MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 24–41 doi: 10.1111/mec.13492

DETECTING SELECTION IN NATURAL POPULATIONS: MAKING SENSE OF GENOME SCANS AND TOWARDS ALTERNATIVE SOLUTIONS

Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective

ANNA-SAPFO MALASPINAS*†

Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012, Bern, Switzerland, †Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark

Abstract

With hundreds of ancient genomes becoming available this year, ancient DNA research has now entered the genomics era. Utilizing the temporal aspect of these new data, we can now address fundamental evolutionary questions such as the characterization of selection processes shaping the genomes. The temporal dimension in the data has spurred the development in the last 10 years of new methods allowing the detection of loci evolving non-neutrally but also the inference of selection coefficients across genomes capitalizing on these time serial data. To guide empirically oriented researchers towards the statistical approach most appropriate for their data, this article reviews several of those methods, discussing their underlying assumptions and the parameter ranges for which they have been developed. While I discuss some methods developed for experimental evolution, the main focus is ancient DNA.

Keywords: adaptation, genomics/proteomics, population genetics - empirical, population genetics - theoretical

Received 1 June 2015; revision received 8 November 2015; accepted 10 November 2015

Time series analysis is widespread in several domains including meteorology, economics and physics (Hamilton 1994). In biology and more specifically population genetics, time series data dates back to the beginning of the field with pioneering work of Wright and Fisher (Fisher & Ford 1947; Wright 1948). Wright and Fisher described and analysed phenotypic times series data from a moth population (*Panaxia dominula*) collected near Oxford, UK. This moth exhibits several wing phenotypes, one of which declined sharply in frequency during the course of the study. Fisher and Wright argued over whether natural selection was needed to explain the change in allele frequency over time. Despite this case study having become a textbook example in population genetics (e. g. Provine 2001), the data are still being (re)analysed and recently several authors favoured the lethal recessive hypothesis, for example (Mathieson & McVean 2013; Foll *et al*. 2014b). The data were collected every year for several decades and represent tens of generations of evolution of the moth population. Indeed a generation of *Panaxia dominula* corresponds to 1 year making it possible to observe time series data across generations within a researcher's lifetime. Because of the need for a very short generation time and either (ideally) a Mendelian trait or molecular data, time series data since have remained largely a property of viral and experimental evolution studies.

The situation changed with the advent of ancient DNA (aDNA, i.e. DNA from extinct and/or long dead organisms). Similar to genomics, the study of aDNA entered a new era in the last 10 years thanks to advances in aDNA preparation and sequencing methods. Previously limited to rather uninformative short segments of mitochondrial DNA, whole-nuclear genomes have now become available from several extinct species, thus providing access to genomewide SNP information at the species level and delivering new insights into deep evolutionary time. To name some recent successes in humans, in 2014 the genome of (or more strictly speaking genomewide data since several of those genomes were shotgun sequenced at a very low depth of coverage) a 45 000-year-old Siberian (Raghavan *et al*. 2014b), 26 ancient Arctic (Raghavan *et al*. 2014a), 13

Correspondence: Anna-Sapfo Malaspinas,
E-mail: sapfo@berkeley.edu

Neolithic Hungarian (Gamba *et al.* 2014), a farmer from Germany and eight hunter gatherers from Sweden and Luxembourg (Lazaridis *et al.* 2014), 11 Scandinavian Stone Age (Skoglund *et al.* 2014), a 12 700-year-old Native American (Rasmussen *et al.* 2014) have become publicly available. These aDNA studies are characterized by a number of common features. In particular, the quality of the DNA remains a challenge owing to post-mortem DNA damage as well as to contamination by microbial and contemporary DNA (e.g. Hofreiter *et al.* 2001). However, the added temporal dimension has proven very fruitful. Indeed, rather than traditional single time point data sets (i.e. sampled at present), having genomic 'snapshots' of the population at several time points, before, during and after the start of a selective sweep, for example, can uniquely illuminate the parameters underlying the adaptive (and demographic) history of a population.

A plethora of methods exist to scan the genome for selective sweeps (for some reviews see, e.g. Crisci *et al.* 2012; Jensen *et al.* 2007; Nielsen 2005; Pool *et al.* 2010), but most were developed for contemporary samples. In the last 10 years, several new methods have been developed capitalizing on the temporal aspect of aDNA data (Bollback *et al.* 2008; Illingworth & Mustonen 2011; Illingworth *et al.* 2012, 2014; Malaspinas *et al.* 2012; Mathieson & McVean 2013; Nishino 2013; de Sousa *et al.* 2013; Feder *et al.* 2014; Foll *et al.* 2014a,b; Lacerda & Seoighe 2014; Steinrücken *et al.* 2014; Ferrer-Admetlla *et al.* 2015; Terhorst *et al.* 2015; Topa *et al.* 2015). The overall goal of this review is to provide a short guide to help choose the appropriate method given a specific application. Tables 1–3 offer a list of features while the text contains more details including the known limitations of each method. To get an idea of the kind of questions one might hope to answer with methods to detect selection with time serial data, I will first review some applications, (i.e. cases where selective processes have been characterized using aDNA), and give some general background on the selection parameters of interest. The idea is to develop some intuition regarding the key parameters involved that ultimately impact the power to detect selection. Please see, for example (Ewens 2004; Wakeley 2008; Nielsen & Slatkin 2013), for a more complete treatment of selection processes. It is worth noting that the field is relatively young such that only one method published thus far has included thorough simulations and a comparison of the power of related methods.

Table 1 Definition of the variables mentioned in the text and in Tables 3 and S1, Supporting information. Inspired by Table 1 in Baldwin-Brown *et al.* (2014).

Variable	Term & description
N_e	Effective population size – number of individuals that successfully reproduce every generation. For the diploid case we assume $2N_e$ chromosomes
s	Selection coefficient – strength of selection, related to the fitness of the selected genotype. Note that s_i refers to the selection coefficient of a specific locus whenever applicable
h	Dominance coefficient – determines whether the phenotype of a heterozygote individual for the mutant allele presents the mutant phenotype
w_{AA}, w_{Aa}, w_{aa}	Fitness effect for each of the three genotypes at a biallelic locus where a is the mutant allele (allele under selection) In most cases discussed here $w_{AA} = 1$, $w_{Aa} = 1 + 2sh$, $w_{aa} = 1 + s$ while $h = 0.5$ (additive or genic selection)
$\gamma = 2N_e s$	Rescaled selection coefficient (diploid case)
m	Migration rate between adjacent demes in a structured population model
t_0	Allele age – time when an allele of interest first appears in the population
x_0	Population allele frequency at the time the mutant allele arises in the population
k	Number of times data is being sampled
$t_1, t_2, \ldots t_k$	Sampling times
$n_1, n_2, \ldots n_k$	Number of chromosomes sampled at each time point
$i_1, i_2, \ldots i_k$	Number of chromosomes carrying the mutant allele (allele under selection)
$x_1, x_2, \ldots, x_k$	Population allele frequency at each sampling time
T	Period of sampling – number of generations between the last and first sampling time
R	Number of replicates – number of independent experimental populations in an experimental evolution study
H	Number of haplotypes at the beginning of an experimental evolution study
ρ	Recombination rate. When indexed refers to a specific locus
U	Mutation rate of new beneficial mutations
u_{Aa}, u_{aA}	Mutation rates of a mutation $A \rightarrow a$ (respectively, $a \rightarrow A$)
θ	Set of parameters to be inferred (θ differs between studies)

Table 2 Characteristics of some methods applicable to aDNA studies.

Study	When to use for aDNA studies	Which aspects of a selection sweep can be characterized	Applied to
Bollback *et al.* (2008)	Weak selection	Selection coefficient	ancient humans, bacteriophage
Malaspinas *et al.* (2012)	Weak selection, when allele age is to be inferred	Selection coefficient and allele age	ancient horses
Mathieson & McVean (2013)	Weak selection and structured population	Selection coefficients and migration rates	moth
Nishino (2013)	Selection scan involving multiple unlinked loci some under selection, some not. Also valid for strong selection	Selection scan	—
Feder *et al.* (2014)	Selection scan involving single loci under selection, also valid for strong selection	Selection scan	bacteriophage, yeast
Lacerda & Seoighe (2014)	Strong selection	Selection coefficient	
Foll *et al.* (2014a,b)	Multiple unlinked loci. Use instead of Mathieson *et al.*, Malaspinas *et al.* and Bollback *et al.* unless the allele age is of interest (Malaspinas *et al.*) or the population is structured (Mathieson *et al.*)	Selection coefficient, dominance coefficient	moth, influenza
Steinrücken *et al.* (2014)	Weak selection, when allele age and dominance coefficient are to be inferred	Selection coefficients, dominance coefficient and allele age	ancient horses

Natural selection and aDNA: some existing applications

A number of studies over the past decade have been published capitalizing on aDNA to characterize selection processes. They differ in (i) the organism under study, (ii) the type of data and (iii) the aspects of the potential selective events that are being characterized. (i) Organisms being investigated include (but are not limited to) maize, humans and horses (see below). In all nonhuman cases, the results are often interpreted in the context of domestication. (ii) Although the data included vary between a few loci (e.g. Jaenicke-Després *et al.* 2003) to several thousands (Mathieson *et al.* 2015), most studies thus far have either been based on specific regions targeted experimentally (captured), or focused on selection at specific loci. Genomic regions are often experimentally captured because they are thought to be functionally important. (iii) Finally, the questions of interest related to selection (that can also be addressed with the methods described below) have included: Is a specific locus under selection? Which loci across the genome could be under selection ('selection scan')? What is (are) the selection coefficient(s)? When did the allele potentially under selection arise in the population? Has there been constant selection? We review some of this work here and concentrate mostly on cases relying on more than one ancient sample (i.e. using time serial allele frequency data).

