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, University of Fribourg, Fribourg, Switzerland, [†]Swiss Institute of Bioinformatics, Fribourg, Switzerland, [‡]Department of Mathematics, University of Fribourg, Fribourg, Switzerland, [§]Paleogenetics Group, University of Mainz, Mainz, Germany

ABSTRACT While genetic diversity can be quantified accurately from high coverage sequencing, it is often desirable to obtain such estimates from low coverage data, either to save costs or because of low DNA quality as observed for ancient samples. Here we introduce a method to accurately infer heterozygosity probabilistically from very low coverage sequences of a single individual. The method relaxes the infinite sites assumption of previous methods, does not require a reference sequence and takes into account both variable sequencing errors and potential post-mortem damage. It is thus also applicable to non-model organisms and ancient genomes. Since error rates as reported by sequencing machines are generally distorted and require recalibration, we also introduce a method to infer accurately recalibration parameter in the presence of post-mortem damage. This method does also not require knowledge about the underlying genome sequence, but instead works from 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 Mb of haploid data is sufficient for accurate recalibration even at average coverages as low as 1-3x. At similar coverages, out method also produces very accurate estimates of heterozygosity down to 10^{-4} within windows of about 1Mb. We further illustrate the usefulness of our approach by inferring genome-wide patterns of diversity for several ancient human samples and found that 3,000-5,000 samples showed diversity patterns comparable to 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 differently 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, Post-mortem damage, Base recalibration, Reference-free

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 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, but also genetic drift and selection, which affected the time that passed since the common ancestor. 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, while comparing heterozygosity between individuals may highlight differences in the demographic histories of populations.

While heterozygosity is readily obtained from high quality genotype calls by counting, it is much harder to infer accurately from low coverage genomes. 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

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 1st April, 2016%
[1]Department of Biology, University of Fribourg, Chemin du Musée 10, 1200 Fribourg, Switzerland, daniel.wegmann@unifr.ch

on a reference genome or by post-mortem 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 jointly estimating it together with heterozygosity from the data itself. Despite the additional parameter, this likelihood based estimator is generally more accurate, even if his implementation is ill-behaved at very low coverages (Lynch 2008).

Here we present a direct extension of the approach by Lynch (2008) that relaxes the assumption of infinitely many sites, does not require base frequencies to be known *a priori* and takes additional biases introduced by PMD fully into account. We achieve this by modeling genotype frequencies using the classic substitution model of Felsenstein (1981), which allows for back mutations, and by modeling PMD explicitly.

We further relax the assumption of a constant error rate. While variable error rates along and between individual reads is a well characterized feature of all current sequencing technologies, the provided estimates of these (the quality 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.* 2015), it requires detailed knowledge of the study species, which is often lacking for non-model organisms.

We circumvent this problem by using a reference-free recalibration approach that makes use of the base-quality information provided by sequencing machines. We rely on haploid sequences such as those from the X-Chromosome in male mammals and integrate 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 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 2x for windows of $\sim$ 1Mb. We further show that a few Mb of data at equally low coverage is sufficient to properly recalibrate distorted quality scores. Finally we use the here developed methods 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

Here we develop a method to estimate heterozygosity from a collection of aligned reads by integrating out the uncertainty of the local genotype as well as the potential effects of post-mortem DNA damage (PMD). Specifically, we are interested in inferring the stationary base frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$, along 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.

To estimate θ, we will extend Felsenstein's model of substitutions (Felsenstein 1981) to account for the uncertainty in the local genotypes. However, we will assume that base-specific rates of sequencing errors and PMD are known constants, motivated by the observation that rates of sequencing errors and PMD can be learned accurately from genome-wide data prior to inferring θ and $\boldsymbol{\pi}$, as we will show below.

Inferring Heterozygosity

Substitution model 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

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

Emission probabilities This model is easily extended to integrate out the uncertainty in observed genotypes. To do so we adopt a model similar to Lynch (2008) and those commonly used for genotype calling (Li 2011, e.g.). We will further closely follow the notation recently introduced by Hofmanová *et al.* (2015).

The observed data d_i at site i shall correspond to what is typically obtained when individual reads of next generation sequencing approaches were mapped to a reference genome. Here we will assume that all sequencing reads were accurately mapped and hence 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 rates of sequencing errors and PMD, which we assume to be known for each observed base. Let us denote these base specific rates by ϵ_{ij} and D_{ij}, respectively, for $j = 1, \ldots, n_i$ and further assume that the sequencing errors and PMD occur independently between reads. The likelihood of the full data at site i is thus given by

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

where $\boldsymbol{\epsilon}_i = \{\epsilon_{i,1}, \ldots, \epsilon_{i,n_i}\}$.

