BEHAVIOR OF POPULAR INDICES OF GENETIC DIVERSITY IN SIMULATED EXPANDING POPULATIONS

A Thesis

by

ADAM M. BYNUM

BS, University of Dayton, 2014

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MARINE BIOLOGY

Texas A&M University-Corpus Christi
Corpus Christi, Texas

May 2022

May 2022

BEHAVIOR OF POPULAR INDICES OF GENETIC DIVERSITY IN SIMULATED EXPANDING POPULATIONS

A Thesis

by

ADAM M. BYNUM

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

Christopher E. Bird, Ph.D.
Chair

Blair Sterba-Boatwright, Ph.D.
Committee Member

David Portnoy, Ph.D.
Committee Member

May 2022

ABSTRACT

Protecting genetic diversity is an integral component of food security, fishery management, and biodiversity conservation, and thus the ability to model and predict the distribution of genetic diversity is valuable. Population genetic theory predicts that genetic diversity will be greatest in the largest populations at mutation-drift equilibrium, implying that efforts to preserve diversity would be best focused on keeping populations as large as feasibly possible. Natural populations, however, are rarely in equilibrium, because their sizes can fluctuate due to a variety of processes, e.g., populations that have had a recent bottleneck or invaded a new habitat. To predict patterns of genetic diversity in natural populations, it has become increasingly important to understand how populations behave in non-equilibrium scenarios. Here we report the effects of mutation rate (μ), initial population size (N_{e0}), and final population size (N_{e1}) on the genetic diversity in expanding populations using a Wright-Fisher forward time model built with SLiM2. Using a 300 bp sequence to simulate modern genome-wide surveys of genetic variation (RAD), a range of naturally occurring mutation rates, and population sizes, multiple models were created to cover a broad portion of parameter space, and six commonly reported measures of genetic diversity estimated. As previously reported, genetic diversity increased with increasing population size given a similar set of circumstances, but there are broad swaths of parameter space where small populations exhibit greater diversity than large populations, making historical context critical in population genetics analysis. Depending on population size and mutation rate, the different diversity indices (nucleotide diversity, gene diversity, number of haplotypes, effective number of haplotypes, number of heterozygotes, and number of substitutions) progressed towards equilibrium at different rates. Furthermore, different diversity indices had different levels of

sensitivity to changes in diversity at different times. To better describe the change in diversity with time, logistic growth models were used to estimate the equilibrium diversity (D_{eq}), initial diversity value (D_0), amount of time required to reach halfway to genetic equilibrium (t_{50eq}), growth parameter (Φ_3), maximum rate of genetic diversity increase (r), and time required to reach 95% of the equilibrium value (t_{95}) in populations that expand from N_{e0} to N_{e1}. We employed both linear and non-linear model fitting and used AIC to identify the best models describing the logistic growth parameters with varying N_{e0}, N_{e1}, and μ. In most cases, the models fit the simulated data well as the relative bias is low, ranging from +/- 3%, but the models did not perform as well when N_{e0}, N_{e1}, and μ, are small, with relative bias as high as 20%. The best models were used to create a tool that estimates the diversity of a population given the time since the onset of expansion, N_{e0}, N_{e1}, and μ. The prediction model performed best when using the N_{e0}, N_{e1}, and μ used in the simulations but could give misleading diversities when interpolating, so a switch was created to restrict the tool to only accept the predefined set of parameter values. This tool can be used to get a rough approximation of how long it will take for genetic diversity to accumulate and determine why there might be deviations from the neutral expectation that large populations have more diversity without running time consuming simulations and subsequent analysis.

DEDICATION

I would like to express my thanks to my advisor, Dr. Chris Bird, for his assistance and patience as we worked through my research. His ability to become excited about results has kept me interested and engaged with my research. I am thrilled with the outcome of all the hard work we put.

My appreciation also extends to the other members of my committee, Dr. Blair Sterba-Boatwright and Dr. David Portnoy. I was able to learn a great deal from both, both in and out of classes and apply what I learned directly into this thesis, from applying ideas to statistical coding.

I would like to extend a big thanks to all the members of the HoBi lab and the help they gave me throughout the years.

The staff making sure the TAMUCC-HPCC is up and running is also deserving of my gratitude. I am thankful that they were quick to answer questions, fix, and update the HPC, which made this project possible. I would like to apologize for filling up the home directory when I first started.

Above all, I would like to thank all my family and friends who supported me throughout this thesis. I would specifically like to thank Cindy Cattey for being supportive of me, especially when I would write code on napkins during dinner.

ACKNOWLEDGEMENTS

The data was processed on the TAMU-CC high performance computing cluster (HPCC) funded by NSF-MRI-CNS-0821475.

Publication supported in part by an Institutional Grant (NA14OAR4170102) to the Texas Sea Grant College Program from the National Sea Grant Office, National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Natural populations fluctuate in size due to a variety of processes which lead to a state of evolutionary disequilibrium. Long-term processes such a glaciation cycles and shorter-term processes (including human activities, Halpern et al., 2008) lead to changes in ecological interactions, range expansions, and gene flow alterations which all affect population sizes. Population size can be categorized by census size (N) and effective size (N_e). Census population size is the number of individuals in a population while effective population size is the number of individuals required for a population to have the same genetic diversity as an ideal population. Census and effective population sizes are not necessarily equal, and differences are affected by such factors as nonrandom mating, harems, fluctuating population sizes.

Many populations, including both plants and animals, are decreasing due to anthropogenic forces including overexploitation (Butchart et al., 2010; Hutchings, 2000), habitat destruction (Rothschild et al., 1994; Flockhart, 2015), climate change (Møller et al., 2008), and the detrimental effects of invasive species (Fritts & Rodda, 1998; Clavero & García-Berthou, 2005). On the other hand, populations of invasive species and others released from competition and/or predation pressure may be expanding (Benning et al., 2002; Hulme, 2009). These changes in population size can lead to rapid evolutionary changes (Charlesworth, 2009), where evolution is defined as the change in allele frequencies through time (Dobzhansky, 1937). As effective population size declines, genetic drift (Fisher 1922, Wright 1931) becomes more rapid and reduces genetic variation (Ellstrand & Elam, 1993; Willoughby et al., 2015). In expanding populations, mutations accumulate, and drive increases in genetic diversity (Frankham et al., 2014). While many analytical procedures used by researchers assume that natural populations are

in mutation-drift equilibrium, the past 40 years of population genetic research with DNA sequencing has revealed that most natural populations are in a state of disequilibrium (Gibbons et al., 2000; Allentoft & O'Brien 2010; Potts et al., 2010; Magurran, 2013) and it is important to understand how diversity changes in these populations through time.

Within-group genetic diversity metrics are intended to quantify the amount of standing genetic variation within a sample. There are several measures of genetic diversity, and each quantifies a different aspect of diversity (Jost, 2008). In its simplest form, genetic diversity can be represented as the number of variants also known as zero-order measures. Common measures include number of segregating or polymorphic sites (*S*), haplotypes (*k*), and heterozygotes (*H*; Hartl & Clark, 2007). Second order measures of diversity represent both numbers of variants and how common they are, disproportionately weighting the most common variants (Jost, 2008). These include gene diversity, ($H_{e,}$ Nei,1973) and the observed heterozygosity. Related metrics of diversity include the mean number of pair-wise mismatches (Π; Hartl & Clark, 2007), where a mismatch is a segregating site between two DNA sequences, and nucleotide diversity (Nei & Li, 1979; Excoffier & Lischer, 2010). These descriptive statistics of genetic diversity provide valuable insights into population variation which are related to evolutionary potential.

Standing genetic diversity is generally related to the ability of a species to adapt on ecological time scales to the accelerating changes in selective landscapes (Bell & Collins, 2008; Pauls et al., 2013). Elimination of alleles can be detrimental to population persistence because a rare allele conferring no fitness advantage may become advantageous in a new selective landscape (Barrett & Schluter, 2008). Consequently, preventing the loss of genetic diversity associated with small population should be a prominent goal of natural resource management (McNeely et al., 1990; Reed & Frankham, 2003). Ultimately, smaller populations support less

genetic diversity, which decreases adaptive capacity and persistence in a changing environment (Soulé, 1976; Frankham, 1996; Pauls et al., 2013; Romiguier et al., 2014) risk spiraling down the extinction vortex as detrimental factors, such as increased inbreeding and buildup of deleterious alleles, compound (Gilpin & Soulé, 1986).

Based upon the relative rates at which diversity can be lost due to genetic drift (Maruyama & Fuerst, 1985; Allendorf, 1986) and accumulated due to mutation (Nei et al., 1975; Maruyama & Fuerst, 1984), it is expected that the accumulation of diversity through mutation alone will take a longer time than the loss of diversity via drift. Mutation rate and the loss of mutations via genetic drift are largely responsible for the rate at which genetic diversity is accrued following a population expansion (Nei et al. 1975, Maruyama & Fuerst 1984). As populations become larger, more mutations are generated, and genetic drift becomes weaker leading to fewermutations becoming extinct. Thus, larger populations at equilibrium are expected to have greater genetic diversity than small populations:$H_0 = \frac{4N\mu}{4N\mu+1}$ (Nei et al. 1975) where H_0 is the original heterozygosity. Because background mutation rates tend to be low ($\mu=10^{-10} – 10^{-7}$) it may take thousands of generations to reintroduce lost genetic diversity (Lacy, 1987), and mutation rates are generally deemed to be too slow to produce adaptively advantageous variants on short ecological time scales (Wright, 1932; Lacy, 1987). Mutation rates, however, do vary among loci (Chakraborty et al., 1997), causing diversity to accrue faster in some loci than in others.

Interestingly, there is variability in the sensitivity of different measures of genetic diversity to expansions in population size. Using the infinite alleles model, Nei et al. (1975) found that the number of alleles in a population increased more rapidly than the gene diversity. Nei & Li (1976) have also shown that the expected number of alleles increases more quickly

than gene diversity (Nei & Li, 1976), and Maruyama and Fuerst (1984) came to a similar conclusion using a different numerical method. Following these studies, Tajima (1989) used the infinite sites model to demonstrate that *S* increases more rapidly in an expanding population than Π and concluded that the difference in the responsiveness was analogous to that previously found for the number of alleles and gene diversity. Differences among the rates of accumulation can be attributed to how new mutations are low in frequency when they first enter a population. Measures of genetic diversity that are insensitive to rare alleles, such as nucleotide diversity, increase at a slower rate than those which are sensitive to rare alleles such as number of heterozygotes, assuming a constant μ.

With μ remaining constant, the time required for a population to reach genetic equilibrium is directly dependent on the change in population size. Additionally, smaller populations reach genetic equilibrium faster (Maruyama & Fuerst, 1984). There are few studies that we know of, however, that compare the full suite of genetic diversity indices, both site and allele-based, within the same context (Aris-Brosou & Excoffier, 1996).

We built upon these previous studies by directly comparing the responses of six indices of genetic diversity in expanding populations using forward-time population genetic simulations of next-generation restriction-site-associated DNA sequences (RADseq) for ease of comparisons between simulated results and data collected from natural populations. We tested for differences in the rate of accumulation in different measures of genetic diversity, time until genetic equilibrium is reached, and differences in the final equilibrium value. A predictive model was made to describe the accumulation of genetic diversity and its performance was evaluated.

CHAPTER II

METHODS

Demographic Population Expansion Model

A logistic population growth model was chosen to accurately portray natural populations and the carrying capacity (Hastings, 2013):

$$\frac{dN}{dT} = r_{max}\frac{(K-N)}{K}N \quad (1)$$

where r_{max} is the growth rate, K is carrying capacity, and N is the number of individuals in the population (Equation 1). To model rapidly increasing populations, a growth rate of 0.3 was used as a reasonable rate of expansion seen in natural populations of both invertebrate and vertebrate populations,including *Tisbe battagliai* and *Macropus spp*. respectively. (Sibly & Hone, 2002).

