Title: Evaluation of site frequency spectrum-based demographic inference methods for use in conservation contexts

Running title: SFS-based demographic methods in conservation

Authors: Isobel Walcott[1], Robyn E Shaw [1], Richard P Duncan[1], Peta Hill[1], Bernd Gruber[1]

Institutional affiliations:

1. Centre for Conservation Ecology and Genomics, University of Canberra, Bruce ACT

Contact information:

Corresponding author: Isobel Walcott, Isobel.Walcott@canberra.edu.au

Abstract

Genomic methods for inferring historical effective population size (N_e) trajectories offer valuable tools for conservation, yet their reliability under conditions typical of conservation datasets—small sample sizes, reduced-representation SNP data, and recent demographic change—remains poorly characterised. We evaluated the performance of two widely used site frequency spectrum (SFS)–based methods, Stairway Plot 2 and Epos, for reconstructing recent demographic histories relevant to conservation management. Using forward-time simulations in SLiM, we generated 609 unique population trajectories across four demographic scenarios (decline, expansion, stability, and bottleneck) with 20 replicates each, varying sample sizes (20–200 individuals) and number of loci (1,000–50,000 SNPs). Simulated genetic data were provided to both inference methods, and outputs were assessed for computational performance, trajectory reconstruction accuracy, temporal reliability, and N_e estimation error. Both methods reliably detected sustained declines and expansions, with correct trajectory reconstruction exceeding 80% overall. Epos was computationally faster and better at identifying stable trajectories, while Stairway Plot 2 was more accurate for N_e estimation and bottleneck detection. Both methods produced inflated N_e estimates in the most recent ~15 generations, and bottleneck scenarios were consistently difficult to reconstruct. Increasing sample size improved inference more than increasing SNP density. SFS-based demographic inference can effectively identify directional population trends under realistic conservation conditions but should not be relied upon for contemporary N_e estimation or short-lived bottleneck detection. We recommend

prioritising individual sampling over marker density, trimming recent estimates, and integrating SFS-based results with complementary methods such as linkage disequilibrium–based estimators for robust conservation decision-making.

Keywords

Introduction

Global environmental change is driving widespread shifts in species distributions, with some populations experiencing rapid declines while others undergo expansion (Parmesan and Yohe, 2003; Cahill *et al.*, 2013). In this context, a central goal of conservation biology is to understand how population sizes have changed in the recent past, because such changes often signal elevated extinction risk or reveal opportunities for targeted management and recovery. However, we often lack data on past population trends, limiting our ability to determine the likely drivers of change and to identify effective conservation strategies. Genomic demographic inference offers a way to fill this critical gap: by analysing genomic signatures left by past demographic events, these methods can reconstruct population trajectories using genomic data from present-day individuals (Li and Durbin, 2011; Liu and Fu, 2020).

Genomic demographic inference methods focus on reconstructing changes in effective population size (N_e), defined as the size of an idealised population that would experience

the same amount of genetic drift as the real one (Wright, 1931; Crow and Kimura, 1970). Effective population size is a useful measure to reconstruct because it is usually correlated with actual population size (N), and because it is directly linked to rates of inbreeding and loss of genetic diversity (Mergeay *et al.*, 2026), making it a long-standing benchmark in conservation genetics.

Several methods have been developed to reconstruct past N_e trajectories, each relying on different types of genomic data and enabling inference across different time scales. Broadly, three classes of approaches are used: site frequency spectrum (SFS)-based methods, Sequentially Markovian Coalescent (SMC) methods (including PSMC, MSMC, MSMC2, and SMC++), and linkage disequilibrium (LD)-based approaches. Together, these complementary approaches allow population histories to be reconstructed across a wide range of timescales using different data types.

SMC methods infer demographic history from patterns of coalescence along whole genomes, using heterozygosity and recombination patterns to estimate how ancestral lineages merge through time (Li and Durbin, 2011). The original Pairwise Sequentially Markovian Coalescent (PSMC) approach uses a single diploid genome and is particularly powerful for reconstructing deep demographic history (Li and Durbin, 2011). Extensions to this framework, including Multisample Sequentially Markovian Coalescent methods (e.g. MSMC, MSMC2, SMC++), incorporate multiple phased or unphased genomes and improve resolution at intermediate timescales (Schiffels and Durbin, 2014; Malaspinas *et al.*, 2016; Terhorst *et al.*, 2017). However, these methods require high-quality reference genomes and high-coverage whole-genome resequencing (Nadachowska-

79 Brzyska *et al.*, 2021). Such data are often unavailable for non-model or threatened
80 species (Bland *et al.*, 2015a; Borgelt *et al.*, 2022; Liu *et al.*, 2022), limiting their routine
81 use in conservation contexts.

82 LD-based methods (e.g. GONE), in contrast, can estimate very recent effective
83 population sizes by exploiting how LD decays with recombination distance across the
84 genome (Waples and Do, 2008; Santiago *et al.*, 2020). Because LD reflects the strength
85 of recent genetic drift, these approaches can detect demographic changes within the
86 past few to tens of generations. Their utility, however, depends on access to accurately
87 mapped, dense SNP data or whole-genome sequence panels, which require high-quality
88 reference genomes. This requirement also limits the routine application of LD methods
89 in conservation settings.

90 Site Frequency Spectrum (SFS)–based methods (e.g. Stairway Plot 2, dadi, fastsimcoal2,
91 epos) use allele frequency distributions across many individuals to infer coalescent N_e
92 over genealogical timescales that can extend to about $4N_e$ generations (Gutenkunst *et*
93 *al.*, 2009; Wakeley, 2009; Terhorst and Song, 2015; Liu and Fu, 2020; Lynch *et al.*, 2020;
94 Gao, Keinan and Nielsen, 2021). The SFS summarises genome-wide variation as the
95 distribution of allele frequencies across sampled individuals (Beichman, Huerta-
96 Sanchez and Lohmueller, 2018), and its shape carries a clear signature of underlying
97 demographic processes. Under a constant population size in an idealised Wright-Fisher
98 population, the SFS follows an exponential-like decay with many rare alleles; population
99 decline produces a deficit of rare alleles; population expansion generates an excess of
100 rare alleles; and bottlenecks produce a mixed pattern, with a dip at low frequencies but

enrichment at intermediate frequencies. Thus, the SFS provides a compact, information-rich summary of genome-wide variation that can be directly linked to historical demographic trajectories.

SFS-based methods are especially well suited to conservation genomics because they rely on population-level SNP data rather than complete genomes. Reduced-representation sequencing approaches, such as restriction-site associated DNA sequencing (RAD-seq), double-digest RAD-seq (ddRAD), and genotyping-by-sequencing (GBS), routinely generate thousands of genome-wide SNPs without the need for a high-quality reference genome, making SFS-based inference feasible for a wide range of conservation and evolutionary applications (Davey *et al.*, 2011; Kilian *et al.*, 2012; Peterson *et al.*, 2012). Recent methodological developments also allow the use of folded SFS, which does not require knowledge of the ancestral allele state, further expanding the applicability of the approach (Liu and Fu, 2020; Lynch *et al.*, 2020).

SFS-based approaches are flexible in the types of demographic trajectories they can detect, including patterns of stability, expansion, decline, and bottlenecks (Beichman, Huerta-Sanchez and Lohmueller, 2018). Some implementations, such as Stairway Plot 2 and Epos, do not require users to specify a hypothesised trajectory in advance for the model to test, allowing data-driven reconstruction of population trajectories (Liu and Fu, 2020; Lynch *et al.*, 2020). This flexibility makes SFS-based methods particularly useful for conservation applications: they work with readily available genomic data types, can reconstruct population trajectories over timescales relevant to management, and can be applied broadly across taxa without the need for reference genomes.

Evaluations of SFS-based demographic inference have primarily focused on theoretical identifiability limits or on large, high-resolution genomic datasets (Terhorst and Song, 2015; Reid and Pinsky, 2022), rather than on the moderate sample sizes and marker densities typical of most conservation genomic studies. In this paper, we use forward-time genomic simulations tailored to conservation-relevant demographic scenarios to assess the performance of two widely used SFS-based inference methods, Epos and Stairway Plot 2. Our simulations were designed to reflect conditions typical of applied conservation work by varying the number of present-day individuals sampled (10-200), the number SNPs available per individual (1000 – 50000), and by modelling population change within the last 250 generations as either stable, expanding, declining, or undergoing a bottleneck. We focused on comparing the two methods in terms of: (1) computational efficiency; (2) accuracy in reconstructing the correct demographic trajectory; (3) the relative influence of sample size and number of loci on performance; and (4) accuracy of N_e estimates across recent timescales. By systematically varying demographic histories, sample sizes, and number of loci, we aimed to identify the practical limits, biases, and data requirements of SFS-based inference in applied conservation genomics.

Materials and Methods

Simulation of trajectories

Simulations provide a powerful framework for genomic analysis because they can generate realistic genetic datasets with known parameters and population histories. In this study, we used simulations to generate populations with predefined demographic trajectories and compared these known histories with the outputs of demographic inference methods across a range of conditions (Table 1; Figure S1). We used SLiM v.3 (Haller and Messer, 2019) to generate forward-time simulations through the slimr package (Dinnage *et al.*, 2022) within the R environment (R Core Team, 2024).

All simulations assumed a biallelic, diploid chromosome system with a standard recombination rate of 1.0×10^{-8}. We used a mutation rate of 1.0×10^{-8}, which is biologically plausible, although on the lower end (Bergeron et al., 2023), and improves computational efficiency by facilitating the generation of the required number of SNPs. Genomes were 5.0×10^{8} base pairs in length, partitioned into five chromosomes each of 1.0×10^{8} base pairs long. Simulated populations were initialised with initial population size held constant for 2000 generations before demographic changes occurred. A Wright-Fisher model with non-overlapping generations with a one-year generation time was used in all simulations. Because one year corresponded to one generation in our simulations, all temporal outputs are represented as years before present.

To ensure the simulations reflected realistic population changes, logarithmic decay models were used to generate decline trajectories and exponential growth models were used for expansions, with both models (decline followed by expansion) used for bottleneck trajectories.