Let us first develop the emission probability $\mathbb{P}(d_{ij}|g_i, \epsilon_{ij})$ for the case of no PMD ($D_{ij} = 0$). Following Lynch (2008) and commonly used approaches (Li 2011, e.g.), 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

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

Post-mortem DNA damage We will now extend this model with the possibility of PMD. The most common form of PMD is C deamination, which leads to a $C \to T$ transition on the affected strand and a $G \to A$ 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. Here we will develop our model for this form of PMD following the formulation of Skoglund *et al.* (2014), but we note that it is readily extended to incorporate other forms of PMD as well.

We feel that the rationale of our approach is best explained with a specific example. Consider $d_{i,j} = T$ 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

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

where D_{ij} denotes the probability that a $C \to T$ 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. Since we consider both ϵ_{ij} and D_{ij} to be known constants and in an effort to unburden the notation, we will refer to the emission probabilities simply as $\mathbb{P}(d_i|g_i)$ in the following.

Inference using EM-Algorithm Assuming sites to be independent, the full likelihood of our model is given by

$$\begin{aligned} L(\theta, \boldsymbol{\pi}) &= \mathbb{P}(\boldsymbol{d}|\theta, \boldsymbol{\pi}) = \prod_{i=1}^{I} \mathbb{P}(d_i|\theta, \boldsymbol{\pi}) \\ &= \prod_{i=1}^{I} \sum_{g} \mathbb{P}(d_i|g_i = g)\mathbb{P}(g_i = g|\theta, \boldsymbol{\pi}), \end{aligned}$$

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

To find the maximum likelihood estimate (MLE) of the model parameters θ and $\boldsymbol{\pi}$, we will adopt an Expectation-Maximization (EM) algorithm. The complete likelihood 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

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

The expected complete data log-likelihood is calculated as

$$\begin{aligned} &Q(\theta, \boldsymbol{\pi}; \theta', \boldsymbol{\pi}') = \mathbb{E}\left[l_c(\theta, \boldsymbol{\pi}; \boldsymbol{d}, \boldsymbol{g}) \mid \boldsymbol{d}; \theta', \boldsymbol{\pi}'\right] \\ &= \sum_{i=1}^{I} \sum_{g} \left[\log \mathbb{P}(d_i|g) + \log \mathbb{P}(g|\theta, \boldsymbol{\pi})\right] \mathbb{P}(g|d_i; \theta', \boldsymbol{\pi}') \end{aligned}$$

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

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

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

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

Let us write out Q_2 explicitly:

$$\begin{aligned} Q_2(\theta, \boldsymbol{\pi}; \theta', \boldsymbol{\pi}') &= \sum_{k} P_{kk} \left[\log \pi_k + \log(e^{-\theta} + \pi_k(1 - e^{-\theta}))\right] \\ &+ \sum_{k} \sum_{l \neq k} P_{kl} \left[\log \pi_k + \log \pi_l + \log(1 - e^{-\theta})\right]. \end{aligned}$$

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

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

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

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

We have to set these equations to zero and solve for π_k, θ, and μ.

With the parameter transformation $\rho = e^{-\theta}/(1 - e^{-\theta})$, the equations can be rewritten as ($k = 1, \ldots, 4$):

$$\begin{aligned} F_k(\boldsymbol{\pi}, \rho, \mu) &:= P_{kk}\left(1 + \frac{\pi_k}{\rho + \pi_k}\right) + 2\sum_{l \neq k} P_{kl} - \mu\pi_k = 0, \\ F_5(\boldsymbol{\pi}, \rho, \mu) &:= I - \sum_k \frac{P_{kk}(\rho + 1)}{\rho + \pi_k} = 0, \\ F_6(\boldsymbol{\pi}, \rho, \mu) &:= \sum_k \pi_k - 1 = 0. \end{aligned} \quad (4)$$

To streamline the notation, we will rename our variables:

$$(\pi_1, \ldots, \rho, \mu) \to \mathbf{x} = (x_1, \ldots, x_5, x_6).$$

We will solve the above system

$$\mathbf{F}(\mathbf{x}) = \mathbf{0} \quad (5)$$

with the Newton-Ralphson method. We thus need to determine the 6×6 Jacobian matrix $J_{ij} = \partial F_i / \partial x_j$. These are the non-zeros entries of the Jacobian where $k = 1, \ldots, 4$:

$$\begin{aligned} J_{kk} &= \frac{P_{kk} x_5}{(x_k + x_5)^2} - x_6, \\ J_{5k} &= (x_5 + 1) \sum_{l=1}^{4} \frac{P_{ll}}{(x_l + x_5)^2}, \\ J_{6k} &= 1, \\ J_{k5} &= -\frac{P_{kk} x_k}{(x_k + x_5)^2}, \\ J_{k6} &= -x_k, \\ J_{55} &= \sum_{l=1}^{4} \frac{P_{ll}(1 - x_l)}{(x_l + x_5)^2}. \end{aligned} \tag{6}$$

We can now approximate the zero of (5) with the iteration

$$\mathbf{x}_{new} = \mathbf{x}_{old} - \mathbf{J}^{-1}(\mathbf{x}_{old}) \mathbf{F}(\mathbf{x}_{old}). \tag{7}$$

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 $\theta = -\log(x_5/(1 - x_5))$.