Maize

One of the first such publications is a maize study that appeared in 2003 (Jaenicke-Després *et al.* 2003). Maize is known to have derived from the wild grass teosinte in Southern Mexico perhaps around 4000 years BP [see (Jaenicke-Després *et al.* 2003) and citations therein]. Samples spanning several thousands of years and typed at three genes involved in the plant's architecture were sequenced in (Jaenicke-Després *et al.* 2003). The authors of that study demonstrate that alleles typical of contemporary maize, and in low frequency in the related wild grass teosinte, were present as early as ~2500 years BP. Selection is characterized qualitatively and it is argued that farmers indirectly selected alleles that relate to plant morphology and to biochemical properties of the starch, since the allele frequency in those early samples is high. These results are based on a limited number of loci and (3) a limited number of samples (11) from a limited number of sites (2). Da Fonseca *et al.* (2015) studied 32 archaeological samples of maize spanning 6000 years and characterized 348 genes shown to be functionally important. This study is unique in characterizing selection for several samples taken through time from the exact same location and considering genomewide data. Using a number of statistics traditionally applied to modern data, including PBS (Yi *et al.* 2010) and Tajima's D (Tajima 1989), they first identify a number of candidate loci under selection. They single out

Table 3 Main characteristics of each study reviewed.

Study	Population genetics model and approximation	Some underlying assumptions	Single/multiple locus	Linkage*	$\theta^{\dagger}$	Drift	Sampling noise
Bollback *et al.* (2008)	WF model, diffusion approx., numerical approach	$N_e \to \infty, s \sim \mathcal{O}(\frac{1}{N_e})$, $x_0 \sim$ *uniform*(1)	single	unlinked	N_e, s	Finite N_e (large)	Yes
Illingworth *et al.* (2012)	Quasi-deterministic	$N_e \to \infty$	single	variable (ρ)	s_i, ρ_i	Infinite N_e	Yes
Malaspinas *et al.* (2012)	WF model, diffusion approx., one-step process	$N_e \to \infty, s \sim \mathcal{O}(\frac{1}{N_e})$	single	unlinked	N_e, s, t_0	Finite N_e (large)	Yes
de Sousa *et al.* (2013)	WF model		single[‡]	complete	$f(s)$[§], U	Finite N_e	No
Mathieson & McVean (2013)	WF model, Gaussian approx., s small	$0 < x_i < 1$[¶]	single	unlinked	m, s	Finite N_e (large)	Yes
Nishino (2013)	WF model, Gaussian approx.	$0 < x_i < 1$[¶]	multiple	unlinked	[N_e], [s]**	Finite N_e (large)	No
Feder *et al.* (2014)	Moran model. Gaussian approx.	$0 < x_i < 1$[¶]	single	unlinked	[N_e], [s]**	Finite N_e (large)	No[††]
Baldwin-Brown *et al.* (2014)	WF model		multiple	variable (ρ)	NA[‡‡]	Finite N_e	No
Foll *et al.* (2014a,b)	WF model		multiple	unlinked	N_e, s_i, h_i	Finite N_e	Yes
Lacerda & Seoighe (2014)	WF model, diffusion approx. WF model, Gaussian approx., WF model, Delta method	$N_e \to \infty, s \sim \mathcal{O}(\frac{1}{N_e}), \gamma = 2N_e s \sim \mathcal{O}(1)$ $N_e \to \infty, s \to 0, \gamma \to \infty$ $N_e \to \infty$	single	unlinked	N_e, s	Finite N_e	Yes[§§]
Steinrücken *et al.* (2014)	WF model, diffusion approx., spectral decomp.	$N_e \to \infty, s \sim \mathcal{O}(\frac{1}{N_e})$	single	unlinked	N_e, s, h, t_0, u, v	Finite N_e (large)	Yes
Terhorst *et al.* (2015)	WF model, Gaussian approx.	$0 < x_i < 1$[¶]	multiple	variable (ρ)	N_e, s_i, h_i, ρ_i	Finite N_e (large)	Yes
Topa *et al.* (2015)	WF model, Gaussian approx.	$0 < x_i < 1$[¶]	single	unlinked	[N_e], s	Finite N_e (large)	Yes

*For single-locus-based methods the sites are either unlinked or in complete linkage with the one of interest.
[†]In square brackets are the parameters that could be estimated but are not a major focus.
[‡]A neutral allele is used as a marker for the evolution of the beneficial allele(s) because all markers are fully linked.
[§]Distribution of beneficial selection coefficients.
[¶]The frequency of the allele under selections is assumed to remain 'far away' from the boundaries.
**This method is a statistical test to detect selected vs. neutral loci. It can be extended to infer those parameters.
[††]Simulations with finite samples were used to test the method's performance.
[‡‡]Simulation study.
[§§]Limited sampling noise since $n_i \sim N_e$.

the genes with the largest allele frequency differences, including some related to drought tolerance and sugar content and infer that selection occurred at specific times based on the timing of those allele frequency changes.

Humans

Three types of traits have been the primary focal point in aDNA studies of humans. First, traits related to pigmentation (including hair colour, skin complexion and eye colour); second, traits linked with diet (e.g. lactose intolerance) and third, traits related to pathogens such as HIV. Those traits have in common that they have been characterized at least partially from a molecular perspective, and have been hypothesized to be under selection. One of the first studies to report allele frequency changes through time in humans appeared in 2005 (Hummel *et al.* 2005). The authors characterize the CCR5-Δ 32 deletion, shown to confer resistance, or even close to complete immunity (in homozygote state) to HIV infection. It has been hypothesized that this allele – having appeared before the HIV outbreak in humans in the 20th century – has conferred an advantage against a different pathogen. Hummel *et al.* find that the deletion was already present 2000 years ago in Germany and Italy alike. They argue that plague was not a major selective force on the allele given that it was not found in higher frequency in healthy vs. plague-infected individuals. However, the authors do not attempt in this study to characterize selection statistically. Bollback *et al.* (2008) re-analyse the data and show that there is no need to invoke selection to explain the observed change in allele frequency for this locus (but see also Foll *et al.* 2014b).

Three more studies based on humans have appeared over the last 2 years, demonstrating the increased pace in aDNA data production. In January 2014, Sverrisdottir *et al.* (2014) consider a mutation upstream of the lactase gene that is strongly associated with lactase persistence in Europeans. They investigate the calcium assimilation hypothesis which posits that fresh milk supplements the lack of vitamin D and calcium observed in Northern Europeans. To do so, they type the aforementioned mutation in eight Neolithic Iberian individuals (i.e. individuals presumably living at a latitude with sufficient UVB light, thus presumably producing enough vitamin D). They find that the derived allele is absent from all eight individuals. Combining their ancient data with modern data, and arguing for population continuity, they find, through population genetics simulations, that natural selection is needed to explain the pattern observed in those eight individuals arguing against the 'calcium assimilation hypothesis'. Their simulations account for the uncertainty of the age of each sample. They estimate that a selection coefficient of around 2% would explain the pattern in frequency change observed. Later that year, in April 2014, another study (Wilde *et al.* 2014) in humans used allele frequencies over time to characterize the potential selection acting on genes (or rather SNPs) known to be involved in pigmentation pathways (HERC2, SLC45A2, TYR). Wilde *et al.* sequence 63 samples from Ukraine ranging from 6500 to 5000 years ago and compare them to modern samples. They find for all three cases an increase in frequency of the derived allele and reject the hypothesis that those SNPs have been evolving neutrally. They estimate a selection coefficient on the order of 2–10% (i.e. strong selection, see below) and conclude that selection has acted in Europeans for the past 5000 years on pigmentation genes. In October 2014, Gamba *et al.* (2014) sequenced genomewide data of 13 individuals from Hungary spanning 5000 years at a depth of coverage between 0.1× and 21×. They infer the phenotype of each individual based on their (sometimes imputed) genotypes for hair and eye colour considering three SNPs in the three genes SLC24A5, SLC45A2 (both having swept to fixation in Europeans) and TYRP1. They first note that the derived allele appears much later in SLC24A5 than in SLC45A2. The authors therefore suggest that the associated selective sweeps must have occurred at different times for those two genes. The derived allele also appears rather late in TYRP1, while the hair and eye colours are predicted to be lighter in recent times. Finally, they consider an allele that has been linked to lactase persistence. It had been previously hypothesized that lactase persistence was first selected starting approximately 5500 years BC possibly in association with the Neolithic LBK culture. However, the derived allele in this case appears at only 1000 years BC. The authors do not attempt to estimate the selection coefficients of those candidate SNPs.

