Can ancient DNA and other forms of time-sampled data aid in the inference of negative frequency dependent selection?

Vivak Soni

Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA

vsoni11@asu.edu

Abstract

Negative frequency-dependent selection (NFDS) is commonly viewed as the most efficacious form of balancing selection. Despite this, inferring NFDS remains challenging, and questions remain as to its relative importance in maintaining genetic variation in populations. Recent advances in both sequencing and genotyping technologies has resulted in a considerable increase in the number of publicly available human ancient DNA datasets, creating new opportunities for development of methods for the inference of NFDS from time-sampled data. In this perspective, I present three brief simulation studies to show how time-sampled data can aid improve inference power. First, I show how multiple time points can help us distinguish between recent NFDS and partial selective sweeps based on allele frequency trajectories. I then demonstrate how selective effects can be distinguished from population history based on changes in genetic variation and the site frequency spectrum over time. Finally, I apply an approximate Bayesian computation approach to compare the power of multiple and single time point datasets in estimating the time for which NFDS has been shaping variation for. Thus, I show how data from multiple timepoints can facilitate the generation of new methodological approaches for better inference of NFDS.

Introduction

Before advances in genomics allowed direct measurement of genetic variation across populations, a key debate in population genetics revolved around whether such variation was expected to be minimal or substantial. This controversy, known as the *classical/balanced debate* [1,2], largely focused on the role of selection—either purifying selection reducing variation or balancing selection preserving it [3]. However, even as next-generation sequencing later revealed abundant polymorphism in natural populations, the explanation shifted away from widespread balancing selection. Instead, the Neutral Theory of Molecular Evolution [4,5] gained prominence, proposing that much of the observed variation consisted of neutral alleles drifting randomly toward fixation or loss. This view has since been widely supported by empirical evidence [6]. Today, the extent to which balancing selection maintains genetic variation remains an open question [7,8], due in part to the challenges in detection, particularly on timescales that are neither extremely recent nor extremely long term (i.e. >25N generations, where N is the population size [9]). Indeed, estimates of the number of sites experiencing balancing selection have been limited to functional regions and are subject to a high level of uncertainty [10,11]. Despite evidence suggesting its role is more limited than once thought, balancing selection has been linked to a number of sites implicated in important functions, including sex determination [12], plant self-incompatibility [13-15], and the major histocompatibility complex (MHC) in vertebrates [16-18].

The term balancing selection encapsulates a number of selective mechanisms that maintain genetic variation, including negative frequency-dependent selection (NFDS), heterozygote advantage and spatial or temporal heterogeneity. Of

these, NFDS - the process by which the relative fitness of a variant is inversely proportional to its frequency in a population ([19], and see [20,21] for the first analytical treatments) – is commonly viewed as the most efficacious selective process maintaining balanced polymorphisms [22-27]. NFDS preserves genetic polymorphism in populations by favoring rare variants, which gain a selective advantage over common ones. Thus, rare alleles increase in frequency and persist rather than being lost from the population.

Although NFDS has been observed experimentally (e.g.[28.29]) and from analyses of phenotypic changes (e.g. [30-32]), distinguishing NFDS from other forms of balancing selection using population genomic data remains a challenge. This is because multiple forms of balancing selection leave similar genomic signatures [18] over differing timescales. The initial phase following the introduction of the selected mutation involves a rapid increase in frequency over time in the population, as the selected allele sweeps to its equilibrium frequency. Such partial sweep patterns can result in an excess of intermediate frequency alleles and extended linkage disequilibrium (LD – the correlation among alleles from different loci), as well as extended haplotype structure (see reviews of [33,34]) and weaker genetic structure [35]. These signals are fleeting however, as recombination breaks up haplotype and LD patterns, potentially resulting in reduced detection power [9,12,36,37] (and see reviews of 38,39]). At this point there is what Soni and Jensen [9] describe as a “large temporal gap”, before there is power to detect balancing selection using site frequency spectrum (SFS) based approaches. It is only once the balanced mutation has been segregating for considerable evolutionary time (>25*N* generations) that we have strong power to detect NFDS using SFS-based methods, as new mutations

have accrued on the balanced haplotype, resulting in a skew in the SFS toward intermediate frequency alleles and an excess of variation in the neighbourhood of the selected locus. Finally, trans-species polymorphisms can be an informative signal of NFDS in species whose expected coalescent time is predated by their divergence time [40,41].

Despite these numerous signatures, a number of other neutral and selective processes can result in highly similar genomic patterns and thereby confound inference (see Table 1 of [38]). Thus, given the temporal gap, the confounding effects of other population genetic processes, and that observed levels of variation in numerous populations have been explained without positive or balancing selection (e.g. [42-46]), inference of NFDS remains challenging. However, time-sampled datasets in which we have access to genomic information over multiple time points - not just the present day – have the potential to improve inference of NFDS and other forms of balancing selection. Although time-sampled data of viral genomes has been available for some time (e.g.[47-49]), recent advances in both sequencing and genotyping technologies, as well as in protocols for handling degraded DNA from archaeological material (also known as ancient DNA, or aDNA), has resulted in a considerable increase in the number of publicly available human aDNA datasets, as exemplified by the Allen Ancient DNA Resource (AADR), a database of 12,761 ancient genomes, ranging across 10 time points, from hunter gatherers through to antiquity [50]. This glut of new data creates new opportunities for the development of methods for the inference of NFDS. In this perspective I explore these opportunities, identifying signals of NFDS that can be inferred across multiple time points. I demonstrate the potential increase in inference power across multiple time points via

simulation, which will hopefully yield fruitful avenues for future studies on empirical time sampled data.

