Inferring Heterozygosity from Ancient and Low Coverage Genomes

Athanasios Kousathanas,*,† Christoph Leuenberger,‡ Vivian Link,*,† Christian Sell,§ Joachim Burger,§ and Daniel Wegmann*,†,1

*Department of Biology and Biochemistry and ‡Department of Mathematics, University of Fribourg, 1700, Switzerland, †Statistical and Computational Evolutionary Biology Group, Swiss Institute of Bioinformatics, Fribourg, 1700, Switzerland, and §Paleogenetics Group, University of Mainz, 55122, Germany

ORCID ID: 0000-0003-2866-6739 (D.W.)

ABSTRACT While genetic diversity can be quantified accurately from high coverage sequencing data, it is often desirable to obtain such estimates from data with low coverage, either to save costs or because of low DNA quality, as is observed for ancient samples. Here, we introduce a method to accurately infer heterozygosity probabilistically from sequences with average coverage $< 1 \times$ of a single individual. The method relaxes the infinite sites assumption of previous methods, does not require a reference sequence, except for the initial alignment of the sequencing data, and takes into account both variable sequencing errors and potential postmortem damage. It is thus also applicable to nonmodel organisms and ancient genomes. Since error rates as reported by sequencing machines are generally distorted and require recalibration, we also introduce a method to accurately infer recalibration parameters in the presence of postmortem damage. This method does not require knowledge about the underlying genome sequence, but instead works with haploid data (*e.g.*, from the X-chromosome from mammalian males) and integrates over the unknown genotypes. Using extensive simulations we show that a few megabasepairs of haploid data are sufficient for accurate recalibration, even at average coverages as low as $1 \times$. At similar coverages, our method also produces very accurate estimates of heterozygosity down to 10^{-4} within windows of about 1 Mbp. We further illustrate the usefulness of our approach by inferring genome-wide patterns of diversity for several ancient human samples, and we found that 3000–5000-year-old samples showed diversity patterns comparable to those of modern humans. In contrast, two European hunter-gatherer samples exhibited not only considerably lower levels of diversity than modern samples, but also highly distinct distributions of diversity along their genomes. Interestingly, these distributions were also very different between the two samples, supporting earlier conclusions of a highly diverse and structured population in Europe prior to the arrival of farming.

KEYWORDS heterozygosity; low coverage; ancient DNA; postmortem damage; base recalibration

THE genetic diversity at a particular location in the genome is the result of its evolutionary past. Comparing the genetic diversity between individuals or regions of the genome thus gives insight into differences in their respective evolutionary histories. For a diploid individual, the heterozygosity of a genomic region (the fraction of sites in a region at which the individual carries two different alleles) is the result of mutations that occurred since the two alleles shared a common ancestor. It is thus a function of the local mutation rate and the time to the most recent common ancestor, which, in turn, is reflective of the demographic and selective past at that locus. Variation in local mutation rates and, due to recombination, also in the strength of selection and genetic drift leads to variable diversity across the genome. Comparing heterozygosity between regions can thus identify locations that were affected differently by selection, or those with an increased mutation rate. Comparing heterozygosity between individuals may further highlight differences in the demographic histories of populations or different levels of inbreeding, which may lead to long runs of homozygosity.

While heterozygosity is readily obtained from high quality genotype calls by counting, it is much harder to infer

accurately from low coverage genomes (*i.e.*, genomes sequenced at low depth). This is primarily due to a substantial probability of observing only one of the two alleles and to sequencing errors, which occur at rates orders of magnitude higher than the expected heterozygosity in many species, including humans. Additional biases may be introduced by relying on a reference genome or by postmortem DNA damage (PMD) when working with ancient DNA.

A natural way of circumventing these issues is to infer genetic diversity probabilistically by taking many of the mentioned issues into account, and several such methods have been developed over the past decade. Johnson and Slatkin (2006), for instance, developed a method for estimating the population scaled mutation rate $\theta = 4N\mu$, where N is the population size, and μ the mutation rate, from large metagenomic data sets in the presence of sequencing errors. Shortly after, multiple moment based estimators were introduced to infer heterozygosity from a single individual (Hellmann *et al.* 2008; Jiang *et al.* 2008). Lynch (2008) then introduced a likelihood-based estimator that relaxed the assumption of a known error rate by estimating it jointly with heterozygosity from the data itself. Despite the additional parameter, this likelihood-based estimator is generally more accurate, even if this implementation is ill-behaved at very low coverages (Lynch 2008).

Several methods to infer genetic diversity that relax the assumption of a constant error rate have since been proposed. These methods commonly make use of error rate estimates provided by the sequencing machines (the quality scores) to calculate the genotype likelihoods then used in the inference. A particularly frequently used approach infers the site frequency spectrum (SFS) while integrating over genotype uncertainty at individual sites (Nielsen *et al.* 2012). If applied to data from a single individual, the SFS is a direct estimation of heterozygosity. However, the current implementation of this approach in the software package ANGSD (Korneliussen 2014) requires *a priori* knowledge of the two potentially observable alleles at each site. While these may be inferred accurately in the case of multiple samples with decent coverage, such an inference from low coverage data are likely biased.

Here, we present a direct extension of these previous approaches to infer heterozygosity within a region that relaxes the assumption of infinitely many sites, does not require any *a priori* knowledge of the underlying alleles, and takes additional biases introduced by PMD fully into account. We achieve this by explicitly modeling PMD using an extension of the likelihood framework proposed by Maruki and Lynch (2015) to infer site allele frequencies by integrating over all possible genotypes, and by modeling genotype frequencies using the classic substitution model of Felsenstein (1981), which allows for back mutations.

A major hurdle for all approaches relying on quality scores provided by sequencing machines is that these scores are not reliable and must be recalibrated, particularly when coverage is low. This is commonly achieved by learning error rates from sites assumed *a priori* to be invariant, for instance, by masking polymorphic sites, repetitive elements, and large structural variants (DePristo *et al.* 2011). While we have extended this approach to tolerate PMD (Hofmanová *et al.* 2016), it requires detailed knowledge of the study species, which is often lacking for nonmodel organisms. We propose here to circumvent this problem by using a reference-free recalibration approach that relies on haploid sequences such as those from the X-chromosome in male mammals. Our approach does not require any *a priori* information on the underlying sequence, as it integrates over all possible but hidden genotypes while taking PMD and covariates such as position in read or read context into account. This renders our approach essentially free of reference biases, since the reference is only required for aligning raw reads by mapping and current mapping techniques tolerate a sequence divergence of up to 10% (*e.g.*, Lunter and Goodson 2011).

Using computer simulations, we show that our method reliably estimates local genetic diversity in single, diploid individuals even with average coverage below $2 \times$ for windows of ~ 1 Mbp. We further show that a few megabasepairs of data at equally low coverage are sufficient to properly recalibrate distorted quality scores. Finally, we use the methods here developed to infer the genome-wide pattern of diversity for several ancient and modern human samples. We found that these patterns differ between European and African samples, but that samples from a few thousand years ago cluster well with modern samples. In contrast, European hunter-gatherer individuals differ strongly from modern Europeans, but also from each other, illustrating the high diversity that existed in Europe before the Neolithic transition.

Theory

Inferring heterozygosity

Here we develop a method to estimate local heterozygosity in a genomic window from a collection of aligned reads by integrating out the uncertainty of the local genotype, and by taking the potential effects of PMD into account. Specifically, we are interested in inferring the stationary base frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$, together with the rate of substitutions $\theta = 2T\mu$, along the genealogy connecting the two alleles of an individual within a genomic region. Here, T corresponds to the time to the most recent common ancestor of the two lineages, and μ to the mutation rate per base pair per generation. Notably, it is not possible to infer T and μ independently, and we therefore only attempt to estimate the compound substitution rate θ from the data.

We extend Felsenstein's model of substitutions (Felsenstein 1981) to infer θ while accounting for the uncertainty in the genotypes in the region of interest. Let us denote the hidden genotype at site i by g_i, where g_i consists of a pair of nucleotides

kl with $k, l = A, G, C, T$. Under the substitution model, the probability of observing a specific genotype $g_i = kl$ given the base frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$, and the substitution rate θ, is given by

$$\mathbb{P}(g_i = kl | \theta, \boldsymbol{\pi}) = \begin{cases} \pi_k\left(e^{-\theta} + \pi_k\left(1 - e^{-\theta}\right)\right) & \text{if } k = l, \\ \pi_k \pi_l\left(1 - e^{-\theta}\right) & \text{if } k \neq l. \end{cases} \quad (1)$$

To integrate out the uncertainty in observed genotypes, we adopt a model similar to Lynch (2008), and those commonly used for genotype calling (*e.g.*, Li 2011). The subsequent notation will closely follow the notation we recently introduced (Hofmanová *et al.* 2016).

The observed data d_i at site i shall correspond to what is typically obtained when individual reads of next generation sequencing approaches are mapped to a reference genome. Here, we will assume that all sequencing reads were accurately mapped and that reads with low mapping qualities have been filtered out. The data d_i obtained at site i thus consists of a list of n_i observed bases $d_i = \{d_{i1}, \ldots, d_{in_i}\}, d_{ij} = A, C, G, T$.

We chose to model the observed data, d_i, at site i as a function of the underlying genotype, g_i, as well as the base-specific rates of sequencing errors ϵ_{ij} for $j = 1, \ldots, n_i$. We assume here that these rates are known (*e.g.*, through quality scores provided by the sequencing machine). Assuming further that sequencing errors occur independently, the likelihood of the full data at site, i, is given by

$$\mathbb{P}(d_i | g_i, \boldsymbol{\epsilon_i}) = \prod_{j=1}^{n_i} \mathbb{P}(d_{ij} | g_i, \epsilon_{ij}),$$

where $\boldsymbol{\epsilon_i} = \{\epsilon_{i1}, \ldots, \epsilon_{in_i}\}$.

