years (*in prep*).

Identifying ROH can also be a starting point for powerful downstream applications: ROH consist of only a single haplotype (the main signal for our method) and is therefore perfectly phased, a prerequisite for powerful methods relying on haplotype copying (e.g Lawson et al., 2012) or tree reconstruction (e.g. Kelleher et al., 2019; Speidel et al., 2019). Moreover, long ROH could be used to estimate contamination and error rates, an important task in ancient DNA studies (Furtwängler et al., 2018), as ROH lack heterozygotes and thus one can identify heterozygous reads within ROH that must originate from contamination or genotyping error, similar to estimating contamination from the hemizygous X chromosomes in males (Korneliussen et al., 2014). More broadly, in many plants and animal species ROH is more prevalent than in humans (due to natural inbreeding or human breeding in domesticates), and low coverage genetic data is a widely applied cost-effective strategy to generate genetic data. Questions about inbreeding are central to a wide range of disciplines across evolutionary biology, such as conservation biology or plant evolution, and new methods building upon the core ideas of our approach could provide useful tools for studying a wide range of natural populations.

References

Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, 97(3):404–418, 2015.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, 2014.

Qiaomei Fu, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, Nadin Rohland, Iosif Lazaridis, Birgit Nickel, et al. An early modern human from romania with a recent neanderthal ancestor. *Nature*, 524(7564):216, 2015.

Anja Furtwängler, Ella Reiter, Gunnar U Neumann, Inga Siebke, Noah Steuri, Albert Hafner, Sandra Lösch, Nils Anthes, Verena J Schuenemann, and Johannes Krause. Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. *Scientific Reports*, 8(1):14075, 2018.

Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343 (6172):747–751, 2014.

Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, 2014.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), 2012.

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

Joseph H Marcus, Cosimo Posth, Harald Ringbauer, Luca Lai, Robin Skeates, Carlo Sidore, Jessica Beckett, Anja Furtwängler, Anna Olivieri, Charleston WK Chiang, et al. Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. *Nature Communications*, 11(1):1–14, 2020.

Vagheesh Narasimhan, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751, 2016.

Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), 2009.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

Fernando Racimo, Gabriel Renaud, and Montgomery Slatkin. Joint estimation of contamination, error and demography for nuclear dna from ancient humans. *PLoS Genetics*, 12 (4), 2016.

Fernando Racimo, Martin Sikora, Marc Vander Linden, Hannes Schroeder, and Carles Lalueza-Fox. Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nature Reviews Genetics*, pages 1–12, 2020.

Peter Ralph and Graham Coop. The geography of recent genetic ancestry across europe. *PLoS Biology*, 11(5):e1001555, 2013.

Gabriel Renaud, Kristian Hanghøj, Thorfinn Sand Korneliussen, Eske Willerslev, and Ludovic Orlando. Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples. *Genetics*, pages genetics–302057, 2019.

Carina M Schlebusch, Pontus Skoglund, Per Sjödin, Lucie M Gattepaille, Dena Hernandez, Flora Jay, Sen Li, Michael De Jongh, Andrew Singleton, Michael GB Blum, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338(6105):374–379, 2012.

Martin Sikora, Andaine Seguin-Orlando, Vitor C Sousa, Anders Albrechtsen, Thorfinn Korneliussen, Amy Ko, Simon Rasmussen, Isabelle Dupanloup, Philip R Nigst, Marjolein D Bosch, et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*, 358(6363):659–662, 2017.

Pontus Skoglund and Iain Mathieson. Ancient genomics of modern humans: The first decade. *Annual Review of Genomics and Human Genetics*, 19:381–404, 2018.

Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

Author Contributions

We annotate author contributions using the CRediT Taxonomy labels (https://casrai.org/credit/). Where multiple individuals serve in the same role, the degree of contribution is specified as 'lead', 'equal', or 'supporting'.

- Conceptualization (Design of study) – lead: HR; supporting: JN, MS
- Software – lead: HR; supporting: MS
- Formal Analysis – HR
- Investigation – HR
- Data Curation – HR
- Writing (original draft preparation) – lead: HR supporting: JN, MS
- Writing (review and editing) – input from all authors
- Supervision – equal: JN, MS
- Project Administration – equal: JN, MS
- Funding Acquisition – JN

Code Availability

A Python package implementing the method is available on the Python Package Index (https://pypi.org/project/hapROH/) and can be installed via *pip*.

Data Availability

All human data used in our analysis are publicly available. The processed reference panel that we used in our tests (haplotypes from 1000 Genomes dataset down-sampled to biallelic SNPs at 1240k sites) is available upon request.