Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples

Renaud, Gabriel; Hanghøj, Kristian; Korneliussen, Thorfinn Sand; Willerslev, Eske; Orlando, Ludovic

Published in:
Genetics (Print)

DOI:
10.1534/genetics.119.302057

Publication date:
2019

Document version
Early version, also known as pre-print

Citation for published version (APA):
Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples

Gabriel Renaud∗,1, Kristian Hanghøj∗†, Thorfinn Sand Korneliussen∗, Eske Willerslev∗†‡§ and Ludovic Orlando∗†

∗Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350K Copenhagen, Denmark, †Laboratoire d’Anthropobiologie Moléculaire et d’Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, 31000 Toulouse, France, ‡Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK., §Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK., ∗∗

ABSTRACT Both the total amount and the distribution of heterozygous sites within individual genomes are informative about the genetic diversity of the population they belong to. As ancient genomes are often characterized by the presence of post-mortem damage and sequenced at low coverage, detecting true heterozygous sites in such samples can be complicated. In addition, large runs of homozygosity found in the genomes of particularly inbred individuals and of domestic animals can skew estimates of genome-wide heterozygosity rates. Current computational tools aimed at estimating runs of homozygosity and genome-wide heterozygosity levels are generally sensitive to such limitations. Here, we introduce ROHan, a probabilistic method which substantially improves the estimate of heterozygosity rates both genome-wide and for genomic windows. It combines a local Bayesian model and a Hidden Markov Model at the genome-wide level and can work both on modern and ancient samples. We show that our algorithm outperforms currently available methods for predicting heterozygosity rates for ancient samples. Specifically, ROHan can delineate large runs of homozygosity (at megabase scales) and produce a reliable confidence interval for the genome-wide rate of heterozygosity outside of such regions from modern genomes with a depth of coverage as low as 5-6X and down to 7-8X for substantially damaged ancient samples. We apply ROHan to a series of modern and ancient genomes previously published and revise available estimates of heterozygosity for humans, chimpanzees and horses.

KEYWORDS inbreeding, heterozygosity, effective population size

Introduction

In diploid organisms, single nucleotide differences observed between paternal and maternal chromosomes are called heterozygous sites. As the history underlying both chromosomes can be viewed under a coalescence process, heterozygous sites result from mutations which occurred in the genealogy, backwards in time. The number of neutral polymorphic sites segregating in a given population both depends on the average coalescence, which itself depends on the effective population size (Ne), and the mutation rate (µ) (Kimura 1969). Consequently, parameters such as the Watterson’s θ, where $\theta = 4N_e\mu$ for diploid organisms (Watterson 1975), are essential in population genetics and have been widely used to infer past population demographies.

If both parents are unrelated, the number of heterozygous sites at equilibrium is expected to be $\frac{\theta}{2\pi}$, which is $\approx \theta$ for small values of θ (Watterson 1975). Several tools have been released to infer the number of heterozygous sites at equilibrium, also referred to as the heterozygosity, from either raw sequence alignment files (Haubold et al. 2010; Korneliussen et al. 2014), multiple sequence alignments (Adams et al. 2018; Gronau et al. 2011) and SNP arrays (Purcell et al. 2007; Yang et al. 2011; Browning and Browning 2015; Szpiech et al. 2017). The underlying methodology has been recently reviewed by (Yengo et al. 2017).

However, if both parents are related, large stretches of the offspring genome will be identical by descent (IBD). At such loci, no or very little heterozygous sites will be found, resulting in the presence of runs of homozygosity (ROH). Such ROHs can be informative about an individual’s demographic history (Ceballos et al. 2018). The total length of such genomic loci depends on the type of inbreeding (Wright 1922) and the length of such regions depends on how far back in the genealogy the inbreeding event took place (Fisher 1954; Keller et al. 2011), considering
that recombinations reduce the length of ROHs with time. A number of statistical packages have been released to investigate the impact of inbreeding on individual fitness (Stoffel et al. 2016).

Inbreeding can be due to small group size, cultural practice (Alvarez et al. 2009) as well as reproductive management procedures such as those underpinning domestic livestock (Wiener and Wilkinson 2011). ROHs have therefore been detected in a number of domestic animals, including sheep (Purfield et al. 2017), cattle (Purfield et al. 2012), pigs (Bosse et al. 2012) and donkeys (Renaud et al. 2018). A first class of methods aimed at the detection of ROH in modern samples have relied on pre-called genotypes (McQuillan et al. 2008; Pemberton et al. 2012) and allele frequencies for the population of interest (Narasimhan et al. 2016). Even without the additional layer of complexity represented by the uncertainty in calling genotypes, available methods show limitations and generally require ad hoc tuning of their parameters to fit the properties of the data at hand (Howrigan et al. 2011). A second limitation is the use of allele frequencies. Such information is not always available especially for rare breeds or remote populations. Elevated drift or distant split times between the population providing the allele frequencies and the sample often make such information inapplicable. Another class of methods have relied on weighted-likelihood methods using genotype data for several individuals (Blant et al. 2017).

In recent years, methodological advances in ancient DNA (aDNA) research have opened access to the complete genome sequence of ancient human individuals (Llamas et al. 2017b), domesticates (Frantz et al. 2016; Gauntt et al. 2018), pathogens (Rasmussen et al. 2015), and extinct species (Miller et al. 2008; Green et al. 2010; Reich et al. 2010). Ancient genomes provide time-stamped genetic snapshots which are instrumental for understanding how the genetic makeup of contemporary species came to be. However, aDNA molecules are generally poorly preserved and co-extracted together with a large fraction of genetic material from environmental microbes (Der Sarkissian et al. 2014). This results in relatively low amount of endogenous molecules, which makes the recovery of high coverage genomes for ancient individuals often prohibitively expensive (Hofreiter et al. 2015). As a consequence, the vast majority of the ancient genomes currently available have only been sequenced to low coverage (Marciniak and Perry 2017).

Inferring heterozygosity on the basis of low sequence coverage data is difficult but several methods have been proposed to do so (Bryc et al. 2013). In addition to coverage limitation, the presence of post-mortem damage, which introduces nucleotide mis-incorporations, and potential contamination either stemming from microbial sources or present-day humans (Llamas et al. 2017a), make heterozygosity estimates in ancient samples particularly difficult. Despite these limitations, a few methods have been developed to address the problem of inferring heterozygosity for ancient samples (Kousathanas et al. 2017). Additionally, other methods have leveraged the power of allele frequencies or recombination maps to predict IDB tracks and infer runs of homozygosity in ancient samples (Vieira et al. 2016). However, the necessary allele frequencies are not always available for past populations or populations poorly represented by public datasets. Furthermore, drift in the lineage of the reference panel or in the sample might skew allele frequencies. Finally, the presence of long and prevalent ROHs can drive down the genome-wide estimate of heterozygosity (Prüfer et al. 2014).