Populations were created with different initial (N_{e0}) and final population sizes (N_{e1}). We assumed that all individuals had equal reproductive rates, and thus, the population sizes are both census (N) and effective (N_e). Simulated populations were diploid, reproduced in discrete generations, and evolved according to the classic Wright-Fisher model (Fisher, 1923; Wright, 1931, Hartl & Clark, 2007). Similar to Nei et al. (1975), the simulations began with either N_{e0} = 2 or N_{e0} =100 to simulate a population starting with either minimal diversity or a population that has a low level of diversity, respectively. Following Eq 1, populations expanded to one of five final population sizes (N_{e1} = 100, 500, 1000, 5000, or 10000). The values of N_{e1} were chosen because they are realistic effective population sizes of species (Hasuer & Carvallho 2008The combination of N_{e0} and N_{e1} that did not result in an expansion (i.e., N_{e0} = 100 and N_{e1} = 100) was omitted.

Population Genetic Model

The modeled DNA within each population included 192-102,400 loci which were unlinked and conformed to the infinite-sites mutation model (Kimura, 1969). Each allele, in each locus, in each individual began with 300 adenosines, and one of five mutation rates ($\mu = 1x10^{-4}$, $1x10^{-5}$, $1x10^{-6}$, $1x10^{-7}$, or $1x10^{-8}$ per base pair per generation) was assigned to each locus, spanning the range of that commonly observed in sexually reproducing eukaryotes (Lynch, 2010). All treatments, demographic and genetic, were fully factorially crossed (see Simulations below). Haplotype sequences were 300 bp to simulate the most common type of data generated double digest restriction site associated DNA sequencing (ddRAD, Peterson et al., 2012), which is popular in both model and non-model species. For the sake of simplicity, the only possible allelic states were either "A" or "T", and it was possible for each to mutate to the other. We did not differentiate between alleles with the same state that originated from different mutational events, thus allowing for homoplasy, as would be the case with empirical data.

Simulations

To simulate expanding populations, a forward-time Wright-Fisher population genetic simulator, Selection on Linked Mutations (SLiM 2, Haller & Messer, 2016), was employed. We simulated populations that began with either no variation (N_{e0}=2) or a small amount of variation (N_{e0}=100). Populations starting with 100 individuals were allowed to reach mutation-drift equilibrium prior to expansion by allowing 1,000 generations of no growth (r=0). Because our simulations were computationally intensive, we ran simulations on the TAMUCC high performance computing cluster, but even then, there were limitations on the number of generations and sampling times. During population growth, each population was sampled 124 times, from 0 to 10^5 - 10^6 generations after the beginning of expansion. These sampling times

were chosen to cover the time between the onset of population growth and mutation-drift equilibrium while balancing the constraints on the amount of time required to run the simulations. Samples consisted of 100 diploid individuals, each with unlinked and selectively neutral 300 bp loci. Each treatment combination was replicated 192 to 102,400 times (see loci in Table S1) to obtain better estimates of population genetic parameters which are subject to the variation driven by genetic drift. The replicates are effectively 300 bp loci. The SLiM2 scripts and scripts for all subsequent data processing and analysis described from here forward can be found at https://github.com/abynum91/BynumThesisScripts.

Data Processing

A custom bash (Free Software Foundation, 2007) script which employed GNU Parallel (Tange, 2011) was used to (1) convert all variant call format (VCF; Danecek et. al., 2011) output from SLiM2 to Arlequin format (Excoffier & Lischer, 2010), (2) run Arlequin, and (3) convert results from Arlequin's xml output files to tidy tab-delimited files (Wickham et al., 2019). Arlequin was used to calculate nucleotide diversity (Nei & Li, 1979), number of polymorphic sites, number of substitutions, number of heterozygotes, number of homozygotes, gene diversity, and number of haplotypes. We additionally used the Arlequin output to the effective number of haplotypes (Jost, 2008).

Data Analysis

Results from Arlequin were analyzed in R (R Core Team, 2020) and plots were created using the package ggplot2 (Wickham, 2016).

Logistic growth models were used to estimate the rate of increase in genetic diversity, the amount of time to reach equilibrium, and the amount of genetic diversity at equilibrium. The logistic regression model is:

$$D = \frac{D_{\text{eq}} - D_0}{1 + e^{\frac{t_{50\text{eq}} - t}{\Phi_3}}} + D_0 \quad (2)$$

where D_0 is the genetic diversity (D) at the start of the population expansion, D_{eq} is D at equilibrium (Fig 1), $t_{50\text{eq}}$ is the time required for $D = 0.5\left(D_{eq} - D_0\right) = 0.5D_\Delta$, Φ_3 is the difference between the time required for $D = \frac{D_\Delta}{1+e^{-1}} \approx 0.73D_\Delta$ and $t_{50\text{eq}}$, and t is number of generations (Fig 1A; Pinheiro et al., 2016). Logistic regression models were fit to each measure of genetic diversity using a nonlinear (weighted) least-squares (nls) model with the self-startup SSlogis in R (R Core Team, 2013). For the models with $N_{e0} = 100$, D_0 was estimated as the lowest observed genetic diversity for a given treatment and subtracted from all observations of genetic diversity within the same treatment prior to fitting the model, which assumed $D_0 = 0$. The times to reach x% of equilibrium ($t_{x\text{eq}}$), where x=5|95, were calculated by solving Eq 2 for t given $y = xD_{eq}$ (Fig 1). The maximum observed rate of increase in genetic diversity (r) per generation was defined as the slope of the best fit logistic model at $t = t_{50\text{eq}}$, and can be derived by taking the derivative of Eq 2 and setting $t = t_{50\text{eq}}$ and $D_0 = 0$:

$$\frac{dD}{dt} = \left(\frac{D_{\text{eq}}}{1 + e^{\frac{t_{50\text{eq}} - t}{\Phi_3}}}\right)^{-2} \left(e^{\frac{t_{50\text{eq}} - t}{\Phi_3}}\right)\left(\frac{1}{\Phi_3}\right),\ r = \frac{\left(1+e^0\right)^{-2}\left(e^0\right)}{\Phi_3} = \frac{D_{eq}}{4\Phi_3} \quad (3)$$

To assess the fits of the logistic models to the data, we created scatterplots (Fig S2) of residuals versus fits to identify deviations from model assumptions of randomly distributed residuals and calculated residual standard error (*RSE*), achieved convergence tolerance (*ACT*), and the standard error (*SE*) for each parameter estimate. The *RSE* is an estimate of the standard deviation of the distribution of residuals, which are the difference between the model-predicted and simulated mean values of genetic diversity. Thus, the smaller the *RSE* are relative to simulated values of genetic diversity, the better the logistic model fits the data.

We also estimated of the logistic model parameters D_{eq}, t_{50eq}, and Φ_3 directly from the simulation data to test for bias in the estimates from the nls model. For the estimation of D_{eq}, we identified the mean diversity values that were at (or very near) equilibrium using an algorithm that started by considering all data points for a particular simulation treatment combination after $log_{10}t_{85eq}$ (estimated from logistic model) and comparing their skewness (R package e1071) to a sample from $log_{10}t_{85eq}$ + 0.1 after removing outliers (Q1-1.5IQR and Q3+1.5IQR) from both. Skewness is expected to trend towards zero as equilibrium is approached. If the skewness of the second sample was closer to zero, then $log_{10}\ t_{eq}$ was iteratively increased by 0.1 until the sample with lowest skewness value was identified, and the mean and standard deviation of the D_{eq} estimates were recorded. t_{50eq} was estimated as the time at which the mean diversity value across loci was $0.5D_{eq}$. This was accomplished by sampling the data points near t_{50eq} $\left(D = 0.5D_{\mathrm{eq}} \pm 0.1D_{\mathrm{eq}}\right)$, fitting a linear model to diversity versus $log_{10}t$ using the lm and summary R commands to generate the equation for the best fit line, and plugging in $0.5D_{eq}$ and solving for t. A similar procedure was followed for estimating Φ_3 except the data sampled were $D = 0.73D_{\mathrm{eq}} \pm 0.1D_{\mathrm{eq}}$ to obtain t_{73eq} from which t_{50eq} was subtracted (refer to the description of Eq. 2).

To directly test for differences among different genetic diversity measures, genetic diversity values were standardized relative to their equilibrium values, D_{eq}, using the following equation:

$$D_{\mathrm{s_{ij}}} = \frac{D_{\mathrm{ij}}}{D_{\mathrm{eq.i}}} \tag{4}$$

where D_s (range: 0-1), is the standardized diversity value for observation j from diversity index i, D_{ij} is any single genetic diversity value, and $D_{\mathrm{eq.i}}$ is the equilibrium value for diversity index i given a particular combination of N_{e0}, N_{e1}, and μ. To test the global model of the effects of Ne_0, N_{e1}, and μ, on r_{std} (the standardized rate of increase and essentially the same as Φ_3 because D_{eq} =

1, see Eq 3), t_{50eq}, t_{05eq}, and t_{95eq}, we employed the following linear model using the lm function in R:

$$y \sim N_{e0} * N_{e1} * \mu \quad (5)$$

where all individual sources of variation and interactions were included in the model. Because each diversity index was independently modeled as described above (LM.1-4, Table 1) and the simplest model explained most of the variation, we decided that it was not important to investigate more complex models that included D_s , μ and N_{e1} were transformed, as necessary, to satisfy the assumptions of least squares multiple regression.

Predictive Model

Systematic variation in the logistic model parameters (D_{eq}, t_{50eq}, Φ_3) would enable straight-forward interpolation of any combination of simulation parameters within the range investigated. We endeavored to generate a predictive model (PhiPhinder.r) that accepts the inputs of N_{e0}, N_{e1}, and μ, and returns $D_{0.i}$, $D_{eq.i}$, $t_{50eq.i}$, and Φ_{3i}, thereby enabling the parameterization of a logistic model describing the change in diversity over time. To test the relationships of the response variables associated with the accumulation of genetic diversity in an expanding population (D_{eq}, t_{50eq}, Φ_3) with the initial population size (N_{e0}), final population size (N_{e1}), and mutation rates (μ), we employed multiple linear regression models with and without quadratic terms using the lm R command (Table 1). Variables were transformed, as necessary, to satisfy the assumptions of least squares multiple regression. The contribution of the predictors to the linear models were evaluated using *t*-statistics and *p* values generated by the summary R function. The adjusted R^2 parameter from the summary R command was used to quantify the amount of variation explained by the model.

To better estimate D_{eq}, which most strongly affects the inferred diversity trajectories, we additionally compared the performance of the linear models to nonlinear models fit using R packages drc (Ritz et al., 2015) and aomisc (Onofri & Garcia 2021, see Table 1). Explicitly, two levels of models were used. First the relationships between D_{eq} and μ were modeled independently for each combination of N_{e0} and N_{e1}. Second, the relationships between nonlinear parameter estimates themselves and N_{e1} were modeled for each N_{e0} (see Table 1). The performances of the models were either described using the AIC and BIC R commands or by directly calculating AIC and BIC so that the linear and nested nonlinear models could be compared. AIC was favored over BIC when they disagreed because we were more interested in obtaining a good fit than minimizing the number of parameters required to obtain the fit. The best models were used to parameterize the PhiPhinder.r script described earlier in this paragraph.

CHAPTER III

RESULTS

Performance of Logistic Models

The logistic model was generally effective at describing the increase in diversity indices with time in an expanding population (Figs 2; S1; Table S1). In all 45 sets of simulation parameters (μ, N_{e0}, N_{e1}) for all diversity indices, the nls function converged (ACT < 9.8E-6), and the *RSE* were small (0.01 – 7.0% of adjusted equilibrium diversity value, D_{Δ}; Table S1), indicating that an adequate number of loci were sampled to control variation. As evidenced by lower *RSE* values, the best model fits were associated with greater values of N_{e0}, N_{e1}, and μ (Table S1). The lowest values of N_{e0} and N_{e1} resulted in the greatest *RSE* (relative to equilibrium diversity) for $\mu = 10^{-8}$ due to the high variability associated with strong genetic drift and computational constraints on the number of loci we could simulate in a reasonable amount of time (up to 102,400, 300 bp loci).

