Simulation-based evaluation of methods, data types, and temporal sampling schemes for detecting recent population declines

Brendan N. Reid[1]*, Malin L. Pinsky[1]

[1]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA.
*E-mail: br450@rutgers.edu

Abstract

Understanding recent population trends is critical to quantifying species vulnerability and implementing effective management strategies. To evaluate the accuracy of genomic methods for quantifying recent declines (beginning <120 generations ago), we simulated genomic data using forward-time methods (SLiM) coupled with coalescent simulations (msprime) under a number of demographic scenarios. We evaluated both site frequency spectrum (SFS)-based methods (momi2, Stairway Plot) and methods that employ linkage disequilibrium information (NeEstimator, GONE) with a range of sampling schemes (contemporary-only samples, sampling two time points, and serial sampling) and data types (RAD-like data and whole-genome sequencing). GONE and momi2 performed best overall, with >80% power to detect severe declines with large sample sizes. Two-sample and serial sampling schemes could accurately reconstruct changes in population size, and serial sampling was particularly valuable for making accurate inference when genotyping errors or minor allele frequency cutoffs distort the SFS or under model mis-specification. However, sampling only contemporary individuals provided reliable inferences about contemporary size and size change using either site frequency or linkage-based methods, especially when large sample sizes or whole genomes from contemporary populations were available. These findings provide a guide for researchers designing genomics studies to evaluate recent demographic declines.

Introduction

Human impacts on wild populations have steadily intensified over the course of the Holocene, and multiple lines of evidence suggest the current era is heading towards a mass extinction event (Ceballos et al. 2015). For species of conservation concern, understanding recent population trends (on the scale of the past several decades or centuries) is critical to quantifying species vulnerability and implementing effective management strategies. Rapidly decreasing population sizes increase the risk of local extirpation or complete extinction in the near future (Caughley 1994). Populations with small population sizes are also at risk of losing genetic diversity and adaptive potential due to genetic drift (Franklin 1980, Lande & Barrowclough1987) and can experience declining average fitness due to inbreeding depression (Hedrick & Kalinowski 2000, Keller and Waller 2002). Accurately estimating current and past population size is therefore a priority for conservation biologists and managers. Unfortunately, baseline data on census population sizes (N_c) over the past few centuries are often missing or unreliable for species of conservation concern (Alagona et al. 2012, McClenachan et al. 2012).

Genomic methods offer a promising alternative to direct census data for inferring effective population size (N_e) over time. N_e determines the rate of inbreeding and genetic drift over time, and although it is not directly substitutable for N_c (usually, N_e is smaller than N_c, sometimes by several orders of magnitude), it is an important determinant of the rate of

evolutionary change (Waples 2022). Changes in N_e leave an imprint on genomic diversity within a population in two important ways. First, the demographic history of a population influences the distribution of allele frequencies in that population, also known as the site frequency spectrum (SFS). A population undergoing a demographic expansion, for example, will contain more rare alleles relative to a population with a constant size, while a population undergoing a demographic decline will contain fewer rare alleles (Griffiths and Tavaré 1998, Beichman et al. 2018). Because of this connection between the distribution of allele frequencies and demographic history, the expected SFS can be constructed for a given population history using coalescent simulations in which the probability of sampled individuals sharing a common ancestor is determined by the population size over time (Excoffier et al. 2013). Computationally efficient approximations of coalescent expectations can be obtained using diffusion models (Gutenkunst et al. 2009) or stochastic models (Kamm et al. 2019) which use continuous time rather than discrete generations, and population parameters can be estimated based on the observed SFS using likelihood methods. The effect of demographic change on the SFS accumulates over time and is dependent on both the mutation rate and the coalescence rate. This means that the signal of ancient demographic processes is easier to detect than the signal of recent change, especially in large populations where coalescence is less frequent, and larger sample sizes will be necessary to detect recent changes (Beichman et al. 2018).

Importantly, SFS methods assume that the loci used to construct the SFS are independent and unlinked. Nonrandom associations between loci can occur for multiple reasons, including physical linkage among loci on the same chromosome and genetic drift in finite populations (Hill 1981). The latter means that patterns of linkage disequilibrium (LD) across loci will also be shaped by demographic history. Multiple methods have been developed to infer population size from patterns of LD. For a sample of physically unlinked loci, LD should be close to zero in an infinite population, and the amount of "excess" LD can be used to estimate N_e at a particular time point. This method is most accurate when the population size is small and the sample size is large (close to the true N_e; Waples 2006). For loci inherited on the same chromosome, the frequency of recombination and thus the amount of LD depends both on cumulative genetic drift and the frequency of recombination between the two loci, meaning that LD for loci with different linkage distances will reflect population size at different points in that population's history. Recombination frequency will usually increase with increasing physical distance on the chromosome, although recombination rate can vary throughout the genome (Peñalba & Wolf 2020). Thus, the pattern of linkage disequilibrium across the genome, combined with a linkage map of known recombination rates, can be used to infer changes in N_e over time (Hayes et al. 2003). Recombination events are more frequent for loci with weaker physical linkage, and since recombination rates for weakly linked loci can be much higher than mutation or coalescence rates, LD data potentially contain more information for inferring recent changes in population size than SFS data alone (Hayes et al. 2003, Santiago et al. 2020). Critically, obtaining detailed linkage information requires the existence of an accurate reference genome for the organism of interest, which may not be available in many cases.

One promising avenue for inferring recent changes in N_e is by comparing genetic patterns in historical and modern samples. Advances in obtaining genetic material from museum specimens or other historical samples has made the acquisition of both baseline and contemporary genetic data (henceforth "temporal data") a possibility (Nielsen and Hansen 2008, Bi et al. 2013, Habel et al. 2014, Walender et al. 2017, Diez-del-Molino et al. 2018, Oosting et al. 2019). For example, temporal RADseq data from salamanders has been used to

accurately reconstruct known recent declines and expansions (Nunziata et al. 2017). In widespread species, such as Atlantic salmon, genomic signatures of population decline from targeted sequencing of historical and contemporary samples have also been used to identify which populations have recently declined and to infer the drivers of these declines (Lehnert et al., 2019).

In addition to temporal sampling, the type of genomic data and the availability of reference genomes can also influence the quality of inference for N_e. Conservation genomics practitioners often use techniques that sample a moderate number of loci across the genome such as RADseq (Andrews et al. 2016) or targeted sequencing (Meek and Larson 2019). Without a reference genome, these data can provide information on the SFS but are anonymous with regard to linkage information. Whole-genome sequencing is now within reach as well and can greatly increase the scope and precision of inference possible from conservation genomic studies (Brandies et al. 2019). Chromosome-level assemblies are the gold standard for providing linkage information; however, draft genomes with incomplete linkage information may be sufficient for making demographic inferences with some methods (Patton et al. 2019).

While several recent studies have used simulations to compare performance of methods for inferring recent population history under a given sampling scheme for either RADseq (Nunziata and Weisrock 2018) or whole-genome data (Patton et al. 2019), the amount of precision gained by adding historical genomic data compared to using only contemporary data remains unclear. The existing simulation studies have also examined performance under somewhat limited ranges of past and present population sizes, timings of population decline, and generation times. Failure to account for ancient population events, such as Pleistocene expansion or contraction, may also affect inferences made under SFS methods (Momigliano et al. 2021, Hoey et al. in press). The lingering uncertainty from all of these potential sources can make it difficult for researchers to make objective decisions regarding how to best spend limited research funds to generate data that will yield the highest-quality inferences regarding recent demographic history.

To help guide study design for researchers interested in recent demographic inference using genomic data, we compare here the performance of four inference methods and three temporal sampling schemes across simulated reduced-representation and whole-genome datasets representing scenarios of recent population stability or decline. We aim to answer the following primary questions: (1) Which methods provide the most accuracy and precision for identifying population declines using contemporary data?; and (2) How do historical genomic data alter the accuracy of demographic inference? By evaluating the accuracy of different study designs and inference methods, we provide concrete recommendations for conservation biologists interested in reconstructing the recent demographic history of a diverse array of potential study organisms.