Here, we introduce ROHan, a method to jointly estimate the local and global heterozygosity rates as well as long ROHs. This method is suitable for both modern and ancient DNA samples at various levels of coverage. Our method relies on a maximum weighted likelihood method to first estimate the rate of heterozygosity locally. It then applies a HMM to simultaneously identify regions in ROHs and compute Watterson’s $\theta$ for regions that were identified as non-ROH. Our method operates on aligned DNA fragments in BAM format on an individual basis. It does not require allele frequencies or any information provided by the reference genome, and only makes use of the sequence data underlying a given sample. The source code is available at https://www.dropbox.com/s/kmn4hfnj29w/k/w?/ROHan-master.zip?dl=0.

Using genomic simulations incorporating aDNA damage, and investigating the effect of coverage, population size and inbreeding, we show that ROHan is more accurate and robust than previous methods aimed at inferring rates of heterozygosity. We demonstrate that ROHan can infer global and local rate of heterozygosity for modern samples with coverage as low as 5-6X and in ancient samples as low as 7-8X even in the presence of substantial damage. For inbred samples, our method can correctly identify ROHs. Masking such regions provides more accurate estimates of global rates of heterozygosity genome-wide than current methods not aided by external allele frequencies.

We also tested ROHan on modern and ancient empirical samples for both human and non-human species. Specifically, we used our methodology on a dozen low-coverage samples from the 1000 Genomes project Phase III (Genomes Project Consortium et al. 2015) and show that our estimates are consistent with the ones presented by the Simons Genome Diversity Project (Mallick et al. 2016) for similar populations sequenced to higher coverage. We also provide heterozygosity estimates for a range of ancient humans spanning a whole range of post-mortem damage and coverage. Additionally, applying our methodology to individual chimpanzee genomes, we obtain more consistent estimates than those reported in the original publication (De Manuel et al. 2016). Finally, we ran ROHan on several horse samples, both modern and ancient, and confirm that the endangered Przewalski’s horses have drastically higher rates of ROH and lower rates of genome-wide heterozygosity. Their Eneolithic direct ancestors showed larger genetic diversity and were not found to be inbred.

Materials and Methods

Our method proceeds in three steps. It first estimates genome-wide coverage (step 1), then estimates local rates of heterozygosity using a user-specified genomic window size (step 2) and finally runs an HMM over the local rate of heterozygosity to simultaneously identify regions in ROH and genome-wide $\theta$ (step 3). This section presents the underlying probabilistic model as well as our simulation framework.

Computational Model

The first step is to get an estimate of the genome-wide coverage from the average per base coverage a few genomic loci. The coverage found at a single site is further used in step 2 in order to weight its contribution to the likelihood function by comparing with the genome-wide coverage. Further details about the coverage correction are found in the text below.

In subsequent sections, as we consider aDNA as well as modern DNA, we use the word fragment to describe individual sequences aligned against a reference genome. We use this word
to distinguish them from reads which represent the raw data as obtained from the sequencing instrument. For ultra-fragmented aDNA fragments, it is crucial to reconstruct the original DNA fragment given the raw reads by removing sequencing adapters at the ends and potentially merging overlapping mates (Kircher 2012).

We first detail how we obtain the local rates of heterozygosity and follow by presenting the HMM model.

Let us define the following variables:

**Data:**
- $b$: any DNA base such $b \in \{A, C, G, T\}$
- $b_i$: a DNA base post-deamination
- $b_j$: the ancestral base
- $b_{ij}$: either a derived base if a mutation occurred or equal to the ancestral otherwise
- $h$: heterozygosity rate
- $\theta$: Watterson’s theta

**G:** all possible 16 genotypes \( \{A, C, G, T\} \)

**$\mathbb{D}_i$:** the entire data over a genomic window such that \( \mathbb{D} = \bigcup \mathbb{D}_i \)

**$\mathbb{D}_i$:** set of all bases at genomic position $i$ such that $\mathbb{D}_i = \bigcup d_{ij}$

**$\kappa$:** the ratio of transitions over transversions

**$C_i$:** coverage at site $i$

**Probabilistic events:**
- $M$: a mismapping event on a specific fragment
- $D$: a deamination event
- $E$: a sequencing error event for a specific base on a specific fragment

We consider 16 distinct genotypes instead of 10. For instance, we consider $b_3 = A, b_4 = C$ to be a distinct genotype from $b_3 = C, b_4 = A$. The use of 16 genotypes has previously been suggested in the literature to account for indels (Luo et al. 2017).

**Local estimates of heterozygosity** For a given genomic window, we seek to find $\hat{h}$ that satisfies the following:

$$
\hat{h} = \arg\max_h (P[h|\mathbb{D}])
$$

(1)

The marginal likelihood for $\mathbb{D}$ is not useful for the optimization. By applying a uniform prior on $h$, we have:

$$
P[h|\mathbb{D}] \propto P[\mathbb{D}|h]
$$

(2)

By assuming that site $i$ represents an independent observation, we use a weighted log-likelihood approach (Hadi and Luceno 1997) to estimate the total log-likelihood:

$$
\log(P[\mathbb{D}|h]) = \sum_i w_i \log (P[\mathbb{D}_i|h])
$$

(3)

where $w_i$ is the weight depending on coverage at site $i$. This weight is aimed at mitigating the impact of potential copy number variations. The exact derivation of such weights are detailed in the Supplementary Section S.7. Finally, the likelihood of observing the data $\mathbb{D}_i$ at site $i$ is given by marginalizing over each 16 genotypes:

$$
P[\mathbb{D}_i|h] = \sum_{G \in \mathbb{G}} P[\mathbb{D}_i|G]P[G|h]
$$

(4)

$P[G|h]$ is the prior on the genotype given the heterozygosity rate. The term $P[\mathbb{D}_i|G]$ is the genotype likelihood. Both are defined in the following sections.

**Genotype prior** To compute $P[G|h]$, the prior probability on the genotype given heterozygosity rate $h$, we consider two possibilities:

1. $G$ is homozygous with probability $(1-h)$ such that $b_a = b, b_d = b$. The probability that the $G$ is homozygous is given by:

$$
P[G = \{b_a = b, b_d = b\}] = P[b_a = b](1-h)
$$

(5)

where $P[b_a = b]$ is simply $f_b$, representing the genomic frequency of occurrence the base $b$ in the genome. For humans, this is $f_A = f_T \approx 0.3$ and $f_C = f_G \approx 0.2$.