Following Maruki and Lynch (2015), and commonly used approaches (*e.g.*, Li 2011), we will assume that a sequencing read is equally likely to cover any of the two alleles of an individual, and that sequencing errors may result in any of the alternative bases with equal probability, $\epsilon_{ij}/3$. The probability of observing a base, d_{ij}, given the underlying genotype, $g_i = kl$, is then given by

$$\mathbb{P}(d_{ij} | g_i = kl, \epsilon_{ij}) = \begin{cases} 1 - \epsilon_{ij} & \text{if } k = l = d_{ij} \\ \dfrac{\epsilon_{ij}}{3} & \text{if } k \neq d_{ij}, l \neq d_{ij} \\ \dfrac{1}{2} - \dfrac{\epsilon_{ij}}{3} & \text{if } k \neq l, k = d_{ij} \text{ or } l = d_{ij} \end{cases}.$$

Assuming sites to be independent, the full likelihood of our model is given by

$$L(\theta, \boldsymbol{\pi}) = \mathbb{P}(\boldsymbol{d} | \theta, \boldsymbol{\pi}) = \prod_{i=1}^{I} \mathbb{P}(d_i | \theta, \boldsymbol{\pi})$$

$$= \prod_{i=1}^{I} \sum_{g} \mathbb{P}(d_i | g_i = g) \mathbb{P}(g_i = g | \theta, \boldsymbol{\pi}),$$

where the sum runs over all combinations $g = AA, AG, \ldots, TT$.

PMD damage: We will next extend this model with the possibility of PMD. The most common form of PMD is C deamination, which leads to a $C \rightarrow T$ transition on the affected strand and a $G \rightarrow A$ transition on the complimentary strand (*e.g.*, Briggs and Stenzel 2007). These deaminations do not occur randomly along the whole read, but are observed much more frequently at the beginning of a read. This is due to fragment ends being more often single-stranded, and thus subject to a much higher rate of deamination. Consequently, the rates of PMD, while specific to the sample and the sequencing protocol used, generally decay roughly exponentially with distance from the ends of the read Skoglund *et al.* (2014). Since ancient DNA is highly fragmented, one read can often cover an entire DNA molecule, and hence $C \rightarrow T$ and $G \rightarrow A$ transitions may be seen in a single read, but they are accumulated at opposite ends.

Here, we will develop our model for this form of PMD following the formulation of Skoglund *et al.* (2014) and Hofmanová *et al.* (2016), but we note that it is readily extendable to incorporate other forms of PMD as well. We feel that the rationale of the approach taken is best explained with a specific example. Consider $d_{ij} = T$, given the underlying genotype $g_i = CT$. There are three possible ways to obtain a T: (i) by sequencing an allele T without error, (ii) by sequencing an allele C affected by PMD without error, (iii) or by sequencing an allele C not affected by PMD with error. We thus have

$$\mathbb{P}(d_{ij} = T | g_i = CT, \epsilon_{ij}, D_{ij})$$
$$= \frac{1}{2}\left((1 - \epsilon_{ij}) + D_{ij}(1 - \epsilon_{ij}) + (1 - D_{ij})\frac{\epsilon_{ij}}{3}\right) \quad (2)$$

where D_{ij} denotes the probability that a $C \rightarrow T$ PMD occurred at the base of read j covering site i.

The emission probabilities for all combinations of d_{ij} and g_i derived following the same logic are found in the *Appendix*.

Inference using expectation-maximization: In this section, we will detail how to find the maximum likelihood estimate (MLE) of the model parameters θ and $\boldsymbol{\pi}$ in a genomic window of I sites using an expectation-maximization (EM) algorithm (Dempster *et al.* 1977). For this, we will make the assumption that base-specific sequencing error rates, ϵ_{ij}, and rates of PMD, D_{ij}, are given constants. For the cases in which they are not known *a priori*, we show below how they can be learned accurately from genome-wide data prior to inferring θ and $\boldsymbol{\pi}$. In an effort to unburden the notation, we will thus refer to the emission probabilities simply as $\mathbb{P}(d_i | g_i)$ in the following.

The relevant property to develop an EM algorithm is the complete data likelihood, which, in the case of our model, is given by

$$L_c(\theta, \boldsymbol{\pi}; \boldsymbol{d}, \boldsymbol{g}) = \prod_{i=1}^{I} \mathbb{P}(g_i, d_i | \theta, \boldsymbol{\pi}) = \prod_{i=1}^{I} \mathbb{P}(d_i | g_i) \mathbb{P}(g_i | \theta, \boldsymbol{\pi})$$

and thus the complete data log-likelihood by

$$l_c(\theta, \boldsymbol{\pi}; \boldsymbol{d}, \boldsymbol{g}) = \sum_{i=1}^{I} (\log\mathbb{P}(d_i|g_i) + \log\mathbb{P}(g_i|\theta, \boldsymbol{\pi})).$$

E-step: The expected complete data log-likelihood is calculated as

$$Q(\theta, \boldsymbol{\pi}; \theta', \pi') = \mathbb{E}[l_c(\theta, \boldsymbol{\pi}; \boldsymbol{d}, \boldsymbol{g})|d; \theta', \pi']$$

$$= \sum_{i=1}^{I} \sum_{g} \Big[\log\mathbb{P}(d_i|g)$$

$$+ \log\mathbb{P}(g|\theta, \boldsymbol{\pi})\Big]\mathbb{P}(g|d_i; \theta', \pi')$$

where the sum runs over all combinations $g = AA, AG, \ldots, TT$. Only the second part Q_2 of this sum depends on the parameters $\theta, \boldsymbol{\pi}$. We have

$$Q_2(\theta, \boldsymbol{\pi}; \theta', \pi') = \sum_{i=1}^{I} \sum_{g} \log\mathbb{P}(g|\theta, \boldsymbol{\pi})\mathbb{P}(g|d_i; \theta', \boldsymbol{\pi})'$$

$$= \sum_{g} \log\mathbb{P}(g|\theta, \boldsymbol{\pi}) \sum_{i=1}^{I} \mathbb{P}(g|d_i; \theta', \boldsymbol{\pi})'$$

$$= \sum_{g} P_g \log\mathbb{P}(g|\theta, \boldsymbol{\pi})$$

where we use the shorthand notation $P_g = \sum_{i=1}^{I} \mathbb{P}(g|d_i; \theta', \pi')$. We have by Bayes' Theorem

$$P_g = \sum_{i=1}^{I} \frac{\mathbb{P}(d_i|g)\mathbb{P}(g|\theta', \pi')}{\sum_g \mathbb{P}(d_i|g)\mathbb{P}(g|\theta', \pi')}. \tag{3}$$

Let us write out Q_2 explicitly:

$$Q_2(\theta, \boldsymbol{\pi}; \theta', \pi') = \sum_{k} P_{kk} \Big[\log\pi_k + \log\big(e^{-\theta} + \pi_k(1 - e^{-\theta})\big)\Big]$$

$$+ \sum_{k} \sum_{l \neq k} P_{kl} \Big[\log\pi_k + \log\pi_l$$

$$+ \log\big(1 - e^{-\theta}\big)\Big].$$

M-step: We have to maximize Q_2 subject to the constraint

$$\sum_{k} \pi_k = \pi_A + \pi_G + \pi_C + \pi_T = 1.$$

For this reason, we form the Lagrangian

$$\mathcal{L}(\theta, \boldsymbol{\pi}, \mu) = Q_2(\theta, \boldsymbol{\pi}; \theta', \pi') - \mu\left(\sum_{k} \pi_k - 1\right)$$

where μ is the Lagrange multiplier. We get the following partial derivatives of the Lagrangian:

$$\frac{\partial}{\partial \pi_k} \mathcal{L} = P_{kk}\left(\frac{1}{\pi_k} + \frac{1 - e^{-\theta}}{e^{-\theta} + \pi_k(1 - e^{-\theta})}\right) + \sum_{l \neq k} \frac{2P_{kl}}{\pi_k} - \mu,$$

$$\frac{\partial}{\partial \theta} \mathcal{L} = -e^{-\theta} \sum_{k} \frac{P_{kk}(1 - \pi_k)}{e^{-\theta} + \pi_k(1 - e^{-\theta})}$$

$$+ \frac{e^{-\theta}}{1 - e^{-\theta}} \sum_{k} \sum_{l \neq k} P_{kl},$$

$$\frac{\partial}{\partial \mu} \mathcal{L} = \sum_{k} \pi_k - 1.$$

We have to set these equations to zero and solve for π_k, θ, and μ. Since this is not possible analytically, we will revert to the Newton-Raphson algorithm.

To streamline the notation, we will rename our variables $x_1 = \pi_A, \ldots, x_4 = \pi_T$, $x_5 = e^{-\theta}/(1 - e^{-\theta})$, and $x_6 = \mu$. With these variable the above system can be simplified to a system

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \tag{4}$$

where

$$F_k(x_1, \ldots, x_6) := P_{kk}\left(1 + \frac{x_k}{x_5 + x_k}\right) + 2\sum_{l=1}^{4} P_{kl} - x_6 x_k = 0$$

for $k = 1, \ldots, 4$, and

$$F_5(x_1, \ldots, x_6) := I - \sum_{k=1}^{4} \frac{P_{kk}(x_5 + 1)}{x_5 + x_k} = 0,$$

$$F_6(x_1, \ldots, x_6) := \sum_{k=1}^{4} x_k - 1 = 0.$$

To apply the Newton-Raphson algorithm, we determine the 6×6 Jacobian matrix $J_{ij} = \partial F_i/\partial x_j$. The nonzeros entries of the Jacobian with $k = 1, \ldots, 4$ are:

$$J_{kk} = \frac{P_{kk}x_5}{(x_k + x_5)^2} - x_6,$$

$$J_{5k} = (x_5 + 1) \sum_{l=1}^{4} \frac{P_{ll}}{(x_l + x_5)^2},$$

$$J_{6k} = 1,$$

$$J_{k5} = -\frac{P_{kk}x_k}{(x_k + x_5)^2}, \tag{5}$$

$$J_{k6} = -x_k,$$

$$J_{55} = \sum_{l=1}^{4} \frac{P_{ll}(1 - x_l)}{(x_l + x_5)^2}.$$

We can now approximate the zero of Equation 4 with the iteration

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} - \mathbf{J}^{-1}(\mathbf{x}_{\text{old}})\mathbf{F}(\mathbf{x}_{\text{old}}). \tag{6}$$

