HAL
open science

Combining contemporary and ancient DNA in population genetic and phylogeographic studies

Miguel Navascués, Frantz Depaulis, Brent C. Emerson

To cite this version:

Miguel Navascués, Frantz Depaulis, Brent C. Emerson. Combining contemporary and ancient DNA in population genetic and phylogeographic studies. Molecular Ecology Resources, 2010, 10, pp.760-772. hal-00584160

HAL Id: hal-00584160

https://hal.science/hal-00584160v1

Submitted on 2 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining contemporary and ancient DNA in population genetic and phylogeographic studies

Miguel Navascués[*†], Frantz Depaulis[†] and Brent C. Emerson[‡]
* INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France.
† Laboratoire Ecologie et Evolution, CNRS UMR 7625, UPMC Paris Universitas, Ecole Normale Supérieure, Paris, France.
‡ Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom.

Key words: phylogeography, serial coalescence, heterochronous data, ancient DNA.

Corresponding author: Brent Emerson, Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom ph: 0044 (0)1603 592237. fax: 0044(0)1603 592250, e-mail: b.emerson@uea.ac.uk

Running head: Molecular polymorphism analysis of ancient DNA

The definitive version is available at www.blackwell-synergy.com

Non-standard abbreviations:
aDNA:ancient DNA; WF: Wright Fisher model; IMSM: infinitely many sites (mutational) model; MRCA: Most Recent Common Ancestor; MCMC: Markov chain Monte Carlo; ABC: approximate Bayesian computation.

Abstract

The analysis of ancient DNA in a population genetic or phylogeographic framework is an emerging field, as traditional analytical tools were largely developed for the purpose of analyzing data sampled from a single time point. The analysis of heterochronous sequence data from closed panmictic populations has received attention with Markov chain Monte Carlo (MCMC) approaches, but attributing genetic differences between temporal samples to mutational events between time points requires the consideration of other factors that may also result in genetic differentiation. Geographic effects are an obvious factor for species exhibiting geographic structuring of genetic variation, and departures such as this from a closed panmictic model require researchers to either exploit software developed for the analysis of isochronous data, take advantage of simulation approaches using algorithms developed for heterochronous data, or explore approximate Bayesian computation. Here we review statistical approaches employed and available software for the joint analysis of ancient and modern DNA, and where appropriate we suggest how these may be further developed.

Introduction

Non-contemporaneous, or ancient DNA, is providing biologists with new and exciting opportunities to investigate evolutionary pattern and process over a range of temporal scales, from decades (e.g. Harper *et al.* 2006; Martinez-Cruz *et al.* 2007) to hundreds of thousands of years (e.g. Willerslev *et al.* 2007). For the purposes of this paper we define ancient DNA (hereafter aDNA) as DNA recovered from non-ideal biological material – that is to say material that was not preserved or maintained in a manner typically associated with downstream DNA analysis, for which the host organism is no longer alive. Such material includes subfossil remains (typically bones and teeth), archaeological remains, coproliths, mummies, naturally (i.e. not laboratory) frozen remains, ice cores, sediments, museum and herbarium tissues. DNA extracted from such material is typically low quality, with constraints upon both the amount and integrity of the DNA that can be obtained. Nevertheless, there is an increasing volume of studies successfully obtaining samples of aDNA, and frequently these data are being analyzed in a temporal and geographic context (see Ramakrishnan & Hadly 2009 for a review). Thus aDNA is becoming increasingly accessible for both population genetic and phylogeographic analysis, offering the possibility of using temporal samples of DNA to characterize population history. However, the analysis of such data is an emerging field, as traditional population genetic and phylogeographic tools were largely developed for the purpose of analyzing data sampled from a single time point, or at the most a sampling interval that spanned no more than a few generations. In this paper we review some of the analytical approaches employed for the joint analysis of ancient and modern DNA, and offer some suggestions for future directions. While we focus on studies utilizing aDNA for temporal sampling, it is important to point out that temporal sampling is not restricted to aDNA. Organisms with short generation times such as viruses offer similar opportunities (e.g. Norja *et al.* 2008; Shackleton *et al.* 2006), but without constraints on DNA quality.

Perhaps unsurprisingly, the most widely employed marker for phylogeographic and population genetic analyses with temporal sampling over more distant time scales (from hundreds, to thousands of years) is the mitochondrial genome, in particular the fast evolving control region (e.g. Barnett *et al.* 2009; Lambert *et al.* 2002; Shapiro *et al.* 2004; Valdiosera *et al.* 2007; Valdiosera *et al.* 2008). Given recent evidence that mtDNA is not a reliable marker for demography (Bazin *et al.* 2006), nuclear data would be a welcome addition, and next generation sequencing technologies offer some promise in this direction. Over more recent time scales of decades, microsatellite markers gain greater prominence, where their high variability provide more potential to reveal demographic events over narrower time scales (e.g. Harper *et al.* 2006; Martinez-Cruz *et al.* 2007). However, the utility of microsatellites is not restricted to decadal analyses, and recent efforts have seen the combination of nuclear microsatellites with mtDNA to investigate demographic changes spanning several thousands of years (Keyser-Tracqui *et al.* 2006). Single nucleotide polymorphism (SNP) data offers similar potential and has been used to assess genetic change over a 4,000 year period for cattle (Svensson *et al.* 2007).

As in classical population genetics, the choice of genetic marker for studies incorporating aDNA should be appropriate for the question(s) to be addressed, and in common to all genetic markers are the fundamental population genetic processes that

may shape differentiation between heterochronous samples. Population samples of DNA from different time points may vary from subtle changes in allele frequency through to allele loss or gain, due to the processes of mutation, selection, genetic drift and migration. Recent interest has focused on sampling DNA sequences from "measurably evolving" populations (Drummond *et al.* 2003) for which there are sufficiently long or numerous sampled sequences and a fast mutation rate relative to the available range of sequence sampling times. Sequences from such populations have the potential to enable the analysis of temporal changes in the size, structure and substitution rates of populations. Problems have been pointed out about the reliability of aDNA for the estimation of population history, as postmortem DNA damage may act to inflate diversity estimates (Axelsson *et al.* 2008). Distinguishing what is a genuine mutation and what is the consequence of postmortem degradation is thus important for the reliable estimation of population history. A conservative approach is to remove all sites that might represent false, or post mortem induced degradation. Applying this to a data set of steppe bison (*Bison bison*/*Bison priscus*) previously analyzed by Shapiro *et al*. (2004) results in the removal of signal for population size change observed in the original analyses (Axelsson *et al.* 2008). However, it has been argued that this approach is too conservative, and a subsequent reanalysis implementing a model where sequence variation is the result of a joint process of mutation and postmortem DNA damage is consistent with previous conclusions of population size change over time (Rambaut *et al.* 2009). Clearly there is a need to accommodate mutational artifacts when analyzing aDNA in a population genetic and phylogeographic framework, and recent methods developed to predict errors due to postmortem degradation are a welcome addition (Mateiu & Rannala 2008).

Phylogeographic analysis of temporally sampled DNA sequences provides for the direct quantification of population turnover within species, with particular reference to climatically mediated regional extinction and recolonization (Benton & Emerson 2007; Stewart *et al.* 2009). This is perhaps the most powerful contribution of aDNA to the field of phylogeography, with recent studies suggesting regional population extinction and subsequent recolonization (Barnes *et al.* 2002; Barnett *et al.* 2009; Hofreiter *et al.* 2007; Leonard *et al.* 2007). Temporally sampled DNA sequences also provide the potential for the estimation of substitution rates and divergence times without paleontological calibrations, and there has been much recent interest in the idea that substitution rates within species may be much higher than previously thought (Ho *et al.* 2007; Ho & Larson 2006; Ho *et al.* 2005; Ho *et al.* 2008; Penny 2005). Bayesian Markov chain Monte Carlo (MCMC) analyses of ancient data sets has consistently generated substitution rate estimates exceeding those from the literature. While this acceleration of molecular rates may partially be explained by evolutionary processes (Ho *et al.* 2005) it has been shown that such rate estimates can be biased due to the low information content of aDNA data (Debruyne & Poinar 2009) or demographic model misspecification (Navascués & Emerson 2009). These caveats should be taken into account when estimating molecular rates and in their use to date historical events.