2. $G$ is heterozygous with probability $h$ such that $b_a = b_1, b_d = b_2$ and $b_1 \neq b_2$. The prior on the genotype is therefore the probability that $b_a$ was the ancestral base multiplied by the probability that a specific mutation happened:

$$
P[G = \{b_a = b_1, b_d = b_2\}] = P[b_a = b_1]P[b_1 \rightarrow b_2]h
$$

(6)

The term $[b_1 \rightarrow b_2]$ depends on the type of mutation:

(a) For transitions, we compute the probability of a transition occurring given the transition/transversion ratio:

$$
P[b_1 \rightarrow b_2] = \frac{\kappa_{tv}}{\kappa_{tv} + 1}
$$

(7)

(b) For transversions, as there are 2 transversions from a given ancestral base, we consider each to be equally likely:

$$
P[b_1 \rightarrow b_2] = \frac{1}{2(\kappa_{tv} + 1)}
$$

(8)

**Genotype likelihood** For a given genotype $G$ and heterozygosity rate $h$, the genotype likelihood is computed by assuming that each base at site $i$ represents independent observations. Since coverage at site $i$ is $C_i$:

$$
P[\mathbb{D}_i|G] = \prod_{1 \leq i \leq C_i} P[d_{ij}|G]
$$

(9)

as $P[d_{ij}|G]$ depends the genotype $G = (b_a, b_d)$ we rewrite $P[d_{ij}|G] = P[d_{ij}|b_a, b_d]$. Since we could have sampled from either chromosome with probability $\frac{1}{2}$, this expression is calculated as follows:

$$
P[d_{ij}|G] = P[d_{ij}|b_ab_d] = \frac{1}{2} P[d_{ij}|b_a] + \frac{1}{2} P[d_{ij}|b_d]
$$

(10)

For a given base $b$ (either $b_a$ or $b_d$), the probability of observing $d_{ij}$ depends on whether the fragment to which $d_{ij}$ is mismapped:

$$
P[d_{ij}|b] = (1 - m)P[d_{ij}|b, \neg M] + mP[d_{ij}|b, M]
$$

(11)

where $M$ is the event that a mismapping event occurred on the DNA fragment where $d_{ij}$ is located. $P[d_{ij}|b, M]$ is defined in equation 17. To quantify $M$, we simply use the mapping quality of the read as our simulations confirm this as a reasonable approximation (see Supplementary Methods section S.1.7). If the aDNA fragment is correctly mapped, two potential events can create a mismatch between the sample base $b$ and the observed $d_i$ - a deamination event or a sequencing error. We consider both events to be successive and the position within fragments where potentially both could have occurred (see Figure 1).
We consider the base $b_p$ to be the base after a potential post-mortem deamination reaction. For the Illumina sequencing technology, this base can be constructed as the base on the flow-cell prior to cluster amplification. As this base is a nuisance parameter, we marginalize over it:

$$P[d_{ij}|b_p, \neg M] = \sum_{b_p = \{A, C, G, T\}} P[d_{ij}|b_p]P[b_p|b]\quad(12)$$

the latter term $P[b_p|b]$ is given by the rate of misincorporation due to deamination:

$$P[b_p|b] = \begin{cases} 1 - \sum_{b'} f_{\text{deam}}(b' \rightarrow b_p) & \text{if } b = b_p \\ f_{\text{deam}}(b \rightarrow b_p) & \text{if } b \neq b_p \end{cases}\quad(13)$$

where $f_{\text{deam}}(b \rightarrow b_p)$ is rate of substitution from original base $b$ to $b_p$. These substitutions should generally be 0 if $b \neq C$ unless there is a type of chemical damage which cannot be due to sequencing errors. These rates are given as input by the user and must be as accurate as possible. Given the base $b_p$, the probability of observing $d_{ij}$ depends on whether a sequencing error happened or not:

$$P[d_{ij}|b_p] = (1 - c)P[d_{ij}|b_p, \neg E] + cP[d_{ij}|b_p, E]\quad(14)$$

where $P[d_{ij}|b_p, E]$ is simply defined by the frequency of base substitution for the given sequencing instrument. Users can pick a simple frequency of $\frac{1}{4}$ for all pairs of bases but due to the idiosyncrasies of Illumina sequencers (Nakamura et al. 2011), empirical Illumina base substitution frequencies are supplied with the software. In such cases, this expression simply becomes:

$$P[d_{ij}|b_p, E] = f_{\text{seq}}(b_p \rightarrow d_{ij})\quad(15)$$

where $f_{\text{seq}}(b_p \rightarrow d_{ij})$ is the frequency of substitution from base $b_p$ to base $d_{ij}$ given that a sequencing error has occurred. These frequencies can be obtained using a sequencing run where DNA libraries have been pooled together with a DNA library constructed on a known genome. In the case of the frequencies supplied with the software, such frequencies were computed using Illumina control reads aligned to the PhiX174 genome. Finally, in the absence of a sequencing error, the first term in equation 14 becomes:

$$P[d_{ij}|b_p, \neg E] = \begin{cases} 1 & \text{if } b_p = d_{ij} \\ 0 & \text{if } b_p \neq d_{ij} \end{cases}\quad(16)$$

Thus far, we have assumed that the DNA fragment for base $d_{ij}$ is correctly mapped. In equation 11, the second term accounts for when this fragment is mismapped. In this case, the probability of observing this base is completely independent of $b$:

$$P[d_{ij}|b, M] = f_{d_{ij}}\quad(17)$$

where $f_{d_{ij}}$ is the expected frequency of occurred of $d_{ij}$. This is usually straightforward but aDNA damage can skew these frequencies by decreasing the probability of finding cytosines at the 5’ end due to deamination. For aDNA, these frequencies depend on the position of the fragment and length of the fragment. Further detailed can be found in the Supplementary Methods, section S.1.1.

To decrease runtime at the cost of increased memory usage, rates of base substitution for a given mapping quality and base quality can be precomputed as the probability space is already discretized due to the use of integers to represent quality scores and mapping quality.

As the goal is to find $h$ from equation 1, we use a gradient descent with momentum (Rumelhart et al. 1986) to find the heterozygosity rate with the highest likelihood. ROHan precomputes the genotype likelihoods and computes prior probabilities for all 16 genotypes at each iteration of the gradient descent. For a given genomic window, the error bounds for $h$ is obtained using the following:

$$\frac{1.96}{\sqrt{\frac{\delta^2 P(h|D)}{\delta h}}}\quad(18)$$

We noticed that low-coverage samples consistently yielded underestimates due to heterozygous sites appearing as homozgyous resulting from the limited chance of sampling the other allele. While for sites with high depth of coverage this is unlikely, for a coverage of 2X for instance, this will happen with a probability of $\frac{1}{4}$. A correction factor was applied to the heterozygosity estimates to overcome this limitation. After the optimization has converged for a local estimate of heterozygosity, this estimate is multiplied by this corrective factor to retrieve reliable estimates (see details in Supplementary Method S.1.3).

**Hidden Markov Model** We use a modified 2-state HMM with customized forward and backward algorithms. The first state corresponds to being in a ROH whereas the second corresponds to being in a non-ROH region. We use a single transition parameter $p$ for both states and added features to account for chromosomal start/end and undefined genomic windows. Briefly, we have modified the HMM to not include the probability of transition from a state at the end of a chromosome to the beginning as those are independent. This applies as well for undefined regions. Finally, the log-likelihood in the forward algorithm is weighted by the fraction of sites that are defined in the particular window.

Each state has a parameter $\theta$ corresponding to Watterson’s theta estimate. Given the local heterozygosity estimate, we compute the expected value of segregating sites $S$ in that genomic window by multiplying the estimated heterozygosity rate by the size of the window. For a given small non-recombining locus (NRL), it has been reported in the literature that $S$ should follow an exponential distribution (Watterson 1975). However, a sufficiently large genomic window will be composed of multiple NRLs.