To comprehensively explore the parameter space, we simulated all combinations of key parameters across stable, declining, expanding, and bottleneck scenarios, resulting in 609 unique simulation scenarios (Table 1-2), each replicated 20 times. We focused on population changes within the last 250 years as these are particularly relevant to conservation. For stable, declining, and expanding trajectories, we varied maximum N_e, the magnitude of N_e change, and the rate and timing of change (Table 1). Simulation scenarios were removed where simulations would have finished with fewer individuals than the designated sample size, such as a decline to $N_e = 50$ with an individual sample size of 200. This filtering resulted in a greater number of expansion trajectories relative to decline and stable.

As bottleneck scenarios encompass a wide range of possible parameter combinations, including variation in the timing, severity and duration of both the crash and recovery, we based our bottleneck simulations on empirical examples of documented population declines and recoveries that span different lengths of population suppression and magnitudes of population decline. We chose two well-studied examples:

- Saltwater Crocodile (*Crocodylus porosus*) in the Northern Territory, Australia, which declined with commercial hunting starting in the early 1900s until protections were introduced in the early 1970s that allowed population recovery (Fukuda *et al.*, 2011).
- The Australian Fur Seal (*Arctocephalus pusillus doriferus*), whose southern hemisphere populations have been recovering since the 1980s following severe

depletion by commercial sealing that started in the 1700s (Kirkwood *et al.*, 2010).

To reflect the timeframes in these empirical examples, we simulated bottleneck trajectories over 350 years, longer than the 250-year window we used for the decline, expansion, and stable trajectories. For each species, we modelled both a full recovery and a partial recovery, with the initial population size adjusted accordingly.

For each of the 609 unique simulation scenarios, we generated 20 replicate forward-time genetic simulations using SLiM V.3 via slimr, exporting restriction-site associated genetic samples of present-day (0ybp) individuals. To do this, individuals were randomly sampled without replacement to the specified sample size, with individual-level genetic data extracted, to realistically reflect data obtained for conservation research. Genetic data were extracted as VCF files, which were converted to genlight objects (Jombart, 2008). Each genlight was then subsampled to 1k, 5k, 10k, 20k, and 50k SNPs, resulting in a total of 60,900 possible separate files. However, some simulations produced fewer SNPs than required for a given subsampling level, for example a dataset containing 15k SNPs could only be subsampled at 1k, 5k, and 10k, we recorded the missing datasets as 'simulation failure'. All outputs were parsed to both Epos and Stairway Plot 2 to ensure comparisons were against the same inputs.

Loci with allele counts < 2 were removed to meet filtering requirements across all methods. The resulting genlight objects were then supplied to the demographic inference methods (Epos and Stairway Plot 2) as a folded SFS, reflecting an unphased genome. Because of the code structure we used, outputs from these inference methods

were only accessible once all methods had completed running for a single simulation. Consequently, if any method exceeded the maximum allotted runtime of 24 hours or crashed due to nonconvergence, no outputs from either method were available for downstream analysis. We refer to this outcome as 'inference failure'.

SFS-based demographic inference methods

Demographic inference using Epos (Lynch *et al.*, 2020) and Stairway Plot 2 (Liu and Fu, 2015, 2020) were conducted on the National Computational Infrastructure *Gadi* supercomputer, using an allocation of 125,000 core hours and a maximum wall time of 24 hours. Epos analyses were run with minimum SFS bin size of 1, chromosome length (L) of 5.0×10^{8}, and a mutation rate (μ) of 1.0×10^{-8}. The greedy parameter in Epos, which determines the frequency of breakpoint re-evaluation, was set to the moderately exhaustive level of 'E -2'. Stairway Plot 2 was run with the same minimum bin size, chromosome length, and mutation rate, and random breakpoints were set at 5. For both methods, generation time was set to 1 year and 40 replicates or 'bootstraps' were generated for each inference run, with the mean and confidence intervals across bootstraps provided as the output.

Computational performance

To evaluate the factors limiting the practical use of each inference method, we examined runtime and the rates of inference failure across the methods, sample size, and number

of loci. Runtime was recorded in minutes and only available for successfully completed runs. Rates of inference failure, simulation failure and successful runs were calculated as proportions of all attempted simulations to identify parameter combinations that were unlikely to generate outputs. Because of data availability constraints, these assessments were restricted to decline, expansion and stable trajectories.

Outputs that converged but produced inferred trajectories shorter than 225ybp were considered to have failed to meet the minimum temporal window for valuable outputs and were excluded from further analysis. We acknowledge that this approach was very restrictive however this was decided to ensure that methods could meaningfully capture population sizes prior to the major drivers of demographic change that impact many species of conservation concern globally (urbanisation, habitat fragmentation etc.).

After this filtering, parameter combinations (defined by sample size, number of loci, trajectory type, and method) with fewer than ten successful runs were deemed to have insufficient replication for comparison and were removed. The final dataset comprised 24,374 Epos outputs and 22,818 Stairway Plot 2 outputs across all trajectories (Table 3).

Inference Evaluation

Temporal accuracy

During preliminary analysis, we observed a consistent pattern of inflated N_e estimates in the most recent years, which could be misinterpreted as evidence of a recent expansion. To quantify this bias and identify periods of reduced accuracy, we calculated the root

mean square error (RMSE) between simulated and inferred N_e values for each year less than 50ybp. Mean ± SE of RMSE values were then calculated across replicate runs and grouped by SNP density and sample size to evaluate the reliability of each method in inferring population trajectories in the most recent time period.

Correct reconstruction of simulated trajectory

Decline, expansion, stable trajectories

We investigated whether each method could identify the form (stable, decline, or expansion) of the simulated trajectory independent of their ability to accurately estimate N_e. To do this, we fit a linear regression model to the data from each output and used the resulting slope parameter to classify the trajectory (Figure S2). The minimum slope parameter we specified in simulated declines and expansions was±0.102. Given there was variability in inference outputs, we classified slopes as identifying a decline or expansion if the inference method generated a slope parameter that was greater than half the minimum slope we specified. Hence, slope parameters > 0.05 were classified as declines, < -0.05 as expansions, and values within ± 0.05 were classified as stable trajectories. In addition, to be classified outputs required a slope p-value < 0.

Bottleneck trajectories

To determine whether each method could successfully reconstruct bottleneck trajectories, we divided each run's inferred trajectory into two segments, split at the minimum N_e within 50 years either side of the midpoint between the start of the crash

and end of the recovery. We then evaluated each segment for evidence of a decline followed by an expansion using the same criteria as above. This more prescriptive approach was employed to minimize the risk of false signals given the greater variation inherent in bottleneck outputs due to the larger number of ways they could differ, to focus on the expected bottleneck timeframe while accommodating this variability, and to evaluate timeframes relevant to conservation management. A run was classified as a bottleneck when it showed a significant decline ($p < 0.05$) prior to the minimum N_e point and a significant expansion afterward. The proportion of runs correctly identifying bottlenecks was then assessed across species, SNP density, and sample sizes to identify the conditions under which each method performed best.

Accuracy of N_e estimates

To evaluate the accuracy of N_e estimates, we quantified the error between inferred and simulated values. Accuracy across each run was assessed using root mean square error (RMSE), calculated over a trimmed temporal window of either 15-250ybp for decline, expansion, and stable trajectories, or 15-350ybp for bottleneck trajectories due to N_e inflation in the most recent years 0-15ybp (see 'Temporal accuracy' in the results). RMSE was compared across methods, SNP densities, and sample sizes to assess how these factors influence accuracy.

We then related N_e accuracy to the rate of correct trajectory reconstruction for all trajectory types and both inference methods, providing an integrated view of overall

method performance. This pairing allowed us to identify conditions under which each method produced both accurate population sizes and correct demographic trajectories.

Results

Computational Performance

The number of completed runs varied across population trajectories, sample sizes and number of loci. Simulation failure was the most common cause of missing outputs , occurring in up to 99.7% of runs for a given parameter combination, with the greatest number of failures under high locus counts (≥50k) and small sample sizes (≤20) (Figure S3). Expansion scenarios had the highest rates of simulation failure, likely due to prolonged periods of low population size prior to expansion, which reduced the number of segregating loci. Inference failures ranged from 0% to 10% within parameter combinations and were greatest for stable trajectories. Conversely, successful inference runs were most frequent for stable trajectories (62.8% across all runs), followed by decline (59.2%) and lowest for expansions (49.2%).

Rates of run convergence were highest when fewer loci were simulated (<20k) and when greater numbers of individuals were sampled (>100). Among successful inference runs, runtime patterns differed markedly between methods, with Epos operating substantially faster than Stairway Plot 2 (mean runtimes: Stairway Plot 2 = 62.5 minutes; Epos = 0.423 minutes; Figure 1, Figure S4). When sample sizes were small (<100), Stairway Plot 2 and Epos runtimes did not increase with greater loci.. Although, as sample sizes and loci

increased, Stairway Plot 2 runtimes rose sharply, while Epos runtimes remained consistent across sampling combinations (max: 11 minutes for unidirectional trajectories). Bottleneck scenarios produced substantially longer runtimes (mean bottleneck runtimes: Stairway Plot 2 = 175 minutes, Epos = 1.93 minutes), likely reflecting the higher simulated N_e values in these trajectories, which require greater computational effort to converge.

Inference Evaluation

Temporal accuracy

The ability of methods to accurately reconstruct N_e values (measured by RMSE) varied across time periods, with several windows showing consistently reduced accuracy (Figure 2). For both methods, in almost every time period, bottleneck trajectories had higher RMSE than decline, stable or expanding trajectories, suggesting both methods less accurately recovered bottleneck N_e values. For the other trajectories, mean RMSE values were often substantially higher in the most recent period spanning 0-15ybp, particularly for Epos.

Beyond 0-15ybp, RMSE was often markedly lower under decline, expansion and stable trajectories, indicating that it was only in this most recent period that inference about N_e values was less accurate. This pattern was less pronounced for bottleneck trajectories, which were less accurate across all time windows for both methods.