A brief outline of an efficient implementation of the algorithm is given in the Appendix.

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

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

and the corresponding derivatives are

$$\frac{\partial}{\partial \theta} \mathbb{P}(g_i = kl | \theta, \boldsymbol{\pi}) = \begin{cases} (\pi_k^2 - \pi_k) e^{-\theta} & \text{if } k = l, \\ \pi_k \pi_l e^{-\theta} & \text{if } k \neq l. \end{cases} \tag{8}$$

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

$$\mathcal{I}(\hat{\theta}) = \sum_{i=1}^{I} \mathcal{R}_i (\mathcal{R}_i + 1) \tag{9}$$

where we have set

$$\mathcal{R}_i = \frac{\sum_g \mathbb{P}(d_i | g) \frac{\partial}{\partial \theta} \mathbb{P}(g | \hat{\theta}, \boldsymbol{\pi})}{\sum_g \mathbb{P}(d_i | g) \mathbb{P}(g | \hat{\theta}, \boldsymbol{\pi})}. \tag{10}$$

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

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

Estimating base-specific error rates

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 DePristo *et al.* (2011); McKenna *et al.* (2010). 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, 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 shortcoming: first, they can not be applied to ancient DNA since they do not take PMD into account. Second, both require a reference sequence as well as knowledge on 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.* 2015), we here develop an approach that also integrates over the unknown reference sequence.

To do so, we will assume that there exists a genomic region for which the individual does not show any polymorphism. A good example of such a genomic region are non-homologous sequences from sex chromosomes in heterogametic individuals (e.g. most of the X chromosomes in mammalian males), and we will describe our approach having this type of data in mind. However, we note that our approach is also 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 hidden 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

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

where $\boldsymbol{q_{ij}} = (q_{ij1}, \ldots, q_{ijL})$ is a given external vector of informations 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 non-linear 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

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

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),

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

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.

EM-Algorithm We propose an EM algorithms 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

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

From this we get the expected complete data log-likelihood

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

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

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

where

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

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

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

where we used the abbreviations $B_{ij} = \frac{4}{3}D_{ij} - 1$ and

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

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

$$F_m(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_n} Q_1(\boldsymbol{\beta}; \boldsymbol{\beta}') = \sum_{i,g,j}{}' \frac{B_{ij}}{1 - D_{ij} + B_{ij}\epsilon_{ij}} \frac{\partial \epsilon_{ij}}{\partial \beta_m}, \qquad (14)$$

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

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

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 (14) we get

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

where

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

Putting everything together we obtain

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

The Newton-Ralphson iteration is

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} - \mathbf{J}^{-1}(\boldsymbol{\beta}_{old})\mathbf{F}(\boldsymbol{\beta}_{old}).$$

Estimating Rates of Post-Mortem Damage

As mentioned above, the most common form of PMD is C deamination, which leads to a $C \to T$ transition on the affected strand, and a $G \to A$ transition on the complimentary strand (e.g. Briggs and Stenzel 2007). These deaminations occur more frequently in single stranded DNA, and are therefore observed more frequently close to natural break-points, i.e. at the ends of the DNA fragments. Consequently, the rates of PMD, while specific to the sample and the sequencing protocol used, are generally decaying 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 $C \to T$ and $G \to A$ transitions may be seen in a single read, but are accumulated and opposite ends.

Here we follow Jónsson *et al.* (2013) and estimate PMD rates directly from genome-wide counts of $C \to T$ and $G \to A$ 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 r read at a site with reference base s 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 here develop an approach that takes this divergence into account.

Position specific estimator 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}. We thus have

$$\mathcal{T}_{rsp} \sim B(\mathcal{T}_{r\cdot p}, \mu_{rs}),$$

where $B(\cdot, \cdot)$ is the binomial distribution and

$$\mathcal{T}_{r\cdot p} = \sum_{b \in A,C,G,T} \mathcal{T}_{rbp}.$$

For cells affected by PMD, such as $\mathcal{T}_{CTp}$, we then have

$$\mathcal{T}_{CTp} \sim B(T_{C\cdot p}, \mu_{CT} + (1-\mu_{CT})D_{C\to T}),$$

where $D_{C\to T}$ is the rate of $C \to T$ PMD.