As the sum of exponential distributions is a negative binomial distribution, it has also been suggested in the literature that, for a sufficiently large genomic window, the number of segregating sites follows a negative binomial distribution (Pitters 2017).
Using coalescence simulations, we confirmed that a negative binomial distribution was indeed a better fit than a standard binomial distribution (see S.1.4 in the Supplementary Material).

We construct $S$ along a genomic window of length $L$ to be the sum of exactly $s$ NRLs. For a given global $\theta$, the exponential rate for any single NRL is given by $\theta' = \frac{\theta}{S}$. We obtain the following:

$$P[S|\theta] = \left(\frac{S + \frac{1}{s} - 1}{S}\right) \left(1 - \theta'\right)^\frac{1}{2} \theta^S$$

(19)

To infer the parameters $(\theta, s, p)$ given local estimates of heterozygosity, we use a Markov Chain Monte Carlo (MCMC) approach to obtain point estimates as well as error bounds. Please refer to (Rydén et al. 2008) for a discussion about the use of expectation-maximization vs MCMC for HMMs. The probability of emitting a certain number of segregating sites given the ROH state is obtained by using a small value of $\theta$ to reflect either de novo mutations or potential miscalls. As local estimates of heterozygosity can differ greatly especially at low coverage, we run the MCMC three times, once using the lower bound estimates for $h$, a second time using the point estimates and finally, using the upper bound estimates. Once the 3 MCMC chains have converged, the minimum and maximum values used as the lower and upper bound of the confidence interval. The average of the MCMC running on the mid values is used as the point estimate.

**Simulations**

To test our methodology, we simulated a set of non-inbred and inbred datasets, using the full human chromosome 1 from hg19 as the genomic reference. To avoid gaps, unresolved bases were filled using a second-order Markov chain trained on the human genome. A total of 16 unrelated haploid chromosomes were generated using msprime (Kelleher et al. 2016) to form 8 diploid individuals. We used the recombination map from HapMap phase II (International HapMap Consortium et al. 2007) in msprime to generate a complete human chromosome 1. As msprime does not currently assign actual bases to the segregating sites, we used the base in the human reference as ancestral allele and added mutations with a $K_{tp}$ of 2.1.

A total of 5 different effective population sizes were used (see Table 1). The individual haploid chromosomes were recombined to produce a sexual gamete using the recombination map from HapMap phase II (International HapMap Consortium et al. 2007). The number of recombinations was on par with rates previously reported in the literature (Li 2011). These gametes were combined in a pairwise fashion to create a diploid individual (see Figure 2 for a schematic overview of the non-inbred pedigree). The 16 haploid chromosomes were combined to form 4 grandparents, 4 parents (2 siblings per couple) and finally 2 diploid individuals corresponding to the great-grandchildren of the original 16 haploid chromosomes. These 2 diploid individuals are used as input to gargamnel (Renaud et al. 2016) to simulate DNA sequencing reads with errors. We simulated a coverage of 30X. These initial 30X genomes were then downsampled to evaluate the program performance with low-coverage data.

To simulate post-mortem damage, we used 3 types of aDNA damage profiles: 1) a high rate of post-mortem deamination consistent with the use of a double-stranded DNA protocol for library preparation (Meyer and Kircher 2010) using the ATP2 sample from (Gamba et al. 2014); 2) a medium rate of post-mortem deamination from a double-stranded DNA library protocol from the LaBraña sample described in (Sánchez-Quinto et al. 2012) and 3) a low rate of post-mortem deamination corresponding to the damage found in a single-strand aDNA library (Gansauge and Meyer 2013) using the Ust’-Ishim sample from (Fu et al. 2014). Please see Supplementary Section S.1.6.2 in the Supplementary Material for further details about the substitution rates and patterns considered.

Sequencing reads were simulated using a read length of 125bp in the single-end mode with the sequencing error profile of an Illumina HiSeq2500. To further test the robustness of our model, we also drastically increased the simulated error rates (see Supplementary Section S.1.6.3 for details).

The *in silico* sequencing adapters were trimmed using lee-Hom (Renaud et al. 2014) and mapping was conducted using a customized version of BWA version 0.5.9 3.

To test ROHan’s ability to infer ROHs, we tested 3 scenarios of inbreeding: 1) between siblings ($F_\text{b} = \frac{1}{4}$) between a grandparent and grandchild ($F_\text{g} = \frac{1}{8}$), and; 3) between first cousins ($F_\text{c} = \frac{1}{16}$). Please refer to section S.1.6.1 for the simulated pedigrees in the Supplementary Materials for details.

**Table 1 Simulated values of effective population size ($N_e$) and expected $\theta$**

<table>
<thead>
<tr>
<th>$N_e$</th>
<th>$\mu$</th>
<th>$\theta = 4N_e\mu$</th>
<th>Expected average $h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>$2 \times 10^{-8}$</td>
<td>0.00024</td>
<td>2.399e-04</td>
</tr>
<tr>
<td>5000</td>
<td>$2 \times 10^{-8}$</td>
<td>0.00040</td>
<td>3.998e-04</td>
</tr>
<tr>
<td>7000</td>
<td>$2 \times 10^{-8}$</td>
<td>0.00056</td>
<td>5.597e-04</td>
</tr>
<tr>
<td>9000</td>
<td>$2 \times 10^{-8}$</td>
<td>0.00072</td>
<td>7.195e-04</td>
</tr>
<tr>
<td>12000</td>
<td>$2 \times 10^{-8}$</td>
<td>0.00096</td>
<td>9.591e-04</td>
</tr>
</tbody>
</table>

![Figure 2](https://github.com/mpieva/network-aware-bwa/)

To evaluate the estimate of heterozygosity on a small but substantial chromosomal region due to the demanding computational resources, we subsampled the first 15 Mbp of chromosome 1 and ran ROHan, ATLASv1.0 and ANGSD v0.919-14 (refer to Supplementary Section S.1.5 for the precise commands). For ANGSD, we used the recommended genotype likelihood model ("-GL 1") for estimating $\theta$ (see Supplementary Section S.1.5 for a brief discussion regarding this parameter). We evaluated the robustness of such software to low coverage, aDNA damage,
and various effective population sizes. As the original sequence of the chromosomes used for simulation was available, we could evaluate the number of segregating sites at the local as well as global level.

We also evaluated BCFtools/RoH (Narasimhan et al. 2016) using version 1.4.1 of BCFtools and PLINK (Purcell et al. 2007) v1.90 to assess their accuracy to predict large and medium size ROH compared to ROHan. We simulated an extra 1000 chromosomes in msprime and used the allele frequencies from those. The population providing these 1000 chromosomes was the same as the one from which the 16 haploid chromosomes were taken from to provide an ideal test set. However, to test the robustness of BCFtools/RoH to allele frequencies from more distant populations, we repeated the simulations by joining the population from which the BAM files are generated and the population providing the allele frequency farther back in time, namely at 150k and 500k years ago. The former case would correspond to trying to infer ROHs in an ancient Khoe-San individual and the latter in a Neanderthal individual while using allele frequencies from an Eurasian population.