In visual inspections of the scatterplots of residuals from the logistic models versus time (Fig S2.1-12), there was a small amount of systematic bias which oscillated through time for all diversity indices. For nucleotide diversity (Figs S2.1, S2.7), gene diversity (Figs S2.2, 2.8), and effective number of haplotypes (Figs S2.5, S2.11), the patterns reflect that simulated diversity increased faster than the logistic models immediately after the population expansion (all expansions completed within 62 generations), slower than the logistic models midway to equilibrium, and faster than the logistic models as equilibrium was approached. The number of substitutions (Figs S2.3, S2.9), heterozygotes (Figs S2.4, .10), and haplotypes (Figs S2.6, .12), exhibited the opposite pattern, and thus estimates of parameters such as t_{05eq} and t_{95eq} are expected to be slightly biased below the true values.

The longer the populations were simulated past the time at which equilibrium was approached ($t_{95eq,}$ see Figs S1.1-12, S2.1-12, S3.1-3), the lower the relative bias and the better the estimate of D_{eq}, with a total simulation time of ~ 2 + $log_{10}\, t_{95eq}$ resulting in minimized relative bias. In simulations run for fewer generations after t_{95eq}, the model estimates of D_{eq} are often biased above the true values (up to 1.06x, Fig S3.1.1), which were independently estimated from the simulation results (Table S1). However, in most cases, the upwards relative bias was much lower (mean = 1.02x). Logistic model estimates of t_{50eq} and Φ_3 were biased above or below the true values by a factor of 0.99 -1.03 and 0.76 -1.24x, respectively, depending upon diversity index, mutation rate, and final population size (Figs S3.1.2-3). Another source of bias was the amount of variation in the mean diversity indices, a function of mutation rate and number of loci (Table S1). At low mutation rates, it became computationally unfeasible to simulate proportionately more loci as the mutation rate decreased, which contributes to the greater degree of relative bias observed at lower mutation rates, especially 10^{-8}.

Modeling Logistic Parameters

Most of the variation in the equilibrium diversity value (Φ_1, Fig 3), and the time for diversity to reach 50% of equilibrium (t_{50eq}, Fig 5), given N_{e0}, N_{e1}, and μ, were well described with the best-fit multiple linear regression models (R^2 = 0.9673 – 0.9989; Tables 2, S2). The best models, as determined by BIC, all included μ^2 (LM.2, .4), and those for nucleotide diversity and number of substitutions also included N_{e1}^2 (LM.4). For AIC, which penalizes additional model parameters less severely than BIC, the more complex model (LM.4) was selected for the cases where AIC and BIC were not in agreement. For all best models, there was an interaction of N_{e1} or N_{e1}^2 with μ or μ^2, but no effect of N_{e0} (see Tables S3.2, S3.4).

There were usually positive relationships of D_{eq} with N_{e1} and μ, but there were exceptions where N_{e1} had little effect on D_{eq} when the effective number of haplotypes were near their minimum ($\mu = 10^{-8}$) or the gene diversity and number of heterozygotes became saturated given the sample size ($\mu = 10^{-4}$; Figs 3, S4.1.1; Tables S3.1-4). For t_{50eq}, a similar pattern was observed with respect to N_{e1}, however there were usually negative relationships with μ, where faster μ was associated with faster approaches to equilibrium, and thus lesser t_{50eq} (Figs 4, S4.1.2, S4.3.2; Table S3.1-4). This effect became more pronounced with larger increases in population size.

By contrast, less variation in the logistic parameter Φ_3 were explained by the models (LM.1-4, Table 1) than for the other parameters (Tables 2, S3.2, S3.4; Figs 5, S4.4.3). The Φ_3 for nucleotide diversity had the worst model fit (R^2=0.38; Table S3.2) with gene diversity exhibiting the next poorest fit (R^2=0.82). For the other indices, 83-98% of the variation in Φ_3 was explained by the best models, which again were either LM.2 (with μ^2) or LM.4 (with both μ^2 and N_{e1}^2) (Table 2; Tables S3.2, S3.4).

Comparisons of Common Genetic Diversity Indices

As mutation rate increased, and to a lesser extent as the N_{e1} increased, the different genetic diversity indices exhibited increasingly unique behavior (Figs 2, S1.7-1.11). There were significant differences among the diversity indices in their rates of increase as well as the times to increase to 5% (t_{05eq}), 50% (t_{50eq}), and 95% (t_{95eq}) of their equilibrium values ($p < 0.05$; Table S1-2; Figs 2, 4, 6-7), and the differences varied with μ, N_{e0}, N_{e1}, and their interaction ($p < 0.05$). At $\mu=10^{-8}$, only two patterns of increasing standardized diversity were evident, regardless of N_{e1} (Figs 2A-B, S1.7), with nucleotide diversity, effective number of haplotypes, gene diversity, and number of heterozygotes increasing towards equilibrium at a greater maximum rate (Φ_3, smaller values indicate faster rates; Fig 5A-B; Table S1), but taking longer to reach t_{05eq} and t_{50eq} (Figs

4A-B, 6A-B) than the number of haplotypes and number of substitutions. At $\mu=10^{-6}$, as N_{e1} increased, Φ_3 became increasingly differentiated among the different diversity indices (Figs 5E-F) with the same two aforementioned groupings of indices at N_{e1}=500 and four groupings at N_{e1}=10,000. The Φ_3 for effective number of haplotypes and number of haplotypes decreased (faster) with increasing N_{e1}, while the others increased (slower). The t_{50eq} of all diversity indices increased with N_{e1}, but for gene diversity and the number of heterozygotes t_{50eq} increased more slowly (Figs 4E-F). The t_{05eq} and t_{95eq} also became more differentiated at higher N_{e1} (Figs 6-7E-F) but with different groupings of indices at N_{e1}=10,000 than for Φ_3. This demonstrates the complexity of the relationships here, and there are clearly interactions among N_{e0}, N_{e1}, μ, and D_s (Tables 2, S3.1-4). Indeed, at $\mu=10^{-5}$ and $\mu=10^{-4}$ there are 5-6 different patterns of diversity increase (Figs 2G-J), due to variation in Φ_3, t_{50eq}, t_{05eq} and t_{95eq} (Figs 4-7, S4.1.2, S4.1.5-7; S4.2.2, S4.2.5-7; S4.3.2, S4.3.5-7; S4.4.2, S4.4.5-7).

The differences in the standardized diversity indices at a given generation can be quite distinct. For example, when standardized nucleotide diversity is at 0.05 ($\mu=10^{-4}$, N_{e0}=2, N_{e1}=10,000; Fig 2), the standardized gene diversity, number of heterozygotes and number of haplotypes are above 0.95, approaching equilibrium. These differences in progression towards equilibrium decrease with decreasing μ and N_{e1}, but even at $\mu = 10^{-8}$ and N_{e1}=500, when the standardized effective number of haplotypes is 0.05, the standardized number of substitutions is at ~0.35. Overall, nucleotide diversity is generally the slowest to t_{05eq}, t_{50eq}, and t_{95eq}, while gene diversity and the number of heterozygotes are generally the fastest (Figs 4, 6-7). When N_{e1} is small and/or $\mu <= 10^{-6}$, however, the number of substitutions and haplotypes can be the fastest to t_{05eq} and t_{50eq}, while the effective number of haplotypes can be the slowest to t_{05eq} and t_{50eq}. Because it consistently has the highest Φ_3 (slowest), the number of substitutions can take as long

as nucleotide diversity and the effective number of haplotypes to reach t_{eq95} and Φ_1 when either N_{e1} is small or $\mu \leq 10^{-5}$.

The time to achieve 5% of the equilibrium value of diversity was lowest for the number of substitutions (Fig 6), except at $\mu=10^{-4}$, and $N_{e1}>1000$. For $N_{e1}<=1000$, $t_{05eq} < 10$ generations for the number of substitutions, and $t_{05eq} < 100$ generations even for $N_{e1} = 10{,}000$. At the other end of the spectrum, t_{05eq} = 20-3000 generations for the effective number of haplotypes and nucleotide diversity and was positively related to N_{e1} and maximized at $\mu=10^{-6}$. The time to reach 95% of the equilibrium level of nucleotide diversity (1000-30,000 generations) was also positively related to N_{e1} but negatively related to μ. Due to a limitation on the sample size (n=100), t_{95eq} was shortest for the number of heterozygotes (~100-10,000 generations, Fig 7).

Prediction Model

Generally, as N_{e1} increased, the amount of bias for PhiPhinder.r approached 1.0 for all measures of genetic diversity for D_{eq} (Figs 8, S5-6). Increasing mutation rate resulted in a similar pattern of the relative bias approaching 1.0 except when $N_{e0} = 100$, $\mu = 10^{-4}$, for number of heterozygotes (Fig 8B) and gene diversity (Fig 8D) which had extreme bias estimates of 0.11 and 0.13, respectively. Effective number of haplotypes (Fig 8E, F) has the highest and lowest bias values of 1.75 and 0.001 respectively and a mean of 0.61. Number of heterozygotes (Figs 8A, B), gene diversity (Figs 8C, D), number of substitutions (Figs 8I, J), and nucleotide diversity (Fig 8K, L) have bias values closer to 1.0 with means of 0.84, 0.84, 0.97, and 0.96, respectively. On average PhiPhinder.r overestimates D_{eq} values with an average bias value of 0.81.

CHAPTER IV

DISCUSSION

While the behavior of heterozygosity in expanding populations is well known (e.g. Nei et al., 1975; Maruyama & Fuerst, 1984; Tajima, 1989), it receives relatively little attention when compared to contracting populations, and it can be difficult to interpret the results of modern studies based upon the seminal work because of differences in units or measures of genetic diversity employed. Detailing the accumulation of genetic diversity in expanding populations should be relevant in studies of species invasions, range expansions, and populations recovering from bottlenecks due to anthropogenic activities.

Differences among the diversity indices

There are clearly differences among diversity indices in populations that have expanded, which can be best visualized in their standardized form (Figs 2, S1.7-11), and it should be possible to exploit these differences in population studies. Differences amongst the measures of genetic diversity can be seen even in the lowest values of μ where accumulation of diversity is similar among metrics of the same order (see Fig 2; $q = 0$ for number of haplotypes and substitutions & $q = 1$ for the others; Jost, 2008). Zero-order diversity measures are more sensitive to rare variants because they consist of simple counts of variants where rare and common variants are equally weighted. Since expanding populations have an excess of rare alleles (Maruyama & Fuerst 1984), these measures tend to increase sooner after the population expansion than most second-order measures unless they become saturated, which depends on μ, sample size, and number of nucleotides per locus. Metrics that approach their upper limits at lower μ (number of haplotypes, heterozygotes, gene diversity) generally increase sooner after the expansion as μ increases than those that do not approach saturation (nucleotide diversity).

The emergent result is that the metrics of diversity are differentially informative depending upon time since expansion and μ, where informative is defined as likely to differ between sampling times. For example, if $\mu=10^{-4}$, at ~100 generations after the expansion some metrics will be near equilibrium (number of haplotypes, heterozygotes, gene diversity) and others will be near zero (nucleotide diversity). Clearly, different diversity metrics used individually could result in different conclusions. The maturation of genetic diversity is best described by the differential progression towards equilibrium among the metrics. The windows of maximal informativeness for the diversity metrics are approximately defined as the generations between t_{05eq} and t_{95eq} (Table S1). They are strongly affected by mutation rate and more weakly affected by initial and final population sizes. With an estimate of population size, which could be obtained from genetic data using NeEstimator (Do et al., 2014), and an estimate of μ, observed metrics of diversity can be compared to their equilibrium values. If at least one of the metrics is in the window of informativeness, then it may be possible to detect changes in diversity between samples from different time points. If the initial population is between 2 and 100, it has little effect on the trajectories of diversity increase (Figs. S1.1-11). A more general recommendation is that when testing for differences in diversity among sampling times in expanding populations, it would be wise to employ several diversity metrics because they may exhibit different patterns.