Inferring recent NFDS from population genomic data

Three challenges exist when attempting to infer very recent NFDS. The first is that NFDS is likely often a transient process because this form of selection favours rare alleles, and therefore other evolutionary processes can result in strong fluctuations in allele frequencies and the loss of allelic lineages. For example, Ejsmond and Radwan [51] showed via simulation that population bottlenecks can result in a severe and rapid loss of variation in genes evolving under NFDS. Thus, this form of balancing selection can be fleeting due to the increased probability of stochastic loss of the selected loci at low allele frequencies. The second challenge when detecting NFDS is that the characteristic SFS-based signatures of NFDS – increased variation and a skew in the SFS toward intermediate frequency alleles – are absent until the balanced mutation has been segregating for long enough for the balanced haplotype to accrue variation (>25*N* generations [9]). A mutation under NFDS that has escaped stochastic loss follows the initial trajectory of a partial selective sweep, and distinguishing between these two selective processes is the third challenge when inferring very recent NFDS. A newly introduced beneficial mutation will rapidly increase in frequency, and linked variation may increase in frequency along with it [52]. This process will increase LD and therefore a number of tests of selection have been developed to detect this pattern of increased non-random association between alleles (e.g. the *EHH* [extended haplotype homozygosity –[53]] and *iHS* [integrated haplotype score – [54]] statistics). Though this beneficial trajectory will be affected by details such as the strength of selection

acting on the beneficial mutation, the general trajectories are equivalent for both selective sweeps and NFDS, at least until the balanced mutation reaches its equilibrium frequency, meaning that NFDS and partial sweeps of positive selection cannot be easily distinguished from one another [12,55].

One avenue for addressing this problem is through tracking temporal changes in allele frequencies, and thus distinguishing very recent NFDS from partial sweeps of positive selection. More specifically, allele frequency information from multiple time points might enable researchers to examine how the beneficial mutation trajectories between selective sweeps and NFDS diverge once the balanced mutation reaches its equilibrium frequency. To demonstrate this approach, I ran simulations in SLiM v4.2.3 [56] of a single equilibrium population, with a beneficial mutation occurring on a neutral background, across 100 simulation replicates. I simulated two selective sweep regimes, with population scaled strengths of selection of $2Ns$ = 100 and 1,000 (where N is the population size of 500 individuals, and s is the selective advantage of the mutant allele relative to the wildtype), and a single NFDS regime in which the mutation under NFDS had an equilibrium frequency of 0.5 (see the Methods section for details of simulation framework). I sampled the population every five generations, starting from the introduction of the beneficial allele, until it had been segregating for $1N$ generations. Figure 1a shows the mean frequency of the beneficial mutations across the 100 simulation replicates, whilst Figure 1b and 1c show mean Tajima's D [57] and mean haplotype diversity respectively. These summary statistics were calculated across sliding windows of size 10kb, with a step size of 5kb. From Figure 1a it is clearly visible where the beneficial mutation trajectories diverge. Whilst the allele under NFDS sweeps to its equilibrium frequency and then fluctuates about that

frequency (0.5 here), the selective sweeps continue to increase in frequency toward fixation. This process occurs more rapidly at the higher strength of selection (2*Ns* = 1,000), though it is notable that the allele under NFDS also initially increases in frequency extremely rapidly, due to the extremely strong selection acting on it at very low frequencies. As its frequency increases, the rate of frequency change slows down.

We can see the genomic signatures of these different selective regimes in Figures 1b and 1c. There is an initial increase in Tajima's *D* around the selected locus immediately after the introduction of the beneficial mutation. A more positive Tajima's *D* is caused by a skew in the SFS toward intermediate frequency alleles. Once a beneficial mutation has escaped stochastic loss, it will immediately increase in frequency, taking linked neutral variation with it, causing this resultant skew in the SFS. Likewise, there is an initial increase in haplotype diversity with this increase in allele frequencies, as shown in Figure 1c. However, these patterns diverge as the allele under NFDS reaches its equilibrium frequency. The continuing increase in frequency of the selective sweeps and their linked variation results in an increase in rare and high frequency alleles, and thus skews the SFS toward rare alleles, as evidenced by the reduction in Tajima's *D*. Haplotype diversity is much reduced too, as the selective sweeps remove variation. By contrast, the balanced haplotype maintains variation at the equilibrium frequency, and thus both Tajima's *D* and haplotype diversity remain relatively consistent whilst the balanced mutation is segregating. Indeed, both Tajima's *D* and haplotype diversity will increase as the balanced mutation segregates for longer, as new mutations accrue on the balanced haplotype (see Figure 1 of[9]). Thus, if we have access to population genomic data

that spans the initial phase of the beneficial mutation once it has escaped stochastic loss, we may be able to distinguish between selective sweeps and NFDS.

It is important to note that one cannot simply apply these summary statistics to time-sampled data and identify recent NFDS. For instance, although the temporal evolution of the SFS helps us differentiate between selective sweeps and balancing selection, we know that there is little power to identify very recent balancing selection from the SFS alone [9] (and see reviews of [38,39]). Methods for detecting recent NFDS from time-sampled data would therefore need to harness linkage and haplotype information to identify candidate regions, and only then can allele frequency information be utilized for distinguishing between partial sweeps and NFDS. With careful use of multiple aspects of population genomic data, there is considerable potential for inferring recent NFDS from time-sampled datasets.

Utilizing time-sampled data to account for the biasing affects of population history in NFDS inference

It is well documented that population history may confound the inference of selective processes (e.g. [58-69]). Soni and Jensen [9] showed how a reduction in population size can confound inference of balancing selection due to the skew in the SFS toward intermediate frequency alleles that both processes induce. Concurrently, many segregating variants might become fixed or lost, potentially breaking up the balanced haplotype. To account for population history – as well as other constantly acting evolutionary processes such as purifying and background selection (BGS) effects, and mutation and recombination rate heterogeneity, Johri et al. [70,71] proposed the necessity of an evolutionary baseline model incorporating these

processes prior to inferring comparatively rare or episodic processes such as positive and balancing selection. However, even with such baseline models, Soni and Jensen [9], have shown that power to infer balancing selection will be reduced under certain population histories.