High RMSE in the most recent period was amplified for Epos by large sample sizes (n = 200), particularly with decreasing numbers of loci (1k) (decline: 1.34×10^{5}). For Stairway Plot 2, RMSE in the most recent period was highest under large numbers of loci (50k) and low sample sizes (≤50). Under stable trajectories, variation in RMSE in the most recent period showed no consistent pattern across number of loci and sample size for both methods. However, Epos RMSE declined sharply further back in time and became more accurate than Stairway Plot 2 beyond >10ybp.

While there was often high RMSE in the most recent period, this had generally declined and plateaued by 15ybp across most parameter combinations for both methods. Consequently, in subsequent analyses we used trimmed outputs in which estimates <15ybp were removed to focus on the more reliable portions of the trajectories. For consistency, we also trimmed bottleneck scenarios to exclude values <15ybp before further analysis.

Correct reconstruction of simulated trajectory

We evaluated methods for how well they recovered the overall trajectory shape (decline, stable, expansion, or bottleneck) independent of how accurately they reconstructed N_e values, with outputs assessed as a proportion of the 24,374 Epos outputs and 22,818 Stairway Plot 2 outputs . Epos most frequently identified the correct trajectory (88.1% of runs), closely followed by Stairway Plot 2 (81.2% of runs). The accuracy of both methods improved with increasing numbers of loci (>5k) and individuals (>20) particularly for decline scenarios. Performance was consistently high for expansion scenarios (Epos

355 expansions: 98.6%, Stairway Plot 2 expansions: 99.4%). Under stable scenarios,
356 increasing the number of loci or individuals sampled had no clear effect on
357 reconstruction success for either method. However, Epos was more accurate than
358 Stairway Plot 2 across all stable trajectories (Epos: 73.1%, Stairway Plot 2: 44.8%; Figure
359 3).

360 Reconstructing bottleneck trajectories proved more difficult than the other scenarios,
361 reflecting the greater complexity of bottlenecks. Stairway Plot 2 recovered bottleneck
362 trajectories more consistently than Epos (55.2% versus 31.1% of runs). For both
363 methods, reconstruction success generally improved with more loci and individuals
364 sampled, with sample size exerting the strongest influence (Figure 3). The positive effect
365 of adding more loci diminished once sample sizes reached 100 or more, suggesting that
366 larger numbers of individuals can compensate for more limited genomic sampling.

367

368 *Accuracy of N_e estimates*

369 Stairway Plot 2 produced lower mean RMSE values than Epos under declining, stable,
370 and expanding scenarios (Stairway Plot 2: 141, Epos: 146; Figure 4). For both methods,
371 inference accuracy was generally stable across different numbers of loci and sample
372 sizes, although accuracy slightly improved as these increased. Accuracy was highest
373 under stable scenarios, indicating both methods reproduced simulated N_e values
374 reliably in this scenario even when they often misclassified this trajectory.

Bottleneck scenarios had substantially higher RMSE than the others (Figure 4), although mean RMSE for bottlenecks were lower when using Stairway Plot 2 (6198) than for Epos (7493), as was observed for the other trajectories. For both methods, inference accuracy improved with increasing sample size but not numbers of loci.

Overall Performance

We summarised overall performance by plotting RMSE values against the rate of correct trajectory identification (Figure 5). For both methods, decline and expansion scenarios typically had both low RMSE values and high rates of trajectory identification, indicating that both methods can reliably identify the correct trajectory and accurately estimate N_e values, although Stairway Plot 2 generally performed better than Epos. Stable trajectories also had low RMSE values, indicating accurate estimation of N_e values, but both methods often struggled to correctly identify this trajectory, with Epos performing better than Stairway Plot 2.

For both Epos and Stairway Plot 2, bottleneck scenarios showed lower overall performance than the other trajectories. Generally lower RMSE values and higher rates of correct trajectory identification indicated that Stairway Plot 2 performed better than Epos, but neither could reliably reconstruct bottleneck trajectories. This limitation may be associated with the more prescriptive approach employed in classifying bottleneck trajectories, which would have failed to identify a bottleneck trajectory if it was reconstructed with an error of >50 generations.

Discussion

Here, we assessed the performance of two widely used SFS-based demographic inference methods, Stairway Plot 2 and Epos, under conditions typical of many applied conservation genomic studies. We used forward-time simulations to generate genomic datasets associated with known population trajectories and then examined the ability of both methods to reconstruct those trajectories from a sample of present-day individuals. We evaluated the performance of each method in terms of their computational behaviour, ability to identify the correct trajectory, and accuracy of N_e estimates across a range of sample sizes, loci, and demographic scenarios typical of conservation applications. We focused on small to moderate N_e, reduced-representation genomic datasets, and demographic change occurring over recent timeframes (<250 generations, here equivalent to years before present), features typical of many empirical studies.

Overall, both SFS-based methods performed well at detecting broad patterns of demographic change under many realistic conservation scenarios. Population declines and expansions were generally reconstructed reliably, particularly when individual sample sizes were sufficient. However, inference accuracy declined towards the present, bottlenecks were difficult to detect consistently, and stable demographic histories were sometimes misclassified. These limitations have important implications for how SFS-based demographic reconstructions are interpreted and applied in

conservation decision-making, where interest often centres on recent change, complex histories, and contemporary status.

Performance of SFS-based inference under conservation-relevant conditions

Computational performance differed markedly between methods. The ability of methods to converge was made more challenging in cases with low sample size and large numbers of loci, where SNP frequencies were distributed across shorter spectrums. Convergence time for both Stairway Plot 2 and Epos was influenced by the number of loci, number of individuals sampled, and the underlying demographic scenario. Epos exhibited substantially lower and more consistent runtimes across the parameter space, whereas Stairway Plot 2 runtimes increased considerably with increasing individual sample size and increasing numbers of loci (Stairway Plot 2 took 147 times longer than Epos, on average). These differences are consistent with previous assessments showing the improved computational efficiency of Epos over Stairway Plot 2 (Lynch *et al.*, 2020). From a practical perspective, Epos may therefore be better suited for exploratory analyses, sensitivity testing, and simulation-based assessments, particularly where computational resources are limited or where many alternative models must be evaluated. Where higher density SFSs are present (low sample size, large numbers of SNPs), loci could be subsampled to facilitate convergence.

Analytical performance of the two methods was broadly similar, reflecting their shared theoretical foundations (Lynch *et al.*, 2020). Across decline and expansion scenarios, both Stairway Plot 2 and Epos were generally effective at reconstructing the correct

qualitative demographic trajectory and producing N_e estimates close to simulated values. Inference accuracy improved more with greater numbers of sampled individuals than with increased numbers of loci. Gains plateaued at relatively modest numbers of loci (10k SNPs), whereas increasing the number of individuals continued to improve N_e accuracy, and to a lesser extent, trajectory detection. Similar diminishing returns with increasing marker density have been reported in previous evaluations of SFS-based inference (Reid and Pinsky, 2022). This pattern is consistent with theoretical results showing that the information content of the SFS scales with the number of sampled chromosomes (i.e. gene copies), which in diploid species increases directly with the number of individuals sampled (Terhorst and Song, 2015). Additional loci contribute limited new information once the allele frequency spectrum is well characterised, while larger sample sizes more effectively resolve allele frequency variation.

Reduced-representation approaches that prioritise sampling more individuals at moderate numbers of loci are likely to provide greater inferential benefit than sequencing fewer individuals at very high marker density. This is particularly relevant where conservation resources are limited and trade-offs between sequencing depth, marker number, and sample size are unavoidable (Bertola *et al.*, 2024). However, this is likely to vary depending on the degree of missing data and its distribution (Gargiulo *et al.*, 2024), as well as the presence of population structure (Mazet *et al.*, 2016; Orozco-terWengel, 2016; Chikhi *et al.*, 2018). Moderate levels of missing data primarily reduce the effective number of loci contributing to the SFS, rather than fundamentally altering inference, but non-random missingness, such as allele dropout associated with degraded samples or

locus-specific amplification bias, can distort allele frequency distributions and compromise inference. In species with large or complex genomes, or in studies relying on degraded or historical material, marker selection and filtering therefore require careful consideration to ensure adequate locus overlap and representative sampling of allele frequencies.

Limits to inference in the very recent past

A consistent pattern across many scenarios was reduced accuracy in the demographic trajectories inferred for recent generations. Both Stairway Plot 2 and Epos frequently produced inflated or unstable N_e estimates close to the present, resulting in high error, regardless of sample size or SNP number, although this was more pronounced for Epos. This behaviour has been reported previously for SFS-based methods and reflects fundamental limits on the information content of allele frequency data at very recent timescales (Reid and Pinsky, 2022). Because few coalescent events occur close to the present, inference in this region relies on weak signal and is highly sensitive to noise, sampling variance, and model assumptions.

In our analyses, trimming the most recent ~15 generations substantially improved both N_e accuracy and trajectory classification. This finding has direct implications for conservation applications. While SFS-based methods are well suited to identifying whether populations have experienced recent directional change, they should not be used to estimate contemporary N_e or to infer demographic changes occurring within the

last few generations. Instead, inference in this region should be treated cautiously, or excluded altogether, unless supported by complementary data.

Where contemporary N_e estimates are required for management, such as for assessing near-term inbreeding risk or setting recovery targets, linkage disequilibrium-based estimators or direct demographic data remain essential. A pragmatic approach is therefore to combine methods, using SFS-based inference to characterise historical trends and directional change while anchoring or contextualising the most recent time interval using independent contemporary N_e estimates, for example from NeEstimator (Do *et al.*, 2014). Combining these approaches can also help address practical limitations of demographic inference. In our simulations the total genome length (L) was known, allowing us to specify this parameter directly; however, this information is rarely available in empirical studies. Subsampling loci effectively shortens the sampled genome and can compress both the temporal (x-axis) and N_e (y-axis) scales of inferred trajectories if not corrected. Applying a scaling factor based on an independent estimate of contemporary N_e, such as that obtained from linkage disequilibrium methods, can therefore help calibrate the inferred trajectory so that recent N_e values align with empirically derived estimates. Explicitly recognising the different timescales captured by coalescent and LD-based estimators can reduce over-interpretation of unstable recent estimates and improve the defensibility of management recommendations.