After a few iterations, we get the new estimate for the original parameters by setting $\pi_k = x_k$ for $k = 1, \ldots, 4$, and $\theta = -\log[x_5/(1 - x_5)]$

Confidence intervals: We calculate an approximate confidence interval for θ using the Fisher information. To simplify the calculations, we consider the π_k as constant. The observed Fisher information at the ML value $\hat{\theta}$ is

$$\mathcal{I}(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta}, \boldsymbol{\pi})$$

$$= -\sum_{i=1}^{I} \frac{\partial^2}{\partial \theta^2} \log \left[\sum_g \mathbb{P}(d_i|g_i = g) \mathbb{P}(g_i = g|\hat{\theta}, \boldsymbol{\pi}) \right],$$

and the corresponding derivatives are

$$\frac{\partial}{\partial \theta} \mathbb{P}(g_i = kl|\theta, \boldsymbol{\pi}) = \begin{cases} (\pi_k^2 - \pi_k)e^{-\theta} & \text{if } k = l, \\ \pi_k \pi_l e^{-\theta} & \text{if } k \neq l. \end{cases} \quad (7)$$

Observe that $\frac{\partial^2}{\partial \theta^2} \mathbb{P}(g_i = kl|\theta, \boldsymbol{\pi}) = -\frac{\partial}{\partial \theta} \mathbb{P}(g_i = kl|\theta, \boldsymbol{\pi})$. From this we easily get that

$$\mathcal{I}(\hat{\theta}) = \sum_{i=1}^{I} \mathcal{R}_i (\mathcal{R}_i + 1) \quad (8)$$

where we have set

$$\mathcal{R}_i = \frac{\sum_g \mathbb{P}(d_i|g) \frac{\partial}{\partial \theta} \mathbb{P}(g|\hat{\theta}, \boldsymbol{\pi})}{\sum_g \mathbb{P}(d_i|g) \mathbb{P}(g|\hat{\theta}, \boldsymbol{\pi})}. \quad (9)$$

An approximate $(1 - \alpha)$ confidence interval is now given by

$$\hat{\theta} \pm z_{1-\alpha/2} \, \mathcal{I}(\hat{\theta})^{-1/2}.$$

Estimating rates of PMD

As mentioned earlier, the method above assumes that rates of PMD are known *a priori*. In cases in which they are not known, they can be readily inferred from genome-wide data as we outline in this section.

Following Jónsson *et al.* (2013), we present an approach to estimate PMD rates directly from genome-wide counts of $C \rightarrow T$ and $G \rightarrow A$ transitions as a function of distance within the read. For this, we first build the three-dimensional table $\mathcal{T}$, where each entry $\mathcal{T}_{rsp}$ corresponds to the number of observed bases, s, read at a site with reference base, r, at position p within a read. While these counts depend on the divergence between the sequenced individual and the reference genome used for mapping, we develop here an approach that takes this divergence into account.

Let us denote by μ_{rs} the probability of a true difference between the sequenced individual and the reference, such that the reference has base r and the sequenced individual base s. Since the reference and a sequenced chromosome form a genealogy on which these mutations occurred, it is

safe to assume that $\mu_{rs} = \mu_{sr}$. We will further assume that the observed counts in a cell $\mathcal{T}_{rsp}$ not affected by PMD are a direct function of μ_{rs}.

Since the rate of PMD is generally low far away from the read ends, position-specific estimates may become noisy for these positions, particularly if data are limited. We thus introduce a method to estimate parameters of a model of exponential decay with the position in the read. The use of such a model was first introduced by Skoglund *et al.* (2014), and we implement here a slightly more general version of their function. Specifically, we assume that the probability of observing base T when the reference sequence is a C at position p is given by

$$\mathbb{P}(d_{ij} = T|g_i = C, p) = \mu_{CT} + (1 - \mu_{CT})(a + be^{-cp}),$$

where μ_{CT} again denotes true differences between the individual and the reference.

Note that some of the parameters of this model are non-identifiable. In the *Appendix*, we show how to obtain ML estimates of the parameters of the probability function

$$\mathbb{P}(d_{ij} = T|g_i = C, p) = \alpha + \delta e^{-\gamma p}. \quad (10)$$

using the standard Newton-Raphson algorithm.

To then obtain estimates for the original parameters μ_{CT}, a, b, and c, we assume that $\mu_{CT} = \mu_{TC}$, and obtain a ML estimate

$$\hat{\mu}_{CT} = \hat{\mu}_{TC} = \frac{\sum_p \mathcal{T}_{TCp}}{\sum_p \mathcal{T}_{TTp}}.$$

Then, $a = \frac{\delta}{1 - \hat{\mu}_{CT}}$, $b = \gamma$, and $c = \frac{\alpha - \mu_{CT}}{1 - \mu_{CT}}$. We use the analogous logic to infer PMD patterns for $G \rightarrow A$ damages, but measuring positions from the opposite end of the read.

Estimating base-specific error rates (recalibration)

The challenge of inferring genetic diversity from next-generation sequencing data lies in the fact that the per base error rates are orders of magnitude higher than the expected heterozygosity of many species (Lynch 2008). While this issue can easily be overcome with high coverages, accurate inference from low-coverage data relies on an exact knowledge of base-specific error rates. Crude estimates of these rates are usually directly provided by the sequencing machines themselves. However, these estimates are often inaccurate, and are recommend to be recalibrated for genotype calling (DePristo *et al.* 2011).

The most commonly used approach for recalibration is BQSR (Base Quality Score Recalibration) implemented in GATK (McKenna *et al.* 2010; De-Pristo *et al.* 2011). This approach infers new quality scores by binning the data into groups based on covariates such as the raw quality score, the position in the read, or the sequence context. All bases within such a bin are assumed to share the same error rate,
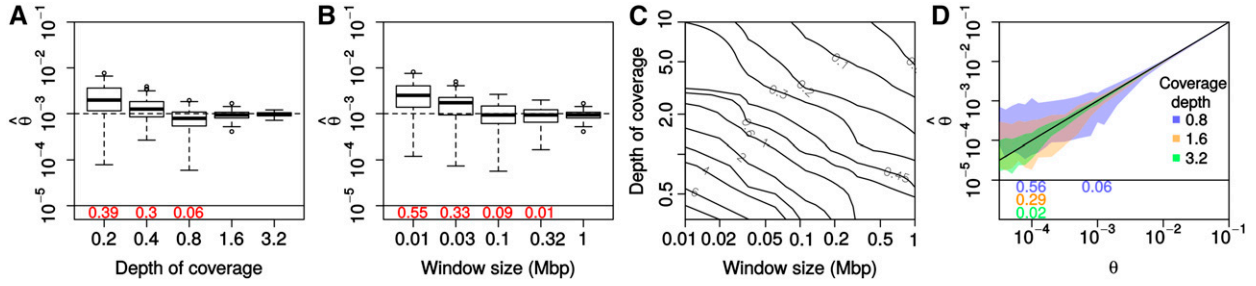
Figure 1 Power to infer θ from low coverage data. Results from sets of 100 simulations with PMD for different average coverage, window size, and true θ values. (A) Estimated $\hat{\theta}$ in windows of 1 Mbp as a function of average coverage. (B) Estimated $\hat{\theta}$ as a function of window size and fixed average coverage of 1×. (C) Accuracy of estimating $\theta = 10^{-3}$ quantified as the median relative error ($|1 - \hat{\theta}/\theta|$) over replicates indicated by contour lines as a function of both coverage and window size. (D) True *vs.* estimated θ for different average coverages (see color legend). Polygons indicate the 95% quantile of estimated $\hat{\theta}$ values among all replicates. The diagonal black line indicates the expectation for perfect estimation. In (A, B, and D), replicates resulting in a $\hat{\theta} < 10^{-5}$ are not shown, but their fraction across replicates are printed below the horizontal black line.

which can be readily inferred if the true underlying sequence is known. As an alternative, Cabanski *et al.* (2012) proposed to fit a logistic regression to the full data where the response variable is the probability of a sequencing error and the explanatory variables are the raw quality scores, and covariates such as position in the read, or base context.

For our purpose, these methods suffer from two shortcomings: first, they cannot be applied to ancient DNA since they do not take PMD into account. Second, both require a reference sequence as well as knowledge of polymorphic positions, such that they can be excluded from the analysis. While we have shown how to extend the BQSR method to ancient DNA (Hofmanová *et al.* 2016), we develop here an approach that also integrates over the unknown reference sequence.

To do so, we will assume the existence a genomic region for which the individual does not show any polymorphism. A good example of such a genomic region are nonhomologous sequences from sex chromosomes in heterogametic individuals, and we will describe our approach with this type of data in mind. However, we note that our approach can also be readily applied to diploid regions that are known to be monomorphic, such as positions that are highly conserved among species or positions retained after filtering out those with high minor allele counts (Cabanski *et al.* 2012).

Model

As above, let us denote the hidden (haploid) genotype at site i by g_i, where g_i is one of the nucleotides A, G, C, T. At each site i, there are n_i reads, and we denote by $d_{ij}, j = 1, \ldots, n_i$, the base of read j covering site i. A sequencing error occurs with probability ϵ_{ij}. These probabilities shall now be given by a model

$$\epsilon_{ij} = \epsilon\left(\boldsymbol{q_{ij}}, \boldsymbol{\beta}\right), \tag{11}$$

where $\boldsymbol{q_{ij}} = (q_{ij1}, \ldots, q_{ijL})$ is a given external vector of information, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_L)$ are the parameters of the model that have to be estimated. While our approach is flexible

regarding the choice of included covariates, we will here consider the raw quality score, the position within the read, the squares of these to account for a nonlinear relationships, and all two-base contexts consisting of the bases of the read at positions $i - 1$ and i.