To improve inference power when inferring NFDS under non-equilibrium population histories, we might attempt to perform genome scans across multiple time points, looking for consistent patterns across time. SFS-based inference approaches (e.g. [10,72]) detect the skew in the SFS generated by balancing selection, and the changes in the amount of genetic variation. If this signal holds in a genomic window across multiple time points, this might provide additional support for balancing selection. To demonstrate how these genomic signals might be maintained across multiple time points, I simulated a neutral region in a single population under an instantaneous population contraction, such that the population size was reduced to 10% of its initial size. I sampled this population multiple times, both prior to the contraction, and up to $1N_{ancestral}$ generations after the population size change, where $N_{ancestral}$ was the initial population size of 10,000 individuals. I also simulated a functional region, modelling purifying and BGS, in which a single mutation under NFDS was introduced $75N_{ancestral}$ generations prior to the population contraction (i.e. such that the balanced mutation had been segregating for long enough for SFS-based methods to have considerable power to detect NFDS). This population was sampled at the same timepoints as the neutral region simulations. I then calculated θ_w and Tajima's *D* (Tajima 1989) across 10kb sliding windows with a 5kb step size for each time point. For full details of the simulation set up, see the Methods and Materials section. Figure 2 shows the changes in θ_w and Tajima's *D* across time,

both under neutrality, and for the functional region. NFDS elevates both statistics, as it increases the level of variation and the relative number of intermediate frequency alleles due to rare and high frequency alleles being lost or fixed respectively. The population contraction results in a reduction in θ_w as variation is lost from the population. The smaller population does not recover to the levels of variation prior to the size change. However, NFDS maintains a level of variation elevated above that of the neutral expectation. Conversely, the elevation in Tajima's *D* due to the population contraction is lost after enough time has passed under neutrality, as the number of rare and high frequency alleles increase in the population. However, this is not the case under NFDS, where the elevation in Tajima's *D* is maintained due to mutations on the balanced haplotype being maintained at intermediate frequencies. Thus, we would expect to see consistent patterns across time points in genome scans, providing increased support for candidate regions under NFDS.

It is important to emphasize the need for an evolutionary baseline model [70,71] to generate a "null" expectation. The example given here would be relevant for humans and other species with coding-sparse genomes, where we have access to non-functional regions distant enough from any coding region to avoid the biasing effects of purifying and BGS [73], thereby modelling neutral processes independently of selective effects. With a population history inferred on these non-functional regions, we can infer the DFE of new mutations on functional regions, conditional on the demography already inferred. Other species such as viruses – in which the entire genome may be functional – would necessitate jointly modelling neutral and selective effects [42]. Once we have our baseline model, we can simulate under this model (sampling at the same time points that we have in our empirical data) and

perform inference of NFDS on this model to infer null thresholds for inference, reasoning that these are the most extreme values of our statistic of interest that we can generate in the absence of NFDS. This process would be necessary for each sampled time point, mirroring our empirical data. Thus, if we identify regions that exceed these values in our empirical data, these are potential candidates for NFDS, and this can be verified across multiple time points.

Inferring the segregation time of a mutation under NFDS

A parameter of interest when performing inference for NFDS is how long the balanced mutation has been segregating for (hereafter T_b), because this parameter provides information on the duration of time for which NFDS has been shaping variation. However, we are generally limited to broad temporal categories based on the detectable signature. For example, Fijarczyk and Babik [38] categorize recent balancing selection as that in which the balanced mutation has been segregating for $<0.4N_e$ generations, whilst intermediate balancing selection is $0.4\text{-}4N_e$ generations old, and ancient balancing selection is $>4N_e$ generations old. Bitarello et al. [39] provide a different but generally compatible schema, with recent balancing selection being $<4N_e$ generations old, long-term balancing selection $>4N_e$ generations old, and ultra long-term balancing selection being $4N_e + T_{div}$ generations ago (i.e. longer than the expected coalescence time between lineages present in species that diverged T_{div} generations ago). Because different ages of selected alleles generate distinct signatures, we are able to infer the timescale on which balancing selection has been acting. A question then remains as to whether we can narrow down this temporal range further.

To answer this question, I used an approximate Bayesian computation (ABC) approach to attempt to infer T_b from simulated data in which a mutation under NFDS had been segregating for $10N_{ancestral}$ generations, with the aim of comparing inference power for a single time point, and multiple time points. I sampled the data at the end of this segregation time (which would be the only time point for the single time point analysis), as well as $0.02N_{ancestral}$ and $0.045N_{ancestral}$ generations prior, with the three time points together forming the multi time point data. These time points were chosen as they fall within the AADR aDNA dataset [50] and are therefore reasonable timescales for which we have population genomic data in humans. This "empirical" data was generated for four different population histories: equilibrium (no population size change); expansion (where the population size instantaneously doubled); contraction (where the population size instantaneously halved); and severe contraction (where the population size was instantaneously reduced by 90%). The Methods and Materials section contains details of how this data was generated. For ABC inference, 3,000 values of T_b were drawn from a uniform distribution with bounds of 1 and 100 $N_{ancesltral}$ and simulated in SLiM v4.2.3 [56]. Summary statistics were calculated across 10kb windows with a 5kb step size, with the mean and standard deviation for each window utilized as a summary statistic for ABC inference. Although any number of summary statistics might be used for inference with ABC, I used Tajima's D [57], θ_w, haplotype diversity, mean r^2, and the number of singletons, given that these statistics together capture different aspects of the data. This created a total of 60 statistics for ABC inference from the single time point data, and 180 statistics for the multiple time point data.

Figure 3 shows the posterior distribution of T_b for the four population histories, both for the single and multiple time point data. In each case, the inference based on a single time point results in a wide posterior distribution. By contrast, the posterior distributions from multiple time points are much more informative, with modal values either over or very close to the true value of T_b of $10N_{ancestral}$ generations. There is a slight but noticeable overestimate of T_b in each case, and this may be explained by the fact that no negative values of T_b were drawn (i.e. prior to the introduction of the mutation under NFDS), and thus the posterior distributions are skewed somewhat toward higher values of T_b. It is clear however, that data from multiple time points yields more informative posteriors (and thus more information on T_b) than data from a single time point, and therefore the increasing prevalence of time sampled data may facilitate more precise estimates of how long a mutation under NFDS has been segregating for.

An important caveat here is that this analysis focuses on a functional region in which a mutation under NFDS is segregating, under the assumption that this functional region has already been identified as a candidate region for NFDS. It is also assumed that due diligence has been performed in terms of inferring an evolutionary baseline model (i.e. the demographic history, DFE, and mutation and recombination rates are known). With real world data, these processes will have to be inferred prior to inference of T_b. It is encouraging however, that inference of T_b from multiple time point data is accurate even when the mutation under NFDS has been segregating for just $10N_{ancestral}$ generations, given that this timeframe sits within the temporal gap between $1N$ and $25N$ generations identified by Soni and Jensen [9]. The drawback of such ABC approaches is that they are computationally

expensive, particularly when exploring a large parameter space and under complex population histories.