Challenges in detecting bottlenecks and stable trajectories

Stable scenarios presented a challenge to both methods, and were sometimes misclassified as weak trends, particularly under low sample sizes. Epos showed somewhat better performance in correctly identifying stability, suggesting that it may be more robust to weak signal in these cases. Counterintuitively, stable trajectories were less likely to be correctly identified with high numbers of loci for both methods (≥20,000). This pattern likely reflects increased model sensitivity rather than reduced information content. As loci numbers increase, the allele frequency spectrum is characterised with greater precision, enabling flexible piecewise demographic models to detect shallow stochastic deviations that may not be biologically meaningful. Minor variation in the SFS therefore may have been over-interpreted as weak demographic change, leading to misclassification of otherwise stable histories. These results indicate that apparent weak trends in inferred trajectories should be interpreted cautiously. Again, it is important to recognise that we did not take the confidence intervals of inference outputs into account, and these may have indicated high uncertainty around weak trends, although RMSE around simulation iterations suggest that these estimates did not vary substantially.

Both methods struggled to reliably detect bottleneck scenarios, particularly when bottlenecks were short-lived or occurred close to the present. In many cases, bottleneck trajectories were misclassified as gradual declines, expansions, or unstable recent behaviour, and were further obscured by the difficulty of resolving short-term demographic events from the SFS and the inflated N_e estimates near the present. Correct detection of a bottleneck did improve with increased numbers of loci and individuals,

however, neither method performed consistently. Higher error around N_e estimates likely reflected the complexity of these simulated population histories. For example, larger effective population sizes were simulated to match real biological systems, which meant relative error could be higher, and the timing had to be correct for both decline and increase. This means that there are multiple dimensions where error can accumulate in bottleneck scenarios. Postaire et al. (2024) also found that Stairway Plot 2 frequently failed to reconstruct simulated expansions, particularly for recent bottlenecks.

Together, these findings highlight that SFS-based inference is most reliable for detecting sustained directional change, and least reliable for identifying short-lived or very recent demographic changes. Overall performance showed that bottlenecks were hardest to detect, followed by stable, decline, and finally, expansion trajectories. Bottleneck detection and discrimination between stability and weak trends should therefore be supported by additional evidence wherever possible, including ecological data, temporal sampling, or alternative genetic estimators. This limitation is especially relevant for conservation, as bottlenecks can have disproportionate genetic consequences, including rapid loss of allelic variation, even when longer-term N_e appears moderate (Allendorf, Hössjer and Ryman, 2024).

Implications for conservation genomics and management

Our findings add to an important body of work clarifying best practice for using demographic inference methods to inform conservation decisions. These methods are well suited to identifying directional demographic change, such as whether populations

have experienced recent declines or expansions, and whether low N_e is likely to be long-standing or the result of recent contraction. In contrast, they are less reliable for estimating absolute N_e close to the present, detecting short-lived bottlenecks, or distinguishing stability from weak demographic trends.

In practice, this means that demographic inference should be framed primarily as a tool for trend detection rather than precise estimation. Interpreting inferred N_e trajectories relative to broad conservation benchmarks, such as whether long-term N_e is likely to fall above or below commonly cited thresholds, is likely to be more informative than focusing on exact N_e values at specific time points. Integrating genomic inference with contemporary N_e estimates, demographic monitoring, or ecological context will provide more robust and defensible guidance for management.

Despite long-standing recognition of the importance of genetic diversity and effective population size, genomic demographic inference has largely remained within the academic domain, for example, for providing accurate inference in the distant past (Nunziata and Weisrock, 2018; Lynch *et al.*, 2020; Reid and Pinsky, 2022). By delineating where these methods perform well, and where caution is required, their outputs can be more meaningfully translated into applied conservation contexts. This may be especially valuable for species assessment and prioritisation. Demographic inference can provide evidence of recent population trajectories for species lacking long-term monitoring data, supporting assessments under frameworks such as the IUCN Red List and national or sub-national listing processes. This is particularly relevant for taxa experiencing cryptic or poorly documented declines (Bland *et al.*, 2015b; Borgelt *et al.*, 2022; Liu *et al.*, 2022).

569 Targeted application of SFS-based inference, interpreted within its known limits, may
570 help identify populations of concern before declines become demographically apparent,
571 enabling earlier and more proactive conservation action.

572

573 *Recommendations and conclusion*

574 Our results demonstrate that SFS-based demographic inference can provide reliable
575 insight into recent population trajectories under conservation-relevant conditions, but
576 only when interpreted within clearly defined limits. These methods are most effective for
577 identifying sustained directional change and distinguishing long-standing low N_e from
578 recent contraction, and less reliable for resolving very recent change or short-term
579 bottlenecks. Rather than functioning as stand-alone diagnostics, they are best
580 embedded within integrated workflows that combine genomic, demographic, and
581 ecological information. Based on our simulations, we propose the following practical
582 guidance:

583 1. Use SFS-based methods to detect directional trends, and be wary of weak trends
584 (pay attention to confidence intervals)

585 2. Treat the most recent generations with caution. The final portion of inferred
586 trajectories is inherently unstable due to limited information content in the SFS.
587 Where contemporary N_e is required, use linkage disequilibrium–based estimators
588 or independent demographic data to complement SFS-based reconstructions.

3. Prioritise sampling more individuals over increasing SNP density. A moderate number of loci is adequate (depending on missing data), and broader individual sampling more effectively improves inference than additional markers.
4. Demographic inference should be integrated with ecological knowledge and complementary genetic analyses, particularly in calibrating axes values to account for the effect of loci subsampling, and results interpreted in light of species' biology and sampling strategy. Bottlenecks can be hard to detect, population structure, migration, and non-random missing data can distort the SFS and generate spurious demographic signals.

By explicitly defining where SFS-based inference performs well and where caution is required, this study helps align methodological capability with conservation decision-making needs. When used within these bounds, SFS-based demographic inference can extend insight into data-limited systems and support earlier identification of populations undergoing sustained decline.

Acknowledgements

The authors wish to thank Russell Dinnage for his support in using the slimr package. This research was conducted under a National Computational Infrastructure adaptor scheme grant.

References

Allendorf, F.W., Hössjer, O. and Ryman, N. (2024) "What does effective population size tell us about loss of allelic variation?," *Evolutionary Applications*, 17(6), p. e13733. Available at: https://doi.org/10.1111/eva.13733.

Beichman, A.C., Huerta-Sanchez, E. and Lohmueller, K.E. (2018) "Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms," *Annual Review of Ecology, Evolution, and Systematics*, 49(1), pp. 433–456. Available at: https://doi.org/10.1146/annurev-ecolsys-110617-062431.

Bergeron, L.A. *et al.* (2023) "Evolution of the germline mutation rate across vertebrates," *Nature*, 615(7951), pp. 285–291. Available at: https://doi.org/10.1038/s41586-023-05752-y.

Bertola, L.D. *et al.* (2024) "A pragmatic approach for integrating molecular tools into biodiversity conservation," *Conservation Science and Practice*, 6(1), p. e13053. Available at: https://doi.org/10.1111/csp2.13053.

Bland, L.M. *et al.* (2015a) "Predicting the conservation status of data-deficient species," *Conservation Biology*, 29(1), pp. 250–259. Available at: https://doi.org/10.1111/cobi.12372.

Bland, L.M. *et al.* (2015b) "Predicting the conservation status of data-deficient species," *Conservation Biology*, 29(1), pp. 250–259. Available at: https://doi.org/10.1111/cobi.12372.

Borgelt, J. *et al.* (2022) "More than half of data deficient species predicted to be threatened by extinction," *Communications Biology*, 5(1), pp. 1–9. Available at: https://doi.org/10.1038/s42003-022-03638-9.

Cahill, A.E. *et al.* (2013) "How does climate change cause extinction?," *Proceedings of the Royal Society B: Biological Sciences*, 280(1750), p. 20121890. Available at: https://doi.org/10.1098/rspb.2012.1890.

Chikhi, L. *et al.* (2018) "The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice," *Heredity*, 120(1), pp. 13–24. Available at: https://doi.org/10.1038/s41437-017-0005-6.

Crow, J.F. (James F. and Kimura, M. (1970) *An introduction to population genetics theory*. New York, Harper & Row. Available at: http://archive.org/details/introductiontopo0000crow (Accessed: March 12, 2026).

Davey, J.W. *et al.* (2011) "Genome-wide genetic marker discovery and genotyping using next-generation sequencing," *Nature Reviews Genetics*, 12(7), pp. 499–510. Available at: https://doi.org/10.1038/nrg3012.

Dinnage, R. *et al.* (2022) "slimr: An R package for integrating data and tailor-made population genomic simulations over space and time." bioRxiv, p. 2021.08.05.455258. Available at: https://doi.org/10.1101/2021.08.05.455258.

645 Do, C. *et al.* (2014) “NeEstimator v2: re-implementation of software for the estimation of
646 contemporary effective population size (Ne) from genetic data.,” *Molecular Ecology Resources*,
647 14(1), pp. 209–214. Available at: https://doi.org/10.1111/1755-0998.12157.

648 Fukuda, Y. *et al.* (2011) “Recovery of saltwater crocodiles following unregulated hunting in tidal
649 rivers of the Northern Territory, Australia,” *The Journal of Wildlife Management*, 75(6), pp. 1253–
650 1266. Available at: https://doi.org/10.1002/jwmg.191.

651 Gao, F., Keinan, A. and Nielsen, R. (2021) “epos: estimation of population size changes from site
652 frequency spectra,” *Bioinformatics*, 37(16), pp. 2386–2388. Available at:
653 https://doi.org/10.1093/bioinformatics/btab113.

654 Gargiulo, R. *et al.* (2024) “Estimation of contemporary effective population size in plant
655 populations: Limitations of genomic datasets,” *Evolutionary Applications*, 17(5), p. e13691.
656 Available at: https://doi.org/10.1111/eva.13691.