Methods

Study outline

To evaluate different methods for estimating contemporary changes in population size, we simulated whole genomes from populations with known histories representing either stability or decline over the past 200 years. We then subsampled these genomes to generate a number of reduced datasets incorporating a smaller number of individuals and/or a random subset of loci distributed throughout the genome. Finally, we applied four estimation methods that have been developed to infer recent declines. Each estimation method can accommodate different data types and temporal schemes and takes different inputs (Table 1). We compared power to detect declines as well as the accuracy and precision for each method.

Simulation scope and approach

We consider a single panmictic population of diploid, dioecious organisms sampled at a contemporary time point—or zero years before present (ybp)—and at several time points in the recent past (t=120, 90, 60, and 30 ybp) corresponding to samples that could be represented in natural history collections or genetic monitoring programs. We express time in years rather than generations so that we can examine more complex demographic scenarios, such as overlapping generations. We do not consider "paleogenomes" from the more distant past in this paper.

We simulate data from this population in two distinct stages (Figure 1). The first stage uses demographically realistic forward-time simulations to generate a set of recent genealogies. The sex of each individual in the population was randomly determined (50/50 chance of being male or female, expected sex ratio = 0.5). Population size in forward simulations was regulated by controlling the number of offspring (N_O) generated in each time step. At each time step, N_O offspring were generated by randomly selecting one male and one female parent with replacement from all breeding-age individuals for each offspring.

We considered two different life history patterns in these simulations. The first pattern (G1) represented an annual organism with a generation time of one year and no overlapping generations (all individuals in the population can breed and all die upon reaching an age of one year). The second (G3) pattern represented an organism with overlapping generations, with age at first breeding equal to one year, an age-specific mortality probability, and a maximum age equal to 8 years. The mortality probabilities and age at first breeding were set such that the mean age of a breeding individual (and thus the mean generation time) was approximately 3 years and the expected number of breeding-age individuals (N_B) was equal to N_O in a stable population. Since for each simulated situation the number of offspring in a given time step equaled the number of breeding age individuals, regardless of population size, the distribution of reproductive success approximately followed a Poisson distribution as expected in an ideal population with random mating, with each parent contributing genes to a mean of two offspring and the expected variance in offspring N_O being approximately two. N_O in each generation should therefore be roughly equivalent to N_e for both life history patterns (Hill 1972).

The forward simulations began 100 years before the first historical sampling time point (i.e., 220 ybp). For the baseline simulations, we set the initial population size ($N_{e,H}$) to either 1000

or 10000. N_O in each subsequent generation either remained stable or began an exponential decline (with $\lambda = N_O$ in the current year / N_O in the following year) at a time point directly after one of the historical sampling points (T_{dec}= 120, 90, 60, or 30 ybp), eventually reaching a contemporary effective population size ($N_{e,C}$) at the final time point (zero ybp). For all declining populations we conducted a set of simulations with $\lambda = 0.99$, resulting in $N_{e,C}/N_{e,H}$ of 0.74, 0.55, 0.40, and 0.30 respectively for each decline scenario. We also conducted one simulation with a recent decline ($T_{dec} = 30$) with $\lambda = 0.95$ and $N_{e,C}/N_{e,H} = 0.21$. This higher λ value was only paired with a recent decline because rapid declines beginning earlier resulted in extremely small population sizes.

We conducted 5 iterations of each forward demographic simulation using SLiM v.3.3.2 (Haller and Messer 2019) and recorded the full pedigree as well as the number of breeders of each sex for each iteration. We then simulated 25 "chromosomes" for each demographic iteration by conducting an independent simulation (with the pedigree fixed to the recorded pedigree) of a single sequence of length 30 Mb and a per-generation recombination rate of 10^{-8} per base per generation (or 1 cM/Mb, which is within the range of recombination rates observed for plants and animals; Stapley et al. 2017). To decrease the computational intensity of chromosome simulations, we did not simulate mutations in SLiM, instead using tree sequence recording and coalescent simulations (see below) for generating polymorphisms (Haller et al. 2019). For tree sequence recording, we recorded 200 individuals at each potential T_{dec} as well as all individuals at 0 ybp.

The second stage of simulations involved simulating genetic data for these genealogies using reverse-time coalescent methods. Simulations were performed using msprime v.0.7.4 (Kelleher & Lohse 2020) in the python package pyslim v.0.501, and populations were projected backward for a number of generations sufficient for all sampled individuals to reach a common ancestor using a coalescent process (i.e. without incorporating the complex life history used in forward-time simulations for G3). The effective population size at the initiation of the simulation was set to the number of breeding-age individuals in the first generation of the forward-time simulation ($N_{e,H}$). Populations in the coalescent simulations either remained stable over time or experienced a 10-fold size change (representing either an ancestral expansion or an ancestral bottleneck) at 10000 generations before present, and they remained at this ancestral population size ($N_{e,A}$) for the remainder of the simulation (until all loci reached coalescence). Ancestral bottlenecks and expansions were only simulated for populations with a larger historic population size ($N_{e,H} = 10000$), for the G1 life history pattern, and for a restricted set of recent demographic scenarios (constant population size, a rapid decline starting 30 ybp, and a slow decline starting 120 ybp). Eighteen demographic scenarios in total were simulated. We simulated data with a recombination rate of 10^{-8} as used in the forward simulations, and we added mutations to simulated chromosomes using a per-generation, per-base mutation rate of 10^{-8} (within the range of mutation rates observed for plants and animals; Lynch 2007). From these two simulation stages, we generated VCF files containing all variable sites for each simulated chromosome from 200 randomly selected individuals at each time point.

Sampling designs

We subsampled from these full datasets to represent realistic constraints of study design choices. We generated datasets with total sample sizes n of 20, 50, 100, and 200 individuals. Temporal sampling schemes can range from a single comparison between a historic baseline and a contemporary sample to a number of samples collected over several time points, as in fisheries monitoring (Hutchinson et al. 2003) and repeated museum collections (Gauthier et al. 2020). As such, for each dataset samples were either all collected from the contemporary timepoint (n samples at 0 ybp; contemporary-only dataset), split evenly between a contemporary and a baseline timepoint ($n/2$ samples at 0 ybp and $n/2$ samples at 120 ybp; two-sample dataset), or split between five serial timepoints ($n/5$ samples at 0, 30, 60, 90, and 120 ybp respectively; serial dataset). For whole-genome datasets we used the three smaller sample sizes (total n = 20, 50, or 100). To create RADseq-like datasets, we used the three larger sample sizes (total n = 50, 100, or 200) and applied an additional filter to keep only SNPs found within a set of randomly placed 150bp loci on each of the 25 chromosomes. For RADseq-like datasets, we used either 400 RADseq loci per chromosome (10,000 total loci) or 2000 loci per chromosome (50,000 total loci). We generated and conducted inference on 2,430 simulated datasets in total.