The concerns raised by Axelsson *et al*. (2008) and the results of Navascués & Emerson (2009) highlight the need for caution when analyzing aDNA in a population genetic or phylogeographic context. Both studies reveal that variation among DNA sequences that originates from processes not accounted for within an underlying evolutionary model may lead to incorrect conclusions regarding evolutionary history.

Axelsson *et al*. (2008) demonstrate that when nucleotide differences between temporal samples arise from postmortem DNA damage they can misleadingly contribute to demographic inference. Navascués & Emerson (2009) demonstrate that when nucleotide differences between temporal samples arise not from mutation, but from other population genetic processes such as genetic drift, immigration or selection, they can bias divergence and substitution rate estimation. The take home message is model violation should be given due consideration as a potential explanation for the result of a model-based analysis. This is not something unique to population genetic/phylogeographic analyses incorporating aDNA (e.g. Becquet & Przeworski 2009; Strasburg & Rieseberg 2009).

New DNA sequencing technologies promise to deliver greater amounts of aDNA sequence information for a greater number of taxa (e.g. Allentoft *et al.* 2009; Briggs *et al.* 2009). While the majority of studies to date have relied on mtDNA sequence variation, nuclear microsatellites and nuclear coding sequence variation is also within reach (e.g. Keyser-Tracqui *et al.* 2006; Krause *et al.* 2007; Lalueza-Fox *et al.* 2009; Lalueza-Fox *et al.* 2008). It is thus both timely and appropriate to evaluate how best to combine ancient DNA with contemporary DNA in a population genetic and phylogeographic context. What follows is a review of the statistical tools available, and suggestions for their future development.

Testing for heterochrony

For the analysis of heterochronous data, one may, as a first step, test if the data show measurable evolution. If they do not, data may be pooled to proceed with further analyses using available tools for isochronous data. PATH-O-GENE (Table 1), is a simple tool that may be used with an input tree, typically a phylogeny reconstructed without assuming a molecular clock, to compute the correlation between tip to root distances and sampling times. This provides associated estimates of the mutation rate and a dated root (the slope and abscissa intercept of the regression) and an output tree assuming a molecular clock, but taking into account sampling times. It should however be noted that rigorous testing of such a correlation would require an additional randomization procedure, since the data do not obey parametric assumptions.

To take into account the potentially confounding effects of factors such as geography (genetic differentiation resulting from spatial rather than temporal separation) it is possible to partition genetic variation through the AMOVA (Excoffier *et al.* 1992) implemented in ARLEQUIN (Table 1). Although originally intended to assess geographic population structure, AMOVA can also be applied to heterochronous data partitioned into sampling times, or time bins when pooling is appropriate. This analysis tests the sampling time effect and estimates the proportion of genetic variation explained by sampling times. Formally, this is equivalent to testing F_{st} between sampling times. Indeed, the F_{st}, commonly used to describe differentiation among populations, is defined as a fixation index and can be seen as a measure of temporal divergence of populations by drift and not only as a (static) distance between equilibrium populations. The AMOVA implemented in Arlequin has been applied to several aDNA data sets (Bramanti *et al.* 2009; Dalen *et al.* 2007; Malmström *et al.* 2009), although these analyses did not allow for the distinction of population effects

from temporal ones. In a different approach, Valdiosera *et al*. (2008) applied the analyses separately on different geological layers to assess geographical structure through time, though the confounding effect of sampling time may not be entirely overcome since some layers may be more heterochronous than others (e.g. Pleistocene *vs* modern). It should be noted that the AMOVA framework was originally designed for contemporaneous (modern) data and presents several limitations for the analysis of heterochronous data within the ARLEQUIN implementation. Ideally factors should be nested, and including both sampling time and geographic location in an analysis would be more appropriate if a set of sampling times (or time bins) is found only in a specific location (or conversely if samples from one population all belong to one chronological layer). An additional constraint is that the available implementation of AMOVA does not quantitatively take into account sampling time differences. Excoffier (2007) notes that a population delimitation issue arises prior to the application of AMOVA and refers to associated attempts through aggregation techniques. This issue of population delimitation also applies to time bins. The qualitative feature of the analyses makes it strictly applicable only for pairwise analyses between two sampling time bins. It would be appropriate for some studies of modern *vs* ancient data if the latter do not span a substantial time range, but frequently this is not the case (e.g. Barnes *et al.* 2002; Lambert *et al.* 2002; Shapiro *et al.* 2004). More quantitatively, it is possible to correlate the pairwise genetic distance with a pairwise time difference through the Mantel test or to both time difference and geographical distance using a partial Mantel test using IBD (Table 1). Finally, there is a generalization of AMOVA (GAMOVA) where sampling time can potentially be treated as a quantitative explanatory cofactor together with other factors (Nievergelt *et al.* 2007), although the procedure has not been applied for this purpose yet.

Beyond these statistical tools to test for heterochrony and assess the proportion of genetic variation explained by sampling time differences, parametric (model based) statistics offer a powerful complement. Liu and Fu (2007) proposed two classes of standardized summary statistics aimed at detecting measurable evolution between two subsets at different times: the first one, D_c, is related to a Nei's (1987) net distance D_a (see Fig. 1). The other class of statistic, T_c, quantifies the excess of polymorphism exclusive to one of the two subsets when compared to the isochronous case (i.e. the null hypothesis being tested). The tests proceed through either (*i*) parametric simulation if it is desirable to take into account the stochasticity of the evolutionary process, or (*ii*) permutation between the subsets, which may partly lessen confounding historical effects and would be appropriate to test if samples could be pooled for subsequent analyses. Both approaches should in general be complementary, the former being generally more robust. Such statistics are largely inspired from coalescent theory (Kingman 1982) which is discussed in the next section, in particular its extension to heterochronous data.

Making inferences under the standard neutral model

The serial coalescent

Coalescent theory models the genetic history of a sample by probabilistic genealogies where the nodes represent most recent common ancestors and where the lengths of the branches - or lineages - are proportional to time (see Wakeley 2008 for a review). The

serial (heterochronous) coalescent is a simple extension of the isochronous standard case (Rodrigo & Felsenstein 1999). Heterochronous data can be described as a list of subsets, each defined by a sampling time and the associated subset of sequences. Events occurring on such a genealogy (e.g. coalescence, mutation, recombination, migration) are limited to lineages that co-occur at the same time point. Thus, at a given time, the number of extant lineages governs the rate of those events. Analytical derivations on the serial coalescent may help to predict the heterochrony effect or to interpret data, to suggest new statistics aimed at investigating heterochrony (see previous section) and to provide necessary corrections for most available statistics so as to enable comparison of data sets with different time sampling schemes.

Fu (2001) has proposed a diversity-based estimator of mutation rates based on serial samples; alternatively this method can be used to estimate sample age when an independent estimate of the mutation rate is available. This approach was subsequently refined by Liu and Fu (2008). For simplicity Liu and Fu (2008) focused on a two sampling time case for both the standard neutral model of Wright-Fisher (WF; Wright 1931), and a model involving any deterministic population size change starting after the first sampling time. From this approach mean, variance and covariance is derived for the number of mutations affecting *i* individuals in a temporally older sampled subset and *j* individuals in a more recent sampling. Although the derivation relates to mutation counts, or segregating sites in the infinitely many site model (IMSM; Watterson 1975), more general cases with other mutational models can be estimated with corrected distances. From these counts commonly used summary statistics are derived (see Fig. 1 for examples) such as the total number of mutations in the sample, and hence the related Watterson's (1975) θ_W estimator of polymorphism as well as Tajima's (1983) π diversity estimator.