657 Gutenkunst, R.N. *et al.* (2009) “Inferring the Joint Demographic History of Multiple Populations
658 from Multidimensional SNP Frequency Data,” *PLOS Genetics*, 5(10), p. e1000695. Available at:
659 https://doi.org/10.1371/journal.pgen.1000695.

660 Haller, B.C. and Messer, P.W. (2019) “SLiM 3: Forward Genetic Simulations Beyond the Wright–
661 Fisher Model,” *Molecular Biology and Evolution*, 36(3), pp. 632–637. Available at:
662 https://doi.org/10.1093/molbev/msy228.

663 Jombart, T. (2008) “adegenet: a R package for the multivariate analysis of genetic markers,”
664 *Bioinformatics*, 24(11), pp. 1403–1405. Available at:
665 https://doi.org/10.1093/bioinformatics/btn129.

666 Kilian, A. *et al.* (2012) “Diversity arrays technology: a generic genome profiling technology on
667 open platforms,” *Methods in Molecular Biology (Clifton, N.J.)*, 888, pp. 67–89. Available at:
668 https://doi.org/10.1007/978-1-61779-870-2_5.

669 Kirkwood, R. *et al.* (2010) “Continued population recovery by Australian fur seals,” *Marine and
670 Freshwater Research*, 61(6), pp. 695–701. Available at: https://doi.org/10.1071/MF09213.

671 Li, H. and Durbin, R. (2011) “Inference of human population history from individual whole-
672 genome sequences,” *Nature*, 475(7357), pp. 493–496. Available at:
673 https://doi.org/10.1038/nature10231.

674 Liu, J. *et al.* (2022) “Undescribed species have higher extinction risk than known species,”
675 *Conservation Letters*, 15(3), p. e12876. Available at: https://doi.org/10.1111/conl.12876.

676 Liu, X. and Fu, Y.-X. (2015) “Exploring population size changes using SNP frequency spectra,”
677 *Nature Genetics*, 47(5), pp. 555–559. Available at: https://doi.org/10.1038/ng.3254.

678 Liu, X. and Fu, Y.-X. (2020) “Stairway Plot 2: demographic history inference with folded SNP
679 frequency spectra,” *Genome Biology*, 21(1), p. 280. Available at:
680 https://doi.org/10.1186/s13059-020-02196-9.

Lynch, M. *et al.* (2020) “Inference of Historical Population-Size Changes with Allele-Frequency Data,” *G3 Genes|Genomes|Genetics*, 10(1), pp. 211–223. Available at: https://doi.org/10.1534/g3.119.400854.

Malaspinas, A.-S. *et al.* (2016) “A genomic history of Aboriginal Australia,” *Nature*, 538(7624), pp. 207–214. Available at: https://doi.org/10.1038/nature18299.

Mazet, O. *et al.* (2016) “On the importance of being structured: instantaneous coalescence rates and human evolution--lessons for ancestral population size inference?,” *Heredity*, 116(4), pp. 362–371. Available at: https://doi.org/10.1038/hdy.2015.104.

Mergeay, J. *et al.* (2026) “The Importance of Effective Population Size in Conservation and Biodiversity Monitoring,” *Evolutionary Applications*, 19(1), p. e70196. Available at: https://doi.org/10.1111/eva.70196.

Nadachowska-Brzyska, K. *et al.* (2021) “Genomic inference of contemporary effective population size in a large island population of collared flycatchers (Ficedula albicollis).,” *Molecular Ecology*, 30(16), pp. 3965–3973. Available at: https://doi.org/10.1111/mec.16025.

Nunziata, S.O. and Weisrock, D.W. (2018) “Estimation of contemporary effective population size and population declines using RAD sequence data,” *Heredity*, 120(3), pp. 196–207. Available at: https://doi.org/10.1038/s41437-017-0037-y.

Orozco-terWengel, P. (2016) “The devil is in the details: the effect of population structure on demographic inference,” *Heredity*, 116(4), pp. 349–350. Available at: https://doi.org/10.1038/hdy.2016.9.

Parmesan, C. and Yohe, G. (2003) “A globally coherent fingerprint of climate change impacts across natural systems,” *Nature*, 421(6918), pp. 37–42. Available at: https://doi.org/10.1038/nature01286.

Peterson, B.K. *et al.* (2012) “Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species,” *PloS One*, 7(5), p. e37135. Available at: https://doi.org/10.1371/journal.pone.0037135.

Postaire, B.D. *et al.* (2024) “Global genetic diversity and historical demography of the Bull Shark,” *Journal of Biogeography*, 51(4), pp. 632–648. Available at: https://doi.org/10.1111/jbi.14774.

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Reid, B.N. and Pinsky, M.L. (2022) “Simulation-Based Evaluation of Methods, Data Types, and Temporal Sampling Schemes for Detecting Recent Population Declines,” *Integrative and Comparative Biology*, 62(6), pp. 1849–1863. Available at: https://doi.org/10.1093/icb/icac144.

Santiago, E. *et al.* (2020) “Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium,” *Molecular Biology and Evolution*, 37(12), pp. 3642–3653. Available at: https://doi.org/10.1093/molbev/msaa169.

717 Schiffels, S. and Durbin, R. (2014) “Inferring human population size and separation history from
718 multiple genome sequences,” *Nature Genetics*, 46(8), pp. 919–925. Available at:
719 https://doi.org/10.1038/ng.3015.

720 Terhorst, J. *et al.* (2017) “Robust and scalable inference of population history from hundreds of
721 unphased whole genomes,” *Nature Genetics*, 49(2), pp. 303–309. Available at:
722 https://doi.org/10.1038/ng.3748.

723 Terhorst, J. and Song, Y.S. (2015) “Fundamental limits on the accuracy of demographic inference
724 based on the sample frequency spectrum,” *Proceedings of the National Academy of Sciences*,
725 112(25), pp. 7677–7682. Available at: https://doi.org/10.1073/pnas.1503717112.

726 Wakeley, J. (2009) “Coalescent theory,” *Roberts & Company*, 834.

727 Waples, R.S. and Do, C. (2008) “ldne: a program for estimating effective population size from
728 data on linkage disequilibrium.,” *Molecular Ecology Resources*, 8(4), pp. 753–756. Available at:
729 https://doi.org/10.1111/j.1755-0998.2007.02061.x.

730 Wright, S. (1931) “EVOLUTION IN MENDELIAN POPULATIONS,” *Genetics*, 16(2), pp. 97–159.
731 Available at: https://doi.org/10.1093/genetics/16.2.97.

732

733

734

735

Data Accessibility and Benefit-Sharing Section

Data Accessibility Statement

Data will be provided upon acceptance.

Benefit-Sharing Statement

Benefits Generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author Contributions

B.G., I.W., R.D., and R.S. designed the research. I.W. ran the simulations with support from P.H., and I.W. performed the data analysis. I.W. led the writing of the original draft. All authors edited and revised the manuscript.

Tables

Table 1. Simulated parameter combinations across decline, expansion and stable scenarios.

	Decline	**Expansion**	**Stable**
Maximum population size N_e	1000, 500, 100, 50 (starting at max Ne)	1000, 500, 100, 50 (ending at max Ne)	1000, 500, 100, 50
Ne magnitude of change	Declines of 50%, 90%, 95%	Expansions of 50%, 90%, 95%	Not applicable
Time taken for N_e change (years)	200, 100, 50, 30	200, 100, 50, 30	Not applicable
When N_e change began (ybp)	200, 100, 50, 30	200, 100, 50, 30	Not applicable
Individuals sampled	200, 100, 50, 20	200, 100, 50, 20	200, 100, 50, 20
Total viable simulation combinations	190	390	13

753 **Table 2.** Bottleneck model timeframes, pre-crash N_e values, post-crash N_e values and
754 classifications

	Saltwater Crocodile	**Australian Fur Seal**
Bottleneck timeframe	53 years	252 years
Pre-crash N_e	7500, 15000	2100, 4200
Crashed N_e	909	872
% N_e change	88%, 94%	58%, 79%
Individuals sampled	200, 100, 50, 20	200, 100, 50, 20
Total simulation combinations	8	8

755

756

Table 3. Final counts of successful runs used for analysis, after removal of runs that inferred <225 ybp and parameter combinations with <10 replicates.

	Epos	*Stairway*
Decline	12,992	12,830
Expansion	9,582	8,197
Stable	465	456
Bottleneck	1,335	1,335
TOTAL	**24,374**	**22,818**