Under the assumption that $\mu_{rs} = \mu_{sr}$, we obtain an ML estimate of $D_{C\to T}$ (and analogously for $D_{G\to A}$) as

$$\hat{D}_{C\to T,p} = \frac{f_{CT,p} - f_{TC,p}}{1 - f_{TC,p}},$$

where

$$f_{rs,p} = \frac{\mathcal{T}_{rsp}}{\mathcal{T}_{r\cdot p}}.$$

We note that this approach may lead to an ML estimate of $D < 0$ when $f_{CT} < f_{TC}$. In this case we set $D = 0$, and our estimator thus corresponds to a maximum *a posteriori* estimate using a uniform prior $\mathbb{P}(D) \sim U[0,1]$.

We further note that this approach assumes that all differences observed between reads and the reference are due to divergence or PMD, but not sequencing errors. While sequencing errors, divergence and PMD can not be jointly inferred, additional insight into the accuracy of our approach is gained by studying the alternative extreme case in which all difference observed between reads and the reference are assumed to be due to PMD or sequencing errors alone. By denoting the genome-wide sequencing error rate by ϵ, the relevant equations become

$$\mathcal{T}T_{TCp} \sim B(T_{T\cdot p}, \epsilon)$$

and

$$\mathcal{T}_{CTp} \sim B(T_{C\cdot p}, D_{C\to T}(1-\epsilon) + (1-D_{C\to T})\epsilon),$$

and the ML estimate

$$\hat{D}_{C\to T} = \frac{f_{CT} - f_{TC}}{1 - 2f_{TC}}.$$

Since average sequencing error rates are on the order of 1%, they dominate the table $\mathcal{T}$ only in cases when f_{TC} are small (on the order of 1%). As a consequence, the error when estimating the rates of PMD due to the omission of the factor of 2 in the denominator is never larger than 1% of the estimated value.

Exponential model Since the rate of PMDs is generally low far away from the read ends, position specific estimates may become noisy for these positions, particularly if data is limited. We thus also 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 will assume that the probability of observing base $\mathcal{T}$ when the reference sequence is a C at position p is given by

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

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

To obtain ML estimates for the parameters of this probability function we again turn to the Newton-Raphson algorithm as shown in the following. However, we note that some of the parameters are non-identifiable, and we thus show here how to obtain estimates for the parameters of the probability function

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

The log likelihood of the data is then 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

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} - \mathbf{J}^{-1}(\boldsymbol{\theta}_{old})\,\mathbf{F}(\boldsymbol{\theta}_{old}). \tag{15}$$

From these estimates we now obtain estimates for our parameters μ_{CT}, a, b and c as follows. First, and under the assumption that $\alpha_{CT} = \alpha_{TC}$, we obtain the ML estimate

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

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

Implementation

All approaches mentioned were implemented in a custom C++ program available at our lab website. We used functions included in the library BamTools for manipulating bam files (Barnett *et al.* 2011).

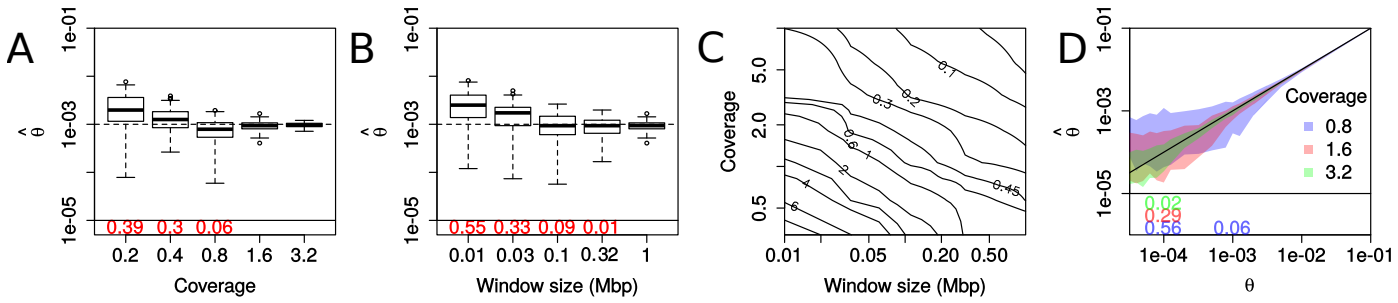

Figure 1 Power to infer θ from low coverage data. Results from sets of 100 simulations with post-mortem damage for different average coverage, window size and true θ values. (A) Estimated $\hat{\theta}$ in windows of 1 Mb as a function of average coverage. (B) Estimated $\hat{\theta}$ as a function of window size and fixed average coverage of 1x. (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 versus 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 panels A,B and D, replicates resulting in a $\hat{\theta} < 10^{-5}$ are not shown, but their percentage across replicates are printed below the horizontal black line.