Testing the standard model

Neutrality test statistics (see Nielsen 2005 for a review) can potentially be corrected for heterochrony within a general time sampling scheme with an arbitrary number of sampling times. Assuming WF and IMSM, Forsberg *et al.* (2005) derived the full probability distribution for the time to the most recent common ancestor of the total data set, the total length of the genealogy and the length of lineages exclusively ancestral to ancient samples (see lineages marked with dotted lines in Fig. 1). Such time quantities (time to MRCA, length of genealogy, length of lineages exclusively ancestral to ancient samples) can be related to the number of segregating sites, the number of mutations affecting only ancient samples (note that this may include some possible fixed differences between ancient and modern samples) and the mutations shared between ancient and modern samples. Thus, it is possible to perform the numerical computation of the probabilities to obtain a given summary statistic value in order to perform a neutrality test.

Although the analytical derivation of means and variances of summary statistics is typically feasible for simple models, deriving their full distribution (as in the work described in the previous paragraph) is generally more challenging. Simulation approaches provide a sensible alternative to test evolutionary models or for parameter inference. Coalescent simulations are commonly used to empirically investigate the distribution of polymorphisms under flexible evolutionary scenarios, to test those scenarios through comparison with data, or make inferences for associated parameters

in a more refined approach such as MCMC and ABC (see latter section). Simulations also allow for flexibility of model specification and provide efficient simulation schemes because one need only consider explicitly the history of a sample, and only at times where events modify that history, such as common ancestry, mutation, recombination or migration. Simulation within the serial coalescent algorithm is a straightforward minor modification of the contemporaneous case (Achaz *et al.* 2004; Anderson *et al.* 2005; Depaulis *et al.* 2009) and provides a powerful tool to assess heterochrony effects, test evolutionary scenarios, design appropriate sampling schemes, and investigate the statistical properties of available methods.

In the context of assessing the effects of heterochrony on neutrality test statistics, coalescent simulation has recently been used to reveal that heterochrony can introduce substantial bias to parameter estimation from summary statistics (Depaulis *et al.* 2009). Heterochrony increases coalescence times by increasing the difference between sampling times, since the ancestral lineage of the youngest individual must first reach the sampling point of the ancient one before the two lineages can coalesce. This leads to lengthening of a genealogy, and the overestimation of polymorphism with classical estimators such as diversity π (Fig. 1), indicating that direct comparison of polymorphism between data sets with sampled from different time points is inappropriate. Using a large aDNA data set for cave bears, heterochrony has been shown to strongly influence the conclusions of several neutrality tests (Depaulis *et al.* 2009). A straightforward correction factor can however be easily implemented for the diversity estimator and between population distances (Depaulis *et al.* 2009). For moderate heterochrony (sampling time differences below N_e generations), a likely scenario for most available aDNA data, terminal branches (i.e. directly leading to sampled tips) tend to be proportionally more affected. This leads to star like trees, an excess of mutations with low frequency in the sample, as revealed by negative Tajima's *D*, and a deficit of associations between mutations (linkage disequilibrium statistics, being sensitive to mutation frequency and the shape of the tree, provide information even in the absence of recombination). The result of this is the spurious mimicry of typical signatures of demographic processes such as population expansion. Both Achaz *et al.* (2004) and Depaulis *et al.* (2009) have shown through simulation that heterochrony can lead to substantial genetic differentiation between sampling points or, more generally, between data sets with different time sampling schemes. A useful consequence of this is that such differentiation may be used to estimate the rate of drift and thus, the effective population size (Achaz *et al.* 2004).

Making inferences under complex models

Likelihood and Bayesian inference using MCMC genealogy sampling

The likelihood of a given parameter value is a function proportional to the probability of the data given the parameter value, $L(\theta|D)=\alpha P(D|\theta)$ (we will ignore the constant of proportionality α from this point). This function is used in statistical modelling to make inferences about the parameter values of a model. One approach to make inferences is to quantify the parameter value that maximizes the likelihood function (maximum likelihood estimate) and use the likelihood function profile around that value to determine the confidence intervals for the estimate. A Bayesian alternative approach is to combine the likelihood with some prior probability (information or

belief prior to the experiment) to obtain the posterior probability distribution, and use this distribution to obtain a point estimate and credible intervals of the model parameter (see Beaumont & Rannala 2004 for further details).

It is possible to calculate the probability of a genealogy given a demographic model, $P(G|\theta)$, by using the coalescent as a model (Felsenstein 1992). Additionally, calculating the probability of the data (genetic state of a sample) given a genealogy, $P(D|G)$ has been established from the field of phylogenetics (Felsenstein 1981). The genealogy of the sample is unknown; thus, in order to derive the likelihood, it is necessary to integrate over all possible genealogies, $L(\theta \mid D) = \sum_G P(D \mid G) P(G \mid \theta)$. This will be the likelihood of full data. In practice, only a selection of random genealogies will be used in this integration, because the number of possible genealogies is infinite (a large number of topologies and an infinite combination of branch lengths). However, most of the possible genealogies will be very unlikely and will contribute little to the estimation of the likelihood. Therefore, it is desirable to sample the genealogies in an efficient way, i.e. favouring plausible genealogies. Currently, the most popular framework to sample genealogies efficiently is the Markov chain Monte Carlo (MCMC) approach, in an implementation originally proposed by Kuhner *et al*. (1995). This method consists of a random walk over the space of the model parameter values and over possible genealogies, and it is designed in a way that it should visit each point of that space proportionally to its likelihood (see Wakeley 2008 for further details and a review of various implementations). This approach has been implemented in several software packages targeting different demographic scenarios (see Kuhner 2009 for a review); among them only BEAST (Table 1) allows for the sampling of genealogies of genes collected at different time points (Drummond *et al.* 2002) or a model of DNA damage for aDNA (Rambaut *et al.* 2009).

The initial interest in developing BEAST was to take advantage of heterochronous data to estimate mutation rates (Drummond *et al.* 2002). In the case of isochronous data, substitution rates are frequently estimated from a phylogeny using fossil or geological data to calibrate node dates. For heterochronous data, it is possible to estimate the number of mutations occurring in the time interval between samples (and, thus, the mutation rate) without using any external data to calibrate the genealogy (see Drummond *et al.* 2003 for further details). This potentially allows one to obtain species-specific or population-specific molecular rate estimates that cannot usually be obtained from phylogenies because of the difficulty of assigning fossils to a particular lineage below the genus level. The MCMC method to estimate the molecular rate implemented in BEAST was first used in the analysis of Adélie penguin (*Pygoscelis adeliae*) modern and ancient DNA samples (Lambert *et al.* 2002) and has subsequently been applied to a number of other aDNA data sets (e.g. Ho *et al.* 2007), usually providing rate estimates higher than those obtained from phylogenies (this has lead to the debate mentioned in the introduction).

Analysis of complex historical change in population size can be addressed in BEAST using the Bayesian skyline plot framework (Drummond *et al.* 2005). This method is based on the classical skyline plot analysis (Pybus *et al.* 2000), in which population size is estimated from consecutive time intervals separated by coalescent events from a given genealogy (in practice, a maximum likelihood estimate of the genealogy). An