Inference on simulated datasets

We applied four different inference methods to the simulated datasets. The methods chosen represent commonly used software packages that use either the SFS or LD to infer current and past population sizes and can incorporate either temporal data, whole-genome data, or both. For conducting inference with temporal data using the SFS, we used the program momi2 (Kamm et al. 2020), a model-based method for demographic inference that can incorporate whole-genome or RAD data. We used pyslim to compute allele counts for each chromosome at each time point, and we combined these counts into aggregate folded SFS for each time point. momi2 assumes a branching tree-like population structure, and to accommodate sampling multiple time points from a single continuous population in momi2, we specified that each SFS was sampled as a "leaf" from a branch at its corresponding sampling time, after which all lineages from that population were shifted to a new branch from which the next sample was taken. For whole genome data, the total number of sites was set to the genome size (750Mb), while for RADlike data the total number of sites was set to 150 times the number of loci (1.5 Mb or 6 Mb) to represent 150 bp RAD loci. For each simulated dataset we fit four different demographic models: (1) a model with a single constant population size parameter, $N_{constant}$; (2) a model with two population size parameters (size at 0 ybp, $N_{contemp}$, and at 120 ybp, $N_{historic}$) and a time parameter specifying the time at which the population began an exponential size change (T_{rc}); (3) a model with an instantaneous ancient size change $\widehat{N}_{\text{ancient}}$ occurring at T_{ac}; and (4) a model including both the ancestral and recent size changes. Potential ranges for $N_{constant}$, $N_{contemp}$, $N_{historic}$, and $N_{ancient}$ were set to 10 – 500,000 individuals (the size change was not assumed to be a decline). The range for T_{rc} was set to between 10-120 ybp and the range for T_{ac} was set to 1,000 and 100,000 ybp. The rate of exponential size change was fully determined by the time of decline and the two population size parameters. We did not constrain recent size changes to be declines, and as such we evaluated whether momi2 inferred the population to be either stable, declining, or expanding. We fit all models to each simulated dataset using the Truncated Newton (TNC) optimizer, and we recorded all parameter estimates as well as the likelihood

and AIC for each model. We retained parameter estimates for population size at 0 ybp $\widehat{N}_{E,H}$ and 120 ybp $\widehat{N}_{E,C}$ for the model with the lowest AIC for each dataset.

For conducting inference using the SFS with whole-genome data, we used Stairway Plot 2 (Liu and Fu 2020). We used the vcf2sfs script (https://github.com/shenglin-liu/vcf2sfs) to compute the folded SFS input. We set the total number of sites (including monomorphic sites) to 750Mb, the number of random breakpoints for each iteration to {7,15,22,28}, the mutation rate per generation to $1x10^{-8}$, and the generation time to 1 year, and we used 67% of the sites for training. We retained the most recent median estimate of population size as $\widehat{N}_{e,C}$. As the time bins for stairway plot estimates can be somewhat irregularly spaced, we used the median population size estimate for the time bin closest to 120 ybp that was closest to this time point a $\widehat{N}_{e,H}$, and we used the 2.5% and 97.5% estimates as confidence intervals.

For conducting inference based on excess LD using RAD-like data, we used NeEstimator2 (Do et al. 2014). The LD method outperformed two other methods implemented in NeEstimator in a previous study (Gilbert and Whitlock 2015). We only used the two-sample scheme for assessing performance of NeEstimator since contemporary-only sampling does not allow for inference of historic size with this method. Before running, we converted vcf files to genepop files using the vcf2genepop.pl script (https://github.com/z0on/2bRAD_denovo/blob/master/vcf2genepop.pl). We used an allele frequency of 0.05 and assumed random mating. We used the point estimates and the used jackknife 95% confidence intervals for population size at 0 ybp and 120 ybp as $\widehat{N}_{e,C}$ and $\widehat{N}_{e,H}$, respectively.

For conducting inference using LD with whole-genome data, we used GONE (Santiago et al. 2020). We used plink (Purcell et al. 2007) to convert the vcf file to ped/map format. We ran GONE using the default parameters (unknown phase, 1 cM/Mb, Haldane correction, 2000 generations, 400 bins, MAF=0, allowing SNPs with zeroes, using all chromosomes, 50000 SNPs/ chromosome, hc = 0.05, 40 reps, 20 threads). We used the estimates for population size at 0 ybp and 120 ybp as $\widehat{N}_{e,C}$ and $\widehat{N}_{e,H}$, respectively. After Santiago et al. 2020, we performed resampling to estimate a confidence interval for $\widehat{N}_{e,C}$ and $\widehat{N}_{e,H}$. Since some datasets contained a relatively small number of SNPs) <300,000 SNPs compared to the datasets in Santiago et al. (2020), we took a random sample of 50,000 SNPs forty times and re-ran the program to generate 95% confidence intervals.

Effects of genotyping errors and allele frequency filters

Some demographic inference methods (including momi2) assume no errors in genotyping and no filtering of genotypes based on allele frequency. Since these conditions are rarely met in practice for non-model species, we performed additional inferences using modified datasets to explore the effects of some potential violations of these assumptions on inference accuracy.

Genotyping errors are more likely to occur for historical samples due to lower coverage and postmortem DNA damage in older samples. For example, Bi et al. (2013) found error rates that were almost fivefold higher (0.19%) in historical samples compared to contemporary error rates (0.04%). Genotyping errors are most likely to create singleton SNP genotypes that would potentially impact site frequency spectrum-based analyses. Minor allele frequency filters are commonly applied to SNP datasets; while this may improve inference for some applications, such as assessing population structure using STRUCTURE-like methods (Linck

and Battey 2019), distorting the site frequency spectrum by removing low-frequency alleles could also negatively impact other analyses (Lou et al. 2021). To assess the potential effect of errors in historical genotypes, we added singletons to the site frequency spectrum in momi2 at three different rates ($1e^{-3}$, $1e^{-4}$, or $1e^{-5}$ singletons per site) for RAD-like datasets and re-ran the momi2 inferences. We also applied a minor allele frequency filter of 0.01 to the data and re-ran momi2 inferences as well.

Evaluating methods

For datasets simulated with a recent decline, we considered the decline to be correctly inferred when a model including a recent decline had the lowest AIC and $\widehat{N}_{e,C} < \widehat{N}_{e,H}$ (for momi2) or when the upper 95% confidence interval for $\widehat{N}_{e,C}$ was lower than the lower 95% confidence interval for $\widehat{N}_{e,H}$ (for the other methods). For datasets simulated with constant population size, we considered the demographic history to be correctly inferred when a model including a constant recent population size had the lowest AIC and $\widehat{N}_{e,C} < \widehat{N}_{e,H}$ (for momi2), or when the 95% confidence intervals for $\widehat{N}_{e,C}$ and $\widehat{N}_{e,H}$ overlapped. Because StairwayPlot and NeEstimator provide 95% confidence intervals for the contemporary and historic population sizes themselves, rather than the difference in population sizes, this test for a decline is somewhat overly conservative (Cumming and Finch 2005).

We evaluated and visualized error using two different approaches. For each method, we calculated mean absolute percentage error for $\widehat{N}_{e,C}$ and $\widehat{N}_{e,H}$ for each simulated dataset *i* over *n* total datasets using the following equation:

$$\frac{100}{n}\sum_{i=1}^{n} \frac{|inferred_i - simulated_i|}{simulated_i}$$

We calculated mean error on an aggregated subset of scenarios, including three demographic scenarios (constant population size, a rapid decline starting 30 ybp, and a slow decline starting 120 ybp), a generation time of 1, and both initial population sizes. We calculated mean error separately for the WGS and 50K RAD loci datasets.

Since this metric uses absolute value and does not convey potential directional biases, we also visualized concordance between true and simulated values by plotting the $\log_{10}$ ratio of the inferred to the simulated value across all simulated demographic scenarios.

For two methods, we assessed alternate metrics of accuracy. Since we fit multiple alternate models in momi2, we evaluated how often the correct model (constant population size for data generated under the constant demographic model, or size change for data generated under the size change model) had the best support (defined as the model with the lowest AICc). Since NeEstimator can return estimates of "infinity" in some situations, we also identified the proportion of simulations for which this occurred for $\widehat{N}_{e,H}$ and $\widehat{N}_{e,C}$. Due to a large number of "infinite" estimates from NeEstimator at n=50 (Supplemental Figure 1), we excluded this sample size from mean error calculations.