In 2015, two studies have included ancient genomewide data from tenths of individuals. Allentoft *et al.* (2015) sequenced the low coverage genomes of 101 ancient Eurasians dating to 3000–1000 BC. The authors highlight four SNPs associated with light skin pigmentation (SLC24A5 and SLC45A2), blue eyes (HERC/OCA2), and lactose intolerance and compute the corresponding allele frequency changes. They find that the SNPs associated with light skin have the largest increase in allele frequency. In contrast, the SNP associated with blue eyes is at intermediate or low frequency during the Bronze Age, while the SNP related to lactase tolerance remains at very low frequency. The authors hypothesize, as did Gamba *et al.* (and see also below), that selection must have started to act much later than previously thought. Finally, in October 2015, Mathieson

et al. (2015) sequenced 300 000 positions (targeted experimentally) in the genome of 83 European samples as old as 8000 years old. They evaluate whether selection acted on ~30 000 SNPs chosen based on their functional importance. To do so, they assume that modern Europeans are a mixture of three ancestral components and look for outlier SNPs in terms of allele frequencies. They identify five loci that reach genomewide 'significance'. Among them, they authors find some 'usual suspects' including the lactase gene but also SLC45A2, HERC/OCA2. They estimate a selection coefficient of 1.5% (95%, CI 1.0–3.4%) for the lactase persistence SNP, while they also find the derived allele only present in the more recent samples. Selection on SLC24A5 and SLC45A2 is estimated to be on the order of 2%. They find weak negative selection on the 'blue eye gene' (HERC2), in disagreement with the Wilde *et al.* study. Two additional genes are identified as outliers, namely the NADSYN1 and FADS1 genes related to fatty acid metabolism. Finally, they also test for selection on complex traits and find a signal of directional selection on height and body mass index, but no evidence for selection on other complex trait tested including waist–hip ratio, type 2 diabetes, inflammatory bowel disease and lipid levels.

Horses

Pigmentation genes in mammals are well characterized functionally compared to other traits. In horses, there have been several studies tracking the changes in allele frequency for variants linked to various coat colours. In 2009 for example, Ludwig *et al.* (2009) type eight coat colour loci for samples ranging from a pre- to a postdomestication period. They demonstrate that the number of inferred horse coat colours increased with time. Using the method by Bollback *et al.* (2008), Ludwig *et al.* found that two loci were likely under selection. Interestingly, this data was re-analysed using two related methods (Malaspinas *et al.* 2012; Steinrücken *et al.* 2014). With the first method, which also infers the allele age, it was found that the change in allele frequencies could be explained by genetic drift alone. Steinrücken *et al.* (2014) also inferred the mode of selection and found that balancing selection best fit the data. More recently, Ludwig *et al.* (2015) reported the changes in allele frequency of the locus encoding the Leopard complex spotting – a specific coat colour in horses. This trait also causes night blindness and the authors used a Bayesian computational approach to estimate the selection coefficient jointly acting on night blindness and the Leopard complex spotting. The spotting trait fluctuates through time and the authors show that the trajectory is compatible with fluctuating selection over a 25 000 year period. The estimated selection coefficients vary between −0.5 and 0.5 suggesting very strong selection. In addition, a number of recent studies on horses performed whole-genome scans using ancient horses as predomesticated 'controls'. For example, to identify the genetic changes associated with the domestication of horses, Schubert *et al.* (2014) (see also Orlando *et al.* 2013) compare the whole genomes of modern domesticated horses to those of two 40 000- and 16 000-year-old horses of the Taymyr peninsula, Russia, predating horse domestication. They apply four different selection scans based on a number of summary statistics including: the rate of nonsynonymous vs. synonymous mutations, Tajima's D, the inferred coalescence rate of alleles in high frequency in domesticated horses ('domesticated alleles') and genomic tracts with a high number of domesticated alleles. These analyses yield a list of 125 gene candidates for positive selection in domesticated horses including genes potentially reflecting the physiological adaptation of horses to utilization by humans: muscular and limb development (ACTA1, C-SKI, MYBPC1), articular junctions (COL22A1) and the cardiac system (ACAD8, BRAF, FANCA, SGCD and CACNA1D). Some other candidates were found to be associated to cognitive functions in relation to social behaviour (GRID1), fear response (VDAC1), learning capabilities and agreeableness (SYNJ2). Der Sarkissian *et al.* (2015), having sequenced the genomes of both modern domesticated and Przewalski's horses, report candidate genes involved in metabolism, cardiac disorders, muscle contraction, reproductive behaviour and signalling pathways, which could be responsible for the striking phenotypic difference between these horses. These genes were identified using a method relying on site frequency spectra (Pavlidis *et al.* 2013) and by identifying regions with unusual F_{st} values.

Selection: some general theoretical background for the single-locus case

Natural selection as an evolutionary force was first put forward by Wallace and Darwin (Wallace 1858; Darwin 2012). Evolution occurs by natural selection when the genetic make-up of individuals confers higher viability, mating success and/or fertility. Let us consider a simplified case where the fitness of diploid individuals in a population is solely determined by their viability and where viability is in turn determined by the genotype at a single locus. Suppose that two alleles are present at a locus in a panmictic population. An allele *a* that arose at time t_0 in the population (also called the mutant allele hereafter) and an ancestral allele *A* that was the

'original' allele in the population. Note that a common assumption is that of no recurrent mutations (i.e. the site can mutate from $A \rightarrow a$ only once). Any individual will be the carrier of either the *AA*, *Aa* or *aa* genotype. Let us denote w_{AA}, w_{Aa} and w_{aa} the fitness of individuals with the *AA*, *Aa* and *aa* genotype, respectively. Assuming random mating and a random union of gametes, then, at generation t, if, y_t is the frequency of the A allele, and x_t the frequency of the a allele, each genotype is present at a frequency of y_t^2, $2x_t\ y_t$ and x_t^2 before selection starts acting. After the onset of selection, only a fraction of the individuals will survive into adulthood. This fraction is determined by the relative fitness of each genotype. Fitness coefficients can be parameterized in a number of equivalent ways because only the relative values come into play. Let us consider a case of selection with the following parameterization: $w_{AA} = 1$, $w_{Aa} = 1 + sh$ and $w_{aa} = 1 + s$, where s is the selection coefficient, h the dominance coefficient and $s > -1$ (such that the fitness coefficients are positive) and $h \in [0, 1]$. We consider the simple case of directional positive selection with $s > 0$. In this case, assuming s is small, and considering the relative fitnesses, then one could say for example that *AA* is less viable than *aa* by an amount s while *Aa* is less viable than *aa* by an amount sh. Debatably, a selection coefficient of 10% corresponds to 'strong selection' while a selection coefficient of 1% or less can be considered 'weak'. Let us assume for what follows that $h = 0.5$ (also called additive selection or genic selection; in this case the heterozygous fitness is the arithmetic average of the homozygous fitnesses). Ignoring the effects of drift and assuming discrete nonoverlapping generations and small s, one can approximate the change in allele frequency by

$$x_{t+1} - x_t = \frac{1}{2} s x_t (1 - x_t)$$

With this notation, x_0 is the frequency of the allele when it first appeared in the population at time $t = t_0 = 0$. The parameter t_0 can also be the allele age (the sampling times will later be denoted t_1, t_2,..., see below). With no genetic drift, or assuming the population is infinite, the frequency of the selected allele will increase in a deterministic fashion at every generation following a sigmoid curve as follows:

$$x_t = x_0 (x_0 + (1 - x_0) e^{-\frac{1}{2} s (t - t_0)})^{-1} \quad (1)$$

Note how in the neutral case and for a case of no genetic drift, the frequency of the allele remains constant ($x_t = x_0$).

We now consider the effect of a finite population. We assume that the individuals reproduce according to a Wright Fisher (WF) population genetics model (see, e.g. Ewens 2004). Every generation is formed from the one preceding it by drawing gametes at random with replacement. The number of individuals reproducing and contributing to the next generation is N_e. Because of drift, the allele frequency can increase or decrease in one generation stochastically. When the allele first arises in the population at time t_0, its frequency is $\frac{1}{2N_e}$, that is a fairly small number for typical effective population sizes. Assuming that the allele escapes loss, then its trajectory will oscillate around the deterministic trajectory with an amplitude that will depend on the population size. The smaller the population size, the larger the 'jumps' between allele frequencies and the jaggedness around the deterministic trajectory. As a result, and because we assumed that the mutation arises only once, in most a alleles will be lost within the first few generations (see Fig. 1). The frequency in the population of an allele arising 1000 generations ago for different combinations of population sizes and selection coefficients is shown in Fig. 2. It is possible to see that the larger the population size, the smaller the effect of genetic drift, and the closer the allele trajectory is to the deterministic trajectory. This illustrates the overall difficulty of the problem, the confounding effects of genetic drift (or demography) and selection. Although the two processes affect trajectories differently (positive selection leading to an overall increase in allele frequency and genetic drift random oscillations around the deterministic trajectory), the population allele frequencies are hard to distinguish, particularly for small

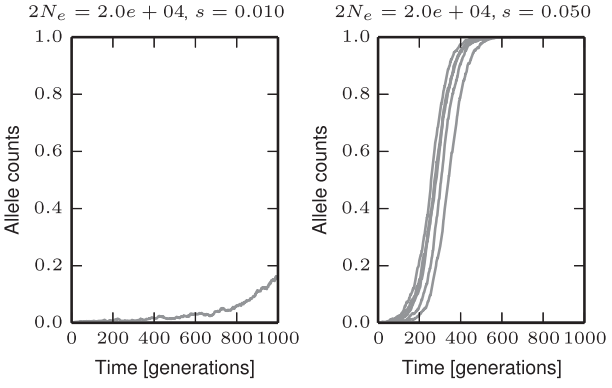

Fig. 1 Simulations (100 replicates) under a WF model of the change in allele frequency over time for $N_e = 10\ 000$ and $s = 0.01$ and $s = 0.05$ for the left and right panel, respectively. We fix the number of generations to 1000 and assume that at the first sampling time the allele frequency was $\frac{1}{2N_e}$. All simulated trajectories are shown in grey. Even in the case of relatively strong selection (5%), only seven replicates out of 100 escape and get eventually fixed instead of lost within the first few generations.