Following Cabanski *et al.* (2012), we impose the logit model

$$\epsilon_{ij}\left(\boldsymbol{q_{ij}}, \boldsymbol{\beta}\right) = \frac{\exp\left[\eta_{ij}(\boldsymbol{\beta})\right]}{1 + \exp\left[\eta_{ij}(\boldsymbol{\beta})\right]} \tag{12}$$

with

$$\eta_{ik}(\boldsymbol{\beta}) = \beta_0 + \sum_{l=1}^{L} q_{ijl}\beta_l.$$

In the case of monomorphic or haploid sites only, the probability of the read vector d_i given the hidden state g_i can be written more generally as

$$\mathbb{P}(d_i | g_i, \boldsymbol{\beta}) = \prod_{j=1}^{n_i}\left[(1 - Dij)(1 - \epsilon_{ij}) + Dij\frac{\epsilon_{ij}}{3}\right], \tag{13}$$

Here, the dependence on the parameters $\boldsymbol{\beta}$ is given by (12),

$$D_{ij} = D\left(d_{ij}, \boldsymbol{q_{ij}}, g_i\right) = \begin{cases} 0 & \text{if } g_i = d_{ij} = A \text{ or } T \\ D_{C \to T}(q_{ij}) & \text{if } g_i = C, d_{ij} = C \\ 1 - D_{C \to T}(q_{ij}) & \text{if } g_i = C, d_{ij} = T \\ 1 - D_{G \to A}(q_{ij}) & \text{if } g_i = G, d_{ij} = A \\ D_{G \to A}(q_{ij}) & \text{if } g_i = G, d_{ij} = G \\ 1 & \text{otherwise} \end{cases},$$

and $D_{C \to T}(q_{ij})$ and $D_{G \to A}(q_{ij})$ refer to the known probability that a $C \to T$ or $G \to A$ PMD occurred at the position covering site i in read j.

Again, this model can be estimated using the EM algorithm with the Newton-Raphson algorithm in the M-step. Details are given in the *Appendix*.

Implementation

All approaches mentioned were implemented in a custom C++ program available at our laboratory website (http://www.unifr.ch/biology/research/wegmann/wegmannsoft). We used functions included in the library BamTools for manipulating bam files (Barnett *et al.* 2011).

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.
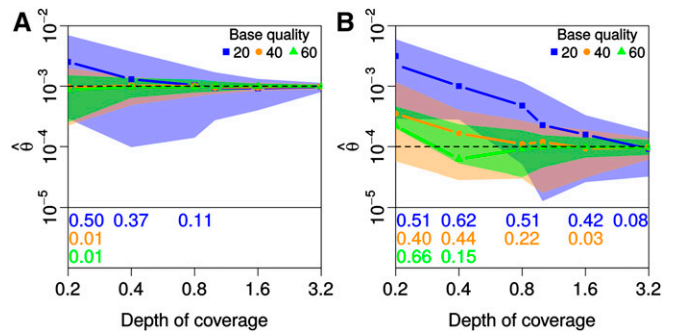
Simulations

Generating simulations

In this section, we illustrate the power and accuracy of our inference approaches with simulations. These were generated using a custom-made *R* script that implements the following steps:

1. The first chromosome of length *L* was simulated using random bases with frequencies $\pi = \{0.25, 0.25, 0.25, 0.25\}$.
2. The second, homologous chromosome was simulated according to the Felsenstein (1981) substitution model (Equation 1) with π and a chosen θ value.
3. Sequencing reads of 100 bases were then generated by copying from one of the two chromosomes with equal probability, and by choosing a starting position uniformly between positions 1 and $L - 99$ until the desired average coverage was reached. All reads copied from the second chromosome were considered to map to the reverse strand.
4. PMD was simulated on all reads with probabilities following an exponential decay with increasing position in the read as proposed by Skoglund *et al.* (2014) to match realistic patterns. Specifically, we simulate PMD at position *p* within the read with probability

$$D = (1-\lambda)^{p-1} p + C,$$

where $\lambda = 0.3$ and $C = 0.01$ for both $C \rightarrow T$ and $G \rightarrow A$ but with *p* counted from the 3′ and 5′ ends, respectively.
5. For each simulated base, a phred-scaled quality score was simulated and sequencing errors were then added with probabilities given by these scores. If not stated otherwise, quality scores were simulated from a normal distribution with mean $\mu_Q = 20$ and SD $\sigma_Q = 4.5$, truncated at zero. When testing our recalibration approach, however, the quality scores were simulated from a uniform distribution $U[5, 60]$, and then transformed according to Equation 12 with coefficients β to obtain the true error rate, with which sequencing errors were simulated.
6. The simulated data were finally used to generate a reference FASTA file containing the first chromosome and a SAM file containing the reads. The latter was then transformed into a BAM file using SAMtools (Li *et al.* 2009).

Figure 2 Effect of sequencing quality on power to estimate θ. Results from sets of 100 simulations to assess the power to estimate θ of 10^{-4} and 10^{-3} for (A) and (B), respectively, for different average base qualities distributed normally with mean 20, 40, or 60, and a SD of 4.5, but truncated at 0. Polygon shapes indicate the 95% confidence interval for estimated $\hat{\theta}$ over all replicates, excluding those resulting in $\hat{\theta} < 10^{-5}$ (the fraction excluded are printed below the horizontal black line). All simulations were conducted with PMD, and the true PMD probability functions were used during the estimation.

Power to infer θ

To check the power of our approach to infer θ from low coverage data, we first simulated data within a 1 Mbp window with a true $\theta = 10^{-3}$ for various coverages. The specific value of $\theta = 10^{-3}$ was chosen to reflect the median heterozygosity in a modern, non-African human individual.

We found the median of our θ estimates across replicates to be very close to the true value, but the variance to be a function of coverage. At low coverage ($<1\times$), θ was often inferred to be zero. This is not surprising, as the information about genetic diversity can come only from sites covered at least twice, which is rare at average coverages $<1\times$. The simulations that did result in an estimate above zero were thus enriched for cases with slightly above average number of polymorphic sites among those covered twice. As soon as average coverage exceeded $1.5\times$, however, our approach estimated θ at 10^{-3} very accurately (Figure 1A).

We next performed simulations with a fixed coverage of $1\times$, but varying the window size (Figure 1). Interestingly, we found that an increase in window size has a positive effect on the estimate accuracy, similarly to an increase in coverage, suggesting that larger windows help to increase accuracy if coverage is very low. To illustrate this effect, we performed simulations at various window sizes and coverages, and recorded the relative estimation error for a series of replicates. As expected, we found the median relative estimation error to be a direct function of the product of window size and coverage (Figure 1C), thus suggesting our method will perform well also at average coverages below $1\times$ if the window size is large enough.

Using the same setting, we also checked the accuracy of the approximate confidence intervals obtained using the Fisher information. For this we inferred θ from 1000 windows of 1 Mbp simulated with θ at 10^{-3} for a coverage of 1.0 and 0.2.
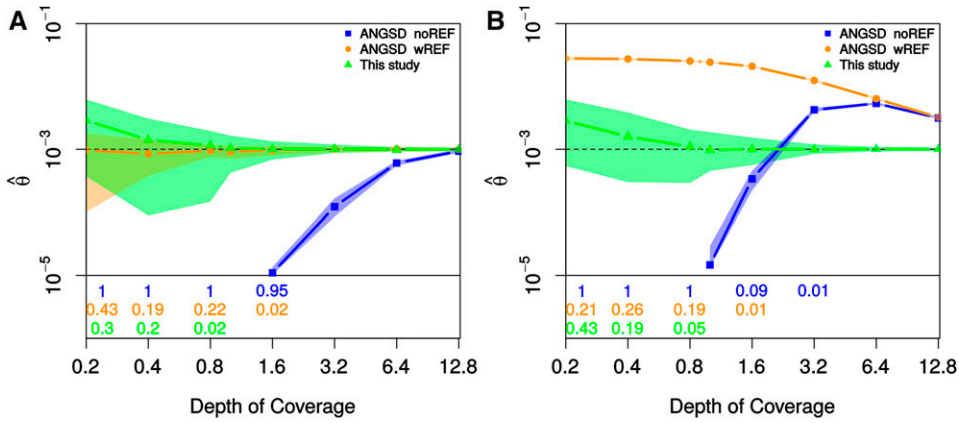
Figure 3 Performance comparison with ANGSD. Results from sets of 100 simulations with (A) or without (B) PMD, comparing the performance of the method presented in this study and ANGSD in inferring θ. ANGSD was run either with the reference sequence provided (ANGSD_wREF) or without (ANGSD_noREF), in which case the major and minor alleles were first inferred in an additional step. For all simulations, we further assumed that base qualities were distributed normally, with $\mu_Q = 20$ and $\sigma_Q = 4.5$, but truncated at 0. Polygon shapes indicate the 95% confidence interval for estimated $\hat{\theta}$, but excluding replicates resulting in $\hat{\theta} < 10^{-5}$. The fraction of replicates excluded are printed below the horizontal black line. When applying our method in simulations conducted with PMD, we provided the true PMD probability patterns during the estimation.

We found the true value to be included in the 95% confidence interval in 93.6 and 98.6%, respectively, suggesting these confidence intervals to be a very accurate reflection of estimation uncertainty.

Using a third set of simulations, we found that, at equal coverage and window size, higher θ values are estimated more accurately than lower values (Figure 1D). This is expected since, in the case of low θ, only few heterozygous sites are present in a given window, rendering the estimate more dependent on the detection of individual sites. Nonetheless, we found our approach to infer $\theta > 10^{-4}$ very accurately in a window of 1 Mbp if the average coverage exceeds $3\times$.

All results above were generated assuming base-specific quality scores to be normally distributed with $\mu_Q = 20$ and SD $\sigma_Q = 4.5$, which is lower than the quality expected with current sequencing approaches (*e.g.*, Utturkar *et al.* 2015). Sequences generated with higher quality will positively affect estimation accuracy. Indeed we found that simulating data with $\mu_Q = 40$ or $\mu_Q = 60$ resulted in much lower estimation error, effectively rendering the estimation of θ feasible even at very low average coverage (Figure 2). For instance, we found that, at an average coverage of $0.8\times$, more than 90% of windows with $\theta = 10^{-4}$ and $\mu_Q = 60$ were estimated within less than half an order of magnitude from the true value. At $\mu_Q = 20$, this accuracy was only reached with an average coverage of $3.2\times$.