Figures

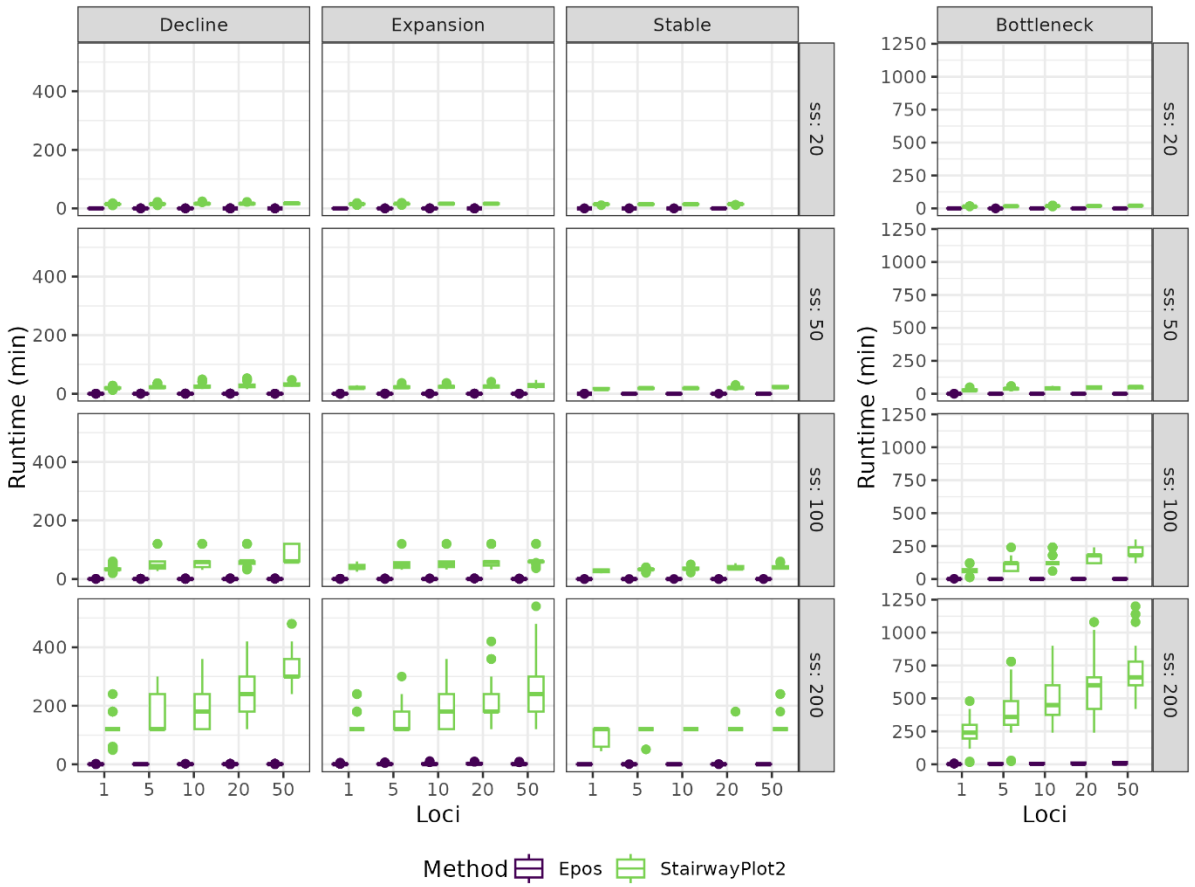


Figure 1. Runtimes (time taken for a method to complete a single inference run) for Epos and Stairway Plot 2, separated by model (columns) and sample size (rows).

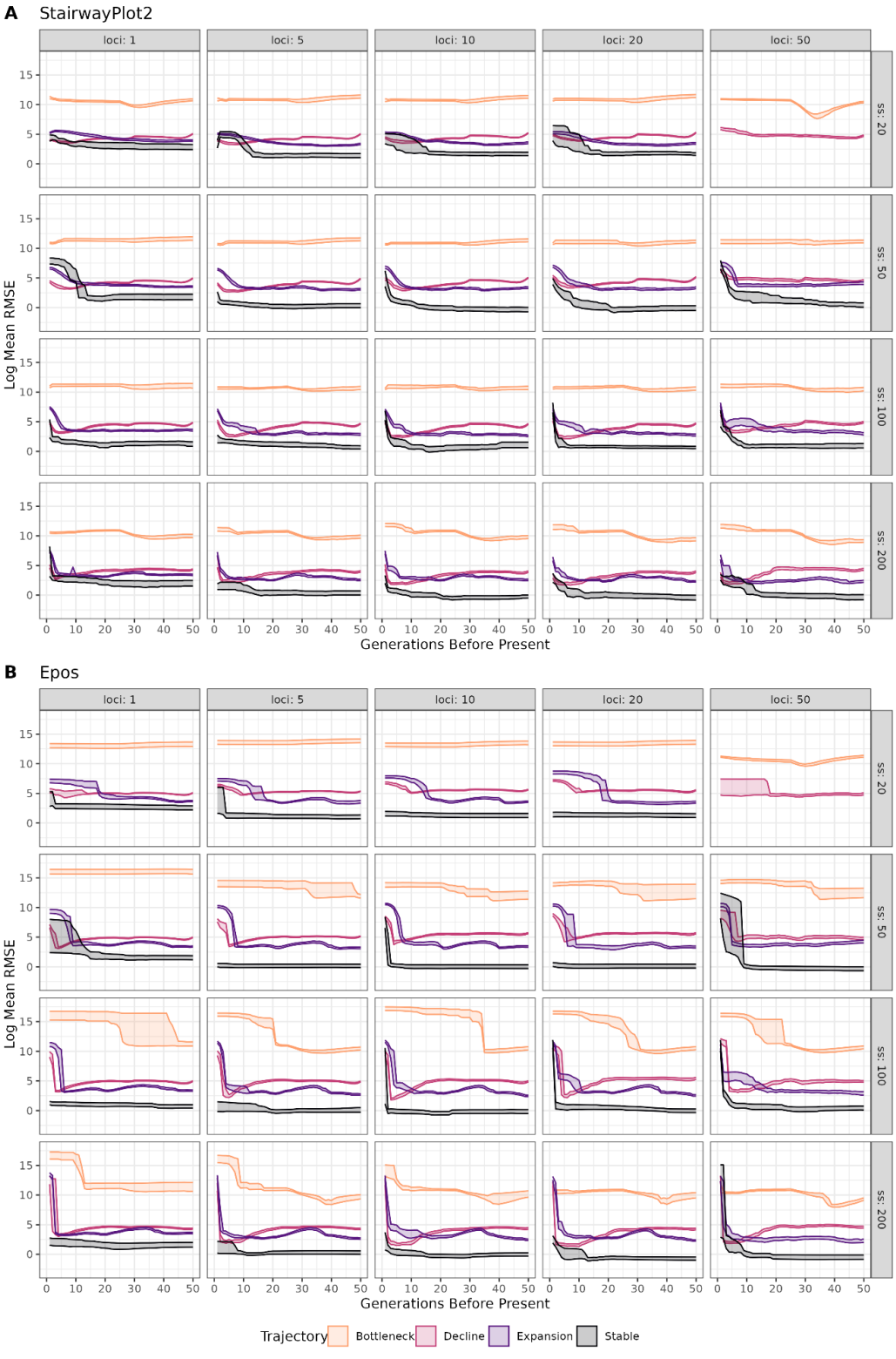
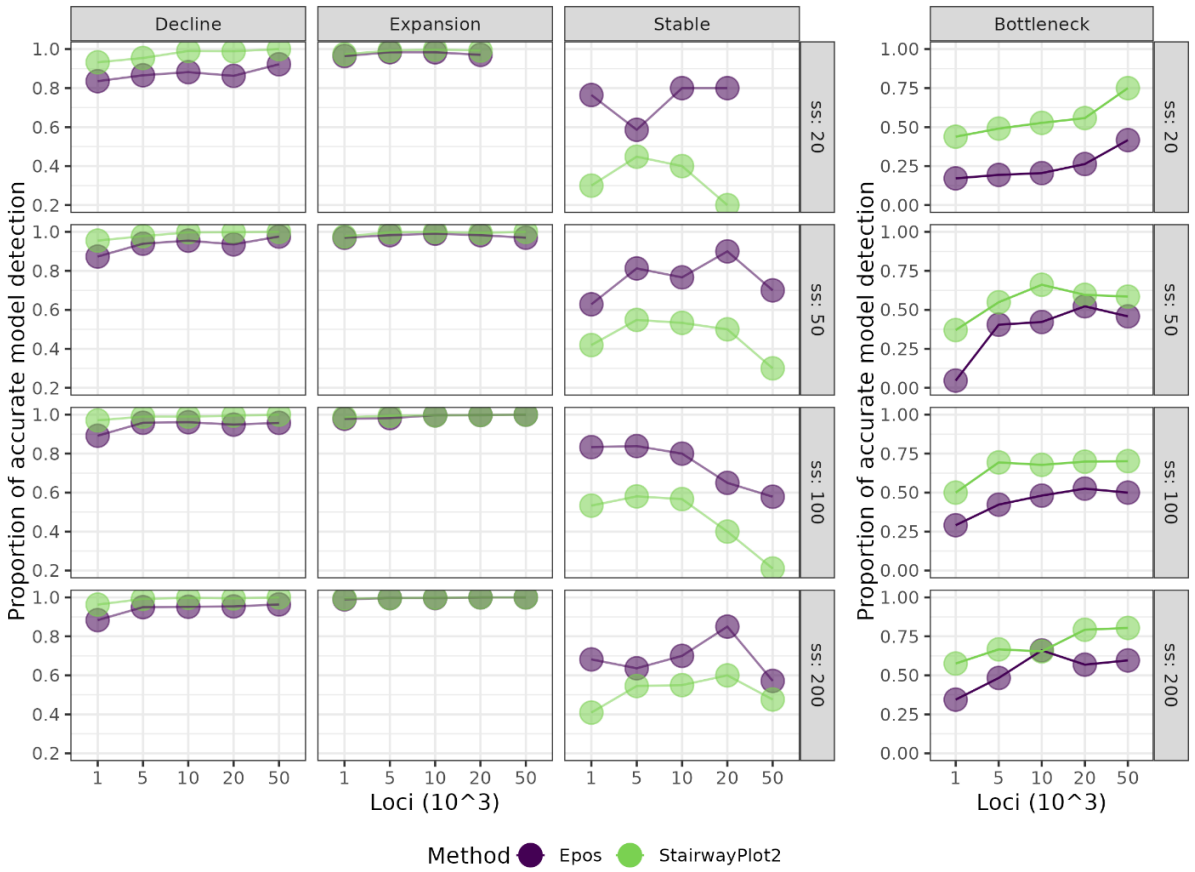


Figure 2. Log RMSE of inference outputs relative to their target simulation values for each generation for 0-50ybp, separated by SNP loci (columns) and sample size (rows). Panel A shows outputs for Stairway Plot 2 and panel B shows outputs for Epos.