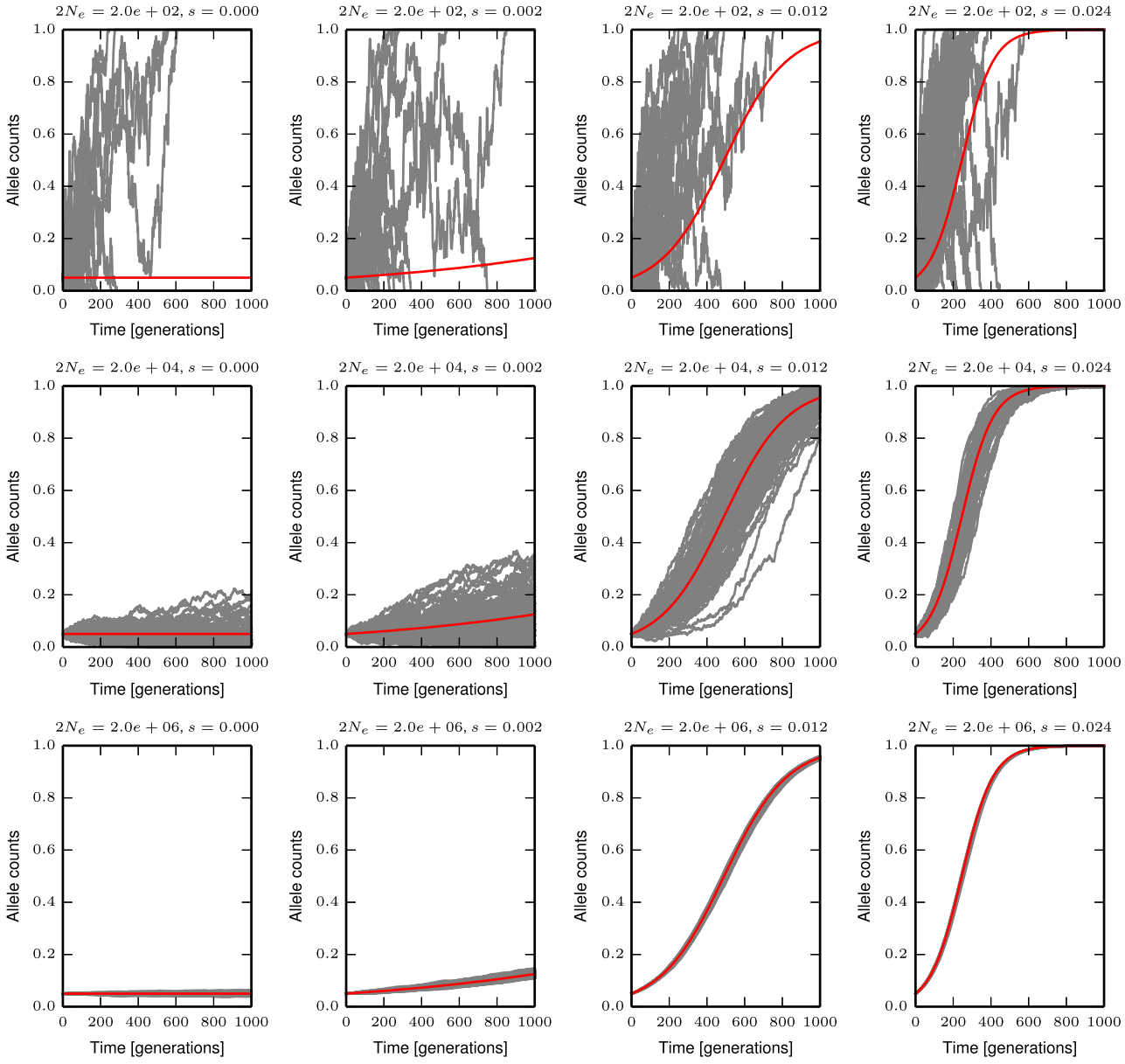

Fig. 2 Simulations (100 replicates) under a WF model of the allele frequency over time for different population sizes and selection coefficients. We fix the number of generations to 1000 and assume that at the first sampling time the allele frequency was 0.05. The top row shows the results for $N_e = 100$, the middle row for $N_e = 10^4$ and the bottom row for $N_e = 10^7$. We simulate data under a neutral case, and for additive selection with $s = 0.0\%$, $s = 0.2\%$, $s = 1.2\%$ and $s = 2.4\%$, for the second, third and fourth column, respectively. All simulated trajectories are shown in grey. The expected deterministic curve (without drift, or assuming an infinite effective size) is shown in red. As expected the oscillations around the red curve – or the drift – increase with smaller N_e.

population sizes (first row of Fig. 2). The problem becomes much simpler for the case of a large population size, as shown in Fig. 2 (last row). So overall, as genetic drift is inversely correlated with the population size, while the effect of (positive) selection is higher for larger s, one can expect that it will be easier to detect the signatures of selection for organisms with large population sizes undergoing strong selection.

Framework and notation

The overall framework is quite similar across methods described below (see Fig. 3 for notation and Table 1 for relevant definitions). We denote X_t a random variable representing the frequency of the allele a through time. Most of the models below will assume that selection is constant from the time the allele arose (or at the first

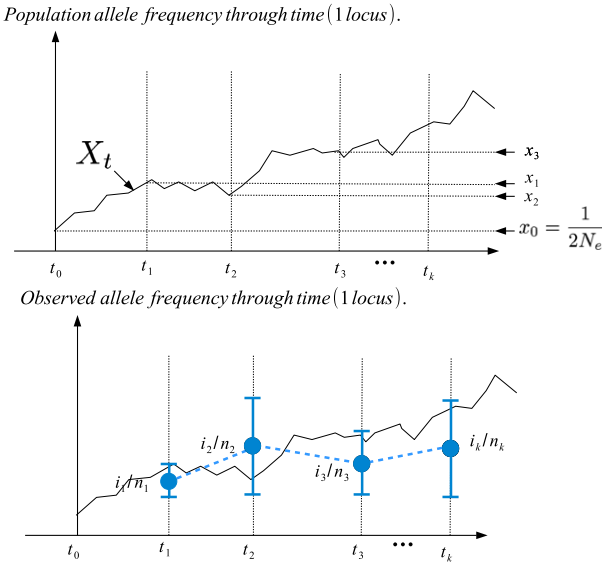

Fig. 3 Schematic of the overall setup applicable to most single-locus methods we discuss in this review. The mutant allele appears at time t_0 and we assume no recurrent mutations. The actual change in allele frequency of the mutant allele through time in the population is shown in black. We generally do not observe the whole trajectory. Instead, we observe the frequency of the mutant allele (in blue) in $n_1, \ldots, n_k$ chromosomes sampled at times $t_1, \ldots, t_k$ (dashed vertical lines). We denote $i_i, \ldots, i_k$ the number of mutant alleles at each sampling time and the vertical bars are potential error bars that relate to the number of chromosomes sampled. Note that some methods (e.g. Feder *et al.* 2014) assume that the number of sampled chromosomes is large enough such that $x_i \sim \frac{i_i}{n_i}$.

sampling time) up to present. The states of X_t are the allelic frequencies ($x_j = \frac{j}{2N_e}$ for $0 \leq j \leq 2N_e$). We will assume that we have samples from k distinct sampling time points. We suppose that $n_1, n_2, \ldots, n_k$ chromosomes were collected among which $i_1, i_2, \ldots, i_k$ are of the a type and that the chromosomes were drawn at times $t_1, t_2, \ldots, t_k$, where time is measured in generations and $t_{k-1} < t_k$.

We do not observe the actual population frequency at every generation in most cases (particularly for the small samples characterizing aDNA). Rather we observe the count of mutant alleles at every sampling time. This adds to the complexity of the model because the observed trajectory oscillates stochastically even more than the underlying population allele frequencies (Fig. 3).

In summary, for the single-locus case, the information to distinguish between neutrality and selection comes from the change in allele frequency through time: an increasing frequency could be indicative of selection where the steepest the extent of the change in frequency informs the degree of the selection coefficient (note that similar data for contemporary populations – i.e. a single frequency measurement in time – provides comparatively little information). Nevertheless, the combination of the finiteness of the population size and of the sampled data results in increased stochasticity in the trajectories, irrespective of s. It is therefore necessary to employ statistical approaches to distinguish between neutrality and selection.

Ideal selection

Even ignoring the effects of genetic drift, some situations will more readily allow for the detection of selection. Feder *et al.* (2014) provide a nice and intuitive argument to compute an 'ideal' selection coefficient. The three columns to the right in Fig. 2 illustrate their point. The allele frequencies plotted in red correspond to the deterministic curve (eqn 1). For all cases of time series data, the experimenter is constrained in terms of the sampling period T. In the case of aDNA, the constraint comes from the availability of samples and the ability to recover DNA from those samples. We consider the case where we are able to recover data at various time points over a thousand generations. In the case of humans, this would mean around 30 000 years ago for a generation time of around 30 [years/generation]. Although only a few '30 000-year-old human genome' have been assembled, it has been shown and its been predicted that DNA can be recovered from such specimens (Allentoft *et al.* 2012). The second column shows a case with $s = 0.2\%$. In this instance, the population allele frequency between all five sampling points (shown as dashed vertical lines in the figure) is quite similar, such that the data are quite uninformative. In this case, noise (due to drift) largely dominates, particularly for small population sizes. The column to the far right shows the case of $s = 2\%$. In this case, selection is 'too strong' because the frequency at the last three sampling points is close to 1. The allele fixes very quickly and several of the samples will provide no additional information about the dynamic. The third column represent the 'ideal selection coefficient', or 's_{power}' of Feder *et al.* (2014). To derive the optimal selection coefficient for a given time interval, one then asks, given a population allele frequency at the first sampling time (x_1), what is the strength of selection needed for the allele to reach a maximal frequency x_{max} (close to 1) at the last sampling time? They derive s_{power} from eqn 1 by setting $x_{max} = x(T)$, where T is the total sampling time:

$$s_{power}(T) = \frac{2}{T}\ln\left(\frac{x_{max}}{1-x_{max}}\frac{1-x_1}{x_1}\right)$$

This provides a baseline to design an experiment in terms of sampling, if one is to use the change in allele

frequencies over time as a summary statistics to infer and detect selection. For example, assuming we manage to sample alleles close to the time they arose in a population of size $N_e = 10^4$, that is $x_1 \approx \frac{1}{2N_e} = 0.00005$ and $x_{max} = 0.99$, then $s_{power}(100) = 0.14$, $s_{power}(1000) = 0.014$, $s_{power}(10\ 000) = 0.0014$. It is often interesting to consider the value of the rescaled selection coefficient $\gamma = 2N_e s$. In this case, $2N_e s_{power}(100) = 2900$, $2N_e s_{power}\ (1000) = 290$, $2N_e s_{power}\ (10\ 000) = 29$. Note that this relation can be extended to the multilocus case, see (Terhorst *et al.* 2015).