Comparison to an existing method

While there is currently no implemented method available to infer heterozygosity in a window or region from a single individual, our method is very similar to those inferring the SFS from multiple individuals (*e.g.*, Nielsen *et al.* 2012; Korneliussen 2014). Indeed, inferring the SFS from a single individual gives a direct estimate of heterozygosity, H, which is related to θ by

$$\mathbb{E}[H] = \sum_{g'} \mathbb{P}\Big(g_i = g' \big| \theta, \boldsymbol{\pi}\Big). \tag{14}$$

The most commonly used approach to infer the SFS while integrating over genotype uncertainty at individual sites (Nielsen *et al.* 2012) is implemented in the software package ANGSD (Korneliussen 2014). This implementation assumes that the two potential alleles are known *a priori* for each site, and hence have to be inferred first using the major-minor option implemented in ANGSD. While the method implemented in ANGSD can be extended to address this limitation, we used our simulation framework to assess its effect on the inference of heterozygosity from low coverage data. In order to make estimates comparable, we report those of ANGSD in terms of θ calculated according to Equation 14 from the fraction of sites reported to be heterozygous in the SFS.

As shown in Figure 3A, ANGSD inferred θ very accurately if the reference allele was provided for each site, and only the second allele needed to be inferred. Making explicit use of this additional information, the estimates were more accurate than those obtained with our approach, which does not make any assumption about the alleles present. However, when no information was given about the observable alleles, and both had to be inferred using the major-minor option of ANGSD prior to the inference of heterozygosity, θ was grossly underestimated whenever coverage was below $10\times$. While we note that ANGSD is designed for applications involving multiple samples, in which case inferring the two observable alleles accurately is much easier; this result illustrates the importance of taking the full genotype uncertainty into account when inferring diversity from low coverage data.

Again using simulations, we next studied the impact of PMD on the inference of heterozygosity (Figure 3B). Unsurprisingly, the diversity estimated using the method implemented in ANGSD that does not take PMD into account was more than one order of magnitude too large. In contrast, our method properly accounts for PMD if its pattern is well characterized. While ANGSD can readily be extended to account for PMD, these results highlight the importance of accounting for PMD for any population genetic inference or comparison involving ancient DNA.
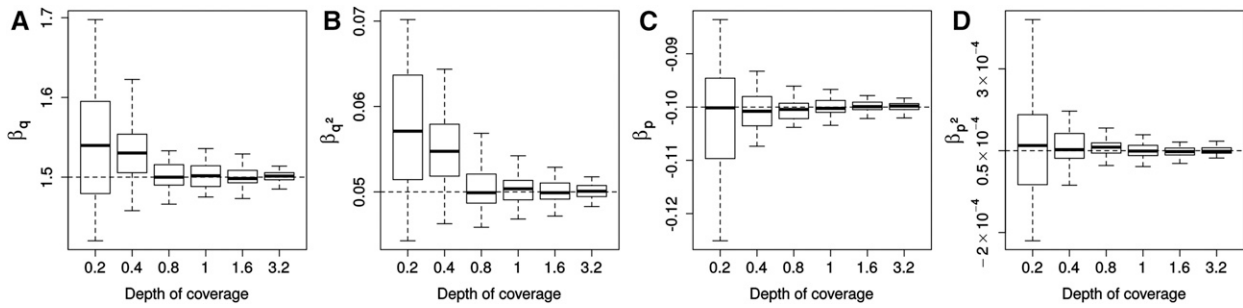
Figure 4 Accuracy in inferring recalibration parameters. Results from sets of 100 simulations are shown where sequence data from a haploid 1 Mbp region was simulated assuming a uniform distribution of observed quality scores ($U[5, 60]$) that were then transformed to true qualities according to Equation 12, with $\beta_q = 1.5$, $\beta_q^2 = 0.05$, $\beta_p = -0.1$, $\beta_q^2 = 5 \cdot 10^{-5}$, and all context coefficients at 1.0. All simulations were conducted with PMD, and the true PMD probability patterns were used during the estimation. The coverage values refer to the diploid regions of the simulated genome.

Accuracy of recalibration

The results discussed so far were all obtained under the assumption that quality scores provided by the sequencing machine are accurate. Unfortunately, this is rarely the case, making recalibration of the quality scores necessary for most applications, and, in particular, when trying to infer genetic diversity from low coverage data. Here, we developed an approach to recalibrate quality scores without prior knowledge of the underlying sequencing information. Instead, we simply assume that a part of the sequence is known to be monomorphic, such as the haploid X-chromosome in mammalian males.

To investigate the power of our approach to infer recalibration parameters, we simulated sequencing reads from a haploid region where the quality scores provided in the SAM files were distorted. We did this by first simulating fake quality scores from a uniform distribution $U[5, 60]$, and then transforming them into true quality scores according to Equation 12. We used the following coefficients: all context coefficients = 1.0, the coefficients for the raw quality score, $\beta_q = 1.5$, the square of the raw quality score, $\beta_{q^2} = 0.05$, the position within the read, $\beta_p = -0.1$, and the square of the position within the read, $\beta_{p^2} = 5 \cdot 10^{-5}$. These values were chosen to reflect a distortion observed in real data from the ancient human samples analyzed in this study (see below). They also result in both a relatively strong distortion, as well as meaningful error rates for the evaluation of our approach.

We found all coefficients to be inferred with high accuracy from a 1 Mbp window with an average coverage above $1\times$ (Figure 4). If the amount of data were much lower than that, estimates were generally less accurate. In particular, we found the coefficients for the quality (β_q and β_{q^2}) to be often slightly overestimated at low coverages, likely because many sequencing errors go undetected since they can only be inferred at sites covered at least twice. However, this bias can be alleviated with larger window sizes if coverage is very low (see below).
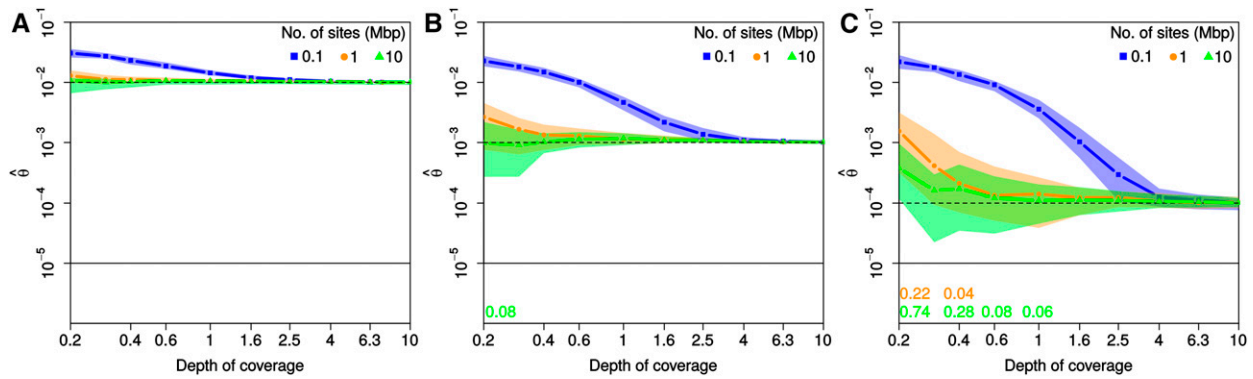
Accuracy of full pipeline

We finally used simulations to assess the accuracy of the full pipeline, that is, when inferring first the pattern of PMD, then

the recalibration coefficients given the inferred PMD pattern, and lastly using the recalibrated quality scores along with the inferred PMD pattern to estimate θ. In these simulations, the distortion of quality scores was, in addition to the four effects included above ($\beta_q = 1.5$, $\beta_q^2 = 0.05$, $\beta_p = -0.1$ and $\beta_q^2 = 5 \cdot 10^{-5}$), also affected by sequence context, in that simulated sequencing errors were 1.5 times more likely to result in a C or G than in an A or T.

Regardless of the true θ value we used, we detected a strong bias toward high values in our estimates whenever very little data were used (Figure 5, 0.1 Mbp). This is a direct result of the overestimation of the quality scores during the recalibration step as reported above, which leads to an overestimation of diversity. Encouragingly, however, this bias is overcome with only slightly more data. Indeed, we found 1 Mbp of data with an average coverage of well below $1\times$ to be sufficient to accurately infer $\theta \gtrsim 10^{-3}$, and of $1\times$ for $\theta = 10^{-4}$. Notably, even lower average coverages were sufficient when data were available for 10 Mb. Finally, we found an average coverage of $4\times$ to be sufficient when conducting recalibration and inference in windows as small as 0.1 Mb. These results thus suggest that our approach may be useful not only for hemizygous individuals with large chunks of haploid DNA (the sex chromosomes), but may also work well in other individuals when using mtDNA, for which high coverage can be obtained, or ultraconserved elements for recalibration.

Application

We illustrate the benefit of our approach by inferring θ for several ancient human male samples, and comparing these estimates to those obtained for several male individuals from the 1000 Genomes Project. For the ancient genomes, we first inferred PMD patterns from chromosome 1 using the exponential model introduced here (Supplemental Material, Figure S1), then used the first 20 Mbp of the X chromosome to perform recalibration individually for each read group, taking the inferred PMD pattern into account (Figure S2). Finally, we used both the inferred PMD patterns, as well as the recalibrated quality scores, to infer θ in windows of

Figure 5 Accuracy in estimating θ using the full pipeline. Results from sets of 50 simulations, each consisting of data from a haploid as well as a diploid region used to conduct recalibration and inference of θ, respectively. The data sets in (A, B, and C) were simulated with different true values of θ, which are indicated with the dashed lines, and were 10^{-2}, 10^{-3} and 10^{-4}, respectively. Each data set was simulated with PMD as well as distorted base quality scores according to Equation 12, with $\beta_q = 1.5$, $\beta_q^2 = 0.05$, $\beta_p = -0.1$, and $\beta_q^2 = 5 \cdot 10^{-5}$. In addition, these simulations also included context effects in that sequencing errors were simulated to result 1.5 times more often in a C or G than in an A or T. The average coverage indicated is for the diploid data, while the haploid data were simulated with half the coverage. Line segments and polygons correspond to the median and the 90% quantile of all estimated $\hat{\theta}$ within the set of simulations, respectively.