Considerations for Empirical Next-Gen Studies of Expanding Populations

In modern studies that employ high-throughput Illumina sequencing (Kuhn et al. 2004), it is common to over-filter the data and remove real diversity in a quest to remove sequencer errors (Kircher et al. 2012, O'Leary et al. 2018) – especially rarer variants, which are most common in expanding populations. Conversely, under-filtering will inflate rare variants, so a balance must

be struck. We recommend that attention be given to retaining as much natural variation as possible while attempting to remove variation caused by procedural error, especially when basing conclusions on rare variants or zero-order diversity metrics (O'Leary et al. 2018). Explicitly, we are advocating against the common practice of (blindly) removing loci with low frequency alternate alleles. This necessitates adequate depth of coverage and the sampling of many cells (genomes) to decipher between errors and rare alleles. The exception to this is for metrics that are relatively insensitive to rare alleles, e.g., nucleotide diversity, gene diversity and effective number of haplotypes, will be less affected by over- or under-filtering because they are not sensitive to rare alleles. The ability to adequately filter the data could affect which diversity metrics are reliable.

It is worthwhile to note that while nucleotide diversity and number substitutions can be calculated with SNP data, haplotypic data is required to calculate the other indices. Software that enables haplotyping of population genomic data such as rad_haplotyper (Hollenbeck, 2017; Willis et al., 2017) will be useful in calculating the indices evaluated here. The inferences made from just nucleotide diversity and number of substitutions, can be limited. Depending on μ, there can be "dead zones" of sensitivity to changes in diversity if only nucleotide diversity and number of substitutions are used. For example, assume a population where N_{e0}=2, N_{e1}=10,000, and $\mu=10^{-6}$, the population will reach genetic equilibrium according to the number of heterozygotes, genetic diversity, and effective number of haplotypes, thousands of generations before either nucleotide diversity or number of substitutions. The one exception here is that SNP data can be used to estimate site-frequency spectra, which can be useful in studying expanding (or contracting) populations (Städler et al., 2009; Keinan & Clark 2012) but is beyond the scope of this work.

It takes a long time for genetic diversity to accumulate and the larger the expansion the longer it takes.

One of the greatest justifications for the conservation of natural populations, beyond that our existence depends upon them, is that diversity can be lost in a short period of time, but it takes a long time for that diversity to replenish. Given middling mutation rates (e.g., $\mu=10^{-6}$), it takes ~3,000-30,000 generations for neutral genetic variation to approach equilibrium values in populations that expand from 2 or 100 to 500-10,000 effective individuals. Perhaps even more sobering is that it can take up to 3,500 generations for nucleotide diversity to increase to 5% of its equilibrium value (N_{e0}=100, Fig. 2F) and up to 161,000 generations to reach 95% in a population expanding to 10,000 effective individuals.

Bridging the gap between simulation and real-world populations, critically endangered species, such as the kākāpō, *Strigops habroptilus,* can require hundreds or thousands of generations to recover genetic diversity, as conservation efforts attempt to expand their populations (Gerrodette et al. 2011 & White et al. 2015). In conservation, the number of variants can be monitored to track the genetic health of a population. While genetic equilibrium may not exist in natural populations, we can use it to measure potential genetic diversity in a population. Here, we can examine the Southern White Rhinoceros, *Ceratotherium simum*. As of January 2020, the estimated population size is around 10,000 adults, which grew from 20-50 adults in the late 19th century (Emslie 2020). Using an estimated μ of 2.5×10^{-8} (Tunstall et al. 2018) and assuming the number of haplotypes to be 1 (based on the D_{eq} value of a population expanding to 50 individuals), we can use PhiPhinder.r to estimate genetic equilibrium for *C. simum.* Using PhiPhinder's results, we can then estimate how long the population will take to reach genetic equilibrium. If the current *C. simum* population is left undisturbed and stabilizes at N_{e1}=10,000;

then an estimated 490 generations (3,920 yrs) are required to recover 5% of the equilibrium number of haplotypes and 61,376 generations (~491,009 yrs) to reach 95% of the equilibrium number of haplotypes, given a generation time of 8 years (Hillman-Smith et al. 1986). Here, we show the stark contrast between how quickly genetic diversity can be lost and the time necessary for a population to regain that genetic diversity. It is important to note that these simulations were done assuming there were no biotic or abiotic factors which results in an even faster accumulation of genetic diversity than in the natural world.

Utility of Models Presented Here

The logistic and linear models presented here can be used to estimate genetic diversity in expanding populations that begin with 2-100 effective individuals and expand to 100-10,000 individuals for the realistic range of mutation rates encountered in nature. It took a substantial amount of computational power to perform the simulations and analysis presented here, and it would require a substantial amount of work to obtain similar results for specific simulations. PhiPhinder.r can be used to estimate logistic model parameters (Eq 2; D_0, D_{eq}, t_{50eq}, Φ_3) given combinations of N_{e0}, N_{e1}, and μ and curves of diversity in expanding populations could be drawn from the output. The R script, PhiPhinder.r, uses the results reported here to generate estimated model parameters for each measure of genetic diversity when given any N_{e0}, N_{e1}, and μ.

Despite very good model fits to the simulated data in most circumstances (Table 2), the performance of PhiPhinder was variable, and it performed well in some areas of parameter space and poorly in others (Fig 8), thus there are some limitations. First, if one wishes to estimate patterns of diversity in populations that exceed the range of N_{e0}, N_{e1}, and μ reported here, they will be better off using another method to predict changes in diversity. PhiPhinder.r has a built-in switch that will give better estimates of D_{eq}, t_{50eq}, and/or Φ_3 if N_{e0}, N_{e1}, and μ, are restricted to the

values reported here. The switch will take the exact values from the estimated model parameter regression rather than attempting to estimate them. Even then, this does not guarantee low bias in the estimates of PhiPhinder.r.

Second, there is a built-in assumption that contig lengths will be 300bp and sample sizes will be 100 diploids. Number of haplotypes and heterozygotes will be sensitive to the sample size and the number of substitutions will be sensitive to the haplotype length, especially when they approach their maximum possible values when N_{e1} and μ are large. The other diversity indices should be more robust to deviations in sample size and haplotype length.

Third, the logistic models used by PhiPhinder.r exhibit systematic bias through time when compared directly with the simulated data (Fig S2), and while the bias was generally below 5% for $\mu >= 10^{-6}$, it did increase to as much as ~30% for $\mu = 10^{-8}$. The degree of bias beyond ~5% was directly a function of the number of loci simulated relative to the number of mutations in the population at equilibrium, and thus low μ and N_{e1} were associated with slightly higher bias when we could not reasonably simulate the number of loci. For N_{e1}=100 and $\mu=10^{-8}$, the amount of computational power available to us was not enough to reasonably control bias.

Fourth, choosing a correct model to fit the logistic model parameters (D_{eq}, t_{50eq}, and Φ_3) with nonlinear regression proved difficult. Each logistic model parameter only had 4 or 5 distinct values, depending on N_{e0}, making it challenging to fit curves. The greatest challenge was towards the extreme values of μ (e.g., 10^{-8}), where model fit was the poorest, presumably due in part to variation. Fitting just one model to all 4 or 5 points led to large errors when estimating the logistic model parameter, especially at the extreme values of μ. The data was split by μ, N_{e0}, and/or N_{e1} to allow multiple models to fit the data better and give better estimates when N_{e0}, N_{e1}, and μ, are restricted to what is reported here. While doing so gives better estimates when input

values are restricted, the estimated values when input values are not restricted can result in poor estimates.

Fifth, choosing the correct models for nonlinear regression proved difficult. While the chosen model, four parameter logistic regression, performed well, it was not the correct curve. Many of the logistic models had difficulty fitting the data around the midpoint. An obvious example is when N_{e0}=2 or 100, N_{e1}=10,000, and $\mu=10^{-6}$, for number of haplotypes (Fig S1.4), for raw or standardized data. A better fitting model would provide better estimates of t_{50eq} and thus better estimates for PhiPhinder.r.

Overall, PhiPhinder.r is a powerful tool that estimates six different genetic diversity measures. The most reasonable estimates when μ, N_{e0}, and N_{e1}, are restricted to what is reported here. The output of PhiPhinder.r can be used to get a rough estimate of how many generations will be required for a population genetic diversity index to reach genetic equilibrium. Given PhiPhinder's restrictions, genetic diversity estimates can help researchers predict the change in genetic diversity metrics in an expanding population given μ, N_{e0}, and N_{e1}. PhiPhinder.r produces estimates in seconds which is significantly faster than using the SLiM2 to Arlequin pipeline to obtain results. If the restrictions of PhiPhinder.r are too great, then the SLiM2 to Arlequin pipeline can prove useful to users wishing to model specific populations. Users can easily change model input parameters, such as μ, N_{e0}, and N_{e1}, and n, to best fit the study population.

Conclusions

While the predictive model performance was somewhat disappointing, the model fitting performance was quite good, and this effort did highlight important principles regarding the behavior of genetic diversity indices in expanding populations. We built upon previous efforts which found that two metrics of genetic diversity approached equilibrium at different rates in expanding populations [heterozygosity and allelic richness (Nei et al., 1975), allelic richness and gene diversity (Nei & Li, 1976; Maruyama & Fuerst, 1984), number of substitutions and gene diversity (Tajima, 1989)], by showing that all six metrics of genetic diversity (nucleotide diversity, number of substitutions, gene diversity, haplotypic richness, effective number of haplotypes, heterozygosity) do so, depending upon population size and mutation rate. It is critical to studies seeking to test for increases in genetic diversity through time to select diversity measures carefully to match the population and time frame being studied to maximize the chances of detecting an increase in diversity. Further, the wealth of simulation results spanning a broad range of realistic parameter space for expanding natural populations can serve as a reference for a broad variety of species and populations.

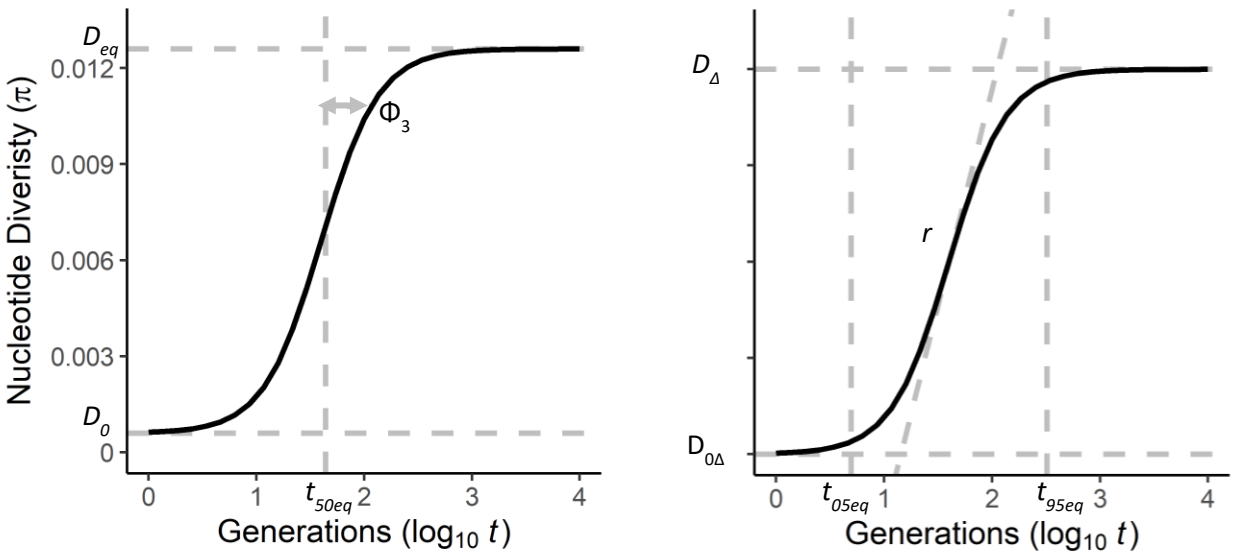

Figure 1. Logistic model of genetic diversity accumulation. (A) A visual depiction of the logistic model (eq. 2) used for describing the change in genetic diversity through time in expanding populations. The variables are described in the text. (B) We reduced the number of model parameters to three by adjusting the starting diversity to be zero ($D_0 - D_0 = D_{0\Delta} = 0$) and D_0 from each observed D value at every sampling time (see eq. 2). The equilibrium diversity value was proportionately decreased ($D_{eq} - D_0 = D_\Delta$). We also used the adjusted model to calculate the rate of increase (r) in genetic diversity at t_{50eq} and the time for genetic diversity to attain both 5% (t_{05eq}) and 95% (t_{95eq}). of its equilibrium value.