Methods to estimate selection from time serial samples

The methods developed to quantify selection are natural extensions of the ones developed to infer the effective population size from time serial data (Krimbas & Tsakas 1971; Pollak 1983; Williamson & Slatkin 1999; Anderson *et al.* 2000; Wang 2001; Berthier *et al.* 2002; Beaumont 2003). Yet in this context, the population size is mostly seen as a 'nuisance parameter' (a parameter that needs to be estimated but that is not the major focus). The methods I will now discuss differ in a number of ways: the underlying population genetics model, the parameters that are being estimated, the assumed parameter ranges, the ascertainment scheme, the initial allele frequency and the expected type of data that are being collected. I first review some key differences in the underlying population genetics model and briefly discuss each method subsequently (see Tables 2–3, Table S1 in Supporting information for a summary of some key aspects of those methods). I will first mention the methods based on one locus and leave the multilocus methods to the end. Whenever possible I discuss the simulations presented in each study (see also Table S1 in Supporting information) and associated parameter ranges. As most methods have not been compared, one can hypothesize that the methods will perform best for the parameter range used for those simulations.

Underlying model – approximation to the WF model

There are two broad categories of models assumed for each method. They can be distinguished by whether the finiteness of the population (or genetic drift) is modelled or not. When drift is modelled, most of the methods are based on the WF model of evolution. When genetic drift is not modelled, the methods are based on quasi-deterministic models that describe the evolution of the population as a set of time-dependent differential equations where the stochasticity arises from the finite number of sequences being sampled only. The latter case applies only to organisms with very large population sizes such as the ones often encountered for microorganisms (see Illingworth *et al.* 2014; Illingworth & Mustonen 2011; Illingworth *et al.* 2012). Although those methods have a limited relevance for application in aDNA, they are very useful to build an intuition on what to expect in idealized situations.

The WF model becomes challenging for large population sizes from a computational perspective. Therefore, several methods we review here consider approximations to the WF model. Those approximations have been shown to be valid when the parameters of interest tend to the limits (0 or ∞), but those are asymptotic results and it is unclear what is large or small enough in practice, or what types of biases one might expect given a certain approximation. Fortunately, in 2014 Lacerda and Seoighe addressed this question specifically for time serial data: they reviewed the most common types of approximations and their applicability – while also proposing a new approximation – for different ranges of population sizes and selection coefficients for time serial data-based inference. They considered three types of approximations: (i) the standard diffusion approximation that assumes the population size N_e is large and the selection coefficient inversely proportional to the population size. More specifically, in this case, $s = \mathcal{O}(\frac{1}{N_{\{e\}}})$ such that, at the limit when $N_e \to \infty$, $N_e s$ is a constant (note that the rescaled selection coefficient $\gamma = N_e s$ is a natural parameterization in this case). (ii) A Gaussian approximation that was developed for cases where the stochastic effects of genetic drift reduce faster than the effects of selection. In this case, it is also assumed that the population size is large but as $N_e \to \infty$, $N_e s \to \infty$, so that the selection coefficient tends to 0 ($s \to 0$) slower than the population size tends to infinity ($N_e \to \infty$). (iii) the Delta method, an approximation introduced by Lacerda and Seoigue, which also assumes that the population size is large but makes no assumption in terms of the selection coefficient.

They show through extensive simulations that the standard diffusion and the Gaussian diffusion perform poorly for strong selection coefficients. Both approximations lead to underestimations of the selection coefficient. In the standard diffusion case, they show that the population size is also overestimated to 'compensate', so that $N_e s$ remains close to the true value. They recommend (i) to use the full WF model for a small population size ($N_e < 5000$) and for any selection (weak or strong), (ii) to use the standard diffusion or the delta method for a large population size and weak selection ($N_e > 5000$ and $|s| < 0.01$, which also translates into $\gamma < 100$) and (iii) to use the Delta method for a large population size and large selection coefficient ($\gamma > 100$). This work implies that the estimates of selection coeffi-

cients from most current methods are likely to be biased if selection is strong.

One-locus

The first method proposed was a likelihood-based approach that co-estimates selection (s) and population size (N_e) from time series data assuming a WF model. It was developed by Bollback *et al.* (2008) and it did set the stage for a number of follow up likelihood methods based on the WF model. It relies on allele frequency data collected at several time points. The WF model is approximated with a standard diffusion process and one single bi-allelic locus is considered. The transition probability of the diffusion process (i.e. the probability of transitioning from one frequency x_t to another x_{t+1}) is calculated using a numerical approach. It is also assumed that the frequency at the first sampling time is uniform. As a result the frequency at the first sampling time does not appear in the likelihood equation (also one of the rationales behind that assumption). Another approach is to assume that the allele frequency is the observed allele frequency, or even that the value is 0.5. Note that it has not been thoroughly investigated when these assumptions actually matters in practice (under which parameter range). Finally, Bollback *et al.* consider two ascertainment schemes: an unconditional process where the allele frequency is not constrained, and the case where the selected mutations are known to reach fixation.

The model is applied to two data sets – human and bacteriophage data – involving two extreme sets of parameter values. The human data are collected over 3000 years, which the authors translate into 150 generations, while the selection coefficient is assumed to take values from -1 to 1. This is equivalent (assuming $N_e \approx 10\,000 = 10^4$ for humans) to a sampling period of $T = 3\,000$ and to a range for the rescaled selection coefficient of $\gamma \in [-2 \times 10^4, 2 \times 10^4]$, that is, very strong negative or positive selection. The bacteriophage data are collected over a similar number of generations, that is 100 generations, but in contrast to the previous example the population size is much larger and assumed to be between 10^7 and 2×10^8. The selection coefficient is assumed to vary from 0 to 5, which corresponds to a range $\gamma \in [0, 10^9]$, that is, even stronger selection. Although the search range includes very strong selection, where the standard diffusion approximation is worse, the inferred parameters are much lower: $s = -0.0005$ (-0.09, 0.01), that is weak selection, for the human case, and strong positive selection for the bacteriophage: $s = 0.4$ (0.4, 0.8).

We proposed an extension of this method in 2012 (Malaspinas *et al.* 2012). As with Bollback *et al.*, we assume a single bi-allelic locus and a WF model of evolution. We also assume that the population allele frequencies are unobserved and consider a hidden Markov model approach. We approximate the WF standard diffusion process with a one-step process, that is a Markov chain where 'jumps' between allele frequencies of at most 'one step' at a time are allowed (to the immediate left or right of the current state). The transition probability is discontinuous at the 0 (when the allele is lost) and 1 (when the allele is fixed) boundaries. Given the discontinuity, finding a valid numerical approximation is challenging and represents a potential difficulty in the Bollback *et al.* implementation. In our case, those 0 and 1 states are included in the one-step process providing a natural approximation. We can co-estimate three parameters, namely the allele age, t_0, the selection coefficient, s, and the population size N_e [and the dominance coefficient (see, e.g. Foll *et al.* 2014a)]. The co-estimation of the allele age seems quite natural and helps avoiding (perhaps unrealistic) assumptions about the allele frequency at the first sampling time point. Note that, as with Bollback *et al.*, in theory, our method is also limited to the case where γ is on the order of 1 (i.e. weak selection). It is shown in practice through extensive simulations by Foll *et al.* (2014b) that, to produce unbiased estimates, our method is limited to small rescaled selection coefficients, as we anticipated and is also demonstrated in (Lacerda & Seoighe 2014) for the standard diffusion approximation. Specifically Foll *et al.* (2014b) show that for $\gamma > 200$ the estimated selection coefficient provided by our method can be strongly biased downwards. Although Foll *et al.* are not able to compute the bias with the Bollback *et al.* approach, intuitively this limitation should apply in this case as well. Hence, as a rule of thumb, to provide unbiased estimates both methods should be used in cases where $\gamma < 200$. Moreover, our method should be used if the age of the allele is of interest.