394 estimate of population size is obtained for each time interval from its length, based on
395 the expected coalescence time for the number of lineages present at in that time
396 interval. The graphical representation of this piecewise demographic model yields a
397 plot that evokes the skyline of a city (Fig. 2b). The Bayesian skyline plot is an
398 extension of this approach with the following modification: (i) time intervals of the
399 piecewise demographic model do not need to be defined by consecutive coalescent
400 events (generalised skyline plot by Strimmer & Pybus 2001); (ii) a prior is used to
401 make population sizes of consecutive time intervals correlated; and (iii) the
402 uncertainty of the genealogy is taken into account with the MCMC algorithm for
403 sampling genealogies (Drummond *et al.* 2005). Although the model parameters are
404 the population sizes at each interval, and the times defining the intervals, posterior
405 probabilities for them are ignored in the output of BEAST. Instead, a plot of the
406 estimated population size and the associated credibility interval as a function of time
407 is presented (Fig. 2c), which is more informative about the past demographic history
408 than the model parameter values. The analysis of heterochronous steppe bison
409 mtDNA data remains an enlightening example of this analysis and its development
410 (Drummond *et al.* 2005; Rambaut *et al.* 2009; Shapiro *et al.* 2004).
411
412 The coalescent model implemented within BEAST was (until recently) that of a single
413 population, and, thus, did not allow one to address the study of spatially structured
414 populations. Potential problems of imposing spatially structured heterochronous data
415 onto such a model for the estimation of mutation rate have been pointed out
416 (Navascués & Emerson 2009), and although it has not been evaluated, similar issues
417 may exist for demographic inference. However, two recent developments have
418 opened the door to phylogeographical analyses with BEAST. First, Lemey *et al.*
419 (2009) have introduced a Bayesian modelling of character evolution for the inference
420 of ancestral states. Using geographical locations as character states, Lemey *et al.*
421 (2009) advocate the inference of posterior probabilities for the location (state) of
422 ancestral nodes and migration events (changes of state). This is a significant
423 improvement over classical phylogeographical studies (e.g. Hofreiter *et al.* 2004;
424 Leonard *et al.* 2000) for two reasons: (i) the uncertainty on the genealogy
425 reconstruction is taken into account (classical phylogeography is frequently based on
426 a single 'phylogeny') and (ii) there is a rigorous statistical interpretation of the results
427 (classical phylogeography is frequenctly reduced to a visual description of the
428 geographical distribution of lineages). Nevertheless, it must be noted that the use of
429 this model of character evolution does not change the model of the coalescent, which
430 remains constrained to that of a single panmictic population and it is not clear whether
431 this would be robust to some population structure scenarios. The second development
432 is the implementation of the multispecies coalescent within BEAST by Heled &
433 Drummond (2009). In this new method (named *BEAST) a set of species is
434 considered to diverge (by successive bifurcations, following a birth-death model)
435 from a common ancestral species. Divergence times and effective population sizes for
436 each contemporary and ancestral species (i.e. the 'species tree') determine the
437 coalescent probabilities used in the sampling of genealogies. The MCMC scheme
438 works by integrating over gene genealogies and over species tree topologies to obtain
439 estimates of the posterior probability distribution of the species tree (gene genealogies
440 have only the role of a 'nuisance parameter' and are ignored in the output). The
441 purpose of the method is to reconstruct the phylogeny from multiple loci. In
442 particular, it addresses the problem of incongruent gene genealogy topologies among
443 loci and with the species tree topology due to incomplete lineage sorting by using a

coalescent model. Although the method is focused on phylogeny reconstruction, it can be applied to a set of populations to determine the 'population tree' and use it to make phylogeographical inferences, i.e. inferring and dating the vicariance or colonization events responsible of the divergence of those populations.

Estimating the likelihood distribution from full data with the MCMC algorithm is in principle the best way to extract most of the information contained within the data. BEAST is currently the only implementation of this method for heterochronous data but unfortunately it only includes single population demographic models, divergence models without migration (*BEAST) and coalescent models without intragenic recombination. Regarding the lack of intragenic recombination in the coalescent model, this has not been considered a problem in the past, as most aDNA studies have targeted mitochondrial DNA (see Ramakrishnan & Hadly 2009 for a review). However, new sequencing technologies are a promising tool for the characterization of nuclear genetic diversity in aDNA (Millar *et al.* 2008) and future studies would likely benefit from the implementation of recombination. These features may be available in future versions or new programs of MCMC coalescent sampling, but efficiently incorporating recombination into MCMC presents many challenges – in particular likelihood surfaces become very rugged and difficult to explore. As an alternative researchers may use alternative methods for the estimation of the likelihood based on summary statistics rather than on full data.

Approximating the likelihood using summary statistics

Estimating the likelihood of full data, $L(\theta|D)$, is computationally intensive, the methods are difficult to implement in a computer program, and these difficulties increase with model complexity and the size of a given data set. Alternatively, it is possible to estimate the likelihood of a subset of the information within the data contained in summary statistics, S: $L(\theta|S)$. This approach takes advantage of the ease of simulating pseudo-samples using the coalescent model and computing summary statistics for those pseudo-samples. The general idea consist in simulating pseudo-samples under a range of parameter values and rejecting those simulations yielding summary statistics very different (using a predetermined threshold) to the summary statistics of the target sample. The proportion of accepted simulations is used as an approximation of the likelihood. This approximation of likelihoods is most frequently used in the Bayesian framework in what has been termed Approximate Bayesian Computation (ABC). Specifically, the model parameter values are taken from some prior distributions for performing the simulations. Simulations are rejected or accepted in the same way as described above. The distribution of parameter values from the accepted simulations is used as an approximation of the posterior distribution (e.g. see Pritchard *et al.* 1999). The estimates obtained from this method are highly dependent on the arbitrarily chosen threshold value for the rejection step: values that are too large will increase the error by accepting pseudo-samples far from real sample; values that are too small will require a prohibitively large number of simulations in order to accept enough simulations for the approximation of the posterior distribution. In order to address this problem, Beaumont *et al.* (2002) proposed an additional regression step after the rejection. In this step a linear regression of the parameter values as a function of the summary statistics is calculated from the accepted simulations. The parameter values are then adjusted using the regression and the posterior distribution is estimated from the adjusted values. This

adjustment is termed the regression algorithm and has been responsible for the recent success of ABC in population genetics because it allows accurate inferences with an affordable number of simulations (Fig. 3).

ABC methods have been little used in aDNA phylogeographic and population genetics. One of the reasons for this could be the lack of user-friendly software to perform all the necessary steps of the analysis (but see DIY ABC; Table 1). However, there are at least three coalescent simulators (SERIAL SIMCOAL, COMPASS and NETRECODON; Table 1) that perform coalescent simulations of heterochronous data. With a minimal knowledge of programming, they can be used for the simulation step of the ABC and their output can be analysed in R (with the functions available from Mark Beaumont or SERIAL SIMCOAL websites, see Table 1) for the rejection and regression steps. This was the procedure followed by Chan *et al.* (2006) to characterize a bottleneck for the rodent *Ctenomys sociabilis* from aDNA and by Ghirotto et al. (in press) to study the genetic continuity of bronze-age and modern human populations of Sardinia; both works using SERIAL SIMCOAL and R to perform the analyses. A simple example for using COMPASS and R for ABC analysis can be found as supplementary material (it will reproduce the analyses presented in Fig. 3).

With the demographic and mutational model flexibility of the currently available coalescent simulators, there is much potential in an ABC approach for the analysis of heterochronous DNA sequence data. This is particularly relevant for addressing models outside the scope of BEAST, such as structured populations with migration or genes with intragenic recombination, for which no other alternative is currently available. However, a note of caution is necessary. With so much flexibility one might be tempted to fit a complex model to data with low information content. As in any model-based inference, special care should to be taken when choosing the proposed model and the posterior and prior distributions should be compared. It is also advisable to evaluate only a few different models and use a procedure of model selection (such as that proposed with an R function by Beaumont (2006)) to estimate the posterior probability for each model.

Model testing and model selection

Estimating the parameters of a single demographic model will not suffice in most study cases for which several demographic models are plausible and should be considered. Thus, it is necessary to be able to test the models or measure the goodness of fit of the models to the data. Coalescent simulation may be used to test particular evolutionary scenarios within a range of parameter values of interest for heterochronous data. Several aDNA studies have made used of SERIAL SIMCOAL to test different demographic hypothesis: genealogic continuity between ancient and modern populations (e.g. Belle et al 2006 and Bramanti *et al.* 2009), demographic changes (e.g. Ramakrishan et al. 2005; Valdiosera *et al.* 2008) or population structure (Fabre et al 2009 (doi: 10.1371/journal.pone.0005151)). The general procedure consist in choosing a set of models, fixing a single value for each parameter of the model (or few combinations of values) and simulating a large number of pseudo-samples for each model. The testing of each model (for a fixed combination of parameter values) is performed by estimating p-values from the distributions of the summary statistic (see Guimaraes et al. 2009 (doi: 10.1093/molbev/msp126) for a test statistic combining the p-values estimated from different summary statistics). An

important drawback of this approach is that it requires a large number of simulations to estimate the p-value for each combination of parameter values considered, which severely limits the range of parameter values that can be explored. In order to select the model with the best fit to the data, Belle et al (2009, doi: 10.1038/hdy.2008.103) propose to rank the models by the number of summary statistic for which the estimated p-value is lower than the threshold for significance. Although this ranking gives an idea of the plausibility of the different models it is a poor measure of goodness of fit because some summary statistics are expected to be correlated and the 'measure' is not quantitative to the departure from the model. Alternatively, Ramakrishnan and Hadley (2009) propose to calculate the Akaike information criterion using a crude approximation of the likelihood based on the estimated p-values. Despite their past popularity, these methods suffer from serious limitations. It is our view that this kind of approach will be superseded by the ABC methods described in the previous section, as they allow to explore each model for a continuous range of parameter values (instead of a single or few fixed values) and offer more statistically rigorous procedures for model selection (see Ghirotto et al. in press) as an example for model selection procedures in ABC).