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 $\boldsymbol{\pi} = \{0.25, 0.25, 0.25, 0.25\}$.

2. The second, homologous chromosome was simulated according to the Felsenstein (1981) substitution model (eq. 1) with $\boldsymbol{\pi}$ 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 $1 - L$ and L until the desired average coverage was reached. All reads copied from the second chromosome were considered to map to the reverse strand.

4. Post-mortem damage (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 $C \to T$ and $G \to A$ 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 μ_Q and standard deviation σ_Q, 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 eq. 12 with coefficients $\boldsymbol{\beta}$ to obtain the true error rate, with which sequencing errors were simulated.

6. The simulated data was 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).

Power to infer Heterozygosity

To check the power of our approach to infer θ from low coverage data, we first simulated data within a 1 Mb 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 (< 1x), θ was often overestimated, or inferred as zero. This is not surprising as the information about genetic diversity can only come from sites covered at least twice, which is rare at average coverages < 1x. As soon as average coverage exceeded 1.5x, however, our approach estimated θ at 10^{-3} very accurately (Fig. 1A).

We next performed simulations with a fixed coverage of 1x, but varying the window size (Fig 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 (Fig 1C), thus suggesting our method to perform well also at average coverages below 1x if the window size is large enough.

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 (Fig. 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 1Mb if the average coverage exceeds 3x.

All results above were generated assuming base-specific quality scores to be normally distributed with $\mu_Q = 20$ and standard deviation $\sigma_Q = 4.5$, which is the minimum quality expected with current sequencing approaches. 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 (2).

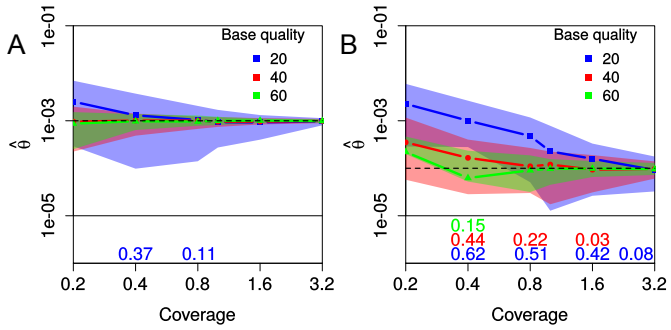

Figure 2 Effect of sequencing quality on power to estimate θ. Results from sets of 100 simulations to assess the power to estimate θ of 10^{-4} and 10^{-3} for panels A and B, respectively, for different average base qualities distributed normally with mean 20, 40 or 60 and a standard deviation of 4.5, 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 percentage excluded across are printed below the horizontal black line). All simulations were conducted with PMD and the true PMD probability functions were used during the estimation.

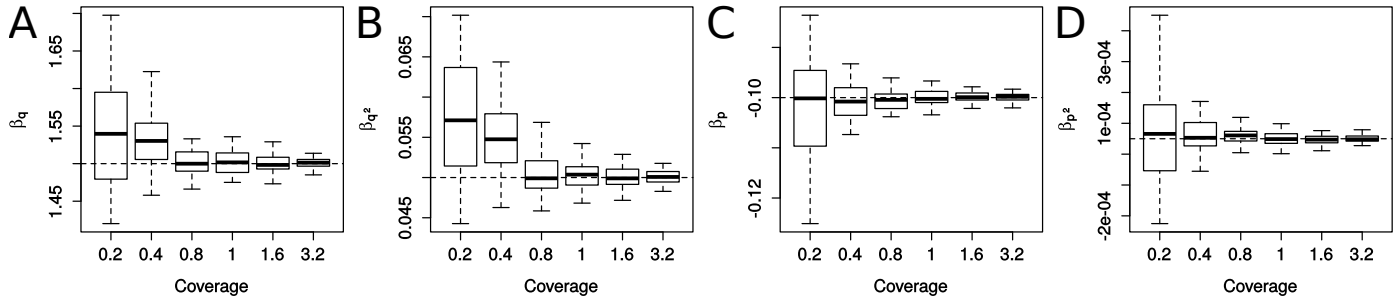

Figure 3 Accuracy in inferring recalibration parameters. Results from sets of 100 simulations are shown where sequence data from a haploid 1Mb region was simulated assuming a uniform distribution of observed quality scores ($U[5,60]$) that were then transformed to true qualities according to eq. 12 with $\beta_q = 1.5$,$\beta_{q^2} = 0.05$,$\beta_p = -0.1$, $\beta_{p^2} = 5 \cdot 10^{-5}$ and all context coefficients at 1.0. All simulations were conducted with PMD and the true PMD probability functions were used during the estimation.

For instance, we found that at an average coverage of 0.8x, 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 3.2x.