1 Mbp in the whole genome, excluding windows closer than 5 Mbp to telomeres or centromeres as defined by the track *Gap* in group *Mapping and Sequencing* in the UCSC Table Browser (Karolchik *et al.* 2008).

The samples that we analyzed this way were (1) two European hunter-gatherer individuals (Jones *et al.* 2015), namely the Mesolithic genome "Kotias" from Kotias Klde cave from Western Georgia (KK1), and the western European Late Upper Paleolithic genome, "Bichon" from Grotte du Bichon, Switzerland (Bich), ~17,700 years old (2) an individual from the Bronze age burial site at Ludas-Varjú-dülö, Hungary (BR2 Gamba *et al.* 2014), and (3) a 4500-year-old male from Mota Cave in the southern Ethiopian highlands (Gallego Llorente and Jones 2015). All these samples had relatively high coverage ($> 10\times$), and thus allowed us to infer fine scaled patterns of heterozygosity along the genomes, even for regions with low diversity ($\theta < 10^{-4}$).

For comparison, we also inferred diversity patterns for nine modern males from three populations that were analyzed as part of the 1000 Genomes Project, phase 3 (alignment files downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/). These were the British males HG00115, HG00116, and HG00117, Tuscan males NA20509, NA20511, and NA20762, and Yoruban males NA18486, NA18519, and NA18522. As shown in Figure 6B, these nine individuals portrayed the expected pattern of higher diversity in African than European individuals, but they also revealed significant variation among individuals of the same population (t-test, $p < 10^{-5}$ in at least two out of three possible comparisons in each population). Larger differences in overall diversity were observed among the ancient samples analyzed. Unsurprisingly, the African sample Mota exhibited the highest diversity of all ancient samples, which was also higher than the diversity observed in modern day Europeans, yet lower than modern day Yorubans. The ancient sample with the second highest diversity was the Bronze Age sample BR2, whose di-

versity falls well within the range of estimates obtained from modern day Europeans. In contrast, the two European hunter-gatherer samples KK1 and Bichon showed much lower diversity than modern day Europeans with their median estimates being 15–25% lower than the median estimates of modern Europeans. These results suggest that while hunter-gatherer populations had much lower diversity, the diversity found in Europeans about 3000 years ago was very comparable to the diversity observed today. This conclusion is in perfect agreement with a temporal trend in the total length of runs of homozygosity (ROH) inferred from imputed genotypes among ancient samples from Hungary that spanned a period from 5700 to 100 BC and also included the sample BR2 (Gamba *et al.* 2014).

The inference of local diversity patterns also allows us to compare the distribution of diversity in the genome between individuals, regardless of the overall level of diversity. This analysis revealed a substantial phylogenetic signal in the distribution of diversity as quantified by Spearman correlations. For instance, the diversity pattern is more strongly correlated among modern Yorubans (pairwise Spearman correlations between 0.55 and 0.60) than between Yorubans and Europeans (0.40–0.50). Similarly, the diversity pattern of the ancient African samples Mota is most strongly correlated with that of modern day Yorubans (0.491–0.537), and much less so with modern day Europeans (0.362–0.433). Interestingly, European samples are more diverse in their patterns than Africans (pairwise Spearman correlations between 0.42 and 0.48), and their correlations do not exceed those obtained when comparing African and European individuals. Nonetheless, hierarchical clustering groups all modern day Europeans together, and also puts the Bronze age sample BR2 at the basis of that clade (Figure 6C).

The lowest pairwise correlations (0.28–0.39) were found for comparisons involving the two European hunter-gatherer samples KK1 and Bich, with the overall lowest being the
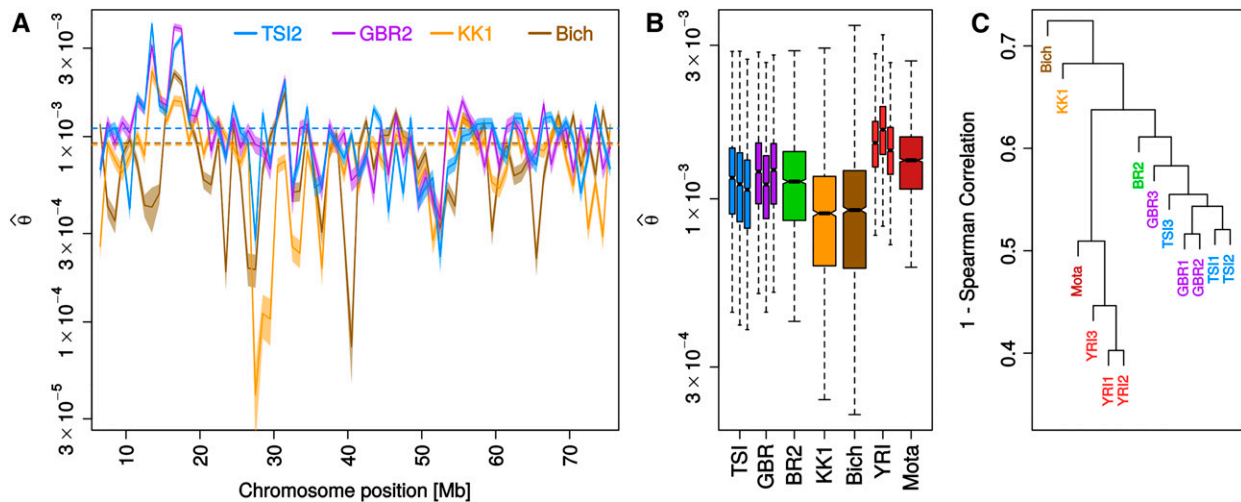
Figure 6 Local diversity in ancient and modern humans. (A) Heterozygosity (θ) inferred in 1 Mb windows along the first 75 Mbp of chromosome 1 (excluding windows closer than 5 Mbp of the telomere) for two modern Europeans (TSI2 and GBR2), and two ancient European hunter-gatherers (KK1 and Bich). Solid lines indicate the MLE estimate, shading indicates the 95% confidence intervals and dashed lines the genome-wide median for each sample. (B) Distribution of estimates $\hat{\theta}$ in 1 Mbp windows across the first 22 chromosomes of each sample. (C) Similarity in the pattern of θ along the genome visualized by hierarchical clustering using $1 -$ Spearman correlation as distance.

correlation between these samples (0.28). This is also illustrated visually when plotting our estimates of the first 75 Mbp of chromosome 1, where we found relatively high concordance in local diversity among the two European samples, but vastly different patterns among the hunter-gatherer samples (Figure 6C). These results suggest that, despite very comparable overall levels of diversity, the distribution of diversity along the genome was very diverse among European hunter-gatherer populations, and very different from the one observed among modern day individuals. Multiple observations support such a conclusion: first, the two samples analyzed here represent two vastly different geographic regions, with one being a sample from Switzerland, and the other from Georgia, and were previously reported to belong to two different clades that split 45,000 years ago, as inferred from genotyping data (Jones *et al.* 2015). Second, the ancestry of modern Europeans traces only partly back to European hunter-gatherers, with early Neolithic people from the Aegean (Hofmanová *et al.* 2016) and Yamnaya steppe herders (Haak *et al.* 2015) contributing the majority of the modern-day genetic makeup. Finally, the two European hunter-gatherer samples both exhibit many, but unique, regions of very low diversity ($\hat{\theta} < 10^{-4}$ in 4% of all windows, *cf.* 0.00–0.03% in all modern Europeans), which do not have particularly low coverage. They are likely the result of small population sizes with some level of consanguinity in the population (Pemberton *et al.* 2012).

Discussion

Quantifying genetic diversity and comparing it between different individuals and populations is fundamental to understanding the evolutionary processes shaping genetic variation. Unfortunately, the inference of heterozygosity is confounded by both sequencing errors, resulting in false diversity as well as the statistical power to identify heterozygous sites, particularly when coverage is low. Several methods have been developed to learn about heterozygosity probabilistically, that is, without the need to first call genotypes. A rather recent such approach (Bryc *et al.* 2013) proposes to leverage data from external reference individuals to obtain an unbiased estimate of the probability that a specific sites is heterozygous. The expected heterozygosity is then estimated from these site-specific estimates. Since this approach requires per site estimates to be accurate, only sites with a coverage of 5× or higher can be included in the analysis, which severely limits the scope of the application.

An alternative is to infer heterozygosity probabilistically from a collection of sites. Among the earliest such methods was a likelihood-based estimator (Lynch 2008), which infers heterozygosity of an individual jointly with the rate sequencing errors. We presented a natural extension of this approach that relaxes the infinite sites assumption and integrates PMD, a particular feature of ancient DNA that is not captured by base quality scores provided by sequencing machines. We achieved this by extending Felsenstein's model of substitutions (Felsenstein 1981) with an explicit model of next-generation sequencing data that incorporates both sequencing errors as well as errors arising from PMD. We have previously used the same PMD error model for variant calling from ancient samples (Hofmanová *et al.* 2016), and we note that it may be used to extend any other inference method based on genotype likelihood to ancient DNA.

Following other recently developed approaches to infer genetic diversity from next-generation sequencing data (*e.g.*, Nielsen *et al.* 2012; Korneliussen 2014; Maruki and Lynch 2015), our method also relaxes the assumption of constant error across all reads by benefiting from the base-specific

quality information provided by current sequencing technologies. Yet since these provided quality scores are often distorted, we also introduced here a method to recalibrate the quality scores for low coverage genomes. In contrast to commonly used methods for recalibration (*e.g.*, McKenna *et al.* 2010; DePristo *et al.* 2011; Cabanski *et al.* 2012), our approach does not require information about the underlying sequence context. It only assumes sites to be monomorphic while integrating over the uncertainty of the sequence itself. Examples of regions known to be monomorphic are the sex chromosomes in hemizygous individuals. But, since we found that our method recalibrates quality scores with high accuracy and reliably, even based on DNA stretches as short as 1 Mbp, we are confident that it will work even on ultraconserved elements or plasmid DNA. Finally, we note that if multiple individuals are sequenced together, they are likely affected by the same distortion of quality scores, and can hence be recalibrated with parameters inferred from a subset of them (*e.g.*, the male samples).