Results

Power to detect declines

Overall, momi2 and GONE exhibited the highest power to correctly detect or reject recent declines. With large sample sizes (n = 200) and RAD-like data, momi2 correctly identified declines for ≥80% of simulated datasets and only performed poorly when declines were recent (30 generation ago) and slow ($\lambda = 0.99$) (Figure 2a). momi2 did not perform as well using WGS data, although it was still ≥80% accurate for detecting slow declines ≥ 90 generations (Figure 2b). Decreasing samples sizes for momi2 generally decreased the power to detect declines (Figure 2, Supplemental Table 1, Supplemental Table 2). GONE exhibited ≥ 90% accuracy for detecting more severe declines ($\lambda = 0.99$ for 120 generations or $\lambda = 0.95$ for 30 generations) with large samples sizes (n = 100). Power to detect declines with GONE decreased for lower sample sizes and for less severe declines (Figure 2b). The other methods (Stairway plot and NeEstimator) generally had lower power compared to momi2 and GONE for similar sample sizes (Figure 2, Supplemental Table 1, Supplemental Table 2).

Accuracy and bias for estimating $N_{e,H}$ and $N_{e,C}$

Mean absolute error for estimating $N_{e,H}$ was lowest for the SFS-based methods (momi2 and the stairway plot; Figure 2a). Mean error for estimating $N_{e,H}$ was somewhat higher for GONE, an LD-based method, than for the SFS-based methods, and the other LD-based method (NeEstimator) displayed substantially higher error for estimating $N_{e,H}$ than any of the other methods (Figure 2a). Mean absolute error tended to be substantially higher overall for estimating $N_{e,C}$ (Figure 2b) than for $N_{e,H}$ across methods. GONE tended to have lower error for estimating $N_{e,C}$ than SFS-based methods, while NeEstimator had comparable or higher error compared to momi2.

Mean absolute error for estimating $N_{e,H}$ did not strongly depend on sample size for SFS-based methods (Figure 2a). Mean error for estimating $N_{e,H}$ decreased with sample size for GONE and actually increased for NeEstimator. For $N_{e,C}$, on the other hand, mean error did tend to decrease with increasing sample sizes across methods (Figure 2b).

Whole-genome and RAD data performed comparably for estimating $N_{e,H}$ with SFS-based methods (Figure 2a). There was no clear difference between the two data types for estimating $N_{e,C}$ as well, and the most accurate estimates were produced by a WGS data (GONE with sample sizes 50-100) and an SFS method (momi2 with sample size of 200).

In general there were no strong directional biases in estimating $N_{e,H}$ except for NeEstimator with a sample size of 100, which produced downward-biased estimates (Supplemental Figure 2a, Supplemental Figure 3). Estimates of $N_{e,C}$ tended to be biased upward when using a serial sampling scheme in momi2, for the rapid decline scenario in momi2, and for some scenarios for the stairway plot, but were otherwise fairly unbiased (Supplemental Figure 2b, Supplemental Figure 4).

Generation time and accuracy

Accuracy of inferences for $N_{e,H}$ based on simulations conducted using a generation time of 3 exhibited similar accuracy for momi2 and GONE overall compared to simulations conducted using a generation time of 1 (Figure 4, Supplemental Figure 6). For the stairway plot, however, $N_{e,H}$ estimates for the longer generation time were less accurate and were biased upward. Estimates for $N_{e,C}$ were biased slightly lower for a generation time of 3 but were otherwise fairly accurate for GONE between the two generation times. However, increasing generation time greatly reduced accuracy for the stairway plot and for momi2 when using contemporary-only data (Figure 4, Supplemental Figure 7).

Effects of ancestral expansions and bottlenecks on accuracy and model selection

Ancestral bottlenecks or expansions did not seem to strongly affect inferences of either $N_{e,H}$ or $N_{e,C}$ made with GONE (Figure 5). For the SFS-based methods, ancestral bottlenecks did not affect estimates of $N_{e,H}$, but ancestral expansions resulted in an upward bias for estimates of $N_{E,H}$ for the stairway plot and in some iterations for momi2. Estimates of $N_{e,C}$ made using whole-genome data with momi2 were somewhat more accurate when an ancestral bottleneck had occurred compared to the constant ancestral size scenario or ancestral expansion scenarios. For RAD data, accuracy for $N_{e,C}$ was highest for the constant size scenario and lowest when an ancestral expansion had occurred. Model selection in momi2 using RAD data was most accurate for the ancestral bottleneck scenario and least accurate for the expansion scenario (Supplemental Figure 8). For WGS data, momi2 again often selected the wrong model for the constant-size scenario with ancestral expansions or bottlenecks (Supplemental Figure 8).

To explore the effects of model misspecification, we also used momi2 to fit models without ancestral size changes to data that did have these changes. In these cases, estimates of $N_{e,H}$ were consistently biased either low (for the ancestral bottleneck scenario) or high (for the ancestral expansion scenario; Supplementary Figure 9). For $N_{e,C}$, accuracy was also reduced somewhat (particularly for the two-sample scenario) and resulted in a slight upward bias for the ancestral bottleneck scenario. In the case of the ancestral expansion scenario, model misspecification did not affect inferences for serial sampling but did result in biases for the other scenarios, particularly the two-sample scenario (Supplementary Figure 9).

Effect of minor allele filtering and singleton errors on momi2

When using contemporary-only data in momi2, minor allele filtering introduced small upward biases in estimated $N_{e,H}$, but a strong downward bias for $N_{e,C}$ (Figure 6). Adding singleton errors also introduced a small upward bias in $N_{e,H}$ but had a much larger effect on $N_{e,C}$, driving a strong downward bias at higher error rates (Figure 6). Adding singleton errors to the contemporary-only dataset at the same rate as the two-sample dataset caused an extremely strong upward bias in estimates of both $N_{e,H}$ and $N_{e,C}$ (Supplementary Figure 10).

Performance of temporal sampling relative to contemporary-only sampling

Two-sample and serial sampling schemes did not show a consistent benefit over contemporary-only sampling in momi2, although these schemes did outperform

contemporary-only sampling in certain cases. When using WGS data, serial and two-sample schemes performed better at identifying more recent slow declines than contemporary-only data (Figure 2b). The two-sample scheme also performed better than the contemporary-only scheme when the generation time was 3 years (Supplementary Figure 6). Finally, the bias produced by minor allele filtering was less pronounced for two-sample data and absent for serially sampled data (Supplementary Figure 11).

Discussion

Researchers interested in estimating population size change over recent time scales now have a larger selection of methodological tools and types of data available to them than ever before. Combined with limited resources, this can lead to difficult choices between generating more data from many contemporary individuals, obtaining historical and modern genomic data to provide a temporal comparison from fewer individuals, or generating whole-genome data and, potentially, a reference genome for their species of interest. While the suite of tools available is constantly expanding, the simulation-based analyses presented here provides guidance for such researchers regarding how certain types of data and analyses perform relative to others and which analyses may be most robust to potential confounding factors.

Inferring recent versus historical population sizes

Inferring historical changes in population size using modern methodologies has generally been considered an easier task than inferring recent changes, with the former possible based on a single whole genome (Li and Durbin 2011) or a reduced-representation data from a small number (>10) of individuals (Beichman et al. 2018). Our results supported this generalization for SFS-based methods. Inference using momi2 and the stairway plot both resulted in extremely low error rates regardless of sample size or data type (whole genome vs reduced representation), though with some caveats that incorrect model specification introduces biases with momi2. Among the linkage-based methods, GONE performed somewhat worse than either SFS-based method for inferring historic size, although performance improved at higher sample sizes and error was still fairly low (usually <25%).