Concluding thoughts

The simulation studies presented here illustrate the potential of time sampled data to radically improve inference of NFDS, as well as other forms of balancing selection from population genomic data. Whereas the signature of a hard selective sweep is fleeting [74,75], balancing selection can persist across both short and long timescales. For examples, Laval et al. [76] found that the haemoglobin β^S sickle mutation – a textbook case of natural selection maintaining a deleterious mutation at high frequencies via balancing selection – initially arose in humans ~22,000 years ago. At the other extreme, Leffler et al. [41] found multiple instances of polymorphisms shared by humans and chimpanzees being maintained by balancing selection, indicating that the balanced mutation has been segregating at least since before the human chimpanzee split time 7-13 million years ago [77]. As I have shown in this study, time sampled data may help researchers to more precisely date how long NFDS has been acting on a mutation, as well as increasing power to tease apart neutral and selective processes.

As with any simulation study however, an important qualification is that this is "perfect" data with many assumptions that will not necessarily hold in natural populations. For example, I have assumed no mutation and recombination heterogeneity, which will impact inference of NFDS, and have modelled a single isolated population, when population structure is a well-known confounder of balancing selection inference. Indeed, hidden structure – where multiple populations

are assumed to represent just a single population – results in a considerably higher false positive rate when performing balancing selection inference [9]. When handling aDNA there are a number of further complications to consider. Though DNA can preserve for hundreds of thousands of years under favorable conditions (e.g [78,79]), genetic material decays progressively through time following the death of an organism, due to cellular repair functions no long functioning postmortem [80,81]. Humidity, temperature, salinity, pH, and microbial growth all influence DNA preservation. The extent and nature of their effect on DNA preservation will vary depending on the archaeological site and stratum in which they are located (see reviews of [82,83]). A further problem is that of DNA contamination. Contamination can occur both from the presence of microbial and environmental sources [84], as well as human DNA introduced during sample extraction or laboratory processing [85-87]. Methods such as that of [87] have been developed to quantify the proportion of present-day DNA contamination in aDNA datasets, and thus any attempt to infer population genetic processes from aDNA datasets must model the effects of DNA damage and contamination where possible, and model the possible inference impact of the uncertainty due to aDNA damage by drawing error rates from an experimentally-informed range to account for the biasing effects such damage and contamination might have on downstream inference.

Thus, by constructing a baseline model that accounts for both the specific nature of the data, as well as the constantly occurring population genetic processes such as genetic drift, purifying and BGS, and mutation and recombination rate heterogeneity, we can harness the extra information from time sampled data to distinguish NFDS from other selective and neutral processes, as well as quantify

how long NFDS has been maintaining variation for. Because NFDS is constantly shaping variation (for example, increasing the skew in the SFS as new mutations accrue on the balanced haplotype), even a small number of closely grouped time points provide increased resolution into detection of this process. The increasing prevalence of aDNA and time sampled datasets in other organisms can spur development of new methods for inference of NFDS and other population genetic processes.

Methods and materials

Generalized simulation framework

All simulations were run forward-in-time using SLiM v.4.2.3 [56], using parameters previously inferred in humans. 100 replicates were simulated for any given parameter combination or simulation regime. All simulations had an initial burn-in time of $10N_{ancestral}$ generations, where $N_{ancestral}$ was the initial population size of 10,000 individuals [88] unless otherwise stated. A single population was simulated with a fixed mutation rate of 2.5×10^{-8} per base pair per generation [89], and a fixed recombination rate of 1×10^{-8} per base pair per generation [90]. Negative frequency-dependent selection was modelled such that the selection coefficient of the balanced mutation, S_{bp}, was dependent on it's frequency in the population: $S_{bp} = F_{eq} - F_{bp}$, where F_{eq} is the equilibrium frequency of the balanced mutation, and F_{bp} is the frequency of the balanced mutation. The dominance coefficient, h of the balanced mutation was 0.5. Simulation replicates in which the balanced mutation failed to establish by reaching a frequency of 0.1 were discarded and restarted.

Selective sweep and NFDS trajectory simulations

A population of $N_{ancestral}$ = 500 individuals was simulated on a 100kb neutral background. This smaller population size was used for computational efficiency. A single beneficial mutation was introduced after the $10N_{ancestral}$ burn-in. For selective sweep simulations, 2 different population scaled strengths of selection were modelled: $2Ns$ = 100 and 1,000, where $N = N_{ancestral}$ and s is the selective advantage of the mutant allele relative to the wildtype. Simulations in which the beneficial mutation failed to fix were discarded and restarted. For NFDS simulations, 3 different equilibrium frequencies were modelled: F_{eq} = [0.1, 0.25, 0.5]. The population was sampled every 5 generations from the introduction of the beneficial mutation until either its fixation, or a further $10N_{ancestral}$ generations had passed.

<u>Population contraction simulations</u>

A single population was simulated, undergoing an instantaneous population contraction $0.1N_{ancestral}$ generations following the burn-in, reducing to $0.1N_{ancestral}$ individuals. The population was sampled at 14 time points, with two prior to the population contraction, one at the time of contraction, and then at 11 time points post-size change, with the final time point $1N_{ancestral}$ generations since the population contraction. Following the approach of Soni and Jensen [73], I separately simulated a 1Mb non-functional region, and a 32,657bp functional region. The functional region consisted of nine exons (of size 1,317bp) and eight introns (of size (1,520bp), flanked by intergenic regions of size 4,322bp. The numbers of introns and exons per functional region were obtained from Sakharkar et al. [91], with mean intron length taken from Hubé and Francastel [92]. Finally, the lengths of exons and intergenic regions were averages estimated from Ensembl's GRCh38.p14 dataset [93], obtained from Ensembl release 107 [94]. Mutations in intronic and intergenic regions

were all strictly neutral, whilst exonic mutations were drawn from the discrete distribution of fitness effects (DFE) inferred by Johri et al. [43] in humans. A single mutation under NFDS, with an equilibrium frequency, F_{eq} of 0.5 was introduced into the middle exon following the burn-in period.