Steinrücken *et al.* (2014) proposed an extension of the two methods above. The underlying model is similar to the one we considered (the age of the allele under selection also being one of the parameters that can be inferred), while in this case, the model allows for recurrent mutation. The main difference lies in the approximation of the WF model. Indeed, in a related paper, (Song, Y.S. and Steinrücken 2012) derived an explicit analytical solution to the transition probability for arbitrary selection. This solution is used for this method (Steinrücken *et al.* 2014) instead of the two approximations described above. However, the analytical solution involves an infinite summation such that the WF model is also being approximated in the actual implementation of their method. Their implementation is flexible in terms of the type of selection (e.g. any h value) such that in practice

they also infer all three fitness parameters for diploids (w_{AA}, w_{Aa} and w_{aa}) or the 'selection scheme' (e.g. genic selection vs. recessive selection etc). Being also based on a standard diffusion approximation, this method should be biased for large selection coefficient. In theory, one of the main advantage of this method over our method or Bollback *et al.*'s method is that the mutation rate can also be inferred. The difference in practice likely lies in the actual implementation, the running times vs. accuracy and the ease of use of each of the methods, aspects that have not yet been tested.

Mathieson & McVean (2013) propose another extension of Bollback *et al.*'s method. Specifically, they incorporate population subdivision in their model, a unique feature across methods, noting that populations are generally structured. This enables them to consider spatially varying selection. They consider the case of a two dimensional lattice with $D \times D$ demes and constant migration m between neighbouring demes. This model is of course more complex in terms of the number of parameters because it allows for specific selection coefficients for each deme. To overcome this challenge, they assume that the population size is known, that the allele frequencies never get close to loss or fixation, that the population size is large and they rely on a Gaussian approximation. To account specifically for the more complex structured case, they ignore for example the contribution of s and m to the variance of the change in allele frequencies. They also assume that that they can ignore s^2 terms. The Gaussian approximation and the latter assumption both imply this method will also be valid/unbiased for weak selection. Finally, they apply an efficient expectation maximization algorithm to infer the parameters. This leads to a method that is more attractive in terms of computational time than the methods discussed above (see also Foll *et al.* 2014b).

They make a number of interesting observations through simulations. First considering the case with a single deme, they find that increasing the sample size and the sampling density decreases the error on the estimate of the selection coefficient, while observing a 'saturation effect' (e.g. 100 and 1000 samples seem to result in similar absolute error on the selection coefficient). As expected by Lacerda and Seoighe (Lacerda & Seoighe 2014), increasing the true selection coefficient leads to an underestimation of this parameter. In the structured case, they also observe that high absolute values of s lead to an over and underestimation for negative and positive selection, respectively. The migration parameter is shown to be difficult to estimate. This is especially true when migration is high, in which case estimates of s and m are worse. The estimates are also worse for a small initial frequency presumably because of the risk of loss, violating their assumption of no loss and no fixation. Overall, they conclude that they better estimate the selection coefficient s than the migration rate m, even when fixing the population size. Their power to estimate m comes from the fluctuations around the equilibrium value. Yet, when N_e is large, the fluctuations are too small, while it seems that when N_e is small, the fluctuations due to the finiteness of the sample size confound the estimates. In summary, the method of Mathieson and McVean is likely to produce estimates that are biased downwards if selection is high similar to the two methods above. However, it is computationally more attractive and is the only method that incorporates migration, even though the power to co-estimate migration, selection and population size is likely to be quite low in practice.

Feder *et al.* (2014) develope statistical tests to identify loci under selection that also builds on Bollback *et al.*'s one-locus bi-allelic model. To construct confidence intervals (CI), Bollback *et al.* (2008) relied on the asymptotic result that the ratio of the maximum likelihoods [the likelihood ratio statistic (LRS)] comparing neutral vs. selected models follows a χ^2 distribution. This result is true asymptotically. Feder *et al.* show that, given the finite number of samples one is bound to collect, this result does not hold in practice. More specifically they show that assuming a χ^2 distribution leads to more false positives (i.e. one would conclude there is selection more often than one should) under a realistic sampling scheme. They develop two tests: the empirical likelihood ratio test (ELRT) and the frequency increment test (FIT).

To model the allele frequencies, they assume a Moran model of evolution [closely related to the WF model, see, e.g. (Ewens 2004) for a description] and proceed to approximate it as a deterministic trajectory with added noise. They model noise using a Gaussian process and mention that their approximation is valid when N_e is large and for any selection (i.e. this approximation is different from the one discussed above). This allows them to compute the transition probabilities analytically. Their approximation is not valid close to the boundaries, that is the allele should not get fixed or lost. They assume no sampling noise, that is the frequency of the derived allele in the population is assumed to be the observed allele frequency ($x_j = \frac{i_j}{n_j}$). As noted by Terhorst *et al.* (2015), although they do not apply their method to infer parameters, they could use this approximation and their likelihood framework to do so.

The principle of the first test, the ELRT, is that the population size is estimated assuming neutrality, and the estimated population size is used to simulate data under a neutral WF model. The distribution is then set as the null distribution of the LRS. They show that this

approach provides a very good approximation to the true LRS distribution except when selection is very strong. In this case, the estimate of N_e is small leading to a high probability of fixation and therefore their Gaussian approximation is incorrect. For the FIT, they derive a statistic that depends on the successive changes in allele frequency. They showed that this statistics is distributed according to a Student's t-distribution (a good approximation if the allele frequency does not approach 0 or 1). This approach is computationally very efficient, but the test appears overly conservative, rejecting neutrality in too few cases.

They test their method with simulations and note that the FIT has slightly more power to reject the null hypothesis of neutrality while the power of the test increases with the number of sampling time points. They also investigate the properties of the two tests when there is only a finite number of samples (i.e. a more realistic scheme). They show that the overall power is decreased and both tests become overly conservative. In particular, they show that the power to detect selection is lower than 30% for both tests when a sample size of a 100 is considered compared to 60–80% with infinite sampling. The assumption of infinite sampling (or even of 100 samples) is particularly unrealistic for aDNA where it is hard both to obtain sufficient samples and recover DNA from those samples. However, despite the expected drop in power, it is still likely that this method is not less powerful than the methods described above to detect loci under selection. One should therefore consider the proposed statistics even with low sample sizes.

Lacerda & Seoighe (2014) derive a new approximation to the WF model that is specifically meant for the case of large selection coefficients remaining valid for large effective population sizes. This method is computationally efficient because the transition probabilities involve a Gaussian approximation and can be computed analytically. They test their approximation through simulations but also implement a method similar to the one just described to infer the selection coefficient and the population size. They sample chromosomes for $T = 20$ generations every generation or every fourth generation. They consider very large sample sizes of 1000 and 10 000 (while $N_e = 1000$). They find that they can estimate the selection coefficient with a narrow confidence interval, which becomes only marginally larger for 1000 vs. 10 000 samples or by sampling every generation or every fourth generation. On the other hand, the confidence interval of the population size estimates tends to be very large when the sample size is 1000 and meaningless when both the sampling frequency and the number of samples is decreased. Note that Bollback *et al.* and our method are also unable to co-estimate the population size in a meaningful way (the confidence intervals tend to be very large for realistic sampling schemes). As for Feder *et al.*'s method above, the expected sampling is quite unrealistic for aDNA. However this does not imply that the method has lower power than the ones discussed above.

So far the methods I have discussed are based on a likelihood approach. However, a number of methods consider a Bayesian framework, and more specifically an approximate Bayesian computation (ABC) approach. Those include the work of de Sousa *et al.* (2013) Foll *et al.* (2014a,b) and Sverrisdóttir *et al.* (2014). Sousa *et al.* developed a method focused on experimental evolution data. In particular, they consider two sets of parameters. The rate of new beneficial alleles U and the distribution of selective effects. They assume the selective effects $f(s)$ follow a (gamma) distribution. Allele frequencies of the neutral marker (a proxy for the frequency of the allele under selection) and overall fitness (through, e.g. relative growth experiment) are measured. They simulate data under a WF model with a population size that repeatedly undergoes growth followed by bottlenecks, typical of experimental evolution data. Importantly, the population is clonal and at the beginning of the experiment is polymorphic at a single neutral site with frequency 0.5. From there on, beneficial-only mutations can arise with a rate U and each mutation has a selective effect drawn from a given distribution. They show that they are able to co-estimate the rate of the beneficial mutations and the distribution of fitness effect. Their performance is better with low mutation rate and lower mean selection coefficient and when the distribution of fitness effects has a small variance. This method is probably unrealistic for applications to aDNA. First, the notion of controlled replicates is irrelevant and it is hard to imagine one would be able to collect the required overall fitness data. Yet the idea of the approach is indeed useful. For example, it is feasible that one could collect populations with a recent shared common ancestor experiencing different selective pressures.

Topa *et al.* (2015) present a Bayesian model for single-locus time series data obtained by next generation sequencing. Their method is tailored for experimental evolution data sets of small population sizes, (i.e. they incorporate genetic drift and do not assume that the whole population is sampled). They assume explicitly that the experiment has been replicated and that the draw of alleles follows a binomial distribution, while directional selection is modelled by a Gaussian approximation – the allele cannot be lost or fixed during the sampling period. The allele frequency observations are fitted into a linear regression that can be either time

dependent, if the site is undergoing selection, or not, if the site is neutral. The selected trajectories are expected to be consistent across the replicates. They show through WF simulations that their method performs better with more replicates with higher effective population size, and larger number of founder haplotypes. Moderate and strong selection (0.05–0.1) is easier to detect than weak (0.01) and very strong selection ($s > 0.2$), reflecting the idea discussed above that for a specific sampling set up a certain range of selection coefficient increases the power. In principle, this method could be applicable to aDNA. However, one would expect a single replicate and therefore limited power.

Multilocus

More recently, a number of methods have been proposed that are not limited to single loci. Nishino (Nishino 2013) for example expands on Feder *et al.*'s approach by building a new statistics, the FITR, that incorporates a number R of neutral loci. This statistics, related to the FIT, is a function of the ratio of the change in allele frequency between two time points between the selected locus and the R reference loci. The underlying assumption is that all R loci are independent of the selected locus. Between two time points, the statistics follows the Student's t-distribution and it is hence straightforward to obtain the significance levels. Note that a very appealing feature is that the statistics is independent of the population size. The power of the FIT and the FITR is tested through simulations under several scenarios including cases with a population experiencing a bottleneck or growth event. Partially linked data is also simulated to test for departure from the underlying model.