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 for instance 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 eq. 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 decent error rates for the evaluation of our approach.

We found all coefficients to be inferred with high accuracy from a 1Mb window with an average coverage above 1x (Fig. 3). If the amount of data was 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 $\beta_{p^2} = 5 \cdot 10^{-5}$), 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 in our estimates whenever very little data was used (Fig. 4).

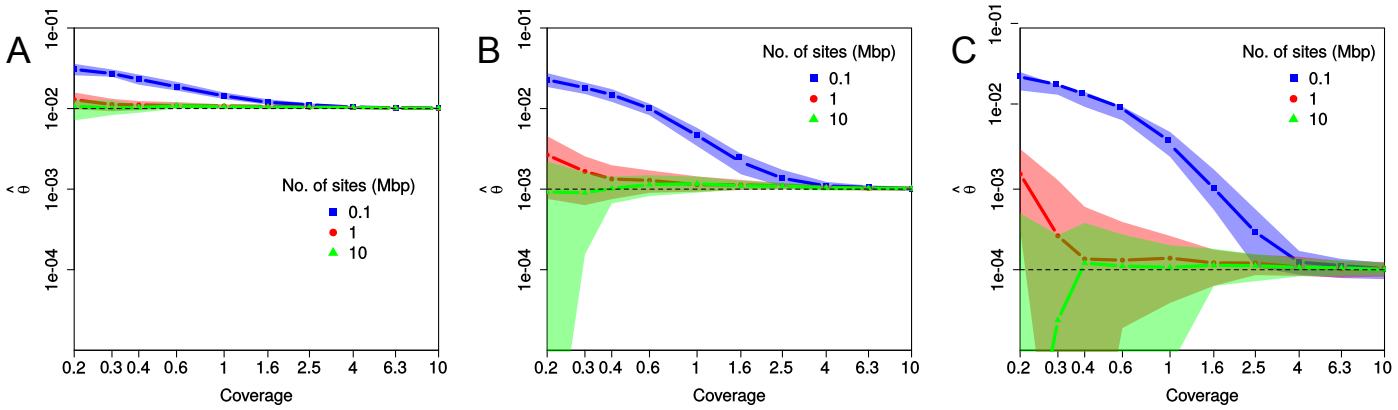

Figure 4 Accuracy in estimating θ using the full pipeline. Results from sets of 50 simulations each consisting of data from a haploid as well as a diploid region used to conduct recalibration and inference of θ, respectively. The data sets in panels A, B and C were simulated with different true values 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 eq. 12 with $\beta_q = 1.5, \beta_q^2 = 0.05, \beta_p = -0.1$, $\beta_q^2 = 5 \cdot 10^{-5}$. 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 was 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.

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 1Mb of data with an average coverage of well below 1x to be sufficient to accurately infer $\theta \geq 10^{-3}$ and pf 1x for $\theta = 10^{-4}$. Notably, even lower average coverages were sufficient when data was available for 10Mb. Finally,we found an average coverage of 4x to be sufficient when conducting recalibration and inference in windows as small as 0.1Mb. 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 or ultra conserved 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 using the exponential model introduced here, then used the first 20Mb of the X chromosome to perform recalibration individually for each read group, taking the inferred PMD pattern into account. Finally, we used both the inferred PMD patterns as well as the recalibrated quality scores to infer θ in windows of 1Mb in the whole genome, excluding windows closer than 5Mb 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 Palaeolithic genome, "Bichon" from Grotte du Bichon, Switzerland (Bich), approximately 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 years old male from Mota Cave in the southern Ethiopian highlands (Mota Gallego Llorente and Jones 2015). All these samples had relatively high coverage (>10x) and thus allowed us to infer fine scaled patterns of heterozygous 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 downlaoded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/). These were the British males HG00115, HG00116 and HG00117, the Tuscany males NA20509, NA20511 and NA20762, and the Yoruban males NA18486, NA18519 and NA18522. As shown in Fig. 5A, these nine individuals portray 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 2 out of 3 possible comparisons in each population). Larger differences in overall diversity was 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 diversity 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 thus suggest that while hunter-gatherer populations had much lower diversity, the diversity found in Europeans about 3,000 years ago was very comparable to the diversity observed today. This conclusion is in perfect agreement with with a temporal trend in the total length of runs of homozygoisty (ROH) inferred from imputed genotypes among ancient samples from Hungary that spanned a period from 5,700 - 1,000 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