Tajima's *D* [57] and and θ_w were calculated across 10kb sliding windows with a step size of 5kb, using the python implementation of libsequence (version 1.8.3, [95]), with mean values calculated across all windows and simulation replicates.

Approximate Bayesian computation (ABC) inference

To generate simulated empirical data for ABC inference, a single population of size $N_{ancestral}$ = 10,000 individuals was simulated for 100 replicates. A single functional region was simulated, using the same genomic structure and DFE as outlined in the section titled "Population contraction simulations". A mutation under NFDS was introduced immediately after the burn-in, and $10N_{ancestral}$ generations prior to sampling. Four population histories were simulated, with ABC inference performed on each separately: population equilibrium ($N_{current} = N_{ancestral}$, where $N_{current}$ is the population size at time of sampling); population expansion ($N_{current} = 2N_{ancestral}$); population contraction ($N_{current} = 0.5N_{ancestral}$); and severe population contraction ($N_{current} = 0.1N_{ancestral}$). The instantaneous size change occurred $0.01N_{ancestral}$ generations prior to sampling.

ABC inference was used to infer a single parameter of interest, T_b, the segregation time of the mutation under NFDS. Values for this parameter were drawn from a uniform distribution with a range between $1N_{ancestral}$ and $100N_{ancestral}$. A total of 3,000

different values of T_b were simulated for 100 replicates. The mean and standard deviation of the number of singletons, Tajima's *D* [57], θ_w, haplotype diversity, and mean r^2 were used for ABC inference, all calculated across 10kb sliding windows with a 5kb step size, using the python implementation of libsequence (version 1.8.3,[95]), with mean values calculated across all windows and simulation replicates, with each of the six genomic windows treated as a separate summary statistic. For the single time point inference, the population was sampled at the termination of the simulation, T_b generations after the simulation burn-in, yielding a total of 60 summary statistics. For multiple time point inference, the population was sampled $0.045N_{ancestral}$ and $0.02N_{ancestral}$ generations prior to the termination of the simulation, as well as at its termination, yielding a total of 180 summary statistics.

The "neural net" regression method with the default parameters provided by the R package "abc" [96] was used to generate posterior distributions. A 100-fold cross validation analysis was performed in order to determine the performance and accuracy of inference for tolerance values of 0.05, 0.08, and 0.1, with a tolerance of 0.08 identified as the most accurate. This value was employed for inference of final parameter values, meaning that 8% of all simulations were accepted by the ABC to estimate the posterior probability of parameter estimates. Inference was performed 50 times, with the mean of the weighted medians of the posterior estimates taken to determine point estimate of T_b.

Data availability

All scripts to generate and analyze simulated data are available at the GitHub repository: https://github.com/vivaksoni/NFDS_time_sampled_data

Acknowledgements

I am grateful to Jeffrey D. Jensen (Arizona State University) for valuable discussions relating to population genetic inference, and feedback on this manuscript.

References

1. Lewontin RC. The genetic basis of evolutionary change. New York: Columbia University Press; 1974. 346 p. (Columbia biological series).

2. Crow JF. Muller, Dobzhansky, and overdominance. J Hist Biol. 1987;20(3):351–80.

3. Dobzhansky T. 1955. A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol. 20:1–15.

4. Kimura M. Evolutionary rate at the molecular level. Nature. 1968 Feb;217(5129):624–626

5. Kimura M. The Neutral Theory of Molecular Evolution. 1st ed. Cambridge University Press; 1983.

6. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018: COMMENTARY. Evolution. 2019 Jan;73(1):111–4.

7. Gillespie JH. 1991. The causes of molecular evolution. New York: Oxford University Press.

8. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005 Dec;39(1):197–218.

9. Soni V, Jensen JD. Temporal challenges in detecting balancing selection from population genomic data. G3: Genes, Genomes, Genetics. 2024 Jun 5;14(6):jkae069.

10. Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al. Signatures of long-term balancing selection in human genomes. Genome Biology and Evolution. 2018 Mar 1;10(3):939–55.

11. Soni V, Vos M, Eyre-Walker A. A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. PLoS Biol. 2022 Jun 2;20(6):e3001645.

12. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006 Apr 28;2(4):e64.

13. Lawrence M. Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. Annals of Botany. 2000 Mar;85:221–6.

14. Castric V, Vekemans X. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Molecular Ecology. 2004 Oct;13(10):2873–89.

15. Goldberg EE, Kohn JR, Lande R, Robertson KA, Smith SA, Igić B. Species selection maintains self-incompatibility. Science. 2010 Oct 22;330(6003):493–5.

16. Kelley J, Walter L, Trowsdale J. Comparative genomics of major histocompatibility complexes. Immunogenetics. 2005 Jan;56(10):683–95.

17. Piertney SB, Oliver MK. The evolutionary ecology of the major histocompatibility complex. Heredity. 2006 Jan 1;96(1):7–21.

18. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci. 2010 Apr 7;277(1684):979–88.

19. Wright S. The distribution of self-sterility alleles in populations. Genetics. 1939 Jun 15;24(4):538–52.

20. Clarke BC. The evolution of genetic diversity. Proc R Soc Lond B. 1979 Sep 21;205(1161):453–74.

21. Levin BR. Frequency-dependent selection in bacterial populations. Phil Trans R Soc Lond B. 1988 Jul 6;319(1196):459–72.

22. Hedrick PW. Maintenance of genetic variation with a frequency-dependent selection model as compared to the overdominant model. Genetics. 1972 Dec 1;72(4):771–5.

23. Hedrick PW. What is the evidence for heterozygote advantage selection? Trends in Ecology & Evolution. 2012 Dec;27(12):698–704.

24. Ayala FJ, Campbell CA. Frequency-dependent selection. Annu Rev Ecol Syst. 1974 Nov;5(1):115–38.

25. Turelli M, Barton NH. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and *G* × *E* interactions. Genetics. 2004 Feb;166(2):1053–79.

26. Fitzpatrick MJ, Feder E, Rowe L, Sokolowski MB. Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. Nature. 2007 May;447(7141):210–

27. Kazancıoğlu E, Arnqvist G. The maintenance of mitochondrial genetic variation by negative frequency-dependent selection. Ecology Letters. 2014 Jan;17(1):22–7.