When enough R loci are considered ($R \in \{10, 20\}$), the FITR performs better than the FIT, especially in the case of a strong bottleneck or rapid growth. In the latter case, the FIT is too conservative and has decreased power. Interestingly and perhaps counter-intuitively, the power of the method saturates with increasing sampling density for FITR. This result differs from what has been discussed earlier, but it might be explained by the fact that the statistic does not depend on the population size, such that, as long as the number of samples is sufficient to capture the increase in the change in allele frequency, added points provide no additional information. In contrast, the power increases when the sampling period T is increased. This might be related to the choice of selection coefficient to simulate the data. One caveat of the method is that one has to select loci that are supposedly neutral. Nishino shows that if the reference alleles are instead under selection, the power is decreased. This is especially true for selection on the reference loci of similar strength than the focal locus. This result is quite intuitive because the FITR is a measure of how different the focal locus is to the reference loci. The power also decreases with fewer sampled chromosomes (n_i). Note though that the number of samples considered varied from 100 to 10 000 for an effective population of $N_e = 10\,000$, a relative ratio which is quite high for aDNA. For most simulations, the reference loci are assumed to be at frequency of 0.5 at the first sampling point, which seems unrealistic. A more realistic approach, as discussed above would require modelling an extra parameter, namely the allele age. Nishino tested the effect of violating this assumption and finds that the rejection rates are affected when the frequency is low, presumably because of how reference alleles get lost under such circumstances and become uninformative thereafter. Finally, Nishino shows that as long as the reference alleles are at linkage equilibrium at the first sampling time, some linkage between the R loci and the focal locus has little impact. This method can be readily applied to aDNA in situations where genomewide data has been collected. The one caveat is that, as said above, it requires predefining which loci evolve neutrally and which are potentially under selection.

Foll *et al.* (2014a,b) developed a method based on ABC (WFABC) to co-estimate the effective population size – assumed to be shared for all L loci – and the selection coefficients where each locus has its own selection coefficient. Their method was originally developed to analyse time series influenza data (Foll *et al.* 2014a). Their approach proceeds in two steps. First, they estimate the distribution of the population size assuming no selection and given time serial data. That distribution is then used in the second step to simulate WF trajectories for each locus with an initial allele frequency matching the observed data. The simulations with the smallest distance to the observed data are then retained. One then gets a posterior estimate of the selection coefficient for each locus and the population size. Loci are therefore assumed to be independent. An important feature of this model is that any ascertainment scheme can be implemented. They simulate data to test their method, and it is essentially the only study to date that includes a comparison with previous methods. In particular, they compare their method with that of Bollback *et al.* (2008), Malaspinas *et al.* (2012) and Mathieson & McVean (2013). As with most of the other methods, they observe that more samples and sampling time points improve the accuracy. To compare their method, they fix the population size to a given value (since by design their method is better to estimate N_e).

They show that their method performs better than all other methods for large selection coefficients, remaining unbiased. In this case, both Mathieson & McVean's (2013) and Malaspinas *et al.* (2012) underestimate the selection coefficient, as predicted by Lacerda & Seoighe (2014). For small selection coefficients, Mathieson and McVean's (2013) method performs better – in the sense of a smaller confidence interval – when no ascertainment is assumed but again worse for a case where only alleles above a certain frequency are kept. Note that Bollback *et al.*'s numerical approximation seems to be unstable, the likelihood function exhibiting several local maxima. As a result, it seems that all simulations lead to the inferred s to be 0. Malaspinas *et al.*'s method seems to perform better when selection is weak (smaller confidence interval and less bias). They also show that the accuracy improves as N_e increases for all methods. Finally they demonstrate that their method is also computationally more efficient. It seems therefore fair to conclude the Foll *et al.*'s method should be used in cases where the selection coefficient is large, when genomewide data is available, or when the ascertainment is known precisely. Even for weak selection, Foll *et al.*'s method should be used when the effective population size or the candidate mutations are not known. Mathieson and McVean's method should be used when one is interested in inferring the migration rate, while Malaspinas *et al.*'s method should be used in the case when one seeks to infer the allele age as well (but see also Steinrücken *et al.* 2014).

I will conclude this method overview by briefly mentioning a few methods that have been proposed to tackle experimental evolution data sets, such as evolve and resequence datasets in *Drosophila*. Terhorst *et al.* (2015) approximate a multilocus WF model using a deterministic path with added Gaussian noise, similar to what has been proposed by Feder *et al.* (2014) but extended to multiple loci. To compute the likelihood, they simplify the computation by considering WF models for a small number of loci at a time. Their method is unique in accounting for the effects of linkage between loci and assuming genetic drift. They test their method through simulations and report that their power to detect selection is higher for larger selection and lower for a high number of starting founding lineages (H). Note that the notion of founding lineages is specific to experimental evolution where one can control for the composition of the starting population. Presumably this is explained by the fact that many sites are segregating at low frequency and potentially getting lost by drift for high H. Increasing the population size allows one to detect lower levels of selection, as expected, because the confounding effect of genetic drift diminishes (see also Fig. 2). While their test has less power to detect sites under selection with a high number of founding lineages, increasing H helps locating the site under selection. They note therefore that there is a trade-off between having enough founding lineages to infer the target site and not too many to be able to detect selection all together. They also observe that using a multilocus model leads to only limited improvement in accuracy. In fact, only when the number of chromosomes sampled is finite/small (low depth) does their multilocus approach improve the estimate, by decreasing its variance. They also note that their method can estimate hs but not h alone, while recombination can be estimated for strong hotspots. Finally, they mention that excluding alleles whose frequency is close to 0 or 1 increases the power. Thus, this is the only method to date for tackling linked loci and genetic drift in a time serial context. However, their method is implemented for pooled data, a type of data unlikely to become common in aDNA.

Illingworth & Mustonen (2011) have developed several related methods to quantify selection for large populations evolving quasi-deterministically. Evolution is modelled by a set of differential equations describing – in its simplest form – two types of mutations: the 'drivers' – mutations that are under selection – and passengers – mutations that are linked to the drivers. Their methods are applicable to organisms with fairly large population sizes (e.g. yeasts), where the stochasticity of the observations arise from the finite sampling in chromosomes. They expect selection to be strong and the initial allele frequencies to be high. The method is computationally very appealing in terms of speed. However, the authors show through simulations that a fairly large population size is needed to avoid a high rate of false positives ($> 10^7$) making it unpractical for most aDNA studies.

Finally, I conclude this section by mentioning a simulation study by Baldwin-Brown *et al.* (2014). Although the authors do not present a method for parameter inference, their results allow one to build intuition on the parameters affecting the power of time series analysis, in that they present a very thorough simulation framework to design time series experiment to achieve maximal power. They consider an experimental evolution setting, and more specifically an E&R experiment. They show that the parameters that primarily affect the results are the selection coefficient, the population size, the number of replicate populations and the number of founding haplotypes. They demonstrate in particular that for a selection coefficient of $s = 0.005$, their power to detect regions containing selected SNPs is essentially null, even for their largest population size (1000) and a low number of founding lineages. As reported by Terhorst *et al.* (2015), the power to localize accurately the SNP of interest increases with a larger number of

founding lineages (H). The number of generations T where the system undergoes selection is also a factor influencing power but in a more complex way, presumably because of the loss of power when the allele is fixed or lost. Indeed intermediate values of T seem to provide better results. They conclude with a set of recommendations in terms of experimental design. For example, to detect more than 80% of the causative regions, and to detect $s = 0.05$ and above, researchers should consider 25 replicates, evolve their populations for 500 generations and consider a population size of 1000. If the idea is also to localize the SNP of interest, then one should consider at least 500 starting haplotypes for $s >= 0.01$ and $r > 25$ and a population size of 1000 and above. However, with their set up, $s = 0.005$ seem all together unachievable. These ideal conditions remain quite unrealistic for aDNA. These recommendations therefore suggest that – unless selection is very high – it will be very difficult to detect regions under selection and even harder to localize SNPs of interest.

Practical considerations and future directions

Evolution is driven by four fundamental processes, namely genetic drift, migration, natural selection and mutation. Despite having been described for over a century, their relative importance is yet to be precisely determined. Time serial data provide a direct window into the past, and yields great promise to address this challenge. With recent advances in ancient genomics, population level data collected at several time points separated by several generations are now accessible for higher organisms. In this review, I have discussed some existing methods and their limitations to characterize selection acting at the genomic level. I discussed some of the key parameters that come into play to maximize the power to detect selection. Specifically, it is easier to detect selection for populations with large N_e ($\sim 10^7$, to minimize the effect of genetic drift), making it all the more challenging to distinguish selection from genetic drift for organisms with small population sizes, including most mammals with $N_e \sim 100, 10\,000$. Moreover, and somewhat counter-intuitively, although in principle strong selection is easier to detect than weak selection, for a specific sampling period (in generations), there will be a selection coefficient above which power decreases. Increasing the number of time points and the number of sampled chromosomes generally increases power as well, while in some cases it has been shown that it is better to increase the number of sampling time points rather than the number of samples per time point if the total number of samples is kept constant.