Concluding remarks

Ancient DNA is being incorporated in population genetic and phylogeographic analyses with increasing frequency. As such, careful consideration is required by researchers regarding available statistical approaches, and the appropriateness of these for the data to be analysed. Approaches are available to assess the influence of evolutionary change on sequences across sampling intervals, and to control for other variables, such as geographic location, that may contribute to genetic differentiation among samples of sequences. While these approaches currently present some limitations, it would seem that these could probably be overcome with further development of existing tools. The analyses available to researchers are perhaps less limited than they might first appear, but to increase flexibility researchers are required to explore options beyond stand alone analytical packages. For the analysis of single populations conforming to panmixia, a full-data likelihood approach as implemented in BEAST offers a versatile tool to make demographic inferences, and recent developments open the door for phylogeographic analysis with multiple population demographic inference (Heled & Drummond 2009; Lemey *et al.* 2009). For aDNA studies where sampled populations deviate from models available within BEAST (e.g. structured populations, gene flow, recombination), summary statistics and their approximation in a Bayesian framework provide for an alternative approach. However for all model-based analyses researchers are advised to take a cautious approach to the interpretation of their results by comparing posterior and prior parameter values, considering model violation, and the potential consequences of lack of information within heterochronously sampled data. We echo previous calls for the need for rigor in aDNA analysis (Cooper & Poinar 2000), but extend this call to the downstream analysis of this data for historical inference.

589 **References**

590 Achaz G, Palmer S, Kearney M, *et al.* (2004) A robust measure of HIV-1
591 population turnover within chronically infected individuals. *Mol Biol*
592 *Evol*, **21**, 1902-1912.
593 Allentoft ME, Schuster SC, Holdaway RN, *et al.* (2009) Identification of
594 microsatellites from an extinct moa species using high-throughput
595 (454) sequence data. *Biotechniques*, **46**, 195-200.
596 Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a
597 population genetics model for data from multiple populations and
598 points in time. *Bioinformatics*, **21**, 1733-1734.
599 Axelsson E, Willerslev E, Gilbert MTP, Nielsen R (2008) The effect of ancient
600 DNA damage on inferences of demographic histories. *Molecular Biology*
601 *and Evolution*, **25**, 2181-2187.
602 Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A (2002) Dynamics of
603 Pleistocene population extinctions in Beringian brown bears. *Science*,
604 **295**, 2267-2270.
605 Barnett R, Shapiro B, Barnes I, *et al.* (2009) Phylogeography of lions (Panthera
606 leo ssp.) reveals three distinct taxa and a late Pleistocene reduction in
607 genetic diversity. *Molecular Ecology*, **18**, 1668-1677.
608 Barton NH, Depaulis F, Etheridge AM (2002) Neutral evolution in spatially
609 continuous populations. *Theoretical Population Biology*, **61**, 31-48.
610 Bazin E, Glemin S, Galtier N (2006) Population size does not influence
611 mitochondrial genetic diversity in animals. *Science*, **312**, 570-572.
612 Beaumont M, Rannala B (2004) The Bayesian revolution in genetics. *Nature*
613 *Reviews Genetics*, **5**, 251-261.
614 Beaumont MA, Matsumura S, Forster P, Renfrew C (2006) Joint determination
615 of topology, divergence time, and immigration in population trees. In:
616 *Simulation, Genetics, and Human Prehistory*. McDonald Institute for
617 Archaeological Research, Cambridge.
618 Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian
619 computation in population genetics. *Genetics*, **162**, 2025–2035-2025–
620 2035.
621 Becquet C, Przeworski M (2009) Learning about modes of speciation by
622 computational approaches. *Evolution*, **63**, 2547-2562.
623 Benton MJ, Emerson BC (2007) How did life become so diverse? The
624 dynamics of diversification according to the fossil record and
625 molecular phylogenetics. *Palaeontology*, **50**, 1-19.
626 Bramanti B, Thomas MG, Haak W, *et al.* (2009) Genetic discontinuity between
627 local hunter-gatherers and central Europe's first farmers. *Science*, **326**,
628 137-140.
629 Briggs AW, Good JM, Green RE, *et al.* (2009) Targeted Retrieval and Analysis
630 of Five Neandertal mtDNA Genomes. *Science*, **325**, 318-321.

631 Chan YL, Anderson CNK, Hadly EA (2006) Bayesian estimation of the timing
632 and severity of a population bottleneck from ancient DNA. *PLoS*
633 *Genetics*, **2**, 451-460.
634 Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. *Science*,
635 **289**, 1139-1139.
636 Dalen L, Nystrom V, Valdiosera C, *et al.* (2007) Ancient DNA reveals lack of
637 postglacial habitat tracking in the arctic fox. *Proceedings of the National*
638 *Academy of Science, USA*, **104**, 6726-6729.
639 Debruyne R, Poinar HN (2009) Time dependency of molecular rates in ancient
640 DNA data sets, a sampling artifact? *Systematic Biology*, **58**, 348-359.
641 Depaulis F, Orlando L, Hänni C (2009) Using classical population genetic
642 tools with heterochronous data: time matters! *PLoS ONE*, **in press**.
643 Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating
644 mutation parameters, population history and genealogy
645 simultaneously from temporally spaced sequence data. *Genetics*, **161**,
646 1307-1320.
647 Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003)
648 Measurably evolving populations. *Trends in Ecology and Evolution*, **18**,
649 481-488.
650 Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent
651 inference of past population dynamics from molecular sequences.
652 *Molecular Biology and Evolution*, **22**, 1185-1192.
653 Epperson BK (1999) Gene Genealogies in Geographically Structured
654 Populations. *Genetics*, **152**, 797-806.
655 Excoffier L (2007) Analysis of population subdivision. In: *Handbook of*
656 *statistical genetics* (eds. Balding DJ, Bishop M, Cannings C). Wiley,
657 Chichester.
658 Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance
659 inferred from metric distances among DNA haplotypes: application to
660 human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.
661 Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum
662 likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376-368–376.
663 Felsenstein J (1992) Estimating effective population size from samples of
664 sequences: inefficiency of pairwise and segregating sites as compared
665 to phylogenetic estimates. *Genetics Research*, **59**, 139–147-139–147.
666 Forsberg R, Drummond AJ, Hein J (2005) Tree measures and the number of
667 segregating sites in time-structured population samples. *BMC Genetics*,
668 **6**, doi:10.1186/1471-2156-1186-1135.
669 Fu YX (2001) Estimating mutation rate and generation time from longitudinal
670 samples of DNA sequences. *Molecular Biology and Evolution*, **18**, 620-626.
671 Harper GL, Maclean N, Goulson D (2006) Analysis of museum specimens
672 suggests extreme genetic drift in the adonis blue butterfly