28. Kurbalija Novičić Z, Sayadi A, Jelić M, Arnqvist G. Negative frequency dependent selection contributes to the maintenance of a global polymorphism in mitochondrial DNA. BMC Evol Biol. 2020 Dec;20(1):20.

29. Turner CB, Buskirk SW, Harris KB, Cooper VS. Negative frequency-dependent selection maintains coexisting genotypes during fluctuating selection. Molecular Ecology. 2020 Jan;29(1):138–48.

30. Estévez D, Galindo J, Rolán-Alvarez E. Negative frequency-dependent selection maintains shell banding polymorphisms in two marine snails (*Littorina fabalis* and *Littorina saxatilis*). Ecology and Evolution. 2021 Jun;11(11):6381–90.

31. Madsen T, Stille B, Ujvari B, Bauwens D, Endler JA. Negative frequency-dependent selection on polymorphic color morphs in adders. Current Biology. 2022 Aug;32(15):3385-3388.e3.

32. Christie MR, McNickle GG. Negative frequency dependent selection unites ecology and evolution. Ecology and Evolution. 2023 Jul;13(7):e10327.

33. Crisci JL, Poh YP, Mahajan S, Jensen JD. The impact of equilibrium assumptions on tests of selection. Front Genet. 2013.

34. Charlesworth B, Jensen JD. Effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst. 2021 Nov 2;52(1):177–97.

35. Schierup MH, Vekemans X, Charlesworth D. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet Res. 2000 Aug;76(1):51–62.

36. Wiuf C, Zhao K, Innan H, Nordborg M. The probability and chromosomal extent of trans-specific Polymorphism. Genetics. 2004 Dec 1;168(4):2363–72.

37. Pavlidis P, Metzler D, Stephan W. Selective sweeps in multilocus models of quantitative traits. Genetics. 2012 Sep;192(1):225–39.

38. Fijarczyk A, Babik W. Detecting balancing selection in genomes: limits and prospects. Mol Ecol. 2015 Jul;24(14):3529–45.

39. Bitarello BD, Brandt DYC, Meyer D, Andrés AM. Inferring balancing selection from genome-scale data. Genome Biology and Evolution. 2023 Mar 3;15(3):evad032.

40. Klein J, Sato A, Nagl S, O'hUigín C. Molecular trans-species polymorphism. Annu Rev Ecol Syst. 1998 Nov;29(1):1–21.

41. Leffler EM, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 2013 Mar 29;339(6127):1578–82.

42. Johri P, Charlesworth B, Jensen JD. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics. 2020 May 1;215(1):173–92.

43. Johri P, Pfeifer SP, Jensen JD. Developing an evolutionary baseline model for humans: Jointly inferring purifying selection with population history. Molecular Biology and Evolution. 2023 May 2;40(5):msad100.

44. Soni V, Jensen JD. Inferring demographic and selective histories from population genomic data using a two-step approach in species with coding-sparse genomes: an application to human data. G3: Genes, Genomes, Genetics. 2025 Jan 30;jkaf019.

45. Terbot JW, Soni V, Versoza CJ, Pfeifer SP, Jensen JD. Inferring the demographic history of Aye-Ayes (*Daubentonia madagascariensis*) from high-quality, whole-genome, population-level data. Genome Biology and Evolution. 2025 Jan 6;17(1):evae281.

46. Soni V, Versoza CJ, Vallender EJ, Jensen JD, Pfeifer SP. Accounting for chimerism in demographic inference: reconstructing the history of common marmosets (*Callithrix jacchus*) from high-quality, whole-genome, population-level data. In Press as Molecular Biology and Evolution. 2025.

47. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, et al. Viral dynamics in human immunodeficiency virus type 1 infection. Nature. 1995 Jan;373(6510):117–22.

48. Foll M, Poh YP, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, et al. Influenza virus drug resistance: A time-sampled population genetics perspective. PLoS Genet. 2014 Feb 27;10(2):e1004185.

49. Renzette N, Caffrey DR, Zeldovich KB, Liu P, Gallagher GR, Aiello D, et al. Evolution of the influenza A virus genome during development of oseltamivir resistance *In Vitro*. J Virol. 2014 Jan;88(1):272–81.

50. Mallick S, Reich D. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes [Internet]. Harvard Dataverse; 2023.

51. Ejsmond MJ, Radwan J. MHC diversity in bottlenecked populations: a simulation model. Conserv Genet. 2011 Feb;12(1):129–37.

52. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974;23(1):23–35.

53. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002 Oct 24;419(6909):832–7.

54. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006 Mar 7;4(3):e72.

55. Isildak U, Stella A, Fumagalli M. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. Molecular Ecology Resources. 2021 Nov;21(8):2706–18.

56. Haller BC, Messer PW. SLiM 4: Multispecies eco-evolutionary modelling. The American Naturalist. 2023 May 1;201(5):E127–39.

57. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov;123(3):585–95.

58. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993 Aug;134(4):1289–303.

59. Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res. 1996 Oct;68(2):131–49.

60. Charlesworth B. Background selection 20 years on. Journal of Heredity. 2013;104(2):161–71.

61. Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 2005 Jul;170(3):1401–10.

62. Kaiser VB, Charlesworth B. The effects of deleterious mutations on evolution in non-recombining genomes. Trends in Genetics. 2009 Jan;25(1):9–12.

63. O'Fallon BD, Seger J, Adler FR. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Molecular Biology and Evolution. 2010 May 1;27(5):1162–72.

64. Nicolaisen LE, Desai MM. Distortions in genealogies due to purifying selection and recombination. Genetics. 2013 Sep 1;195(1):221–30.

65. Zeng K. A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity. 2013 Apr;110(4):363–71.

66. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 2016 Jan;25(1):135–41.

67. Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD. The Impact of purifying and background selection on the inference of population history: Problems and prospects. Molecular Biology and Evolution. 2021 Jun 25;38(7):2986–3003.

68. Charlesworth B, Jensen JD. Some complexities in interpreting apparent effects of hitchhiking: A commentary on Gompert et al. (2022). Molecular Ecology. 2022 Sep;31(17):4440–3.