Results

Simulated data

Local heterozygosity estimates We start by evaluating ROHan’s ability to estimate local heterozygosity rates using genomic windows of 1Mb and simulated data. As mentioned above, ROHan computes local estimates of heterozygosity rates which are then be used to infer the genome-wide estimate of \( \theta \). As we have the sequence of the chromosomes used in the simulations, we could compare the estimated rate of heterozygosity to the simulated one. The results in absence of post-mortem DNA damage and for various depth of coverage are found in Figure 3. For coverage equal to 3X, we find that the point estimate is generally inferior to the expected value and that confidence intervals are large. At 5-fold coverage, we find narrower confidence intervals and point estimates closer to the expected value. This trend toward higher precision and accuracy is confirmed when increasing coverage to 9X. It is noteworthy that the first genomic window seems to be consistently underestimated in our experimental framework, probably due to a poor correlation between the reported and true mapping qualities or due to the lower mappability of the region (16% of the first 1Mb were undefined bases whereas 3.7% of the first 15Mb, these undefined bases were replaced by defined bases).

The Supplementary Material (see Supplementary Figures S.9-S.14) provides the results of more extensive simulations, including effective population sizes of 3000 and 9000, coverage variation between 3, 5 and 9X, and for various types of nucleotide misincorporation patterns due to the ancient DNA damage. At \( N_e = 9000 \), we notice large confidence intervals at a coverage of 3X regardless of the damage patterns considered. For samples extremely affected by post-mortem damage (eg the ATP2 sample), ROHan even fails to produce confidence intervals overlapping with the expected value, and often provides underestimates. For this type of damage patterns, the results improve in accuracy with a coverage of 5X. However one can still see the impact of having a high rate of aDNA damage on both precision and accuracy. At 9-fold coverage, there is little difference in terms of accuracy between the sample with heavy damage compared to the ones with either very little or no damage at all. Although the confidence interval obtained generally comprises the expect value, the point estimate recovered is generally slightly underestimated.

Global heterozygosity estimates In ROHan, the local estimates of heterozygosity are used together with an HMM to compute the genome-wide estimate of Watterson’s \( \theta \). We compared the simulated value for the entire 15Mbp of simulated data to the global estimates of \( \theta \) for the same data. This was done for various levels of heterozygosity, aDNA damage and coverage. The results obtained when considering a sample with medium rates of damage associated with a double-stranded DNA library preparation protocol can be found in Figure 4. The remaining results obtained can be found in the Supplementary Material (see Supplementary Figures S.15-S.19).

In general, the only time where the confidence interval did not include the simulated values was at 0.9X for the cases with medium and high rates of aDNA damage associated with a double-stranded DNA library preparation protocol. However, analyses carried out on the basis of 1X-3X coverage data were extremely imprecise. From coverage of 8-fold and above, the point estimate recovered was stable and close to the simulated value (although slightly underestimated), regardless of the amount of aDNA damage considered. Decreasing coverage generally resulted in underestimated values.

We compared our results to those obtained with ATLAS and ANGSD, using the same 15Mbp simulations (see Supplementary Figures S.20-S.24 and Supplementary Figure S.25-S.32). In general, we found that ATLAS undercompensated for either sequencing errors or aDNA damage, which leads to overestimates in the value of \( \theta \). This overcompensation issue is consistent at a coverage of 10X or higher, but can be introduced at lower coverage depending on the aDNA damage level considered. Furthermore, the confidence interval for the point estimate rarely includes the expected value. While ANGSD does not provide confidence intervals, it consistently returns underestimated values in absence of aDNA when coverage is inferior to 10-fold. In the presence of little to moderate aDNA damage, largely overestimated values are returned, however, the recovered estimates converge to the expected value given sufficient coverage (>20X-30X). In the presence of high levels of aDNA damage, ANGSD consistently returns largely overestimated values, regardless of the coverage considered. We found that this effect could be mitigated by disregarding transitions (C,G \( \rightarrow \) T,A transitions are the most prevalent nucleotide mis-incorporation resulting from post-mortem damage (Briggs et al. 2007)) and restricting the analyses to transversions only. For instance, using an effective population size of 9000, using only transversions can help the point estimate converge to the expected value as long as high-coverage data (15X-20X) are provided (see Supplementary Figure S.29).

As some sequencing runs can have very high error rates, we sought to test whether ROHan’s model for sequencing error was robust to elevated sequencing error rates. We repeated the simulation while increasing the rate of sequencing errors to 1.6% which represents a 10 fold increase compared to previous simulations, first in the absence of aDNA damage (see Supplementary Figure S.39). The results indicate that generally, ROHan is sufficiently robust as long as coverage equal to 4X and above are considered. When high sequencing error rates are combined the highest levels of simulated aDNA damage the point values recovered appears consistently underestimated until high-coverage data are available (>20X-25X) but the confidence intervals include the expected value from coverage values above 10X.
Figure 3 Comparison between the simulated local rates of heterozygosity versus the predicted one’s using windows of 1Mbp for simulated data without any aDNA damage patterns at various rates of coverage. The original 30X simulated data was downsampled at 3 (left), 5 (middle) and 10X (right) and the effect on the predicted rate of heterozygosity at a local level was compared to the simulated one. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue cross represent the simulated value.

Figure 4 Global estimate of $\theta$ using simulated data at various coverage with an $N_e$ of 9000 and simulated damage rates from the La Braña sample, showing medium rate of post-mortem damage. The dotted line represents the target rate of heterozygosity.

To further assess how effective ROHan was in handling post-mortem cytosine deamination, which introduces an excess of nucleotide mis-incorporation in the sequencing data), we re-ran ROHan but forcing the model’ probabilities of aDNA damage to zero. Results show that while the estimates retrieved on simulations carried out in absence of aDNA damage were accurate, those carried out in the presence of increasing levels of aDNA damage consistently returned overestimates, regardless of the coverage considered (see Supplementary Figure S.34). This is expected as an underestimate of the error automatically leads to an overestimate of $\theta$.

Finally, we sought to test whether blending 2 different libraries with drastic different rates of aDNA damage has a significant impact on the predicted rates of heterozygosity. Such situations can happen when different molecular tools are used during library preparation (Rohland et al. 2015), and when different extracts from the same individual are used during library preparation (Seguin-Orlando et al. 2014) (see Supplementary Figure S.36). As expected, the measured rates of damage on the new dataset was intermediate between the ones of the original sets (i.e. aDNA data showing the highest damage levels and no damage, respectively). Although the point estimates were consistently underestimated for all coverage considered (~2X~28X), all confidence intervals retrieved intercepted the expected values. Relatively accurate point estimates were obtained from 10X-12X coverage and above.

In terms of runtime, running ROHan on a 5X dataset consisting of the human chromosome 1 (~250Mbp), and using 4 Intel Xeon cores at 3.50GHz took 10m0.1s and about 2G of RAM for the estimate of the local heterozygosity. Running the HMM to map ROHs took 4m39.2s on a single core.