In contrast to inference of historical population dynamics, Beichman et al. (2018) recommended avoiding inference regarding recent demographic events (within the last hundred generations) using whole-genome data for fewer than 10 individuals or reduced-representation data for fewer than 100 individuals. The greater difficulty involved in inferring contemporary population size was reflected in generally much higher error rates for recent population sizes compared to historical sizes for most methods. In line with Beichman et al.'s recommendations, we observed the lowest error rates for the whole-genome data linkage method GONE, particularly when sample sizes were larger than 25, and for momi2 and NeEstimator when using sample sizes greater than 100. The stairway plot generally performed worse than momi2 for inferring recent size except when sample sizes were large (100). However, the stairway plot did outperform a number of other whole genome methods and accurately reconstructed an approximately 100-fold decline over the past hundred years in Tasmanian devils (Patton et al. 2019).

The utility of temporal versus contemporary data

Temporal sampling schemes contain specific information that contemporary-only schemes lack, in the form of both direct information on the genetic composition of past populations (which is leveraged by methods that provide point estimates of population size, such as the

LD method implemented in NeEstimator) as well as information on the magnitude of genetic drift over time (used by the Jorde-Ryman method, Jorde & Ryman 2007). Our results, however, demonstrate that contemporary-only samples contain a substantial amount of information on changes in size over time as well, and it may not be necessary or sufficient to incorporate temporal data in order to accurately infer population sizes. This may be somewhat counterintuitive, as temporal data have been used extensively in the past for inferring population size changes (e.g. Ramakrishnan et al. 2005, Skoglund et al. 2014, Nunziata et al. 2017). Critically, however, when testing a method that can incorporate either contemporary or temporal schemes (momi2), we found that temporal data did not perform noticeably better compared to contemporary-only data when keeping the total number of samples constant and assuming the model was appropriately specified. A possible explanation of this pattern is that the additional information on rare alleles gained from sequencing twice the number of individuals in a contemporary-only sample. These alleles can be particularly informative for inferring recent changes in population size; rare alleles will be lost quickly in a bottleneck (Allendorf 1986).

NeEstimator using temporal data performed noticeably worse than other methods, and performance did not seem to be increased by increasing sample size. In contrast to our results, Nunziata and Weisrock (2017) found NeEstimator to be more accurate for detecting recent size changes, although they only assessed scenarios where the starting population size was on the smaller end of the range used here. The performance of this method is dependent on population size, and as most of our scenarios involved population declines the higher historic population sizes may have affected both. NeEstimator in particular requires a small but substantial proportion of the population to be sampled (~1%; Marandel et al. 2019), and in cases where the population size is on the order of 10,000, our simulated datasets would not have had sufficient numbers of individuals. Sufficient historic sample sizes may be possible to obtain sometimes, but in many cases would be difficult to obtain for many species. It may thus be difficult to use NeEstimator to accurately infer historic population sizes for most populations, unless they were historically very small and isolated.

Confounding factors

Both MAF filtering and the presence of singletons associated with sequencing error can cause extreme biases in estimates of contemporary population size for SFS-based methods. The loss of rare alleles is characteristic of a bottleneck (Allendorf 1986, Garza & Williamson 2001), and the application of a minor allele filter can create the illusion of a severe, recent bottleneck. An excess of rare alleles, on the other hand, is characteristic of a recent expansion (Keinan and Clark 2012), and the introduction of singleton errors could therefore lead to erroneous inference of an expansion. Interestingly, temporal sampling did seem to reduce the bias associated with minor allele filtering in our analyses, possibly because temporal data contain more explicit information on drift. Another explanation for this observation may be that, with smaller sample sizes per time point, applying a dataset-wide MAF cutoff will be less likely to remove truly rare alleles as the observed population allele frequencies are more affected by sampling variation and smaller sample sizes at each time point. We note that although we did not consider other sequencing artifacts that can have a substantial effect on the SFS (i.e. allelic dropout for RAD data; Heller et al. 2021), researchers should be aware of these as additional confounding factors in demographic inference. We also did not examine the effects of singletons on WGS methods. Singletons can be masked in Stairway Plot 2 (Liu and Fu 2020) and for GONE MAF has little effect on estimated population size (Novo et al. 2022), and as such these methods should be less sensitive to singletons and minor allele filtering, respectively. However, future work examining the effect of genotyping error could

be worthwhile, especially since WGS studies in non-model species often use low-coverage WGS for which accurate genotype calling is difficult (Lou et al. 2021).

Historic population sizes are rarely stable over deeper time scales, and as such it is important for demographic inference methods to be robust to these more ancient changes. We simulated a 10-fold expansion or decline similar to a demographic change experienced by many organisms at the time of the last glacial maximum (Hewitt 2004). Encouragingly, the methods we examined seemed to be fairly robust to more ancient changes when inferring recent or historic sizes. Demographic reconstruction methods such as the stairway plot or GONE possess a built-in ability to infer these changes as they attempt to infer the entire demographic history of the population. Care must be taken with methods in which the user specifies the demographic model to fit, such as momi2, since these methods will only include ancient declines if the user includes them in the set of models to assess. If they are not included, then inferences of historic and recent size may be severely biased, as seen in our results. We also note that while we did not include recent population expansion in our set simulated scenarios, the signatures of expansion and decline are opposite (Beichman et al. 2018), and power to distinguish expanding populations from declining populations should be at least as high as power to distinguish expanding populations from stable populations. Accurately detecting recent expansions (in, for example, invasive species) is also highly relevant to conservation and would be a worthwhile avenue for future research.

The genetic signal of demographic change accumulates on the scale of generations, and as such longer generation times (and therefore fewer generations elapsed) could severely reduce accuracy for inferring recent change. We did find lower accuracy for SFS-based methods when we increased generation time to 3 years. Organisms of conservation concern may have much longer generation times, and care should be taken to consider the number of generations since suspected declines. Nunziata and Weisrock (2017) found that declines were detectable when 10-20 generations had elapsed since the start of a decline. GONE seemed to perform fairly well even when only 10 generations had elapsed since the start of a decline, suggesting that using this method with whole genome data may be the best for inferring recent declines when generation times are longer. Before choosing a data type and method, researchers should consider the generation time of their study organisms and the number of generations that have elapsed since suspected changes in population size. If generation time is unknown (as it may be for some non-model species), researchers can attempt to estimate generation time for closely related species. Since the timing of estimated declines is scaled to generation time, uncertainty in generation time will mainly result in uncertainty in the timing of the change rather than the magnitude of change.

A number of factors that we did not examine here could potentially confound the inference of population size. We considered only a single panmictic population for simplicity. Inferring population size is less straightforward in structured populations and populations receiving migration (Orozco-terWengel 2016, Mazet et al. 2016). Model-based analyses can potentially include migration and population structure in their framework, although it would be important to sample all populations in that case. GONE seems to be robust to high levels of gene flow (in which case it infers a metapopulation-level estimate of size), but low levels of migration can distort estimates (Santiago et al. 2020). Researchers should be aware of any potential population structure when applying these methods.

The time scale on which whole genome data are informative for inferring recent change will depend on the frequency of recombination - specifically, recent recombination events are

more likely to occur between alleles that are less tightly linked (McVean 2002). As such, long-range linkage data are necessary for inferring recent demography. While a reference genome is needed for providing the linkage information, our investigations suggest that the reference genome used does not need to be chromosome-scale. Specifically, reducing the size of known linkage groups to 5 centiMorgans did not seem to meaningly affect inference with GONE, suggesting that even when using an incomplete draft genome, this method can provide reliable inference (Supplemental Figure 12). Recombination rate variation could also potentially influence inferences of population size made using GONE. While recombination rate was fully determined by physical distance for out simulated datasets, recombination rate can vary substantially across the genome in real populations (Peñalba & Wolf 2020). When it is possible to construct accurate linkage maps, incorporating these maps in GONE and similar analyses would improve inference of recent population size.