69. Soni V, Johri P, Jensen JD. Evaluating power to detect recurrent selective sweeps under increasingly realistic evolutionary null models. Evolution. 2023 Jul 3;qpad120.

70. Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, et al. Recommendations for improving statistical inference in population genomics. PLoS Biol. 2022 May 31;20(5):e3001669.

71. Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. On the prospect of achieving accurate joint estimation of selection with population history. Genome Biology and Evolution. 2022 Jul 2;14(7):evac088.

72. Cheng X, DeGiorgio M. Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. Molecular Biology and Evolution. 2020 Nov 1;37(11):3267–91.

73. Soni V, Jensen JD. Inferring demographic and selective histories from population genomic data using a two-step approach in species with coding-sparse genomes: an application to human data. G3: Genes, Genomes, Genetics. 2025 Jan 30;jkaf019.

74. Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002 Mar 1;160(3):1179–89.

75. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002 Feb 1;160(2):765.

76. Laval G, Peyrégne S, Zidane N, Harmant C, Renaud F, Patin E, et al. Recent adaptive acquisition by African rainforest hunter-gatherers of the late Pleistocene sickle-cell mutation suggests past differences in malaria exposure. The American Journal of Human Genetics. 2019 Mar;104(3):553–61.

77. Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci USA. 2012 Sep 25;109(39):15716–21.

78. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013 Jul 4;499(7456):74–8.

79. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga JL, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature. 2014 Jan;505(7483):403–6.

80. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993 Apr;362(6422):709–15.

81. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc R Soc B. 2012 Dec 7;279(1748):4724–33.

82. Dabney J, Meyer M, Paabo S. Ancient DNA damage. Cold Spring Harbor Perspectives in Biology. 2013 Jul 1;5(7):a012567–a012567.

83. Garrido Marques A, Rubinacci S, Malaspinas AS, Delaneau O, Sousa Da Mota B. Assessing the impact of post-mortem damage and contamination on imputation performance in ancient DNA. Sci Rep. 2024 Mar 14;14(1):6227.

84. Der Sarkissian C, Ermini L, Jónsson H, Alekseev AN, Crubezy E, Shapiro B, et al. Shotgun microbial profiling of fossil remains. Molecular Ecology. 2014 Apr;23(7):1780–98.

85. Sampietro ML, Gilbert MTP, Lao O, Caramelli D, Lari M, Bertranpetit J, et al. Tracking down human contamination in ancient human teeth. Molecular Biology and Evolution. 2006 Sep 1;23(9):1801–7.

86. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR: Science & Technology of Archaeological Research. 2017 Jan;3(1):1–14.

87. Peyrégne S, Prüfer K. Present-day DNA contamination in ancient DNA datasets. BioEssays. 2020 Sep;42(9):2000081.

88. Takahata N. Allelic genealogy and human evolution. Molecular Biology and Evolution. 1993 Jan;10(1):2–22.

89. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000 Sep;156(1):297–304.

90. Payseur BA, Nachman MW. Microsatellite Variation and Recombination Rate in the Human Genome. Genetics. 2000 Nov 1;156(3):1285–98.

91. Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol. 2004;4(4):387–93.

92. Hubé F, Francastel C. Mammalian Introns: When the junk generates molecular diversity. IJMS. 2015 Feb 20;16(3):4429–52.

93. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

94. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Research. 2022 Jan 7;50(D1):D988–95.

95. Thornton K. libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 2003 Nov 22;19(17):2325–7.

96. Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012 Jun;3(3):475–9.

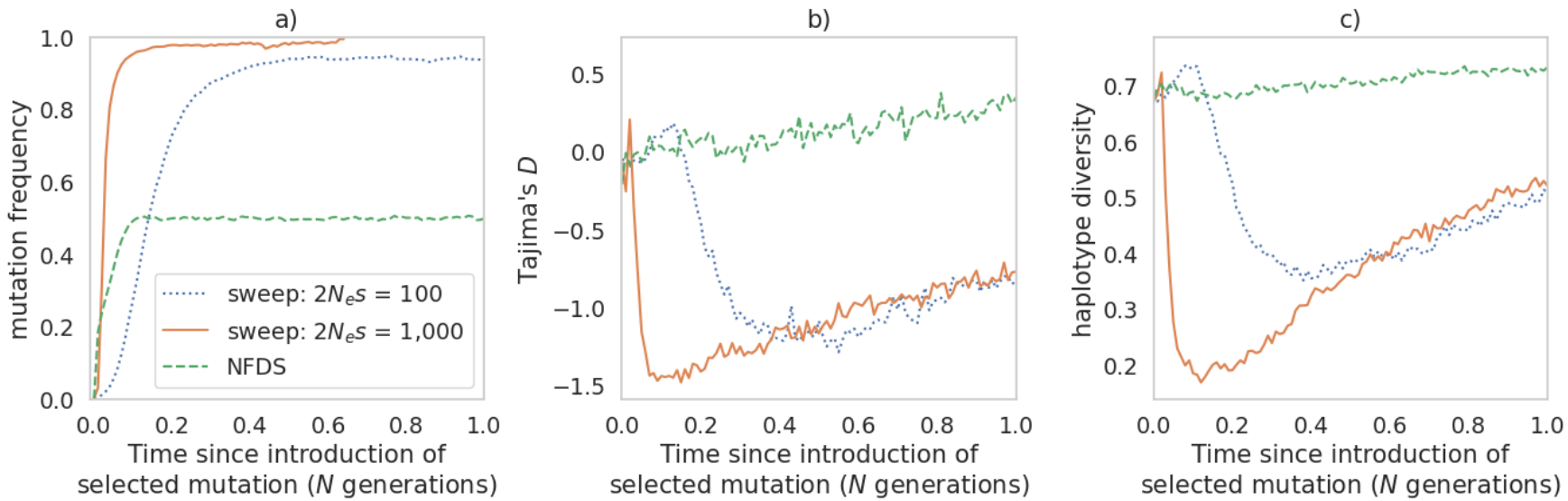

Figure 1: a) Mutation trajectories, **b)** Tajima's *D* and **c)** haplotype diversity for selective sweeps and NFDS in an equilibrium population, on a neutral background. Summary statistics were calculated across sliding 10kb windows, with a 5kb step size. Only the focal window containing the mutation under NFDS is shown. All statistics shown are averages across 100 simulation replicates. Equilibrium frequency, F_{eq} of the mutation under NFDS (green dashed line) is 0.5. The population scaled strength of selection, *2Ns* of selective sweeps are 100 (blue dotted line) and 1,000 (orange line), where *N* is the population size of 500 individuals, and *s* is the selective advantage of the mutant allele relative to the wildtype.