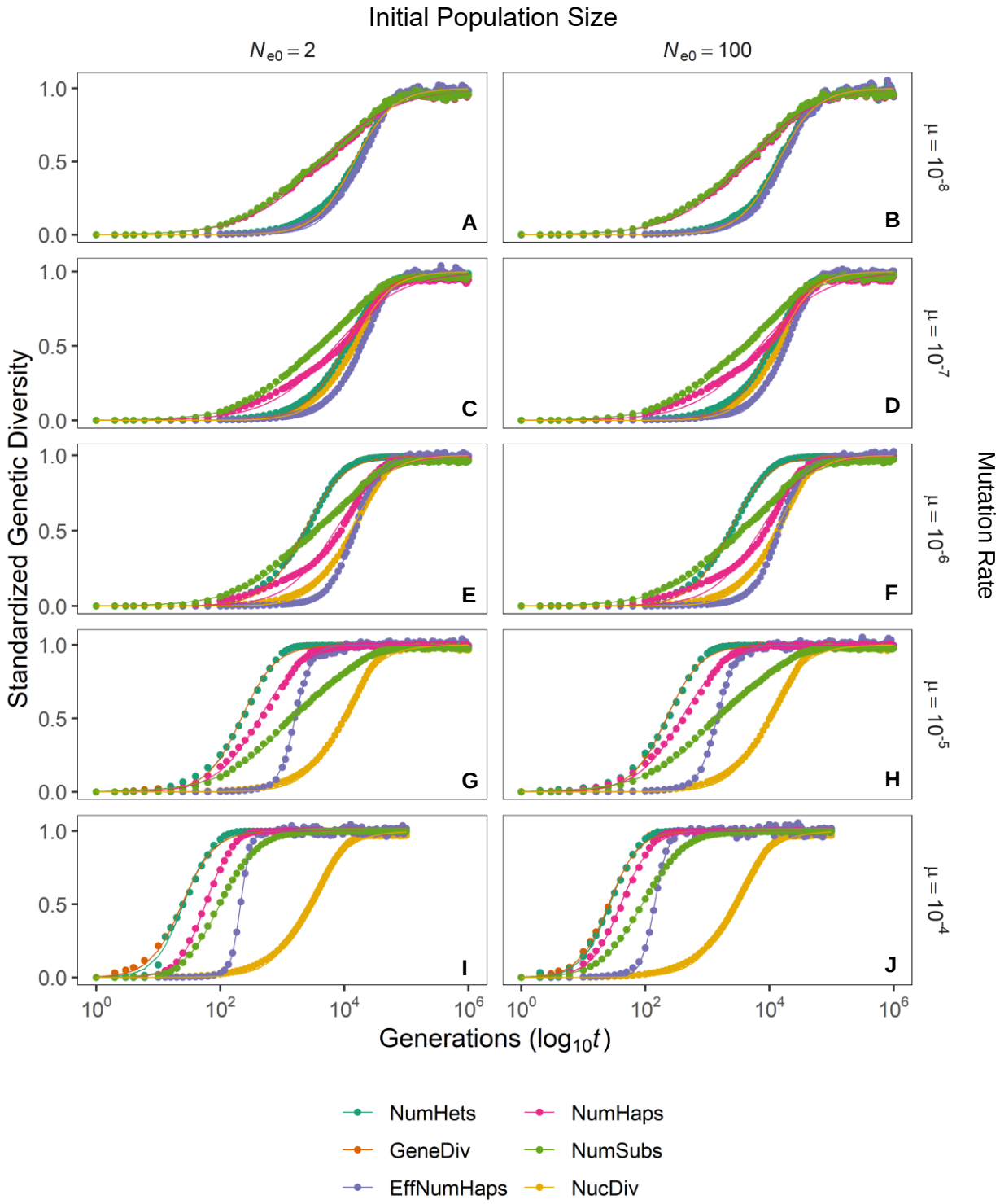

Figure 2. Changes in standardized genetic diversity in expanding populations, $N_{e1}=10^4$. All values have been standardized to a scale of $D_{0std} = 0$ to $D_{eqstd} = 1$. The results from all simulations are represented, with each point being the mean standardized diversity from each time point in each treatment combination. Columns depict initial population sizes; rows depict mutation rate; and colors depict diversity indices. Lines are best-fit logistic models described in Table S1.

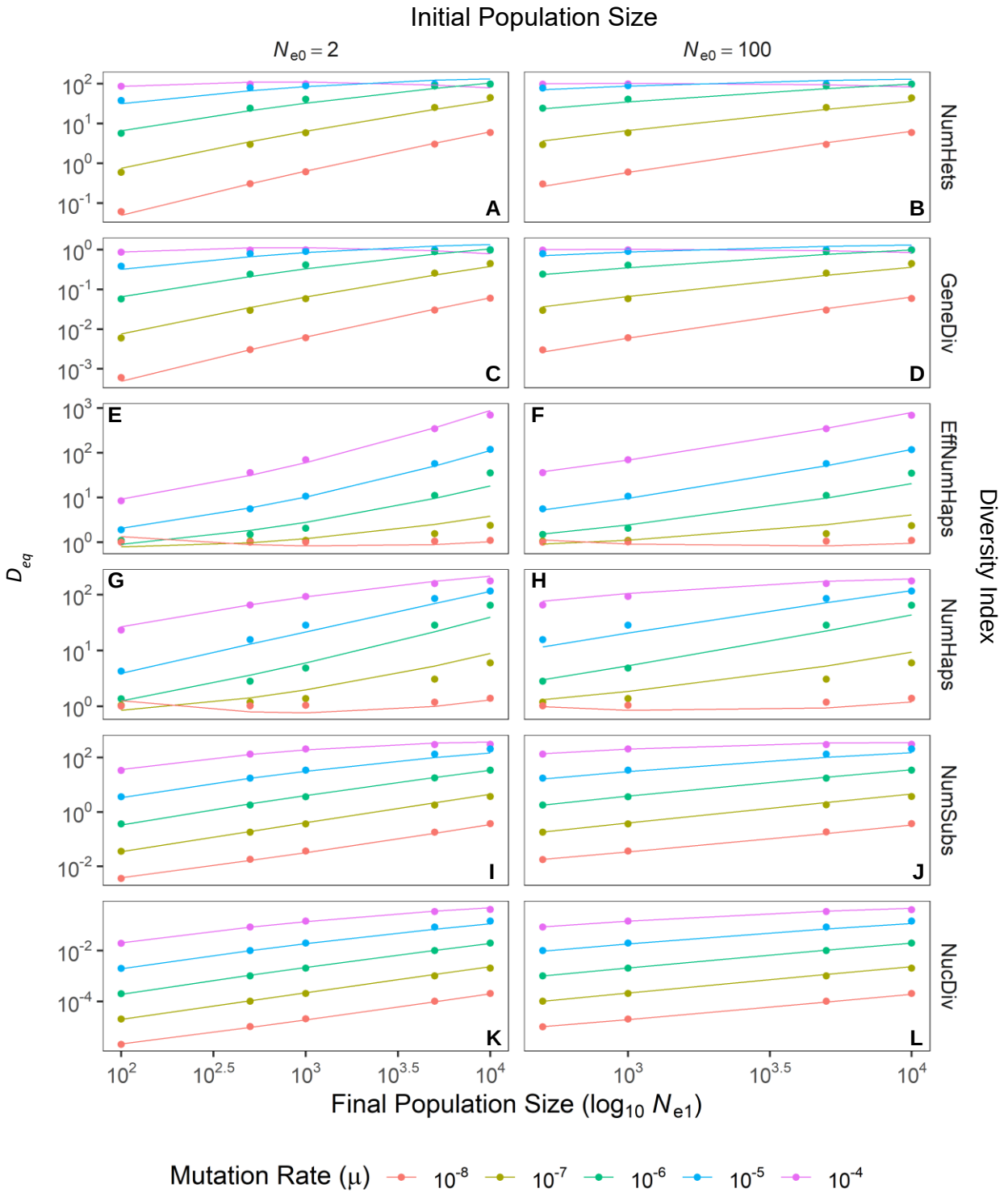

Figure 3. Modelled equilibrium diversity D_{eq} versus N_{e1}, μ, N_{e0} and diversity index. Points represent estimates of D_{eq} from the logistic models. All axes are log_{10} scaled. Columns depict different initial population sizes, and colors depict different final mutation rates (μ). The best fit global linear models (eq. 4) and associated statistics are listed in Table 2.