With the data currently available, estimating population sizes from aDNA has proven virtually impossible for most methods, with quite often the confidence interval spanning the whole range of the parameter search. The use of information from several loci helps in that regard. Moreover, although estimates of selection can be unbiased, the confidence intervals are large and often weak positive selection cannot be distinguished from neutrality. Overall, as most of the time serial data accessible through aDNA will be sampled for populations with small population sizes, and as it is unclear what the mean selection coefficient should be, inferring specific selection coefficients is a challenging task and one can expect that only loci undergoing strong selective sweeps will be detectable with current approaches. In this review, I considered mostly methods based on single or multiple loci through change in allele frequency data. Yet there is potential to modify and apply approaches traditionally implemented for modern data (for a review, Nielsen *et al.* 2007). For example, haplotype-based methods, site frequency-based methods or even tests based on population subdivision could be extended to exploit time serial aspect of the data.

It is important to note that, while most of the discussed methods do mention aDNA as a potential application, and in three cases are directly applied to an aDNA data set, none of the methods are tailored to take into account explicitly two main characteristics of aDNA, namely the high missingness (since the DNA is highly fragmented) and the high error rate (since the DNA is damaged). The underlying assumption is that the data is perfect (no missing data or error). This of course leaves the door open for a number of obvious improvements in terms of method development. The high missingness of the data can be already somewhat alleviated from an experimental perspective using methods relying on the capture of specific regions (as in, e.g. da Fonseca *et al.* (2015)) while accounting for damage should be tackled in the future using genotype likelihoods. Moreover, for the methods relying on single-locus allele frequency data, an easy work around for the high missingness is to consider the observed frequency in the sampled data (i.e. assume that the data missing is not biased in terms of which allele is missing, which is a rather realistic assumption).

The field of ancient population genomics is experiencing an exponential increase in terms of the amount of data available, accompanied by an increase in available tailored statistical methods. While the current methods I have reviewed still have limited power to distinguish genetic drift from selection for realistic sampling schemes, the power will increase as more data and methods become available. I therefore anticipate that, while traditionally limited by biochemical challenges, the aDNA field will soon become a science mostly driven by evolutionary hypotheses.

Acknowledgements

I would like to thank the Swiss National Swiss Foundation (SNFS) for funding, Clio Der Sarkissian, Johannes Krause, Michael Hofreiter, Morten Allentoft, Vitor Sousa, Lars Boshard, Andy Foote and Tom Gilbert for a number of very helpful discussions.

References

Allentoft ME, Collins M, Harker D *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B. Biological Sciences*, **1748**, 4724–4733.

Allentoft ME, Sikora M, Sjögren KG *et al.* (2015) Population genomics of Bronze Age Eurasia. *Nature*, **522**, 167–172.

Anderson EC, Williamson EG, Thompson EA (2000) Monte carlo evaluation of the likelihood for Ne from temporally spaced samples. *Genetics*, **156**, 2109–2118.

Baldwin-Brown JG, Long AD, Thornton KR (2014) The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Molecular Biology and Evolution*, **31**, 1040–1055.

Beaumont MA (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Berthier P, Beaumont MA, Cornuet J-M, Luikart G (2002) Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics*, **160**, 741–751.

Bollback JP, York TL, Nielsen R (2008) Estimation of 2Nes from temporal allele frequency data. *Genetics*, **179**, 497–502.

Crisci JL, Poh Y-P, Bean A, Simkin A, Jensen JD (2012) Recent progress in polymorphism-based population genetic inference. *Journal of Heredity*, **103**, 287–296.

Darwin C (2012) *The Origin of Species*. William Collins, Harper-Collins, UK.

Der Sarkissian C, Ermini L, Schubert M *et al.* (2015) Evolutionary genomics and conservation of the endangered Przewalski's Horse. *Current Biology*, **25**, 2577–2583.

Ewens WJ (2004) *Mathematical Population Genetics 1: Theoretical Introduction*, 2nd edn. Springer, New York.

Feder AF, Kryazhimskiy S, Plotkin JB (2014) Identifying signatures of selection in genetic time series. *Genetics*, **196**, 509–522.

Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D (2015) An Approximate Markov Model for the Wright-Fisher Diffusion. bioRxiv, 030940. URL:http://biorxiv.org/content/early/2015/11/08/030940.full-text.pdf+html.

Fisher RA, Ford EB (1947) The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity*, **1**, 143–174.

Foll M, Poh Y-P, Renzette N *et al.* (2014a) Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genetics*, **10**(2), e1004185.

Foll M, Shim H, Jensen JD (2014b) WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, **15**, 87–98.

da Fonseca RR, Smith BD, Wales N *et al.* (2015) The origin and evolution of maize in the Southwestern United States. *Nature Plants*, **1**, 14003.

Gamba C, Jones ER, Teasdale MD *et al.* (2014) Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, **5**, 5257. doi: 10.1038/ncomms6257.

Hamilton JD (1994) *Time Series Analysis*, 1st edn. Princeton University Press, Princeton, New Jersey.

Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S (2001) Ancient DNA. *Nature Reviews Genetics*, **2**, 353–359.

Hummel S, Schmidt D, Kremeyer B, Herrmann B, Oppermann M (2005) Detection of the CCR5-Delta32 HIV resistance gene in Bronze Age skeletons. *Genes and Immunity*, **6**, 371–374.

Illingworth CJR, Mustonen V (2011) Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*, **189**, 989–1000.

Illingworth CJR, Parts L, Schiffels S, Liti G, Mustonen V (2012) Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular Biology and Evolution*, **29**, 1187–1197.

Illingworth CJR, Fischer A, Mustonen V (2014) Identifying selection in the within-host evolution of influenza using viral sequence data. *PLoS Computational Biology*, **10**, e1003755. doi: 10.1371/journal.pcbi.1003755.

Jaenicke-Després V, Buckler ES, Smith BD *et al.* (2003) Early allelic selection in maize as revealed by ancient DNA. *Science*, **302**(5648), 1206–1208.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends in Genetics*, **23**, 568–577.

Krimbas CB, Tsakas S (1971) The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control-selection or drift? *Evolution*, **25**, 454–460.

Lacerda M, Seoighe C (2014) Population genetics inference for longitudinally-sampled Mutants under strong selection. *Genetics*, **198**, 1237–1250.

Lazaridis I, Patterson N, Mittnik A *et al.* (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513**, 409–413.

Ludwig A, Pruvost M, Reissmann M *et al.* (2009) Coat color variation at the beginning of horse domestication. *Science*, **324**, 485.

Ludwig A, Reissmann M, Benecke N *et al.* (2015) Twenty-five thousand years of fluctuating selection on leopard complex spotting and congenital night blindness in horses. *Philosophical Transactions of the Royal Society of London B. Biological Sciences*, **370**, 20130386.

Malaspinas A-S, Malaspinas O, Evans SN, Slatkin M (2012) Estimating allele age and selection coefficient from time-serial data. *Genetics*, **192**, 599–607.

Mathieson I, McVean G (2013) Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, **193**, 973–984.

Mathieson I, Lazaridis I, Rohland N *et al.* (2015) Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, doi: 10.1038/nature16152. [Epub ahead of print].

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nielsen R, Slatkin M (2013) *An Introduction to Population Genetics: Theory and Applications*, 1st edn. Sinauer Associates Inc, Sunderland, Massachusetts.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Nishino J (2013) Detecting selection using time-series data of allele frequencies with multiple independent reference loci. *G3 (Bethesda)*, **3**, 2151–2161.

Orlando L, Ginolhac A, Zhang G *et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499**, 74–78.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N. (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.

Pollak E (1983) A new method for estimating the effective population size from allele frequency changes. *Genetics*, **104**, 531–548.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Provine WB (2001) *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago.

Raghavan M, DeGiorgio M, Albrechtsen A *et al.* (2014a) The genetic prehistory of the New World Arctic. *Science*, **345**, 1255832.

Raghavan M, Skoglund P, Graf KE *et al.* (2014b) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, **505**, 87–91.

Rasmussen M, Anzick SL, Waters MR *et al.* (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, **506**, 225–229.

Schubert M, Jónsson H, Chang D *et al.* (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, **111**, E5661–E5669.

Skoglund P, Malmstrom H, Omrak A *et al.* (2014) Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*, **344**, 747–750.

de Sousa JAM, Campos PRA, Gordo I (2013) An ABC method for estimating the rate and distribution of effects of beneficial mutations. *Genome Biology and Evolution*, **5**, 794–806.

Song YS, Steinrücken MA (2012) Simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, **190**, 1117–1129.

Steinrücken M, Bhaskar A, Song YS (2014) A novel spectral method for inferring general diploid selection from time series data. *Annals of Applied Statistics*, **8**, 2203–2222.

Sverrisdottir OO, Timpson A, Toombs J *et al.* (2014) Direct estimates of natural selection in iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Molecular Biology and Evolution*, **31**, 975–983.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Terhorst J, Schlötterer C, Song YS (2015) Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genetics*, **11**, e1005069.

Topa H, Jonas A, Kofler R, Kosiol C, Honkela A (2015) Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, **31**, 1762–1770.

Wakeley J (2008) *Coalescent Theory: An Introduction*, 1st edn. Roberts and Company Publishers, Greenwood Village, Colorado.

Wallace AR (1858) On the tendency of varieties to depart indefinitely from the original type *Proceedings of the Linnean Society of London*, **3**, 53–62.

Wang J (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research*, **78**, 243–257.

Wilde S, Timpson A, Kirsanow K *et al.* (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, **111**, 4832–4837.

Williamson EG, Slatkin M (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, **152**, 755–761.

Wright S (1948) On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution*, **2**, 279–294.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

A.-S.M. planned and wrote the review.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1. Range of values for some of the parameters used in the simulations of each study (whenever applicable).