Infer ROHs in inbreed samples We tested ROHan, PLINK and BCFtools/RoH on a simulated chromosome corresponding to human chromosome 1 for various inbreeding scenarios as well as different levels of coverage. For inbreeding scenario 1 (mating between full siblings) and in the absence of aDNA damage, we find that ROHan can accurately estimate the total proportion of the genome in an ROH using windows of 1Mbp for the estimate of the local heterozygosity as long as at least ~5X coverage data are provided (see Figure 5). The results for the remaining inbreeding scenarios indicate similar performance and are presented in the Supplementary Results (see Supplementary Figure S.42).

Expectedly, the ROHs delineated by ROHAN were found to be of uneven sizes due to uneven recombination rate across the chromosome (see Supplementary Figure S.41 for the distribution of segregating sites). Both the centromere region and the last portion of the chromosome were associated to a local depression of the heterozygosity rate and were correctly decoded by RO-Han. The accuracy achieved for different coverage and window sizes for the local estimate of heterozygosity can be found in
Figure 5 Estimates of the proportion of the genome in an ROH as predicted by ROHan on a simulated full chromosome of 250Mbp at various depth of coverage compared to the simulated rate using the original simulated chromosomes from the diploid organism. The proportion of ROH reflects inbreeding between siblings. The dotted line represents the target fraction of the genome in a ROH obtained from the simulated chromosome. Whiskers represent the 95% confidence interval. Both the detection of segregating sites for the computation of the theoretical value as well as ROHan used a window size of 1Mbp.

Comparison to existing tools reveals that PLINK seems to reliably predict large ROHs at a coverage of ∼10X and above but also seems to overpredict some small ROHs, an effect which tends to disappear as coverage increases (see Supplementary Figure S.50). In comparison, the results for BCFtools/RoH for both long and short ROHs seem stable at ∼10X and above but seems to predict fewer small ROHs compared to PLINK (see Supplementary Figure S.51). However, the allele frequencies used for computation were selected to be perfectly known. Therefore, this simulation framework does not assess the method’s performance in the case where allelic frequencies are obtained from a distant population. To test the robustness of BCFtools/RoH to this, we repeated the test using join times between the lineage providing the allele frequency and the simulated chromosomes at 150k and 500k years ago (see Supplementary Figure S.52 and S.54, respectively). At a split time of 150k years ago, large ROHs could be detected at around 5-10X but the signal was too unstable to resolve short ROHs. Using a split time of 500k years, even large ROHs were difficult to delineate, regardless of the coverage considered.

As ROHan requires the user to specify the size of the genomic window used for the estimate of local rates of heterozygosity, we finally sought to evaluate the accuracy of our methodology if different sizes of genomic windows were specified. To achieve this, we ran ROHan on 2 simulated sets, with a simulated Ne of 3000 and 9000 respectively and with window sizes of 100, 250, 500 and 1000kbp. The results for such tests are found in the Supplementary Results section (see Supplementary Figures S.37-S.38). We found that when using smaller windows of 100kbp, confidence intervals tend to be stable around 8-10X. For windows of 250kbp, a coverage of 7-8X and above is recommended whereas for windows of 500kbp, we obtain reliable estimates at coverage of 6-7X and above. Finally, for windows of 1Mbp, confidence intervals seem stable around 5X and above. Due to limited computational resources, it should be noted that these tests were run without added simulated aDNA damage.

Empirical samples
In the following section, we applied our methodology to empirical data, where in contrast to simulations, the correct value of the heterozygosity rate or the location of ROHs are not known in advance. As our methodology is both applicable to ancient and modern samples and human as well as non-human animals, we investigated all four possibilities. Overall, we found that our results mostly agreed with previously reported estimates,
excepting a few cases where the new estimates recovered appear to be more consistent with the literature.

Modern samples

Humans

We first downloaded 26 present-day human genomes in BAM format from 1000 Genomes Project Phase 3 (Genomes Project Consortium et al. 2015), all of which had relatively limited coverage (7.8X on average). We sought to evaluate (1) whether the genome data is indicative of inbreeding in these individuals and (2) whether the genome-wide \( \theta \) estimates recovered from ROHan are compatible with the ones obtained by the Simons Genome Diversity Project (Mallick et al. 2016) which had access to data at a much higher depth of coverage for the same populations (43X on average). We considered various individuals with ancestry from Africa, Eurasia and Indigenous People of the Americas. It is expected that heterozygosity will vary according to drift and that individuals of African ancestry will have the highest heterozygosity rate (Ramachandran et al. 2005).

We found that only four the 26 individuals considered showed signs of minor inbreeding (i.e., approximately 0.03−0.17% of their genome consisted of ROHs; see Table 2). A visualization of the output for a single individual can be found in Figure 6. The results for the remaining individuals are found in the Supplementary Results (see Supplementary Table S.5). We also find that our estimates of \( \theta \) for these low-coverage individuals are consistent with the ones obtained by (Mallick et al. 2016) while using higher coverage genomes. This shows the robustness of our method to samples with lower coverage.

non-Humans

We next considered the high-coverage chimpanzee data from (De Manuel et al. 2016), including 3 animals from 3 different geographical locations in Africa (Western, Central and Eastern) and sequenced at an average coverage of 24.6X (Table 3). In the original publication, heterozygosity rates were reported on an individual basis and Watterson’s \( \theta \) were computed for each of the 3 populations using G-PhoCS (Gronau et al. 2011). Using ROHan, we found little evidence of large ROHs in those samples. Consistently with the original publication, we find Central chimpanzees to have a greater effective population size than the Eastern ones which in turn, have a greater effective population size than the Western chimpanzees. The genome-wide \( \theta \)s reported for each individual by ROHan are consistent with the ones reported by G-PhoCS for their population of origin and significantly larger than those originally reported by (De Manuel et al. 2016). ROHan estimates also appear on par with the estimates produced by another method (ANGSD), which we demonstrated on the basis of simulations to converge to the correct value at equivalent coverage.

Ancient Samples

Humans

We next used publicly available ancient hominin genomes sequenced in various aDNA research centers and encompassing a full range of post-mortem DNA damage to estimate genome-wide \( \theta \) and detect ROHs using ROHan. For comparison purposes, we ran ANGSD with and without including transitions in the calculation. We also report the heterozygosity rate previously measured, if that information was available (see Table 4).

We found a very low rate of heterozygosity for the Vindija Neanderthal 33.19 sample, despite the presence of extensive aDNA damage signatures. Likewise, for the Stuttgart early Neolithic farmer, both ANGSD’s and ROHan’s \( \theta \) estimates are similar to the one obtained in the original publication by (Lazaridis et al. 2014). However, for both the Loschbour and Ust’-Isbash hunter-gatherers, ROHan estimates seem slightly higher than the ones from the original publication. In general, ROHan estimates are consistent with those obtained by ANGSD using transversions only, except when low coverage data are available (e.g. Barcin 31, AndronovoS05 and Wezmeh Cave 1, sequenced at 3.14, 9.47 and 12.74-fold coverage, respectively). For the Wezmeh Cave 1 early Neolithic farmer sample from (Broushaki et al. 2016), the obtained heterozygosity rate using both ANGSD and ROHan are not consistent with the estimates reported by the original publication which were computed using ATLAS. Following the results from our simulations, we can assume that, at an equivalent coverage (∼13X), ANGSD provide underestimates of \( \theta \) while ATLAS provides overestimates. ROHan is expected to return accurate estimates, albeit at the cost of large confidence interval. This is consistent with our observations.