673 (Polyommatus bellargus). *Biological Journal of the Linnean Society*, **88**,
674 447-452.
675 Heled J, Drummond AJ (2009) Bayesian inference of species trees from
676 multilocus data. *Molecular Biology and Evolution*, **in press**.
677 Ho SYW, Kolokotronis S-O, Allaby RG (2007) Elevated substitution rates
678 estimated from ancient DNA sequences. *Biology Letters*, **3**.
679 Ho SYW, Larson G (2006) Molecular clocks: when times are a-changin'. *Trends*
680 *in Genetics*, **22**, 79-83.
681 Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of
682 molecular rate estimates and systematic overestimation of recent
683 divergence times. *Molecular Biology and Evolution*, **22**, 1561-1568.
684 Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B (2008) The effect of
685 inappropriate calibration: Three case studies in molecular ecology.
686 *PLoS ONE*, **3**, e1615.
687 Hofreiter M, Muenzel S, Conard NJ, *et al.* (2007) Sudden replacement of cave
688 bear mitochondrial DNA in the late Pleistocene. *Current Biology*, **17**,
689 R122-R123.
690 Hofreiter M, Rabeder G, Jaenicke-Despres V, *et al.* (2004) Evidence for
691 reproductive isolation between cave bear populations. *Curr Biol*, **14**, 40-
692 43.
693 Keyser-Tracqui C, Crubezy E, Pamzsav H, Varga T, Ludes B (2006)
694 Population origins in Mongolia: Genetic structure analysis of ancient
695 and modern DNA. *American Journal of Physical Anthropology*, **131**, 272-
696 281.
697 Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications*,
698 **13**, 235-248.
699 Krause J, Lalueza-Fox C, Orlando L, *et al.* (2007) The derived FOXP2 variant of
700 modern humans was shared with neandertals. *Current Biology*, **17**,
701 1908-1912.
702 Kuhner MK (2009) Coalescent genealogy samplers: windows into population
703 history. *Trends in Ecology & Evolution*, **24**, 86–93-86–93.
704 Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population
705 size and mutation rate from sequence data using Metropolis-Hastings
706 sampling. *Genetics*, **140**, 1421–1430-1421–1430.
707 Lalueza-Fox C, Gigli E, de la Rasilla M, Fortea J, Rosas A (2009) Bitter taste
708 perception in Neanderthals through the analysis of the *TAS2R38* gene.
709 *Biology Letters*, **doi: 10.1098/rsbl.2009.0532**.
710 Lalueza-Fox C, Gigli E, de la Rasilla M, *et al.* (2008) Genetic characterization of
711 the ABO blood group in Neandertals. *BMC Evolutionary Biology*, **8**.
712 Lambert DM, Ritchie PA, Millar CD, *et al.* (2002) Rates of evolution in ancient
713 DNA from Adelie penguins. *Science*, **295**, 2270-2273.

714 Lemey P, Rambaut A, Drummond A, Suchard MA (2009) Bayesian
715 phylogeography finds its roots. *PLoS Computational Biology*, **5**,
716 e1000520.
717 Leonard JA, Vila C, Fox-Dobbs K, *et al.* (2007) Megafaunal extinctions and the
718 disappearance of a specialized wolf ecomorph. *Current Biology*, **17**,
719 1146-1150.
720 Leonard JA, Wayne RK, Cooper A (2000) Population genetics of Ice Age
721 brown bears. *Proceedings of the National Academy of Sciences of the United*
722 *States of America*, **97**, 1651-1654.
723 Liu X, Fu YX (2007) Test of genetical isochronism for longitudinal samples of
724 DNA sequences. *Genetics*, **176**, 327-342.
725 Liu X, Fu YX (2008) Summary statistics of neutral mutations in longitudinal
726 DNA samples. *Theor Popul Biol*, **74**, 56-67.
727 Malmström H, Gilbert MT, Thomas MG, *et al.* (2009) Ancient DNA reveals
728 lack of continuity between neolithic hunter-gatherers and
729 contemporary Scandinavians. *Curr Biol*, **19**, 1758-1762.
730 Martinez-Cruz B, Godoy JA, Negro JJ (2007) Population fragmentation leads
731 to spatial and temporal genetic structure in the endangered Spanish
732 imperial eagle. *Molecular Ecology*, **16**, 477-486.
733 Mateiu LM, Rannala BH (2008) Bayesian inference of errors in ancient DNA
734 caused by postmortem degradation. *Molecular Biology and Evolution*, **25**,
735 1503-1511.
736 Millar CD, Huynen L, Subramanian S, Mohandesan E, Lambert DM (2008)
737 New developments in ancient genomics. *Trends in Ecology & Evolution*,
738 **23**, 386-393.
739 Navascués M, Emerson BC (2009) Elevated substitution rate estimates from
740 ancient DNA: model violation and bias of Bayesian methods. *Molecular*
741 *Ecology*, **18**, 4390-4397.
742 Nei M (1987) *Molecular evolutionary genetics* Columbia University Press, New
743 York.
744 Nievergelt CM, Libiger O, Schork NJ (2007) Generalized analysis of molecular
745 variance. *PLoS Genetics*, **3**, e51.
746 Norja P, Eis-Hubinger AM, Soderlund-Venermo M, Hedman K, Simmonds P
747 (2008) Rapid sequence change and geographical spread of human
748 parvovirus B19: Comparison of B19 virus evolution in acute and
749 persistent infections. *Journal of Virology*, **82**, 6427-6433.
750 Penny D (2005) Evolutionary biology - relativity for molecular clocks. *Nature*,
751 **436**, 183-184.
752 Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population
753 growth of human Y chromosomes: a study of Y chromosome
754 microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798-1791–
755 1798.

Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**, 1429–1437-1429–1437.

Ramakrishnan U, Hadly EA (2009) Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies. *Molecular Ecology*, **18**, 1310–1330-1310–1330.

Ramakrishnan U, Hadly EA, Mountain JL (2005) Detecting past population bottlenecks using temporal genetic data. *Molecular Ecology*, **14**, 2915-2922.

Rambaut A, Ho SYW, Drummond A, Shapiro B (2009) Accommodating the effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution*, **26**, 245-248.

Rodrigo AG, Felsenstein J (1999) Coalescent Approaches to HIV Population Genetics. In: *The Evolution of HIV* (ed. Crandall K). Johns Hopkins Univ. Press, Baltimore.

Shackleton LA, Rambaut A, Pybus OG, Holmes EC (2006) JC evolution and its association with human populations. *Journal of Virology*, **80**, 9928-9933.

Shapiro B, Drummond AJ, Rambaut A, *et al.* (2004) Rise and fall of the beringian steppe bison. *Science*, **306**, 1561-1565.

Stewart JR, Lister AM, Barnes I, Dalen L (2009) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B-Biological Sciences*, **doi:10.1098/rspb.2009.1272**.

Strasburg JL, Rieseberg LH (2009) How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution*, **in press**.

Strimmer K, Pybus OG (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, **18**, 2298–2305-2298–2305.

Svensson EM, Anderung C, Baubliene J, *et al.* (2007) Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Animal Genetics*, **38**, 378-383.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Valdiosera CE, Garcia N, Anderung C, *et al.* (2007) Staying out in the cold: glacial refugia and mitochondrial DNA phylogeography in ancient European brown bears. *Molecular Ecology*, **16**, 5140-5148.

Valdiosera CE, Garcia-Garitagoitia JL, Garcia N, *et al.* (2008) Surprising migration and population size dynamics in ancient Iberian brown bears (Ursus arctos). *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 5123-5128.

Wakeley J (2008) *Coalescent Theory: An Introduction* Roberts & Company Publishers.

Watterson GA (1975) On the number of segregation sites. *Theoret. Popul. Biol.*, **7**, 256-276.

Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539-1546.