As an illustration, we applied the methods developed here to modern and ancient human samples of various coverage. While our approach to infer heterozygosity incorporates the possibility of PMD, it assumes that the probability of a PMD event occurring is known. We thus also introduce a method to infer these probability functions from raw data, which is robust to divergence between the sample and the reference genome. By inferring PMD patterns for each sample, then the recalibration parameters, and, finally, local diversity in 1 Mbp windows, we found that both ancient and modern African samples exhibited much larger diversity than European individuals. In addition, the diversity inferred from two ancient European hunter-gatherer samples was much lower than that of modern samples, which is likely explained by smaller population sizes. Besides overall differences in diversity, also the pattern of diversity along the genome revealed a strong geographic clustering among modern and ancient samples. The exceptions were the two European hunter-gatherers that showed patterns very different from both modern samples, as well as from one another, further corroborating the view (Jones *et al.* 2015) that these samples represent different and ancient clades that contributed only marginally to the genetic make-up of modern day Europeans.

Acknowledgments

Literature Cited

Barnett, D. W., E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth, 2011   Bamtools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27: 1691–1692.

Briggs, A., and U. Stenzel, 2007   Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. USA 104: 14616–14621.

Bryc, K., N. Patterson, and D. Reich, 2013   A novel approach to estimating heterozygosity from low-coverage genome sequence. Genetics 195: 553–561.

Cabanski, C. R., K. Cavin, C. Bizon, M. D. Wilkerson, J. S. Parker *et al.*, 2012   ReQON: a bioconductor package for recalibrating quality scores from next-generation sequencing data. BMC Bioinformatics 13: 221.

Dempster, A. A., N. N. Laird, and D. D. B. Rubin, 1977   Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39: 1–38.

DePristo, M. a., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011   A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Felsenstein, J., 1981   Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Gallego Llorente, M., E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur *et al.*, 2015   Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350: 820–822.

Gamba, C., E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes *et al.*, 2014   Genome flux and stasis in a five millennium transect of European prehistory. Nat. Commun. 5: 5257.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick *et al.*, 2015   Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522: 207–211.

Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega *et al.*, 2008   Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 18: 1020–1029.

Hofmanová, Z., S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann *et al.*, 2016   Early farmers from across Europe directly descended from Neolithic Aegeans. Proc. Natl. Acad. Sci. USA 113: 6886–6891.

Jiang, R., S. Tavare, and P. Marjoram, 2008   Population genetic inference from resequencing data. Genetics 181: 187–197.

Johnson, P. L. F., and M. Slatkin, 2006   Inference of population genetic parameters in metagenomics: a clean look at messy data Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 16: 1320–1327.

Jones, E. R., G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson *et al.*, 2015   Upper palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Comm.* 6: 1–8

Jónsson, H., A. Ginolhac, M. Schubert, P. L. F. Johnson, and L. Orlando, 2013   mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29: 1682–1684.

Karolchik, D., R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson *et al.*, 2008   The UCSC genome browser database: 2008 update. Nucleic Acids Res. 36: D773–D779.

Korneliussen, T., 2014   ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15: 1–13.

Li, H., 2011   A statistical framework for {SNP} calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009   The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Lunter, G., and M. Goodson, 2011   Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21: 936–939.

Lynch, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25: 2409–2419.

Maruki, T., and M. Lynch, 2015 Genotype-frequency estimation from high-throughput sequencing data. Genetics 201: 473–486.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang, 2012 SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One 7: e37558.

Pemberton, T., D. Absher, M. Feldman, R. Myers, N. Rosenberg *et al.*, 2012 Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. 91: 275–292.

Skoglund, P., B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo *et al.*, 2014 Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. USA 111: 2229–2234.

Utturkar, S. M., D. M. Klingeman, J. M. Bruno-Barcena, M. S. Chinn *et al.*, 2015 Sequence data for Clostridium autoethanogenum using three generations of sequencing technologies. Sci. Data 2: 150014.

Communicating editor: J. Novembre

Appendix

Emission Probabilities in the Presence of PMD

Following Lynch (2008) and commonly used approaches (*e.g.*, Li 2011), we assume here that a sequencing read is equally likely to cover any of the two alleles of an individual, and that sequencing errors may result in any of the alternative bases with equal probability, $\epsilon_{ij}/3$. In the absence of PMD, the probability of observing a base, d_{ij}, given the underlying genotype $g_i = kl$ is then given by

$$\mathbb{P}(d_{ij}|g_i = kl, \epsilon_{ij}) = \begin{cases} 1 - \epsilon_{ij} & \text{if } k = l = d_{ij} \\ \dfrac{\epsilon_{ij}}{3} & \text{if } k \neq d_{ij}, l \neq d_{ij} \\ \dfrac{1}{2} - \dfrac{\epsilon_{ij}}{3} & \text{if } k \neq l, k = d_{ij} \text{ or } l = d_{ij} \end{cases},$$

In ancient DNA, differences between the base observed within a read and the underlying alleles may also be the result of PMD. Following Hofmanová *et al.* (2016), let us denote by $D_{C \to T}(q_{ij})$ and $D_{G \to A}(q_{ij})$, the known probability that a $C \to T$ or $G \to A$ PMD occurred at the position covering site i in read j, respectively. In the presence of PMD, the probability of observing a base, d_{ij}, given the underlying genotype $g_i = kl$ is given by

$$\mathbb{P}(d_{ij}|g_i = kl, \epsilon_{ij}, q_{ij}) = \begin{cases} [1 - D_{G \to A}(q_{ij})]\dfrac{\epsilon_{ij}}{3} + D_{G \to A}(q_{ij})(1 - \epsilon_{ij}) & \text{if } d_{ij} = A, g_i = GG \\[2mm] \dfrac{[1 + D_{G \to A}(q_{ij})](1 - \epsilon_{ij})}{2} + \dfrac{[1 - D_{G \to A}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = A, g_i = AG \\[2mm] \dfrac{D_{G \to A}(q_{ij})(1 - \epsilon_{ij})}{2} + \dfrac{[2 D_{G \to A}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = A, g_i = CG, GT \\[2mm] [1 - D_{C \to T}(q_{ij})](1 - \epsilon_{ij}) + D_{C \to T}(q_{ij})\dfrac{\epsilon_{ij}}{3} & \text{if } d_{ij} = C, g_i = CC \\[2mm] \dfrac{[1 - D_{C \to T}(q_{ij})](1 - \epsilon_{ij})}{2} + \dfrac{[1 + D_{C \to T}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = C, g_i = AC, CG, CT \\[2mm] [1 - D_{G \to A}(q_{ij})](1 - \epsilon_{ij}) + D_{G \to A}(q_{ij})\dfrac{\epsilon_{ij}}{3} & \text{if } d_{ij} = G, g_i = GG \\[2mm] \dfrac{[1 - D_{G \to A}(q_{ij})](1 - \epsilon_{ij})}{2} + \dfrac{[1 + D_{G \to A}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = G, g_i = AG, CG, GT \\[2mm] [1 - D_{C \to T}(q_{ij})]\dfrac{\epsilon_{ij}}{3} + D_{C \to T}(q_{ij})(1 - \epsilon_{ij}) & \text{if } d_{ij} = T, g_i = TT \\[2mm] \dfrac{D_{C \to T}(q_{ij})(1 - \epsilon_{ij})}{2} + \dfrac{[2 - D_{C \to T}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = T, g_i = AC, CG \\[2mm] \dfrac{[1 + D_{C \to T}(q_{ij})](1 - \epsilon_{ij})}{2} + \dfrac{[1 - D_{C \to T}(q_{ij})]\epsilon_{ij}}{6} & \text{if } d_{ij} = T, g_i = CT \\[2mm] 1 - \epsilon_{ij} & \text{if } d_{ij} = A, g_i = AA \text{ or } d_{ij} = T, g_i = TT \\[2mm] \dfrac{1}{2} - \epsilon_{ij} & \text{if } d_{ij} = A, g_i = AC, AT \text{ or } d_{ij} = T, g_i = AT, GT \\[2mm] \dfrac{\epsilon}{3} & \text{otherwise} \end{cases}.$$

Newton-Raphson Algorithm to Infer PMD Patterns

Under the model proposed in Equation 10, the log likelihood of the data are given by

$$l(\alpha, \delta, \gamma) = \sum_p \mathcal{T}_{CTp} \log(\mu + \delta e^{-\alpha p}) + \sum_p \mathcal{T}_{CCp} \log(1 - \mu - \delta e^{-\alpha p}),$$

the gradient vector $\mathbf{F}(\alpha, \delta, \gamma)$ by

$$\mathbf{F}(\alpha, \delta, \gamma) = \begin{bmatrix} \mathbf{F}_\alpha \\ \mathbf{F}_\delta \\ \mathbf{F}_\gamma \end{bmatrix} = \sum_p \frac{\mathcal{T}_{CTp}}{\alpha + \delta e^{-\gamma p}} \begin{bmatrix} 1 \\ e^{-\gamma p} \\ -p\delta e^{-\gamma p} \end{bmatrix} + \sum_p \frac{\mathcal{T}_{CCp}}{1 - \alpha - \delta e^{-\gamma p}} \begin{bmatrix} -1 \\ -e^{-\gamma p} \\ p\delta e^{-\gamma p} \end{bmatrix}$$

and the Jacobian matrix $\mathbf{J}(\alpha, \delta, \gamma)$ by

$$\mathbf{J}(\alpha, \delta, \gamma) = \begin{bmatrix} \mathbf{F}_{\alpha\alpha} & \mathbf{F}_{\alpha\delta} & \mathbf{F}_{\alpha\gamma} \\ \mathbf{F}_{\alpha\delta} & \mathbf{F}_{\delta\delta} & \mathbf{F}_{\delta\gamma} \\ \mathbf{F}_{\alpha\gamma} & \mathbf{F}_{\delta\gamma} & \mathbf{F}_{\gamma\gamma} \end{bmatrix} = \sum_p \frac{n_p}{(\alpha + \delta e^{-\gamma p})^2} \mathbf{J}_p' + \sum_p \frac{N_p - n_p}{(1 - \alpha - \delta e^{-\gamma p})^2} \mathbf{J}_p'',$$

where

$$\mathbf{J}_p' = \begin{bmatrix} -1 & -e^{-\gamma p} & p\delta e^{-\gamma p} \\ -e^{-\gamma p} & -e^{-2\gamma p} & -p\alpha e^{-\gamma p} \\ p\delta e^{-\gamma p} & -p\alpha e^{-\gamma p} & p^2\alpha\delta e^{-\gamma p} \end{bmatrix}$$

and

$$\mathbf{J}_p'' = \begin{bmatrix} -1 & -e^{-\gamma p} & p\delta e^{-\gamma p} \\ -e^{-\gamma p} & -e^{-2\gamma p} & p(1-\alpha)e^{-\gamma p} \\ p\delta e^{-\gamma p} & p(1-\alpha)e^{-\gamma p} & -p^2(1-\alpha)\delta e^{-\gamma p} \end{bmatrix}.$$