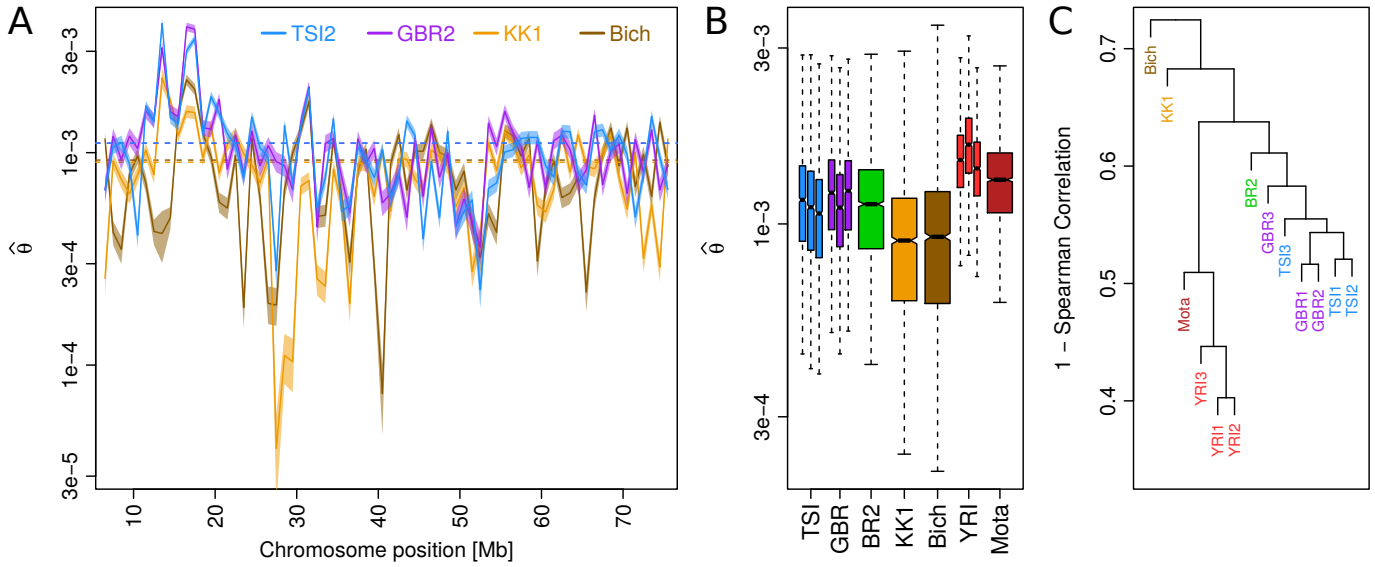

Figure 5 Local diversity in ancient and modern humans. A) Heterozygosity (θ) inferred in 1Mb windows along the first 75 Mb of chromosome 1 (excluding windows closer than 5Mb of the telomere) for two modern Europeans (TSI2 and GBR2) and two ancient European hunter-gatherers (KK1 and Bich). Solid lines indicate the MLE estimate, shades indicated the 95% confidence intervals and dashed lines the genome-wide median for each sample. B) Distribution of estimates $\hat{\theta}$ in 1Mb 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.

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 (Spearman correlations between 0.55 - 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 (Spearman correlations between 0.42 - 0.48) and their pair-wise 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 (Fig. 5B).

The lowest pair-wise correlations (0.28 - 0.39) were found in comparisons involving the two European hunter-gatherer samples KK1 and Bich, with the overall lowest being the correlation between these samples (0.28). This is also illustrated visually when plotting our estimates of the first 75Mb 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 (Fig. 5C). 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 samples in Switzerland and the other in 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.* 2015) and Yamnaya steppe herders (Haak *et al.* 2015) contributing the majority of the modern day genetic make up. 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, compared to 0.00 - 0.03% in all modern Europeans), 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 5x 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 post-mortem damage (PMD), a particular feature of ancient DNA. We then showed that this allows for unbiased estimates a much lower coverage than the original estimator, which was found to be ill-behaved

at coverages below 4x (Lynch 2008).

A downside of our approach is that it assumes sequencing errors to be known, while the original approach estimates the rate of sequencing errors jointly with heterozygosity. This, however, allows us to relax the assumption of constant error across all reads and to benefit from the 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. DePristo *et al.* 2011; McKenna *et al.* 2010; 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 1Mb, we are confident that it will work even on ultra conserved 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 here developed methods 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 two methods to infer these probability functions from raw data that are 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 1Mb 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.

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. Proceedings of the National Academy of Sciences **104**: 14616–14621.

Bryc, K., N. Patterson, and D. Reich, 2013 A novel approach to estimating heterozygosity from low-coverage genome sequence. Genetics **195**: 553–61.

Cabanski, C. R., K. Cavin, C. Bizon, M. D. Wilkerson, J. S. Parker, K. C. Wilhelmsen, C. M. Perou, J. Marron, and D. Hayes, 2012 ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. BMC Bioinformatics **13**: 221.