773



774

775 **Figure 3.** Proportion of outputs where inference methods accurately reconstructed the
776 simulated trajectory, displayed in separate panels by population model (columns) and
777 sample size (rows)

778

779

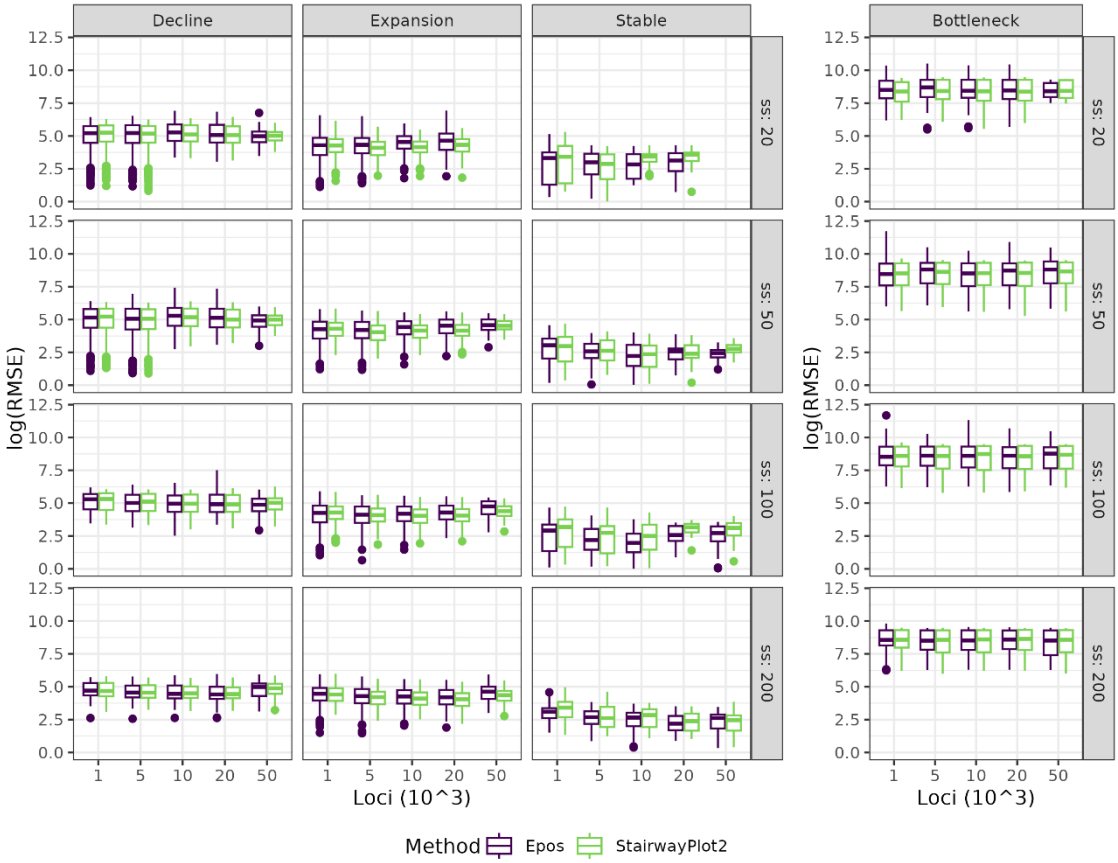
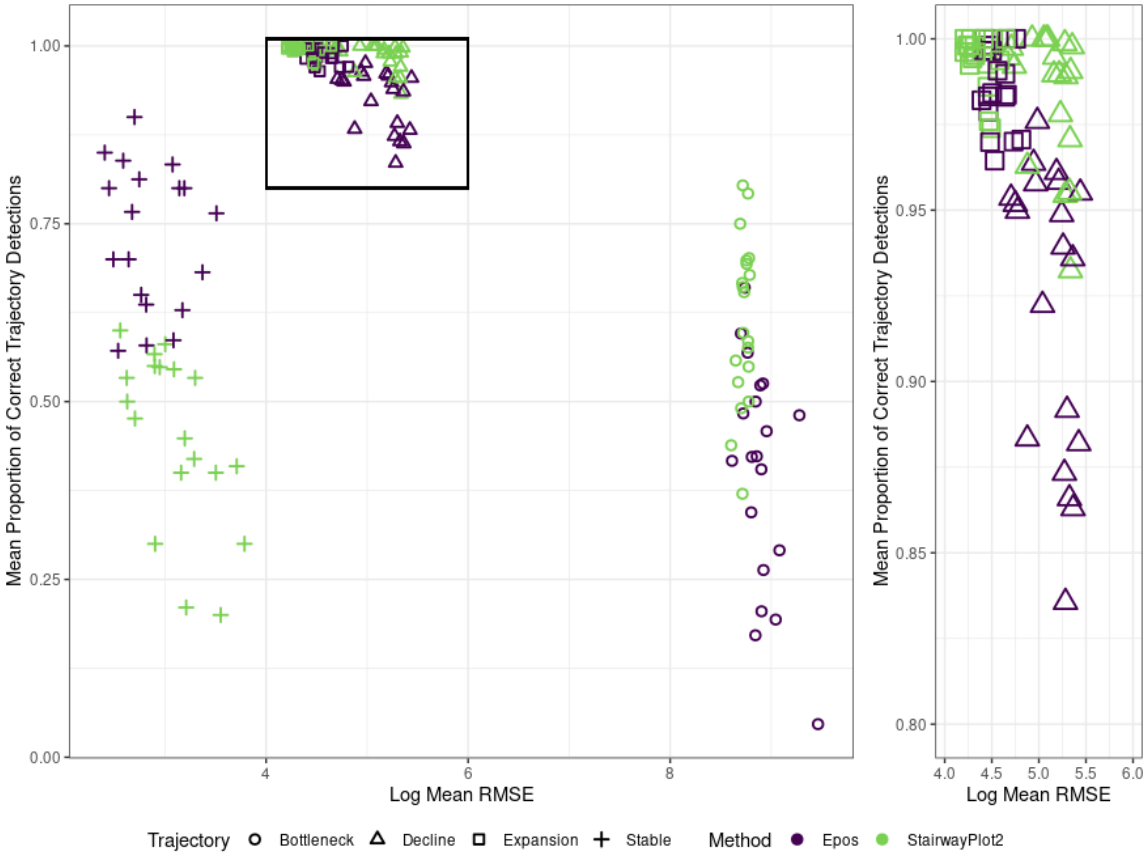


Figure 4. Log RMSE of N_e values of inference runs against their respective target simulations, for all inference methods, displayed in separate panels by population model (columns) and sample size (rows)



786

787 **Figure 5.** Log RMSE against proportion of correct model detection for decline, expansion
788 and stable trajectories and across Epos and Stairway Plot 2 outputs. Left: Full dataset,
789 right: zoom in of cluster in black square.

790

Supporting information

Title: Evaluation of site frequency spectrum-based demographic inference methods for use in conservation contexts

Running title: SFS-based demographic methods in conservation

Authors: Isobel Walcott[1], Robyn E Shaw [1], Richard Duncan[1], Peta Hill[1], Bernd Gruber[1]

Institutional affiliations:

1. Centre for Conservation Ecology and Genomics, University of Canberra, Bruce ACT

Contact information:

Corresponding author: Isobel Walcott, Isobel.Walcott@canberra.edu.au

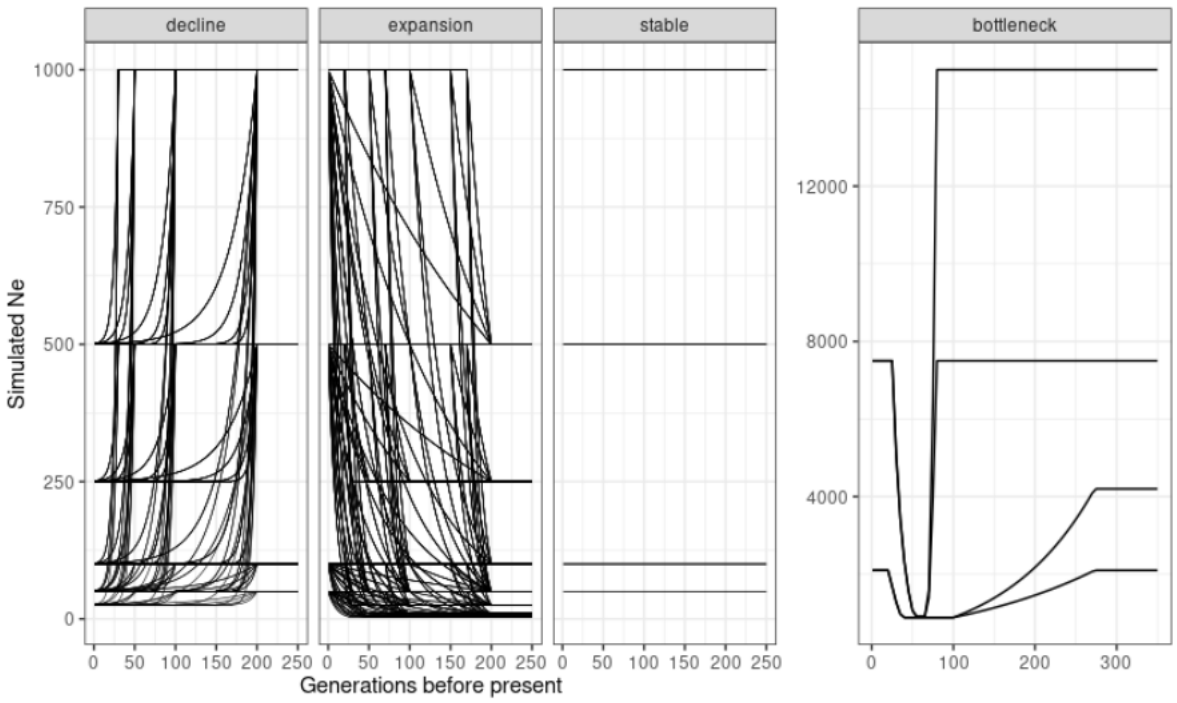
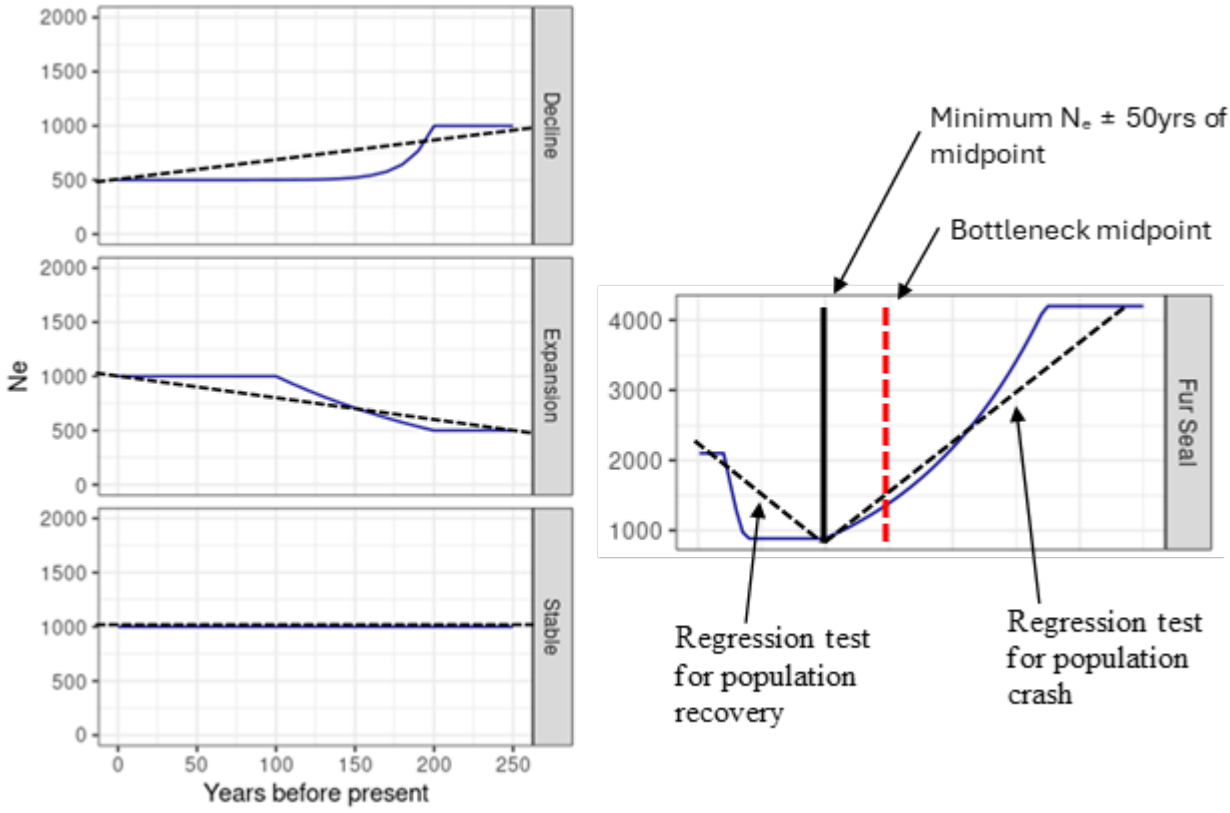