The Newton-Raphson iteration for $\boldsymbol{\theta} = (\alpha, \delta, \gamma)^T$ is given by

$$\theta_{new} = \theta_{old} - J^{-1}(\theta_{old})F(\theta_{old}). \tag{15}$$

EM Algorithm to Infer Base-Specific Error Rates

We propose an EM algorithm for this estimation that is similar to the one above, but assume here that the base frequencies π_g, $g = A, G, C, T$ are known, *i.e.*, can be derived accurately from counting in the region. The complete data log-likelihood of our model is given by

$$l_c(\boldsymbol{\beta}|\boldsymbol{d}, g) = \sum_{i=1}^I \left(\log\mathbb{P}(d_i|g_i, \boldsymbol{\beta}) + \log\pi_{g_i} \right).$$

From this, we get the expected complete data log-likelihood

$$Q\left(\boldsymbol{\beta}; \boldsymbol{\beta}'\right) = \mathbb{E}\left[l_c(\boldsymbol{\beta}|\boldsymbol{d}, g)|\boldsymbol{d}, \boldsymbol{\beta}' \right] = \sum_{i=1}^I \sum_g \left(\log\mathbb{P}(d_i|g, \boldsymbol{\beta}) + \log\pi_g \right) \mathbb{P}\left(g|d_i, \boldsymbol{\beta}'\right).$$

For the M-step, we need only to consider the first part of $Q(\boldsymbol{\beta}; \boldsymbol{\beta}')$:

$$Q_1\left(\boldsymbol{\beta}; \boldsymbol{\beta}'\right) = \sum_{i=1}^I \sum_g \mathbb{P}\left(g|d_i, \boldsymbol{\beta}'\right) \log\mathbb{P}(d_i|g, \boldsymbol{\beta}),$$

where

$$\mathbb{P}\left(g|d_i, \boldsymbol{\beta}'\right) = \frac{\mathbb{P}\left(d_i|g, \boldsymbol{\beta}'\right)\pi_g}{\sum_h \mathbb{P}\left(d_i|h, \boldsymbol{\beta}'\right)\pi_h}$$

by Bayes' formula. From (13), we get more explicitly

$$Q_1\left(\boldsymbol{\beta};\boldsymbol{\beta}'\right) = \sum_{i,g,j}{}' \log\left(1 - D_{ij} + B_{ij}\epsilon_{ij}\right),$$

where we used the abbreviations $B_{ij} = [(4/3)D_{ij} - 1]$ and

$$\sum_{i,g,j}{}' \ldots = \sum_{i=1}^{I} \sum_{g} \mathbb{P}\left(g|d_i,\boldsymbol{\beta}'\right) \sum_{j=1}^{n_i} \ldots .$$

In order to maximize Q_1 for $\boldsymbol{\beta}$, we calculate the gradient vector $F(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta};\boldsymbol{\beta}')$ with components

$$F_m(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}_n} Q_1\left(\boldsymbol{\beta};\boldsymbol{\beta}'\right) = \sum_{i,g,j}{}' \frac{B_{ij}}{1 - D_{ij} + B_{ij}\epsilon_{ij}} \frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\beta}_m}, \tag{16}$$

for $m = 0, \ldots, L$. From (12), we obtain

$$\frac{\partial\epsilon_{ij}}{\partial\beta_m} = \epsilon_{ij}\left(1 - \epsilon_{ij}\right) \frac{\partial\eta_{ij}}{\partial\beta_m}.$$

Observe that $\partial\eta_{ij}/\partial\beta_0 = 1$ and $\partial\eta_{ij}/\partial\beta_m = q_{ijm}$ for $m = 1, \ldots, L$.

We solve $\mathbf{F}(\boldsymbol{\beta}) = (\mathbf{0})$ with the Newton-Ralphson method with the Jacobian matrix $J_{mn} = \partial F_m/\partial\beta_n$. From (16), we get

$$J_{mn}(\boldsymbol{\beta}) = \sum_{i,g,k}{}' \left[\frac{B_{ij}}{1 - D_{ij} + B_{ij}\epsilon_{ij}} \frac{\partial^2\epsilon_{ij}}{\partial\beta_m\partial\beta_n} - \frac{B_{ij}^2}{\left(1 - D_{ij} + B_{ij}\epsilon_{ij}\right)^2} \frac{\partial\epsilon_{ij}}{\partial\beta_m} \frac{\partial\epsilon_{ij}}{\partial\beta_n} \right]$$

where

$$\frac{\partial^2\epsilon_{ij}}{\partial\beta_m\partial\beta_n} = \epsilon_{ij}\left(1 - \epsilon_{ij}\right)\left(1 - 2\epsilon_{ij}\right) \frac{\partial\eta_{ij}}{\partial\beta_m} \frac{\partial\eta_{ij}}{\partial\beta_n}.$$

Putting everything together we obtain

$$J_{mn}(\boldsymbol{\beta}) = \sum_{i,g,k}{}' \frac{B_{ij}\epsilon_{ij}\left(1 - \epsilon_{ij}\right)}{\left(1 - D_{ij} + B_{ij}\epsilon_{ij}\right)^2} \left(\left(1 - D_{ij}\right)\left(1 - 2\epsilon_{ij}\right) - B_{ij}\epsilon_{ij}^2\right) \frac{\partial\eta_{ij}}{\partial\beta_m} \frac{\partial\eta_{ij}}{\partial\beta_n}.$$

The Newton-Ralphson iteration is

$$\boldsymbol{\beta}_{\text{new}} = \boldsymbol{\beta}_{\text{old}} - \mathbf{J}^{-1}(\boldsymbol{\beta}_{\text{old}})\mathbf{F}(\boldsymbol{\beta}_{\text{old}}).$$
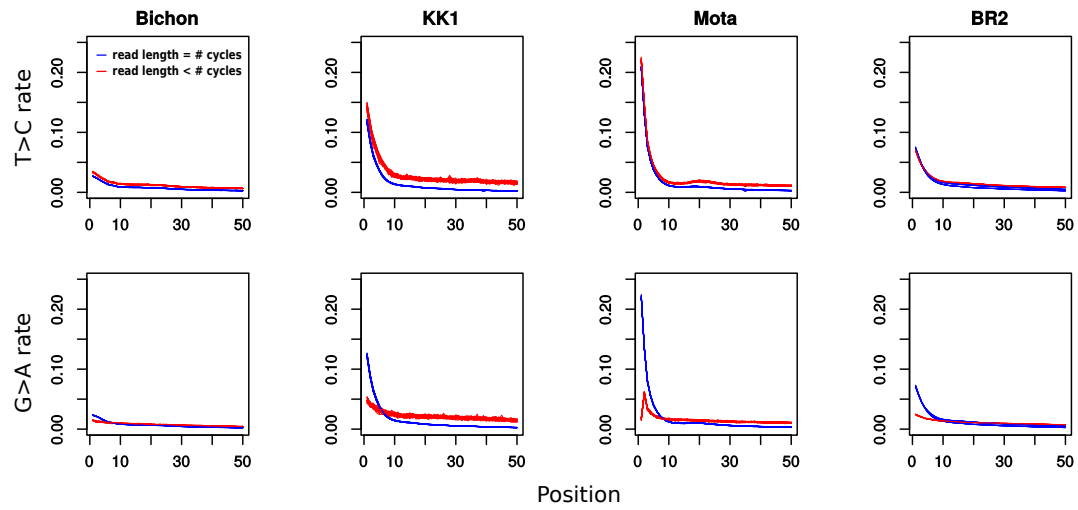
Figure S1: **Post-mortem damage pattern profiles of ancient samples**. Plotted is the frequency of reads showing T at all sites where the reference is C (top panels) and the frequency of reads showing A at all sites where the reference is G (bottom panels) as a function of the distance from the 5' and 3' end of the read, respectively, for each individual and read group. For single end sequencing runs, damage patterns are further shown individually for reads that are shorter (<cycles) or as long (=cycles) as the number of sequencing cycles used.
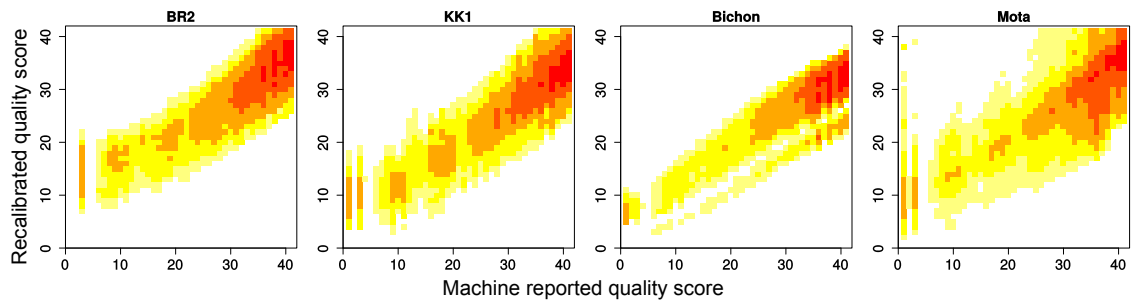
Figure S2: **Effect of Recalibration on Quality Scores.** Shown are the density distributions of the quality transformations as a result of the applied quality recalibration for each sample. It appears that the machine-reported qualities were overall too high, and in particular for the qualities > 30.