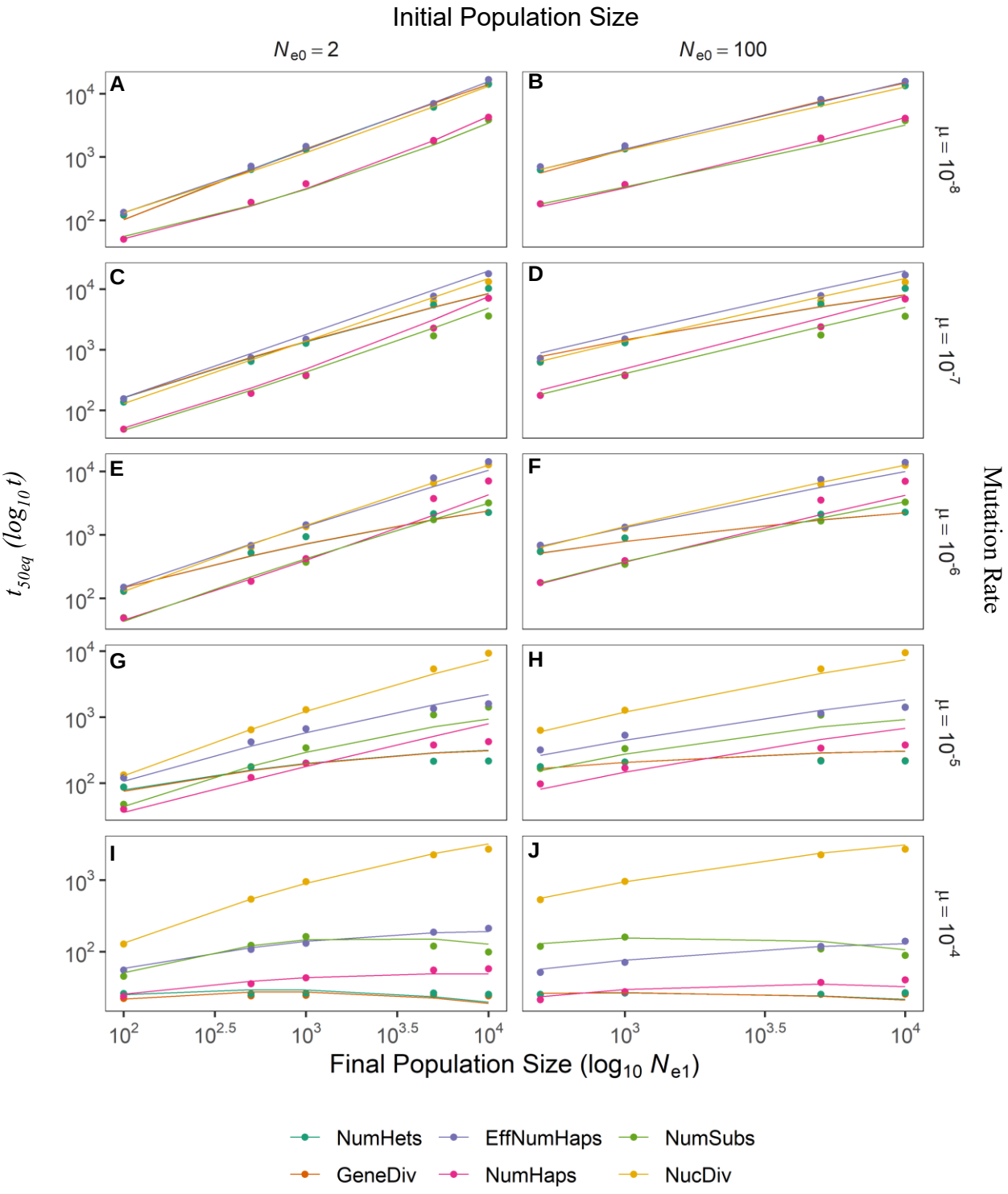

Figure 4. t_{50eq} for different genetic diversity indices in expanding populations. Points represent estimates of t_{50eq} from the logistic models. All axes are log_{10} scaled. Columns depict different initial population sizes, rows depict different mutation rates, and colors depict different mutation rates. The lines are the best fit linear model (Table 2).

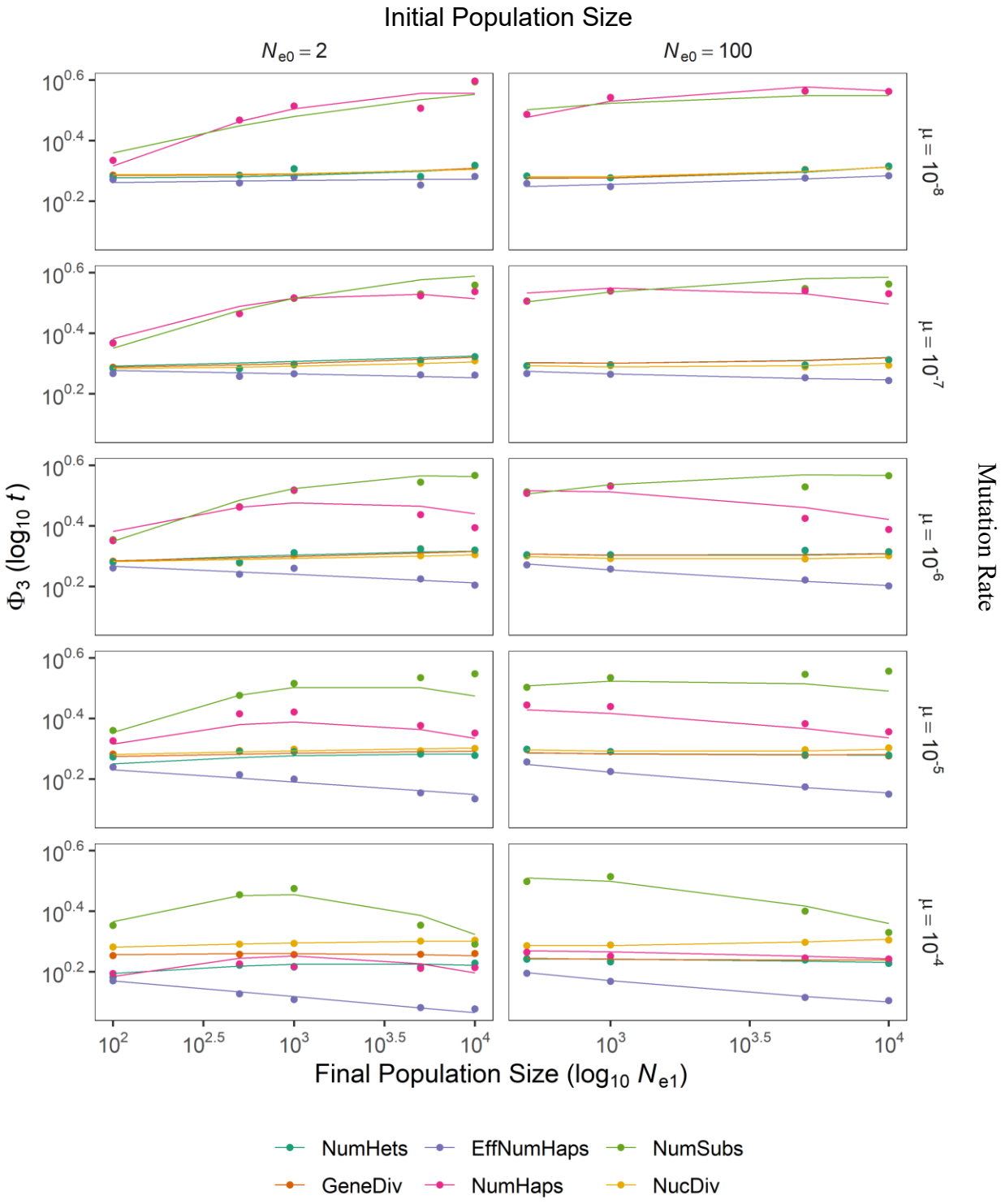

Figure 5. Φ_3 for different genetic diversity indices in expanding populations. Points represent estimates of Φ_3 from the logistic models. All axes are *log*$_{10}$ scaled. Columns depict different initial population sizes, rows are different mutation rates, and colors depict different measures of genetic diversity. Lines are best fit linear models (Table 2).

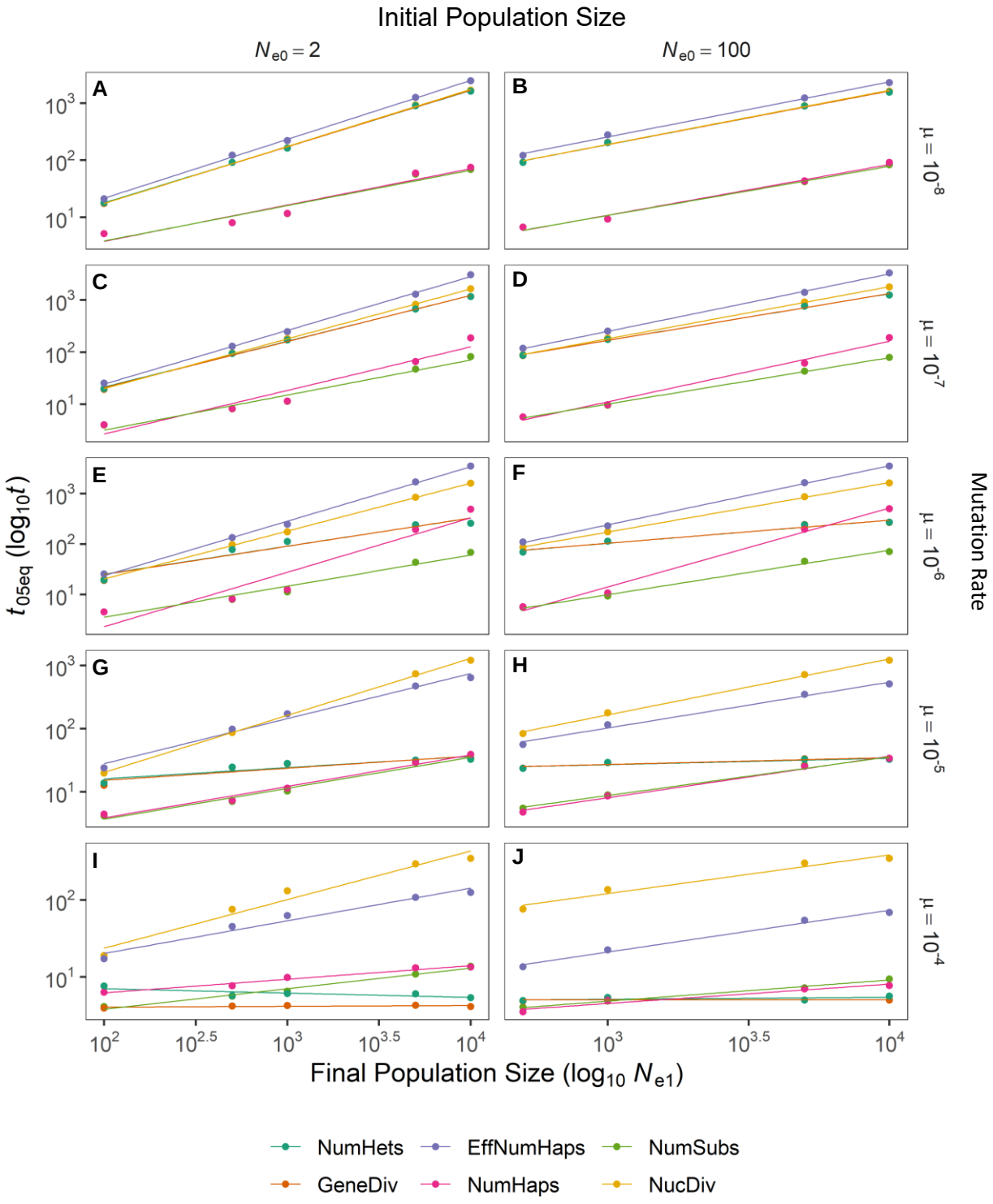

Figure 6. t_{05eq} for different genetic diversity indices in expanding populations. Points represent estimates of t_{05eq} from the logistic models. All axes are log_{10} scaled. Columns depict different initial population sizes, rows depict different mutation rates, and colors depict different measures of genetic diversity. Lines are best fit linear models.