Willerslev E, Cappellini E, Boomsma W, *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science*, **317**, 111-114.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

808 **Figure 1. The coalescent and summary statistics.** Three coalescent trees are
809 presented. Mutations are indicated in bold on the resulting sequence alignments
810 (right). There are S=4 segregating sites in all cases. Lineages exclusively ancestral to
811 ancient DNA samples are marked with a dotted line. **(**A): isochronous case, average
812 difference between sequences (total diversity) π=1.8. (B): moderate heterochrony,
813 mutation 1 is shared between ancient and modern sequences, mutation 2 is exclusive
814 to ancient sequences while mutation 3 and 4 are exclusive to modern sequences;
815 π=2.0; Nei's net distance: D_a=0. (C): large heterochrony, mutations 1, 2 and 3 lead to
816 fixed differences between modern and ancient sequences while mutation 4 is
817 exclusive to modern sequences; π=2.2, D_a=3. However, estimates corrected for
818 heterochrony (Depaulis *et al*. 2009) are: π_h=1.8 (A), π_h=1.54, D_{ah}=-1 (B) and π_h=1.16,
819 D_{ah}=2.33 (C).
820
821 **Figure 2. Skyline plot.** The classical skyline plot analysis uses the reconstructed
822 genealogy of the sample (a) to obtain estimates of the population size for each time
823 interval defined by consecutive coalescent events, resulting in a plot shape similar to
824 the skyline of a city (b). The Bayesian skyline plot (c) implemented in BEAST (Table
825 1) takes into account the uncertainty within the genealogy of the sample; the final
826 (smooth) plot represents the posterior probability density for the population size
827 (typically the median and 95% highest posterior density interval) calculated from the
828 skyline plots (in grey) for the genealogies sampled with the MCMC.
829
830 **Figure 3. Likelihood approximation through summary statistics.** A fictitious data
831 set of ten sequences (five mDNA and five aDNA sampled $0.2\times N$ generations ago)
832 containing 20 segregating sites is analysed. The objective of the analyses is to
833 estimate the parameter $\theta=2N\mu$ of a constant population size model and an IMSM of
834 mutation. Prior and posterior distributions (estimated with rejection and regression
835 algorithms) from an approximate Bayesian computation analysis are presented (50%
836 of simulations rejected from a total of 30 000 simulations; a particularly high rejection
837 threshold has been chosen for a better illustration of the improvement by the
838 regression algorithm). Simulations were performed with COMPASS (Table 1) under
839 the infinite site model and ABC was performed with the R functions of Mark
840 Beaumont (Table 1). See supplementary files for a script in R to reproduce the
841 analyses represented in this figure.

Table 1. Software discussed in the main text for the analysis of ancient DNA.

Program	Purpose	Method	Models	Web	Reference
ARLEQUIN	Test heterochrony & population structure	Nested AMOVA, Mantel test	**Null model:** no effect of time or population structure in genetic diversity	http://cmpg.unibe.ch/software/arlequin3/	Excoffier et al. (2005)
Multivariate Distance Matrix Regression	Test heterochrony & population structure	GAMOVA	**Null model:** no effect of time or population structure in genetic diversity	http://polymorphism.scripps.edu/~cabney/cgi-bin/mmr.cgi	Nievergelt et al. (2007)
PATH-O-GEN	Test heterochrony	Regression analysis		http://tree.bio.ed.ac.uk/software/pathogen/	n.a.
IBD	Test heterochrony & geographic structure	Partial Mantel test	Isolation by distance and by time	http://www.bio.sdsu.edu/pub/andy/IBD.html	Bohonak (2002)
BEAST	Demographic inference	MCMC	**Demography:** Flexible (single population) **Mutation:** Flexible (except SMM) **DNA Damage:** Yes (Rambaut et al. 2009) **Recombination:** No	http://beast.bio.ed.ac.uk/	Drummond and Rambaut (2007)
*BEAST	Coalescent-based phylogenetic inference	MCMC	**Demography:** Population divergence without migration **Mutation:** Flexible (except SMM) **DNA Damage:** Yes (Rambaut et al. 2009) **Recombination:** No	http://beast.bio.ed.ac.uk/	Heled and Drummond (2009)
GENIE	Demographic inference	Classical skyline plot	**Demography:** Flexible (single population) **Recombination:** No	http://evolve.zoo.ox.ac.uk/Evolve/Genie.html	Pybus and Rambaut (2002)
R scripts[1]	ABC		Models depend on external simulator	http://www.rubic.rdg.ac.uk/~mab/	Beaumont et al. (2002)
DIYABC	Demographic inference	ABC	**Demography:** Flexible (except migration) **Mutation:** SMM[2], K80, HKY, TN[3] (beta version) **DNA Damage:** No **Recombination:** No	http://www1.montpellier.inra.fr/CBGP/diyabc/	Cornuet et al. (2008)
BAYESIAN SERIAL SIMCOAL	Coalescent simulation, demographic inference	ABC	**Demography:** Flexible **Mutation:** SMM, K80[3] **DNA Damage:** No **Recombination:** No	http://www.stanford.edu/group/hadlylab/ssc/ (includes R scripts for ABC)	Anderson et al. (2005)
COMPASS	Coalescent simulation		**Demography:** Flexible (single population) **Mutation:** IMSM[4] **DNA Damage:** No **Recombination:** No	http://www.egs.uu.se/evbiol/Research/JakobssonLab/compass.html	Jakobsson (2009)
NETRECODON	Coalescent simulation		**Demography:** Flexible (except admixture)	http://darwin.uvigo.es/software/netrecod	Arenas and

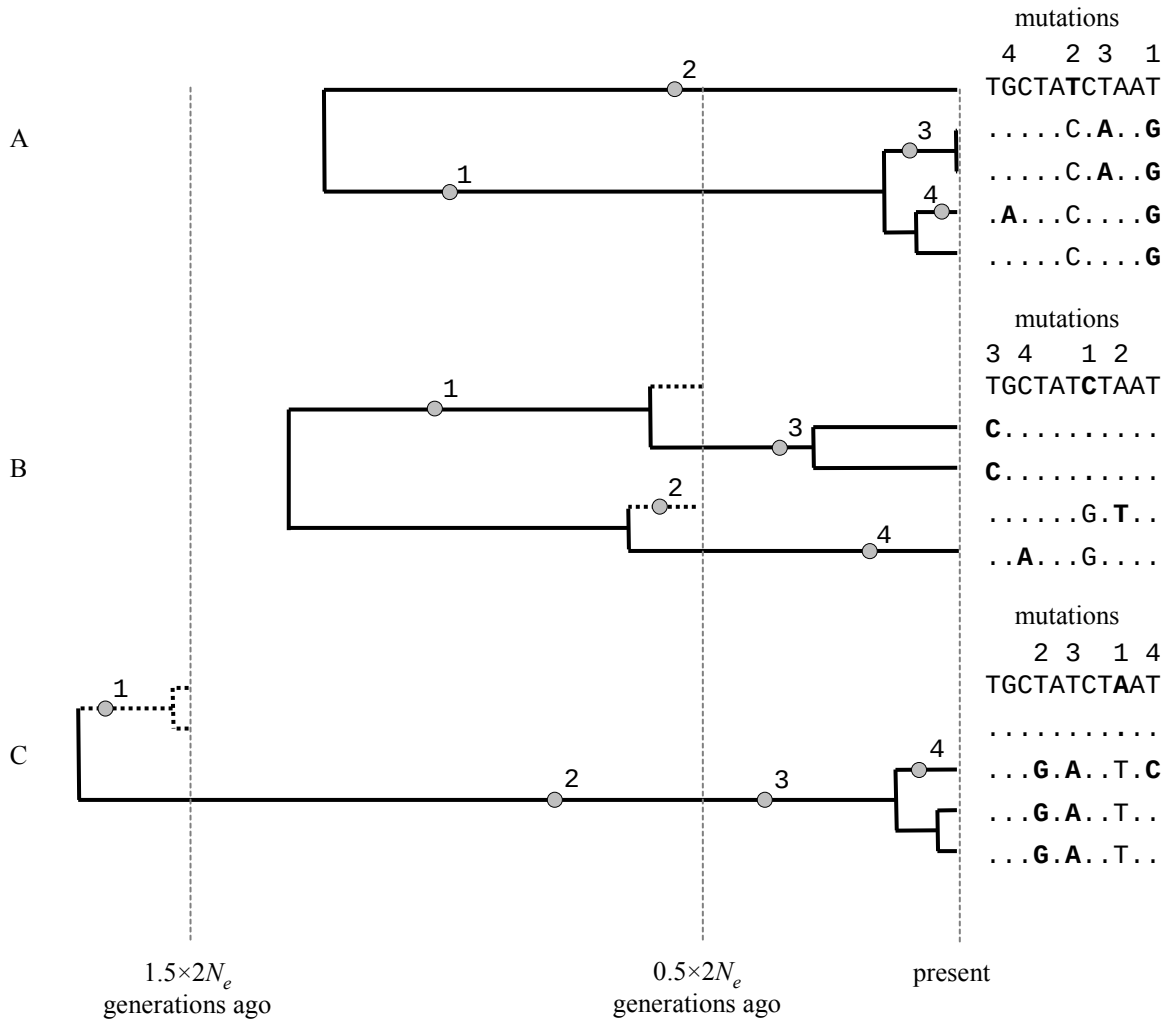
		Mutation: Most DNA models (except IMSM) **DNA Damage:** No **Recombination:** Yes	on.html	Posada (2009)