In a recent review, Marchi et al. (2021) noted that whole genome data may not always be ideal for demographic inferences compared to reduced-representation data, since patterns of variation in whole genome data will be more influenced by non-stationary processes such as variation in recombination rates and selection across the genome that are difficult to model. It will be important to detect and account for these processes whenever possible when using whole genome data. GONE seems to be robust to selection (Novo et al. 2022) and can incorporate observed recombination rates across the genome rather than use a uniform rate (Santiago et al. 2020), meaning that this method could potentially surmount these obstacles presented by genome-scale data.

Recommendations and future directions

Based on our results, we recommend different methods for inferring recent changes in population size depending on the samples and resources available. When contemporary whole-genome sequencing data can be collected from at least 50 samples and a reasonably complete draft genome is available, we recommend the linkage-based methods implemented in GONE. These methods appear powerful and accurate across a wide range of demographic scenarios . In contrast, we recommend the SFS-based momi2 when linkage information and whole-genome data are not available. In particular, we recommend serial sampling with momi2 to help reduce the impacts of model misspecification or genotyping error, both of which are difficult to fully avoid. Care must be taken, however, to ensure that the SFS used for inference with momi2 accurately represents the full SFS (including rare alleles) in the population of interest. While NeEstimator performed relatively poorly in our tests, it could be useful when historical and contemporary samples are available and when an appreciable fraction of the population (1%) has been sampled at each time point.

There are currently gaps in methods that can incorporate whole-genome data with historical samples and in methods that can combine SFS and linkage information. ABC and machine learning methods could bridge this gap (Beichman et al. 2018, Schrider and Kern 2018, Sanchez et al. 2021), and they represent promising approaches for integrating multiple data types. Moving forward, it will be important to evaluate these new methods under a wide range of scenarios and data types to determine how useful they are for inferring recent size changes.

References

Alagona PS, Sandlos J, Wiersma YF. 2012. Past imperfect: using historical ecology and baseline data for conservation and restoration projects in North America. Environmental Philosophy 9:49–70.

Allendorf FW. 1986. Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology 5:181–90.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92.

Beichman AC, Huerta-Sanchez E, Lohmueller KE. 2018. Using genomic data to infer historic population dynamics of nonmodel organisms. Annual Review of Ecology, Evolution, and Systematics 49:433–56.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: next-generation museum population genomics. Molecular Ecology 22:6018–32.

Brandies P, Peel E, Hogg CJ, Belov K. 2019. The value of reference genomes in the conservation of threatened species. Genes 10:846.

Caughley G. 1994. Directions in conservation biology. Journal of Animal Ecology 63:215–44.

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM. Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science Advances 1:e1400253.

Díez-del-Molino D, Sánchez-Barreiro F, Barnes I, Gilbert MTP, Dalén L. 2018. Quantifying temporal genomic erosion in endangered species. Trends in Ecology & Evolution 33:176–85.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources 14:209–14.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLOS Genetics 9:e1003905.

Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. Molecular Ecology 10:305–18.

Griffiths RC, Tavaré S. 1998. The age of a mutation in a general coalescent tree. Communications in Statistics Stochastic Models 14:273–95.

Habel JC, Husemann M, Finger A, Danley PD, Zachos FE. 2014. The relevance of time series in molecular ecology and conservation biology. Biological Reviews 89:484–92.

Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2019. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Molecular Ecology Resources 19:552–66.

Haller BC, Messer PW. 2019. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Molecular Biology and Evolution 36:632–37.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res 13:635–43.

Hedrick PW, Kalinowski ST. 2000. Inbreeding depression in conservation biology. Annual Review of Ecology and Systematics 31:139–62.

Hill WG. 1972. Effective size of populations with overlapping generations. Theoretical Population Biology 3:278–89.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. Genetics Research 38:209–16.

Hoey JA, Able KW, Pinsky ML. 2022. Demographic but not genetic recovery of a once overfished species. Molecular Ecology, in press.

Jorde PE, Ryman N. 2007. Unbiased estimator for genetic drift and effective population size. Genetics 177:927–35.
Kamm J, Terhorst J, Durbin R, Song YS. 2020. Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association 115:1472–87.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740–43.
Kelleher J, Lohse K. 2020. Coalescent simulation with msprime. Methods Mol Biol 2090:191–230.
Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. Trends in Ecology & Evolution 17:230–41.
Lehnert SJ, Kess T, Bentzen P, Kent MP, Lien S, Gilbey J, Clément M, Jeffery NW, Waples RS, Bradbury IR. 2019. Genomic signatures and correlates of widespread population declines in salmon. Nat Commun 10:2996.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–96.
Linck E, Battey CJ. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources 19:639–47.
Liu X, Fu Y-X. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome Biology 21:280.
Lou RN, Jacobs A, Wilder AP, Therkildsen NO. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology 30:5966–93.
Marandel F, Lorance P, Berthelé O, Trenkel VM, Waples RS, Lamy J-B. 2019. Estimating effective population size of large marine populations, is it feasible? Fish and Fisheries 20:189–98.
Marchi N, Schlichta F, Excoffier L. 2021. Demographic inference. Current Biology 31:R276–79.
Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. 2016. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity 116:362–71.
McClenachan L, Ferretti F, Baum JK. 2012. From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems. Conservation Letters 5:349–59.
McVean GAT. 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162:987–91.
Meek MH, Larson WA. 2019. The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. Molecular Ecology Resources 19:795–803.
Momigliano P, Florin A-B, Merilä J. 2021. Biases in demographic modeling affect our understanding of recent divergence. Molecular Biology and Evolution 38:2967–85.
Neel MC, McKelvey K, Ryman N, Lloyd MW, Short Bull R, Allendorf FW, Schwartz MK, Waples RS. 2013. Estimation of effective population size in continuously distributed populations: there goes the neighborhood. Heredity 111:189–99.
Nielsen EE, Hansen MM. 2008. Waking the dead: the value of population genetic analyses of historical samples. Fish and Fisheries 9:450–61.
Novo I, Santiago E, Caballero A. 2022. The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. PLOS Genetics 18:e1009764.
Nunziata SO, Lance SL, Scott DE, Lemmon EM, Weisrock DW. 2017. Genomic data detect corresponding signatures of population size change on an ecological time scale in two salamander species. Molecular Ecology 26:1060–74.

Nunziata SO, Weisrock DW. 2018. Estimation of contemporary effective population size and population declines using RAD sequence data. Heredity 120:196–207.
Oosting T, Star B, Barrett JH, Wellenreuther M, Ritchie PA, Rawlence NJ. 2019. Unlocking the potential of ancient fish DNA in the genomic era. Evolutionary Applications 12:1513–22.
Orozco-terWengel P. 2016. The devil is in the details: the effect of population structure on demographic inference. Heredity 116:349–50.
Patton AH, Margres MJ, Stahlke AR, Hendricks S, Lewallen K, Hamede RK, Ruiz-Aravena M, Ryder O, McCallum HI, Jones ME, Hohenlohe PA, Storfer A. 2019. Contemporary demographic reconstruction methods are robust to genome assembly quality: a case study in Tasmanian devils. Molecular Biology and Evolution 36:2906–21.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81:559–75.
Ramakrishnan U, Hadly EA, Mountain JL. 2005. Detecting past population bottlenecks using temporal genetic data. Molecular Ecology 14:2915–22.
Sanchez T, Cury J, Charpiat G, Jay F. 2021. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources 21:2645–60.
Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A. 2020. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Molecular Biology and Evolution 37:3642–53.
Schrider DR, Kern AD. 2018. Supervised machine learning for population genetics: a new paradigm. Trends in Genetics 34:301–12.
Skoglund P, Sjödin P, Skoglund T, Lascoux M, Jakobsson M. 2014. Investigating population history using temporal genetic differentiation. Molecular Biology and Evolution 31:2516–27.
Waples RS. 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv Genet 7:167.
Willis KJ, Bennett KD, Walker D, Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 359:183–95.