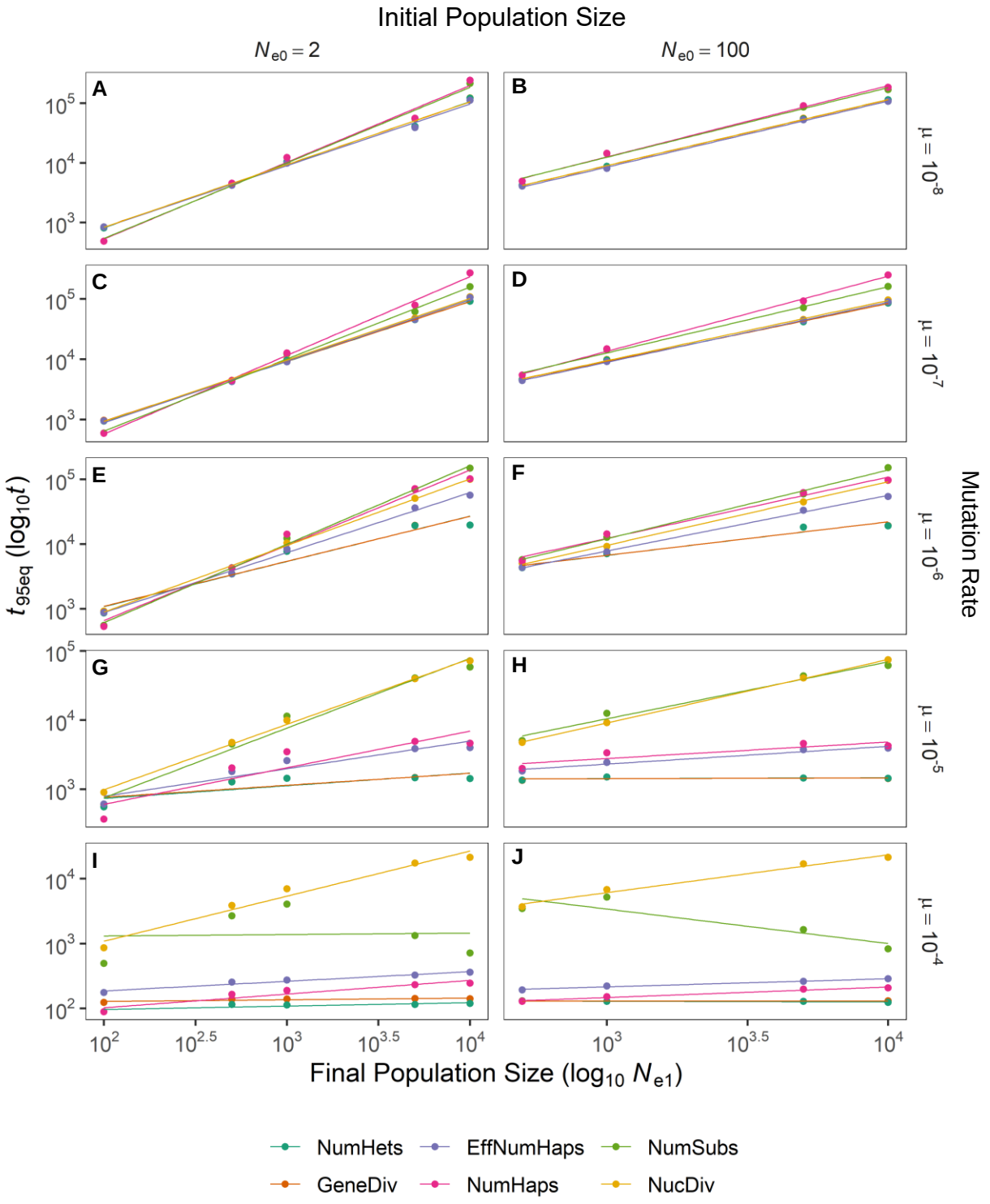

Figure 7. t_{95eq} for different genetic diversity indices in expanding populations. Points represent estimates of t_{95eq} from the logistic models. All axes are log_{10} scaled. Columns depict different initial population sizes, rows depict different mutation rates, and colors depict different measures of genetic diversity. Lines are best fit linear models.

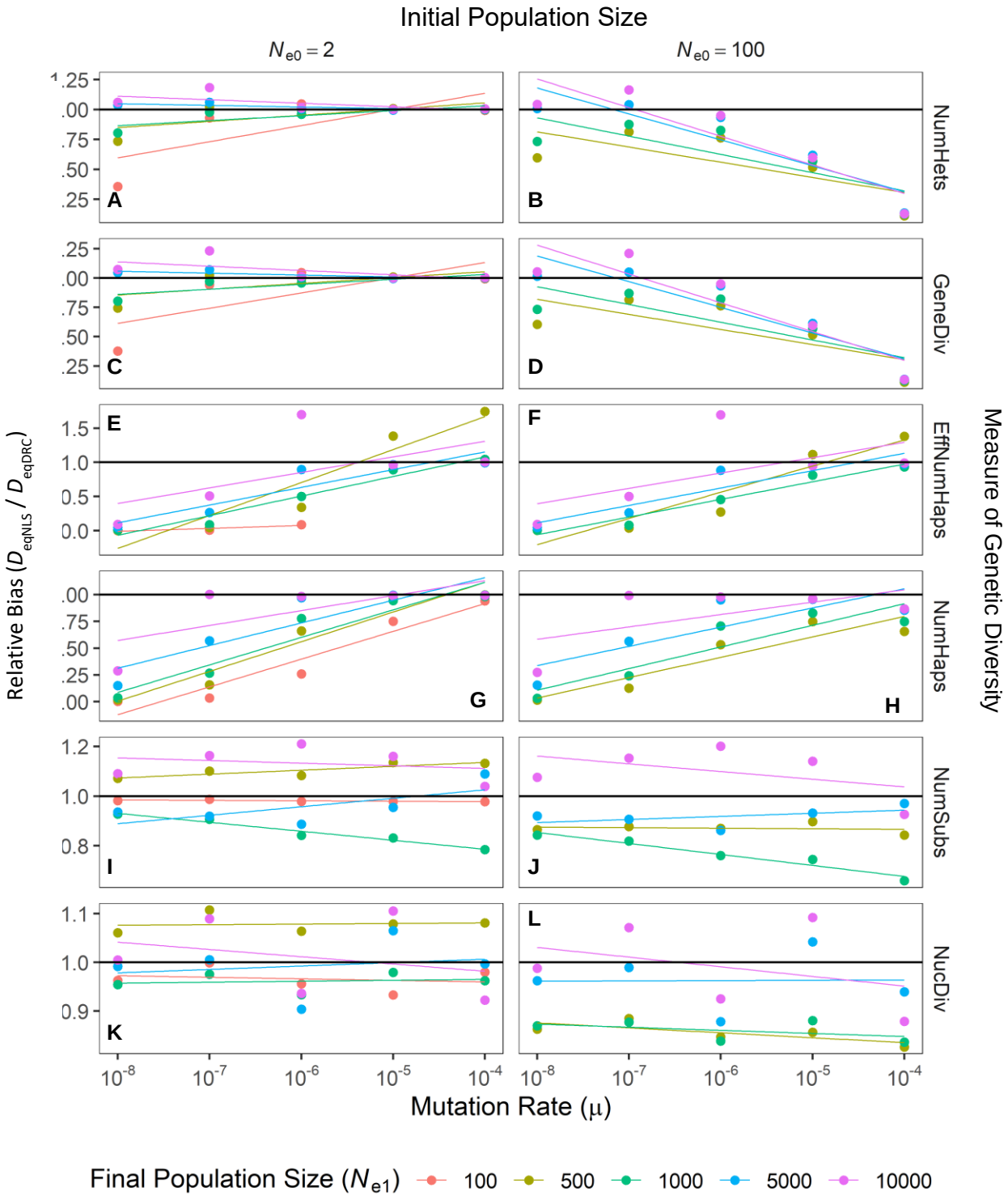

Figure 8. Relative bias in prediction of D_{eq}. The equilibrium diversity values estimated by the PhiPhinder.r script are divided by the logistic model estimated D_{eq} values and plotted against mutation rate, genetic diversity index, and final population size.

Table 1. The functions used in modeling the nonlinear model parameters of D_{eq} and the coefficients of the resulting models. Equations are presented how they are coded in R. The nonlinear models were used to both estimate D_{eq} in response to mutation rate ($x = \mu$), as well as the nonlinear model parameter estimates themselves ($x = a$, x = b, ... etc) in response to N_{e1}. In some instances, a quadratic model (DRC.poly2) was a better estimator of the variation in the nonlinear model parameters in response to N_{e1}.

R Package	Model Name	Function Description	Response Variables	Equation
Base R	LM.4.1	First-order polynomial	D_{eq}, t_{50eq}, Φ_3	$f(x) =$ log10(μ) * log10(N_{e1}) * N_{e0}
	LM.4.2	Second-order polynomial	D_{eq}, t_{50eq}, Φ_3	$f(x) =$ N_{e0} * log10(μ)^2 * log10(N_{e1}) + log10(μ) * log10(N_{e1}) * N_{e0}
	LM.4.3	Second-order polynomial	D_{eq}, t_{50eq}, Φ_3	$f(x) = N_{e0}$ * log10(N_{e1})^2 * log10(μ) + log10(μ) * log10(N_{e1}) * N_{e0}
	LM.4.4			
		Second-order polynomial	D_{eq}, t_{50eq}, Φ_3	$f(x) = N_{e0}$ * log10(N_{e1})^2 * log10(μ) + log10(μ) * log10(N_{e1}) * N_{e0} + N_{e0} * log10(μ)^2 * log10(N_{e1})
aomisc	DRC.poly2	Second-order polynomial	a, b, c, d, g, f	$f(x) = a + bx + cx^2$
aomisc	DRC.powerCurve	Power	D_{eq}, a, b	f(x) = a * x^b
aomisc	DRC.asymReg	Asymptotic	D_{eq}, a, b, c, g	$f(x) = a + (b - a) * (1 - \exp(-cx))$
DRC	AR.3	Three-parameter asymptotic	D_{eq}, a, b, c, g	$f(x) = c + (d - c) * (1 - \exp(-x/g))$
DRC	AR.2	Two-parameter asymptotic	D_{eq}, d, g	$f(x) = d * (1 - \exp(-x/g))$
DRC	LL.3	A three-parameter log-logistic function with a lower limit of 0.	D_{eq}, b, d, g	$f(x) = 0 + (d - 0) / (1 + \exp(b * (log(x) - log(g))))$
DRC	LL.4	A four-parameter log-logistic function	D_{eq}, b, c, d, g	$f(x) = c + (d - c) / (1 + \exp(b * (log(x) - log(g))))$
DRC	LL.5	A five-parameter log-logistic function	D_{eq}, b, c, d, g, f	$f(x) = c + (d - c) / (1 + \exp(b * (log(x) - log(g))))$ ^ f

Table 2. Results from the best models, as determined by AIC, from the multiple linear regression analysis of the relationship of the logistic model parameters (D_{eq}, t_{50eq}, Φ_3) and the rate of diversity increase at t_{50eq} (r) with respect to initial population size (N_{e0}), final population size (N_{e1}), and mutation rate (m) for each of six diversity indices (see LM). *RSE* is the residual standard error, *df* is the degrees of freedom, and *p* is the probability of observing a t-statistic at least as extreme as that reported here by random chance. Mutation rate and final population size were log_{10} transformed to satisfy the assumptions of the linear model. Emphasis indicates whether the information criteria supported the selected model. In the case where BIC did not support the model selected, it supported LM.4. See Figs 3-6 for visualizations of the models and data.