Subsampling the Neanderthal Vindija 33.19 sample down to 1X provided an ideal empirical test case of the robustness of our method, in case it is applied to a difficult sample combining both high levels of aDNA damage and very low rates of heterozygosity. We obtained confidence intervals encompassing the global heterozygosity estimates retrieved from the full data at coverage of 9X and above (see Supplementary Figure S7). However, the point values retrieved for coverage inferior ∼18X were consistently underestimated.

non-Humans

We next sought to evaluate the heterozygosity of one ancient dog from Ireland, which dates back to 4.8k year old and whose genome was sequenced in (Frantz et al. 2016). Raw reads were downloaded from the European Nucleotide Archive (ENA), trimmed using leeHom v.1.1.5 (Renaud et al. 2014) and aligned using BWA (Li and Durbin 2009) 0.5.10. Using ROHan, we obtained an estimate of genome-wide \( \theta \) of 1.29 × 10^{−3} (95% confidence interval: 1.18 × 10^{−3}−1.42 × 10^{−3}). The estimate retrieved in ANGSD when considering all substitution types was more than doubled (\( \theta \approx 2.97 × 10^{−3} \)). However, restricting the analysis to transversions only lowered the \( \theta \) estimate to 0.99 × 10^{−3} as the original sample had extensive damage. In this case, the estimate was obtained by multiplying by K_{ref}+1 (3.1) to obtain a comparable value (including both transitions and transversions). Both the original and ROHan estimates are in agreement with previously reported values of \( \theta \) for modern wolf and dog breeds (Wang et al. 2013).

Finally, we ran ROHan on 13 ancient and 20 modern horse genomes as an example of domestic animals were ROHs could potentially be identified.

Specifically, the 20 modern domestic horses represented a wide range of breeds, including Arabian, Mongolian, Yakutian, Icelandic, and Shetland horses (Leegwater et al. 2016; Jäderkvist et al. 2014; Librado et al. 2015; Der Sarkissian et al. 2015; Metzger et al. 2014; Frischknecht et al. 2015; Do et al. 2014; Kim et al. 2013), as well as 6 endangered Przewalski’s horses. The 11 ancient horses considered spanned a large temporal range, from 2.3kyrs ago to 43 krys ago, and represented both wild horses that lived prior to domestication, Eneolithic early domesticates and Iron Age domesticates (Schubert et al. 2014; Librado et al. 2017; Gaunttiz et al. 2018). Corresponding results are presented
Table 2 Estimated values of $\theta$ for human samples with medium coverage

<table>
<thead>
<tr>
<th>sample ID $^a$</th>
<th>pop. code</th>
<th>coverage (X)</th>
<th>$\theta$ $^b$</th>
<th>$\theta_{low}$</th>
<th>$\theta_{high}$</th>
<th>$\theta$ from SGDP $^c$ in ROH (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG02367</td>
<td>CDX</td>
<td>7.2</td>
<td>8.05949</td>
<td>6.87336</td>
<td>9.30138</td>
<td>7.93714-8.2176</td>
</tr>
<tr>
<td>NA21141</td>
<td>GIH</td>
<td>7.8</td>
<td>9.09886</td>
<td>7.99777</td>
<td>10.2094</td>
<td>8.63594 $^d$</td>
</tr>
<tr>
<td>HG04222</td>
<td>ITU</td>
<td>8.2</td>
<td>9.24757</td>
<td>8.18621</td>
<td>10.3226</td>
<td>8.26589-8.87523</td>
</tr>
<tr>
<td>HG03139</td>
<td>ESN</td>
<td>7.3</td>
<td>11.4978</td>
<td>10.1661</td>
<td>12.8663</td>
<td>10.9226-11.4413</td>
</tr>
</tbody>
</table>

$^a$ Four different individuals from the 1000 Genomes Project Phase III (Genomes Project Consortium et al. 2015) for which minor amounts of long ROHs were detected. The population codes are as follows: CDX: Chinese Dai in Xishuangbanna, China GIH: Gujarati Indian from Houston, Texas ITU: Indian Telugu from the UK ESN: Esan in Nigeria. ROHs inferred on chromosome 11 are plotted in Supplementary Figures S.56

$^b$ The reported values of $\theta$ were multiplied by $10^4$

$^c$ reported $\theta$ from the same population

$^d$ closest from Kashmiri Pandits

Table 3 Estimated values of $\theta$ for chimpanzee samples with high coverage

<table>
<thead>
<tr>
<th>sample ID $^a$</th>
<th>population</th>
<th>coverage (X)</th>
<th>ROHan $\times 10^4$</th>
<th>ANGSD $\times 10^4$ reported heterozygosity $^b$</th>
<th>$\theta$</th>
<th>$\theta_{low}$</th>
<th>$\theta_{high}$</th>
<th>in ROH (%) $^c$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bwambale</td>
<td>Eastern</td>
<td>20.9</td>
<td>18.3731</td>
<td>16.93</td>
<td>19.8412</td>
<td>0-0.0358038</td>
<td>15.0504</td>
<td>12.9 (15.6 $^*$)</td>
<td></td>
</tr>
<tr>
<td>Lara</td>
<td>Central</td>
<td>25.2</td>
<td>19.9676</td>
<td>18.6854</td>
<td>21.4625</td>
<td>0-0.503597</td>
<td>18.0424</td>
<td>14.7 (22.7 $^*$)</td>
<td></td>
</tr>
<tr>
<td>Linda</td>
<td>Western</td>
<td>27.6</td>
<td>8.66859</td>
<td>7.88069</td>
<td>9.51959</td>
<td>0-0.648882</td>
<td>8.04221</td>
<td>6.2 (8.3 $^*$)</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Estimates of genome-wide $\theta$ by ROHan, ANGSD and from the original publication for 3 chimpanzees from Western, Central and Eastern Africa. $^*$ The first number was the heterozygosiy estimate on the individual itself whereas the second number was the estimate for Watterson’s $\theta$ for the population.