Figure S1. Scenarios simulated with varying initial N_e, and magnitude, rate, and timing of Ne change for decline, expansion stable, and bottleneck scenarios. For each individual scenario (line), separate simulations were run for sample size levels. 20 replicate simulations were run for each scenario and sample size combination.



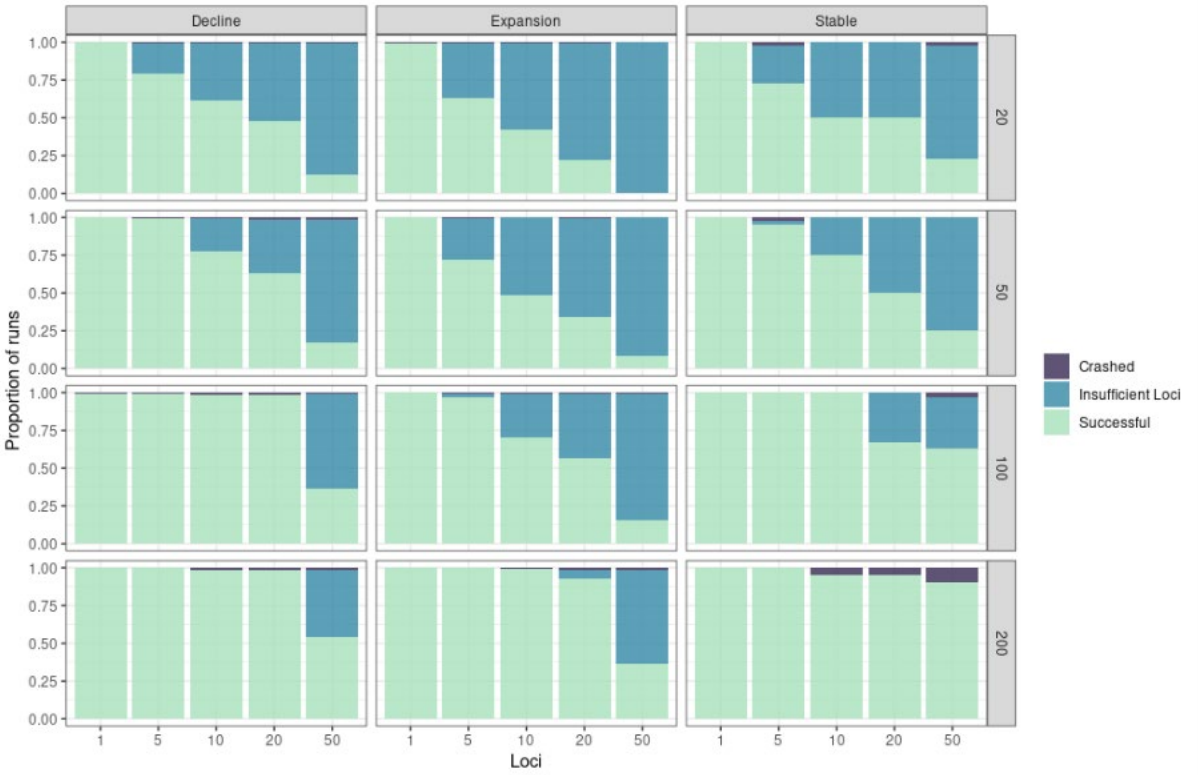
816

817 **Figure S2.** The Trajectory classification process for identifying decline, expansion, stable (left)
818 and bottleneck (right) trajectories. For decline, expansion, and stable trajectories, linear
819 regressions (black dotted line) were fitted across the full inference range to detect directional
820 change in the inferred demographic curve (solid blue line). For bottleneck trajectories, we first
821 identified the midpoint between the start of the crash and the end of the recovery (red dotted
822 line). We then located the minimum Ne value occurring within 50yrs of this midpoint and used it
823 to divide the trajectory into two segments: the expected crash and recovery phases. We fitted
824 separate linear regressions to each segment to test for a significant decline followed by
825 expansion.

826

827

828



829

830 **Figure S3.** Proportion of run successes, simulation failures and inference failures.

831

832

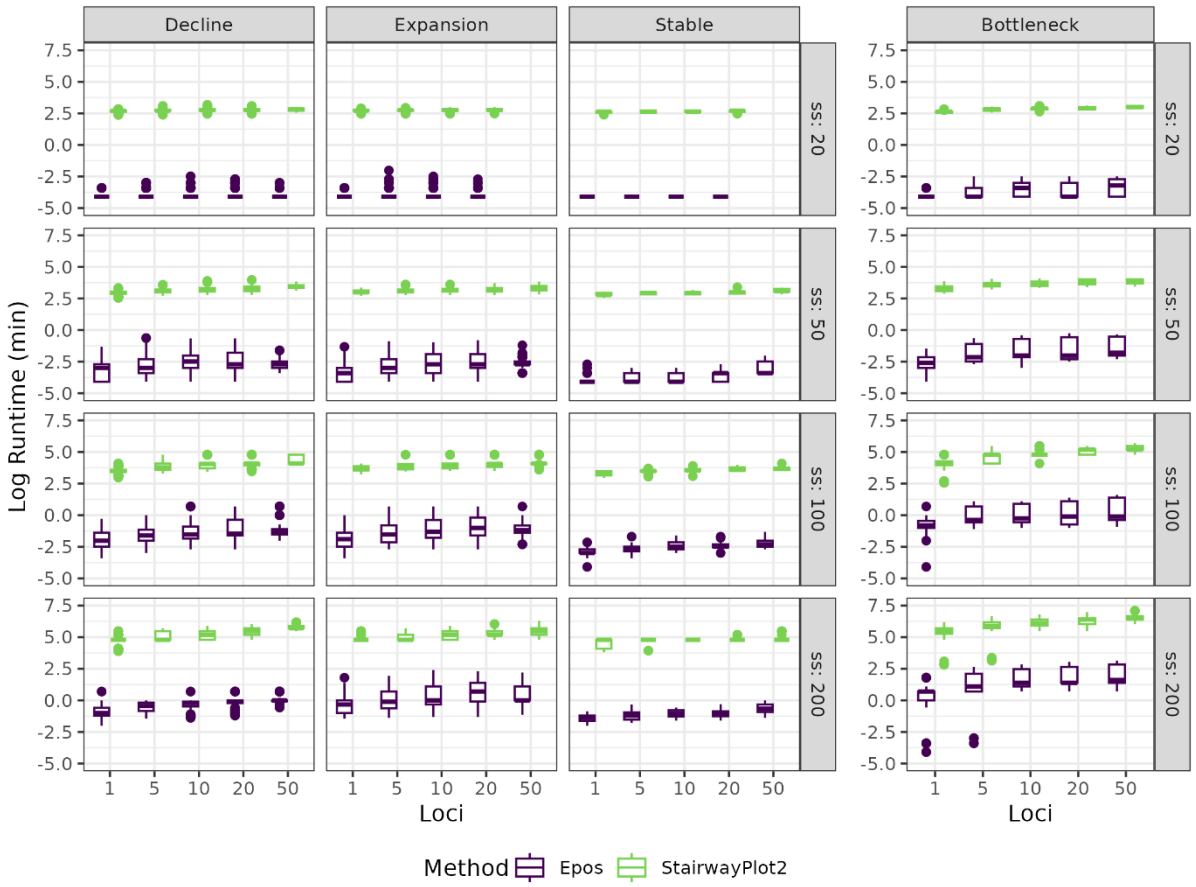


Figure S4. Runtimes (time taken for a method to complete a single inference run, in log minutes) for Epos and Stairway Plot 2, separated by model (columns) and sample size (rows).