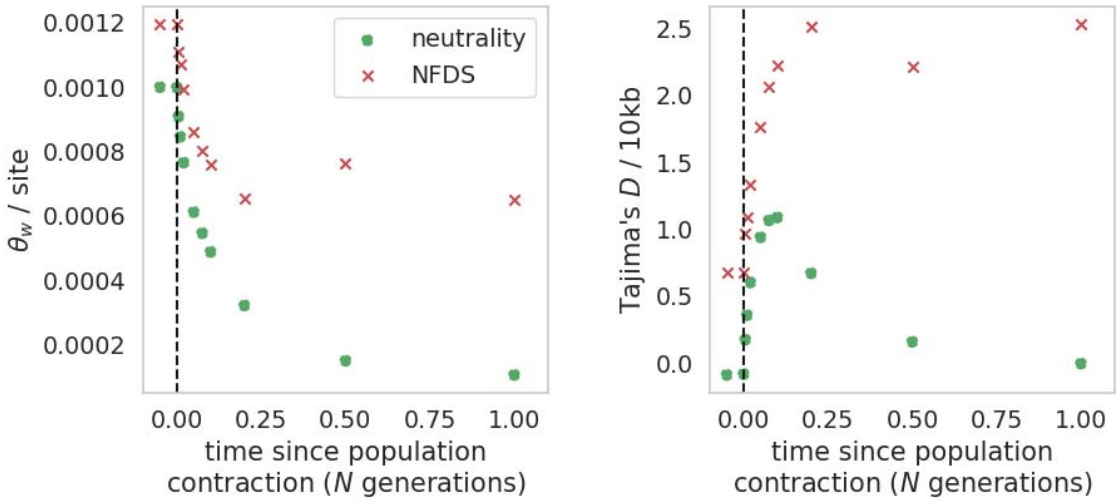

Figure 2: and Tajima's *D* for neutral (green "o"s) and NFDS (red "x"'s) simulations in which an instantaneous population contraction occurred (black dashed line represents time of contraction). Population is reduced to $0.1N_{ancestral}$ individuals, where $N_{ancestral}$ = 10,000. Data points to left of 0 on x axis indicate sampling prior to the population contraction. Equilibrium frequency, F_{eq} of the mutation under NFDS is 0.5. Mutation under NFDS has been segregating for $75N_{ancestral}$ generations at time of population contraction. Summary statistics were calculated across 10kb windows, then averaged across all windows and all 100 simulation replicates.

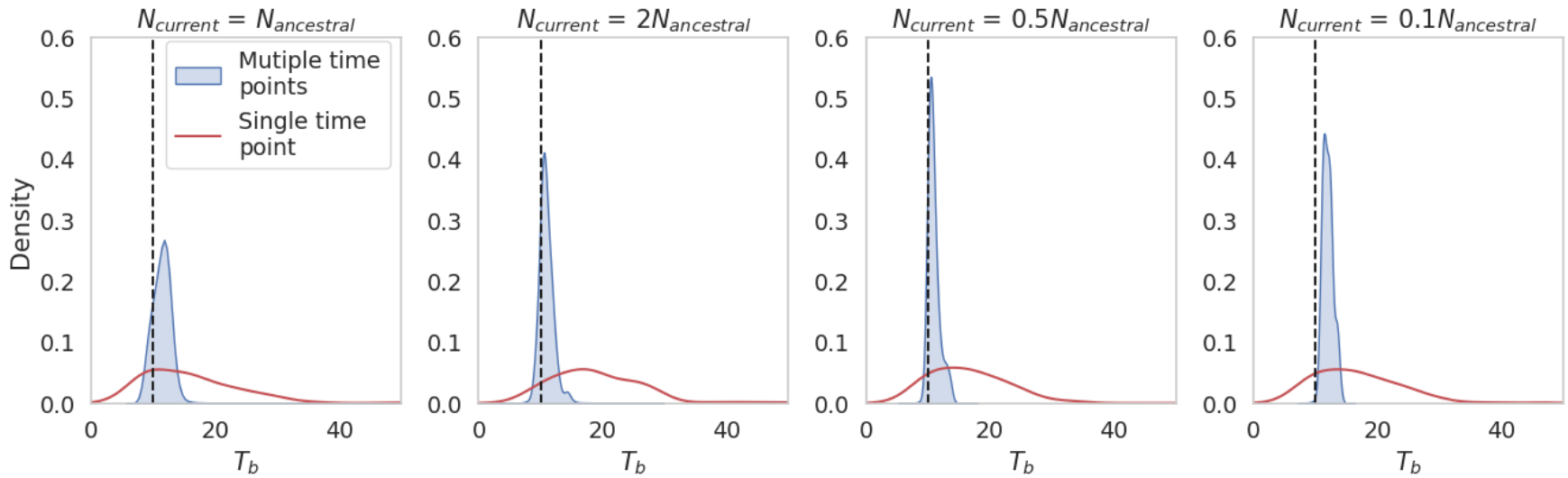

Figure 3: Posterior distributions from ABC inference of the segregation time, T_b of the mutation under NFDS for four different population histories. Black dashed line represents the true value of T_b ($10N_{ancestral}$ generations, where $N_{ancestral}$ is the initial population size of 10,000 individuals). Multiple data points (blue shaded distributions) include sampling at the current day, $0.02N_{ancestral}$ generations prior to current day, and $0.045N_{ancestral}$ generations prior to current day. Single time point data (red unshaded distributions) is for sampling at the current day only only. Four underlying population histories are represented. From left to right: Population equilibrium ($N_{current} = N_{ancestral}$), population expansion ($N_{current} = 2N_{ancestral}$), population contraction ($N_{current} = 0.5N_{ancestral}$), and severe population contraction ($N_{current} = 0.1N_{ancestral}$). All population size changes occurred instantaneously, $0.01N_{ancestral}$ generations prior to the current day.