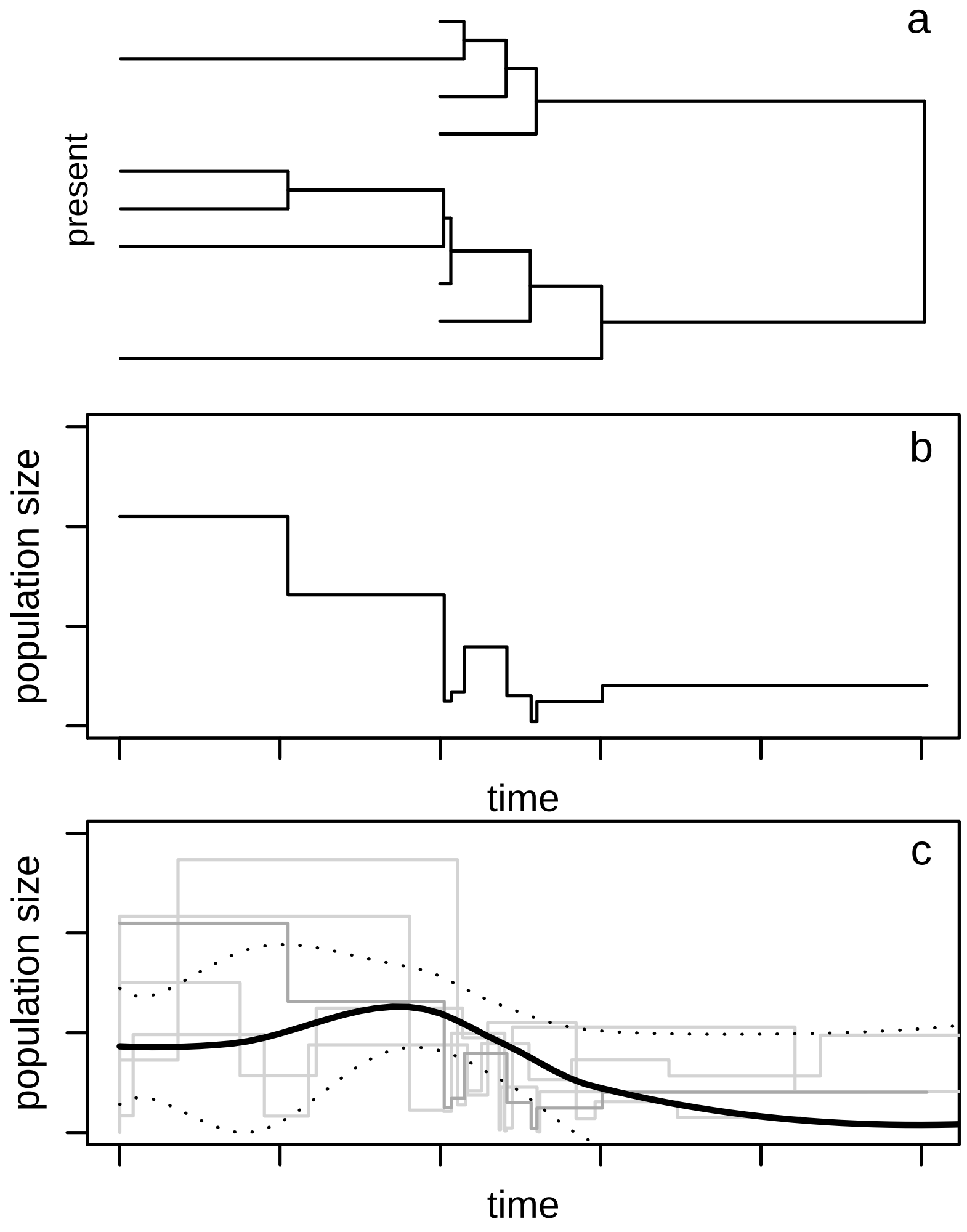
[1] Scripts for the R statistical computing environment (R Development Core Team 2009)

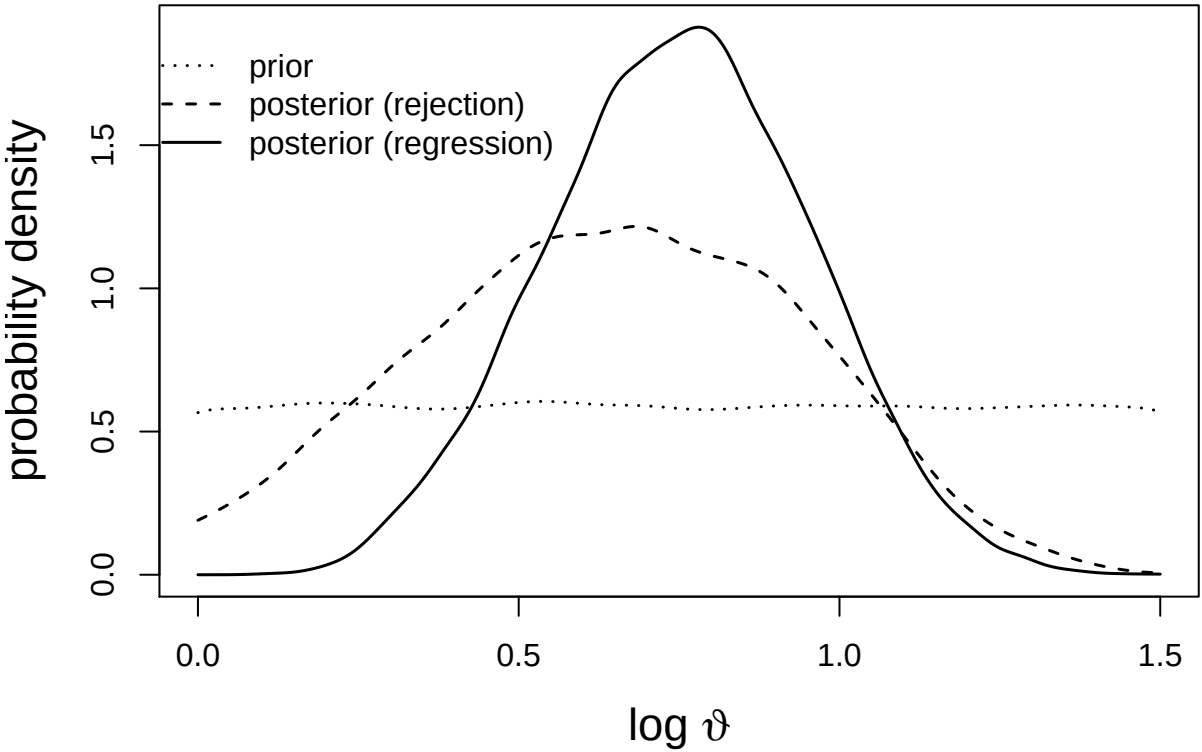
[2] SMM: stepwise mutation model (for microsatellites)

[3] K80: Kimura (1980); HKY: Hasegawa-Kishino-Yano (1985), and TN: Tamura-Nei (1993)

[4] IMSM: infinitely many site model


A
B
C
mutations
4 2 3 1
TGCTATCTAAT
.....C.A..G
.....C.A..G
.A...C....G
.....C....G
mutations
3 4 1 2
TGCTATCTAAT
C.....
C.....
.....G.T..
..A...G....
mutations
2 3 1 4
TGCTATCTAAT
.....
...G.A..T.C
...G.A..T..
...G.A..T..
1.5×2Ne generations ago
0.5×2Ne generations ago
present


a
present
b
population size
time
c
population size
time


prior
posterior (rejection)
posterior (regression)
probability density
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
log ϑ