Data accessibility
Scripts used to conduct simulations and plot results can be found at https://github.com/philippinespire/PIRE_TemporalSims_HPC.

Funding
This work was supported by National Science Foundation grant # 1743711 (The Philippines PIRE Project: Centennial Genetic and Species Transformations in the Epicenter of Marine Biodiversity).

Acknowledgments
The authors would like to thank PIRE project personnel, members of the Pinsky Lab, Anthony Snead, and two anonymous reviewers for their constructive comments during the development of this study and the writing of this manuscript. The authors also acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

Table 1

Software	Inference Method	Data Type	Temporal Scheme	Input Format	Additional inputs
momi2	SFS	WGS or RAD	Contemporary or Temporal	SFS/VCF*	Mutation rate, generation time
Stairway Plot	SFS	WGS	Contemporary Only	SFS/VCF*	Mutation rate, generation time
NeEstimator	LD	RAD	Temporal	Genepop	Chromosome locations (optional)
GONE	LD	WGS	Contemporary Only	PLINK	Recombination map (optional)

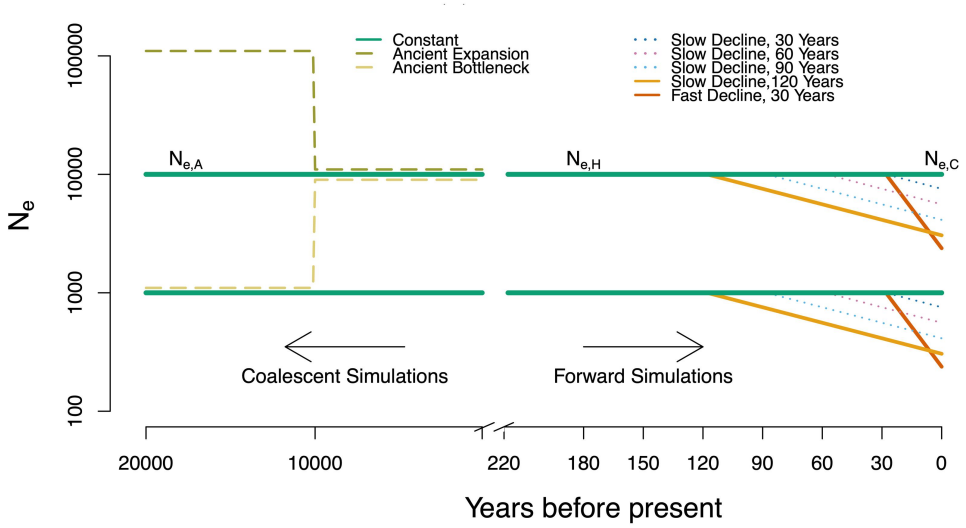

Figure 1. Simulation schemes. Parameters of interest are effective population sizes at three time points: contemporary effective population size (Ne,C), historic effective population size (Ne,H), and ancient effective population size (Ne,A). Solid lines represent simulation scenarios used for mean error calculations (constant size over recent time, a fast decline beginning 30 years before present, and a slow decline beginning 120 years before present). Accuracy and bias were also assessed for three other recent decline scenarios (dotted lines) and two ancestral population size change scenarios (dashed lines).

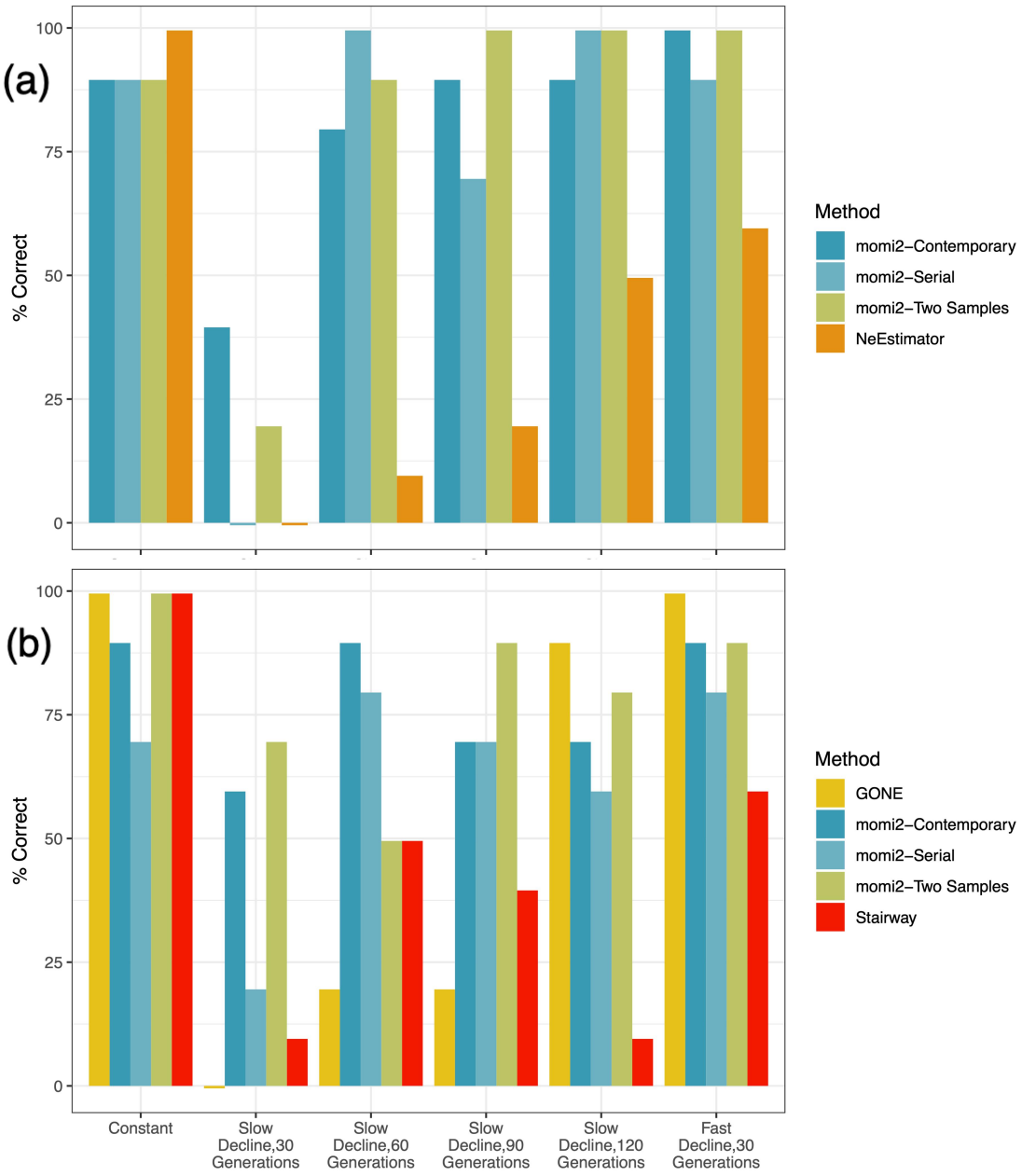

Figure 2. Power for detecting the correct demographic scenario for (a) RAD data and (b) WGS data. Results are shown for the largest sample size for each data type (RAD n = 200, WGS n = 100).