$^b$ from (De Manuel et al. 2016)
## Table 4 Estimated values of $\theta$ for ancient hominin samples

<table>
<thead>
<tr>
<th>sample name</th>
<th>ANGSD $\times 10^4$</th>
<th>ROHan $\times 10^4$</th>
<th>coverage damage</th>
<th>reported</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta$</td>
<td>$\theta_{TV}^b$</td>
<td>$\theta$</td>
<td>$\theta_{low}$</td>
<td>$\theta_{high}$</td>
</tr>
<tr>
<td>Vindija 33.19</td>
<td>3.47660</td>
<td>1.836100</td>
<td>1.97284</td>
<td>1.53055</td>
<td>2.40187</td>
</tr>
<tr>
<td>Barčin 31</td>
<td>13.33030</td>
<td>0.346186</td>
<td>3.69527</td>
<td>1.72332</td>
<td>5.80466</td>
</tr>
<tr>
<td>Andronovo 505</td>
<td>7.52603</td>
<td>2.823350</td>
<td>5.76088</td>
<td>4.82866</td>
<td>6.74453</td>
</tr>
<tr>
<td>Loschbour</td>
<td>5.44105</td>
<td>5.403330</td>
<td>7.51366</td>
<td>6.43011</td>
<td>8.27084</td>
</tr>
<tr>
<td>Wezme Cave 1</td>
<td>6.80335</td>
<td>4.056190</td>
<td>8.07936</td>
<td>6.98759</td>
<td>9.21832</td>
</tr>
<tr>
<td>Stora Karlsö 12</td>
<td>7.50118</td>
<td>7.835080</td>
<td>8.59265</td>
<td>7.96388</td>
<td>9.23319</td>
</tr>
<tr>
<td>Ust'-Ishim</td>
<td>7.66156</td>
<td>7.732150</td>
<td>9.85926</td>
<td>8.74113</td>
<td>10.64050</td>
</tr>
</tbody>
</table>

- $^a$ Rate of C to T substitutions at the 5' end and G to A at the 3' end. For samples that used the single-stranded DNA protocol for library preparation, the rate of C to T is reported at the 3' instead of the G to A.
- $^b$ ANGSD $\theta_{TV}$ is the $\theta$ estimate using only transversions and multiplying the estimate by $\kappa_{tv}+1$ (3.1).
- $^c$ (Prüfer et al. 2017)
- $^d$ (Hofmanová et al. 2016)
- $^e$ (Möhl et al. 2015)
- $^f$ (Prüfer et al. 2017)
- $^g$ (Möhl et al. 2015)
- $^h$ (Prüfer et al. 2017)
- $^i$ (Hofmanová et al. 2016)
- $^j$ (Prüfer et al. 2017)
We have explained our methodology for jointly inferring ROHs in Figure 7.

We found that the individual used while sequencing the horse reference genome (Twilight, (Wade et al. 2009)) showed the largest fraction of the genome within ROHs. This is not surprising as this individual was selected to facilitate the genome assembly due to its extreme inbreeding levels. Likewise, we found that the 6 endangered Przewalski’s horses have a substantial fraction of their genome in ROHs, in line with previous estimates (Der Sarkissian et al. 2015). Even when masking ROHs for estimate genome-wide, the estimates of θ are in the lower range of those estimated in all other samples, be modern or ancient. This is expected for a population founded in the early 20th century from a limited number of only 12-15 founders (Der Sarkissian et al. 2015). Interestingly, Eneolithic early domesticates from the Botai culture (Botai2, Botai6, Botai5) and the Boryl4 archaeological site (Boryl4_PAVH4, Boryl4_PAVH8, Boryl4_PAVH11) have recently been shown to represent the direct ancestors of modern Przewalski’s horses (Gaunitz et al. 2018)).

Our tests with BCFtools/RoH show that having accurate allele frequencies can improve the inference accuracy while delineating ROHs. However, using allele frequencies can also add biases, as the analyzed samples do not necessarily belong to the panel population used for estimating the genome-wide distribution of allele frequencies. The impact of such an ascertainment bias can be especially acute in non-model organisms with coverage inferior to 5X-6X. For ancient samples showing high levels of post-mortem DNA damage, a minimal coverage of 8X-10X is required to retrieve meaningful point estimates. In all simulations, ROHan returned more accurate genome-wide θ estimates than existing tools, especially with limited coverage data. We mentioned that users must supply the desired window size for the local heterozygosity estimates and the sensitivity to short ROHs depends on this window size. The choice of the window size depends on available coverage where higher coverage allows for smaller window sizes for the local estimate of heterozygosity. This also entails that our method is not suited for measuring distant and continuous inbreeding. Another limitation is that we do not account for present-day (human) contamination or exogenous DNA such as microbial contamination in aDNA samples, which can disrupt sequence diversity patterns underlying otherwise long blocks of low heterozygosity. Such situations are expected to reduce the length of inferred ROHs.

Discussion

We have explained our methodology for jointly inferring ROHs and the genome-wide θ for regions flagged outside ROHs. Using simulations, we found that both our model and state-of-the-art methods cannot provide reliable estimates in the presence of limited coverage data and/or post-mortem DNA damage, unless significant amounts of data are available. For modern samples, the point estimate for θ seems to be underestimated for samples...
and non-human animals such as domesticates, where breeds of economical relevance generally retain most of the research focus.

A drawback of our approach entails the use of quality scores as being representative of the true probability of a sequencing error. This would be especially problematic in the presence of batch effects, i.e., when comparing 2 samples not sequenced with the same technology and/or instrument, or if different basecallers were used. Any underestimation of the real probability of error will lead to an overestimate of heterozygosity and vice-versa. This is shown in the analyses that considered that aDNA damage could be present in modern samples, which resulted in a significant drop in the $\theta$ point estimate recovered. This suggests that the presence of even limited amounts of sequencing errors showing signatures similar to those observed in ancient samples can significantly impact estimates. Reciprocally, it follows that overestimating rates of aDNA damage in ancient samples will lead to underestimates of the rate of heterozygosity. Since our methodology expects users to provide rates of damage that exclude potential polymorphic positions and sequencing errors, we recommend caution when comparing ancient samples to modern samples, or to other ancient samples that either have been analyzed using different molecular tools or show drastically different rates of aDNA damage.

Throughout the manuscript, we have assumed that for an aDNA sample, an individual is composed of a single library. This can potentially affect our computations as the rates of aDNA damage are provided by the user and can sometimes represent the average across the genome for all libraries. Ideally, we should allow users to provide read group specific aDNA damage rates. This approach however is likely to require additional RAM as the computation for the nucleotide substitutions are pre-computed and stored for speed as the cost of memory. Other avenues for further improvements of our model include accounting for base compositional bias, such as %GC bias, which can introduce uneven coverage along the genome and potentially skew the weights considered for the likelihoods function. This effect might be magnified in those ancient genomes showing pattern of depth-of-coverage variation on par with nucleosomal protection (Pedersen et al. 2014; Hanghej et al. 2016).

Acknowledgement

We would like to thank Anders Albrechtsen, Martin Sikora, Fernando Racimo, Pablo Librado and Filipe Vieira for help running software, helpful discussions and feedback. We are indebted to Laurent Frantz for questions relating to ancient dog data as well as Martin Kuhlwilm and Marc de Manuel Montero for providing us access to the chimpanzee data. We also would like to acknowledge Ziheng Yang, Jerome Kelleher and Vagheesh Narasimhan for their help running software. The project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 752657. This work was supported by the Danish National Research Foundation (DNRF54); Initiative d’Excellence Chaires d’attractivité, Université de Toulouse (OURASI), and; the Villum Fonden miGENEPI research project. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 681605).

Literature Cited


Gamba, C., E. R. Jones, M. D. Teasdale, R. L. McLaughlin,