DePristo, M. a., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. a. Philippakis, G. del Angel, M. a. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics **43**: 491–8.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution **17**: 368–376.

Gallego Llorente, M. and E. R. Jones, 2015 Ancient Ethiopian Genome Reveals Extensive Eurasian Admixture Throughout the African Continent. Sciencexpress **350**: 1—-7.

Gamba, C., E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T. F. G. Higham, M. Hofreiter, D. G. Bradley, and R. Pinhasi, 2014 Genome flux and stasis in a five millennium transect of European prehistory. Nature Communications **5**: 5257.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. a. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich, 2015 Massive migration from the steppe was a source for Indo-European languages in Europe. Nature .

Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega, A. G. Clark, and R. Nielsen, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome research **18**: 1020–9.

Hofmanová, Z., S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez del Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, K. Kirsanow, L. M. Cassidy, R. Martiniano, M. Strobel, A. Scheu, K. Kotsakis, P. Halstead, S. Triantaphyllou, N. Kyparissi-Apostolika, D.-C. Urem-Kotsou, C. Ziota, F. Adaktylou, S. Gopalan, D. M. Bobo, L. Winkelbach, J. Blöcher, M. Unterländer, C. Leuenberger, Ç. Çilingiroğlu, B. Horejs, F. Gerritsen, S. Shennan, D. G. Bradley, M. Currat, K. Veeramah, D. Wegmann, M. G. Thomas, C. Papageorgopoulou, and J. Burger, 2015 Early farmers from across europe directly descended from neolithic aegeans. bioRxiv .

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 pp. 1320–1327.

Jones, E. R., G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R. L. Mc Laughlin, M. G. Llorente, L. M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Muller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T. F. G. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, and D. G. Bradley, 2015 Upper palaeolithic genomes reveal deep roots of modern eurasians. Nat. Comm. pp. 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 (Oxford, England) **29**: 1682–4.

Karolchik, D., R. M. Kuhn, R. Baertsch, G. P. Barber, H. Claw-

son, M. Diekhans, B. Giardine, R. a. Harte, a. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, a. S. Zweig, D. Haussler, and W. J. Kent, 2008 The UCSC Genome Browser Database: 2008 update. Nucleic acids research **36**: D773–9.

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, N. Homer, G. Marth, G. Abecasis, and R. Durbin, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25**: 2078–9.

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. Molecular Biology and Evolution **25**: 2409–2419.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, 2010 The {Genome} {Analysis} {Toolkit}: {A} {MapReduce} framework for analyzing next-generation {DNA} sequencing data. Genome Research **20**: 1297–1303.

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

Skoglund, P., B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson, 2014 Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proceedings of the National Academy of Sciences **111**: 2229–34.

Appendix

Emission probabilities in the presence of post-mortem damage

Following Lynch (2008) and commonly used approaches (Li 2011, e.g.), 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 post-mortem damage (PMD), the probability of observing a base d_{ij} given the underlying genotype $g_i = kl$ is then given by

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

In ancient DNA, differences between the base observed within a read and the underlying alleles may also be the result of PMD. 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

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

Implementation of the EM algorithm to infer θ

Here we present an efficient implementation of the algorithm to infer θ for a genomic region containing I sites:

1. Calculate the matrix of emission probabilities $\mathbb{P}(d_i|g_i)$ for all positions and genotypes according to the formulas given in the Appendix.
2. Estimate the initial base frequencies π from the base frequencies among all reads in the window.
3. Set the initial θ to the genome-wide expectation.
4. Run the EM algorithm by repeating the following steps until convergence is reached:

(a) Set $\theta' = \theta$, $\boldsymbol{\pi}' = \boldsymbol{\pi}$, $\rho' = e^{-\theta'}/(1 - e^{-\theta'})$ and $\mu' = 0$.

(b) Calculate substitution probabilities $\mathbb{P}(g|\theta', \boldsymbol{\pi}')$ for all genotypes g according to eq. 1 using the current estimates of θ' and $\boldsymbol{\pi}'$.

(c) Calculate $P_g = \sum_{i=1}^{I} \mathbb{P}(g|d_i; \theta', \boldsymbol{\pi}')$ for all genotypes g according to eq. 3.

(d) Find the new estimates of the parameters θ and $\boldsymbol{\pi}$ using the Newton-Ralphson method by setting $\mathbf{x} = \{\boldsymbol{\pi}', \rho', \mu'\}$ and repeating the following steps until convergence:

   i. Calculate vector F according to eq. 4.

   ii. Calculate $\mathbf{J}^{-1}$ according to eq. 6.

   iii. Update $\mathbf{x}$ according to eq. 7.

(e) estimate new parameter estimates as $\pi_k = x_k$ and $\theta = -\log(x_5/(1 + x_5))$.