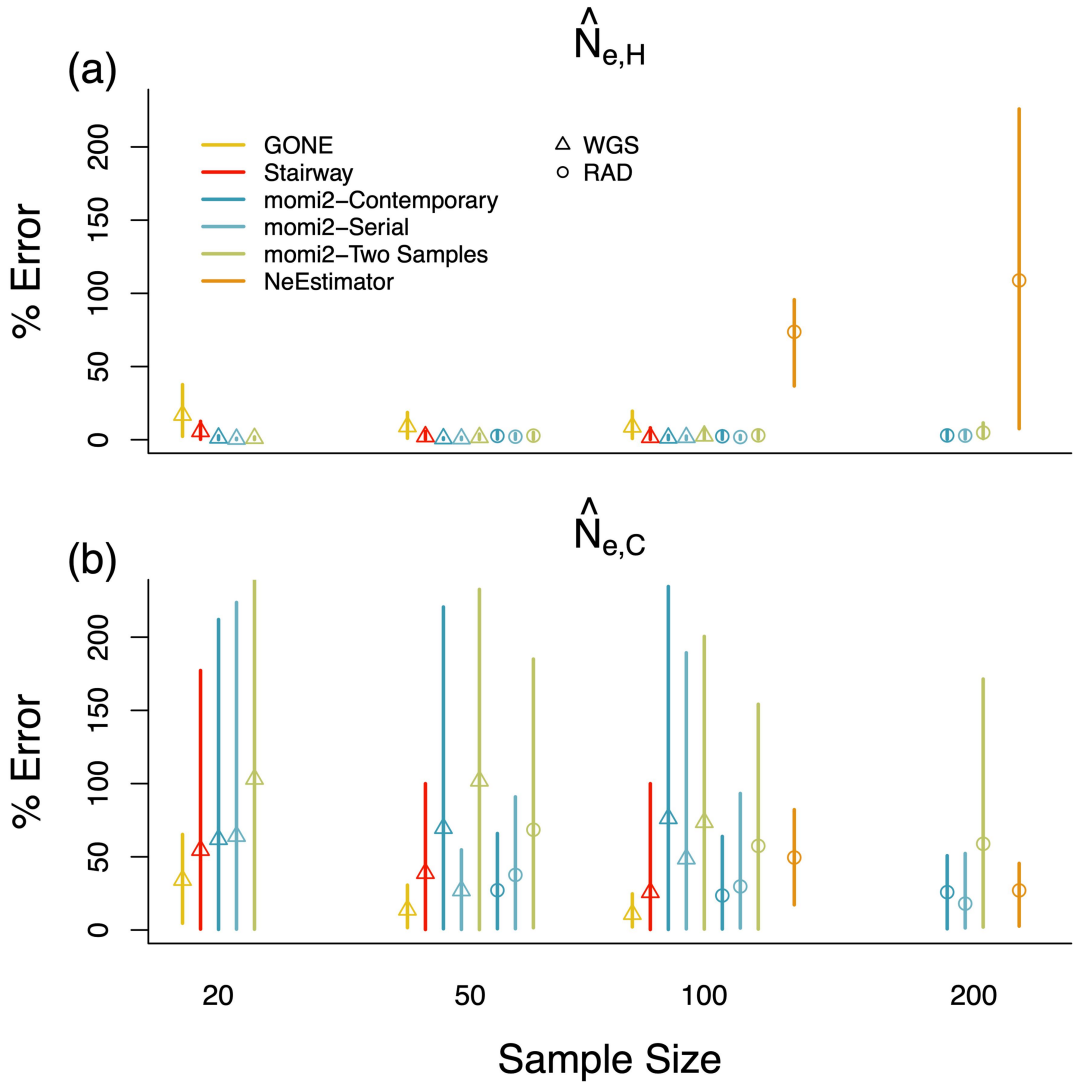

Figure 3. Mean absolute percentage error for estimating (a) historic effective population size (Ne,H) and (b) contemporary effective population size (Ne,C). Error bars show 10%-90% quantiles for each.

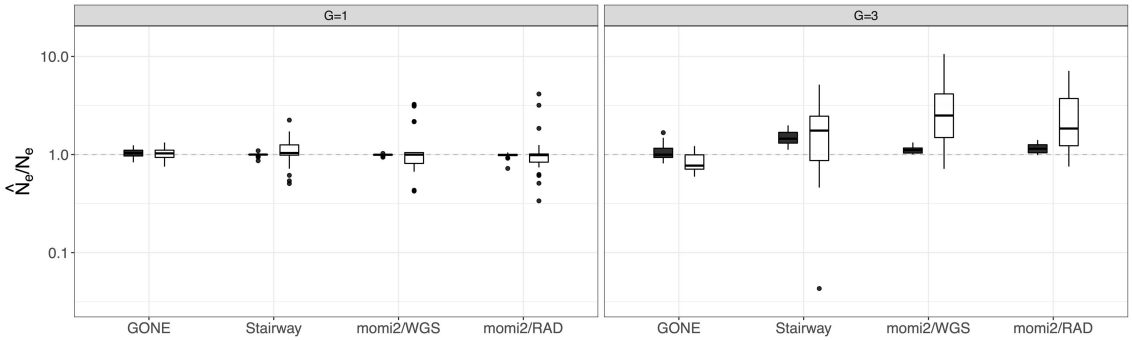

Figure 4. Relationship between generation time and accuracy and precision. Results are shown as box-and-whisker plots for the largest sample size for each data type and for inferences made using contemporary-only data. Results for Ne,H are shown in black and results for Ne,C are shown in white. Ratio of estimated to true Ne is plotted on a log10 scale. Perfect agreement between simulated and estimated values is shown as a 1:1 dotted line.

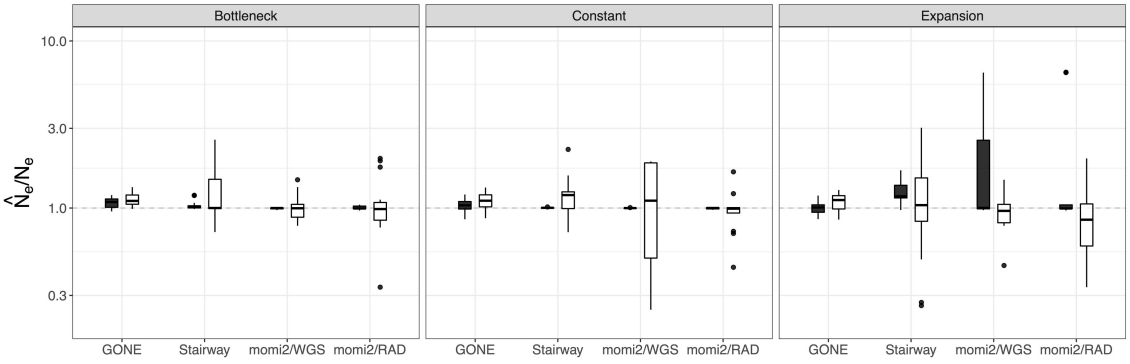

Figure 5. Effects of ancestral expansions and bottlenecks on accuracy and precision. Results are shown for the largest sample size for each data type and for inferences made using contemporary-only data. Results for Ne,H are shown in black and results for Ne,C are shown in white. Ratio of estimated to true Ne is plotted on a log10 scale. Perfect agreement between simulated and estimated values is shown as a 1:1 dotted line. Model selection results for these scenarios are shown in Supplementary Figure 7.

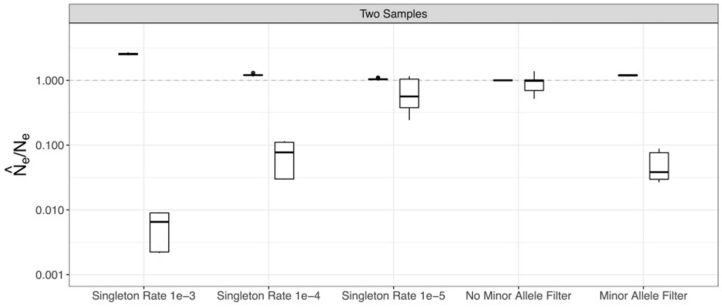

Figure 6. Effect of minor allele filtering and singleton errors on accuracy and precision for momi2. Results for Ne,H are shown in black and results for Ne,C are shown in white. Ratio of estimated to true NE is plotted on a log10 scale. Perfect agreement between simulated and estimated values is shown as a 1:1 dotted line.