Response	Diversity Index	Best Model	R^2	R^2_{adj}	*RSE*	*F*-stat	*p*	*df*	*AIC*	*BIC*
D_0 Linear	EffNumHaps	LM.4	0.9992	0.9990	6.12E-03	6.24E+03	7.94E-55	37	-3.22E+02	-3.05E+02
	GeneDiv	LM.4	0.9954	0.9945	6.18E-03	1.14E+03	3.49E-41	37	-3.21E+02	-3.05E+02
	NucDiv	LM.4	0.9931	0.9918	2.11E-04	7.63E+02	5.38E-38	37	-6.25E+02	-6.09E+02
	NumHaps	LM.4	0.9995	0.9994	7.93E-03	1.02E+04	8.49E-59	37	-2.98E+02	-2.82E+02
	NumHets	LM.4	0.9999	0.9998	8.99E-03	3.66E+04	4.86E-69	37	-2.87E+02	-2.71E+02
	NumSubs	LM.4	0.9992	0.9991	1.38E-02	6.92E+03	1.19E-55	37	-2.48E+02	-2.32E+02
D_{eq} Linear	EffNumHaps	LM.4	0.9829	0.9772	1.36E-01	1.72E+02	7.38E-26	33	-3.95E+01	-1.61E+01
	GeneDiv	LM.4	0.9932	0.9910	8.37E-02	4.40E+02	1.81E-32	33	-8.35E+01	-6.00E+01
	NucDiv	LM.4	0.9993	0.9989	4.85E-02	2.68E+03	2.26E-41	29	-1.30E+02	-9.97E+01
	NumHaps	LM.2	0.9796	0.9728	1.36E-01	1.44E+02	1.36E-24	33	-3.96E+01	-1.61E+01
	NumHets	LM.4	0.9932	0.9910	8.37E-02	4.39E+02	1.82E-32	33	-8.35E+01	-6.00E+01
	NumSubs	LM.4	0.9982	0.9973	7.28E-02	1.09E+03	1.04E-35	29	-9.38E+01	-6.31E+01
D_{eq} Non linear	EffNumHaps †	LL.3	-	-	7.78E-02	-	-	30	-8.84E+01	-5.95E+01
	EffNumHaps ‡	DRC.pwrCurve	-	-	3.57E+00	-	-	35	2.53E+02	2.73E+02
	GeneDiv	LL.4	-	-	2.66E-03	-	-	25	-3.91E+02	-3.53E+02
	NucDiv	LL.4	-	-	2.48E-02	-	-	25	-1.90E+02	-1.52E+02
	NumHaps †	LL.5	-	-	2.06E-02	-	-	20	-2.06E+02	-1.59E+02
	NumHaps ‡	asymReg	-	-	1.03E+01	-	-	30	3.51E+02	3.80E+02
	NumHets	LL.4	-	-	2.48E-01	-	-	25	1.77E+01	5.56E+01
	NumSubs	LL.5	-	-	5.80E-03	-	-	20	-3.20E+02	-2.73E+02
t_{50eq} Linear	EffNumHaps	LM.2	0.9909	0.9879	8.48E-02	3.26E+02	2.36E-30	33	-8.24E+01	-5.89E+01
	GeneDiv	LM.4	0.9924	0.9899	8.69E-02	3.93E+02	1.13E-31	33	-8.01E+01	-5.66E+01
	NucDiv	LM.4	0.9955	0.9932	5.11E-02	4.29E+02	7.13E-30	29	-1.26E+02	-9.50E+01
	NumHaps	LM.2	0.9783	0.9711	1.32E-01	1.35E+02	3.70E-24	33	-4.23E+01	-1.88E+01
	NumHets	LM.2	0.9923	0.9897	8.72E-02	3.85E+02	1.62E-31	33	-7.98E+01	-5.63E+01
	NumSubs	LM.4	0.9802	0.9700	1.05E-01	9.57E+01	1.45E-20	29	-6.06E+01	-2.99E+01
Φ_3, r_{std} Linear	EffNumHaps	LM.2	0.9814	0.9752	9.38E-03	1.58E+02	2.89E-25	33	-2.81E+02	-2.57E+02
	GeneDiv	LM.2	0.8658	0.8210	1.01E-02	1.94E+01	2.52E-11	33	-2.74E+02	-2.50E+02
	NucDiv	LM.2	0.4788	0.3802	7.53E-03	4.86E+00	5.89E-04	37	-3.03E+02	-2.87E+02
	NumHaps	LM.4	0.9605	0.9400	2.84E-02	4.70E+01	2.89E-16	29	-1.79E+02	-1.48E+02
	NumHets	LM.2	0.8955	0.8606	1.26E-02	2.57E+01	4.70E-13	33	-2.54E+02	-2.30E+02
	NumSubs	LM.4	0.8884	0.8307	3.20E-02	1.54E+01	6.17E-10	29	-1.68E+02	-1.37E+02
r Linear	EffNumHaps	LM.2	0.9979	0.9972	8.91E-02	1.44E+03	5.94E-41	33	-7.79E+01	-5.44E+01
	GeneDiv	LM.4	0.9918	0.9891	8.79E-02	3.65E+02	3.81E-31	33	-7.91E+01	-5.56E+01
	NucDiv	LM.4	0.9992	0.9988	5.00E-02	2.51E+03	5.80E-41	29	-1.28E+02	-9.70E+01
	NumHaps	LM.4	0.9967	0.9956	9.30E-02	9.15E+02	1.09E-37	33	-7.40E+01	-5.05E+01
	NumHets	LM.4	0.9922	0.9896	8.71E-02	3.81E+02	1.90E-31	33	-8.00E+01	-5.65E+01
	NumSubs	LM.4	0.9992	0.9987	5.02E-02	2.31E+03	1.94E-40	29	-1.27E+02	-9.65E+01
	† $\mu <= 10^{-6}$	‡ $\mu >= 10^{-6}$								

REFERENCES

Allendorf, F. W. 1986. Genetic drift and the loss of alleles versus heterozygosity. *Zoo biology, 5***,** 181-190.

Allentoft, M. E. & O'brien, J. 2010. Global amphibian declines, loss of genetic diversity and fitness: a review. *Diversity, 2***,** 47-71.

Barrett, R. D. & Schluter, D. 2008. Adaptation from standing genetic variation. *Trends in ecology & evolution, 23***,** 38-44.

Bell, G. & Collins, S. 2008. Adaptation, extinction and global change. *Evolutionary Applications, 1***,** 3-16.

Benning, T. L., et al. 2002. Interactions of climate change with biological invasions and land use in the Hawaiian Islands: modeling the fate of endemic birds using a geographic information system. *Proceedings of the National Academy of Sciences, 99***,** 14246-14249.

Butchart, S. H., et al. 2010. Global biodiversity: indicators of recent declines. *Science, 328***,** 1164-1168.

Chakraborty, R., et al. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences, 94***,** 1041-1046.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics, 10***,** 195-205.

Clavero, M. & García-Berthou, E. 2005. Invasive species are a leading cause of animal extinctions. *Trends in ecology & evolution, 20***,** 110.

Danecek, P., et al. 2011. The variant call format and VCFtools. *Bioinformatics, 27***,** 2156-2158.

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular ecology resources, 14(1), 209-214.

Dobzhansky, T. 1937. Genetic nature of species differences. *The American Naturalist, 71***,** 404-420.

Dobzhansky, T. 1982. *Genetics and the Origin of Species*, Columbia university press.

Ellstrand, N. C. & Elam, D. R. 1993. Population genetic consequences of small population size: implications for plant conservation. *Annual review of Ecology and Systematics, 24***,** 217-242.

Emslie, R. 2020. Ceratotherium simum. The IUCN Red List of Threatened Species 2020: e. T4185A45813880.

Excoffier, L. & Lischer, H. E. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources, 10***,** 564-567.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222***,** 309-368.

Fisher, R. A. 1923. XXI.—on the dominance ratio. *Proceedings of the royal society of Edinburgh, 42***,** 321-341.

Fisher, R. A. 1958. *The genetical theory of natural selection*

Flockhart, D. T., et al. 2015. Unravelling the annual cycle in a migratory animal: breeding-season habitat loss drives population declines of monarch butterflies. *Journal of Animal Ecology, 84***,** 155-165.

Frankham, R. 1996. Relationship of genetic variation to population size in wildlife. *Conservation biology, 10,* 1500-1508.

Frankham, R., et al. 2014. Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation, 170,* 56-63.

Fritts, T. H. & Rodda, G. H. 1998. The role of introduced species in the degradation of island ecosystems: a case history of Guam. *Annual review of Ecology and Systematics, 29,* 113-140.

Gerrodette, T., et al. 2011. A combined visual and acoustic estimate of 2008 abundance, and change in abundance since 1997, for the vaquita, Phocoena sinus. *Marine Mammal Science, 27,* E79-E100.

Gibbons, J. W., et al. 2000. The Global Decline of Reptiles, Déjà Vu Amphibians: Reptile species are declining on a global scale. Six significant threats to reptile populations are habitat loss and degradation, introduced invasive species, environmental pollution, disease, unsustainable use, and global climate change. *BioScience, 50,* 653-666.

Gnu, P. 2007. Free Software Foundation. *Bash (3.2. 48)[Unix shell program]*.

Haller, B. C. & Messer, P. W. 2017. SLiM 2: Flexible, interactive forward genetic simulations. *Molecular biology and evolution, 34,* 230-240.

Halpern, B. S., et al. 2008. A global map of human impact on marine ecosystems. *science, 319,* 948-952.

Hartl, D. L. & Clark, A. G. 2007. Principles of Population Genetics.

Hastings, A. 2013. *Population biology: concepts and models,* Springer Science & Business Media.

Hauser, L. & Carvalho, G. R. 2008. Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries, 9***,** 333-362.

Hillman-Smith, A., et al. 1986. Age estimation of the white rhinoceros (Ceratotherium simum). *Journal of Zoology, 210***,** 355-377.

Hollenbeck, C. M., Portnoy, D. S., Wetzel, D., Sherwood, T. A., Samollow, P. B., & Gold, J. R. (2017). Linkage mapping and comparative genomics of red drum (Sciaenops Ocellatus) using next-generation sequencing. G3: Genes, Genomes, Genetics, 7(3), 843-850.

Hulme, P. E. 2009. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of applied ecology, 46***,** 10-18.

Hutchings, J. A. 2000. Collapse and recovery of marine fishes. *Nature, 406***,** 882-885.

Jost, L. 2008. GST and its relatives do not measure differentiation. *Molecular ecology, 17***,** 4015-4026.

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. science, 336(6082), 740-743.

Kircher, M., et al. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research, 40***,** e3-e3.

Lacy, R. C. 1987. Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection, and population subdivision. *Conservation biology, 1***,** 143-158.

Lynch, M. 2010. Evolution of the mutation rate. *TRENDS in Genetics, 26***,** 345-352.

Magurran, A. E. 2013. *Measuring biological diversity,* John Wiley & Sons.

Maruyama, T. & Fuerst, P. A. 1984. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics, 108***,** 745-763.

Maruyama, T. & Fuerst, P. A. 1985. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics, 111***,** 675-689.

Mcneely, J. A., et al. 1990. *Conserving the world's biological diversity*, International Union for conservation of nature and natural resources.

Møller, A. P., et al. 2008. Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proceedings of the National Academy of Sciences, 105***,** 16195-16200.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, 70***,** 3321-3323.

Nei, M. & Li, W.-H. 1976. The transient distribution of allele frequencies under mutation pressure. *Genetics Research, 28***,** 205-214.

Nei, M. & Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, 76***,** 5269-5273.

Nei, M., et al. 1975. The bottleneck effect and genetic variability in populations. *Evolution***,** 1-10.

O'leary, S. J., et al. 2018. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Wiley Online Library.

Onofri, A. & Garcia, J. (2021). Aomisc. https://github.com/OnofriAndreaPG/aomisc

Pauls, S. U., et al. 2013. The impact of global climate change on genetic diversity within populations and species. *Molecular ecology, 22***,** 925-946.

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***,** e37135.

Pinheiro, J., et al. 2013. nlme: Linear and nonlinear mixed effects models. *R package version, 3***,** 111.

Potts, S. G., et al. 2010. Global pollinator declines: trends, impacts and drivers. *Trends in ecology & evolution, 25***,** 345-353.

Reed, D. H. & Frankham, R. 2003. Correlation between fitness and genetic diversity. *Conservation biology, 17***,** 230-237.

Ritz, C., et al. 2015. Dose-response analysis using R. *PloS one, 10***,** e0146021.

Rochette, N. C. & Catchen, J. M. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols, 12***,** 2640-2659.

Romiguier, J., et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature, 515***,** 261-263.

Rothschild, B. J., et al. 1994. Decline of the Chesapeake Bay oyster population: a century of habitat destruction and overfishing. *Marine Ecology Progress Series***,** 29-39.

Sibly, R. M. & Hone, J. 2002. Population growth rate and its determinants: an overview. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 357***,** 1153-1170.

Soulé, M. E. 1976. Allozyme variation, its determinats in space and time. *Moleculor Evolution***,** 60-77.

Städler, T., Haubold, B., Merino, C., Stephan, W., & Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics, 182(1), 205-216.

Tajima, F. 1989. The effect of change in population size on DNA polymorphism. *Genetics, 123***,** 597-601.

Tange, O. 2011. Gnu parallel-the command-line power tool. *The USENIX Magazine, 36***,** 42-47.

Team, R. C. 2013. R: A language and environment for statistical computing. Vienna, Austria.

Team, R. C. 2013. R: A language and environment for statistical computing.

Tunstall, T., et al. 2018. Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells. *Genome research, 28***,** 780-788.

White, K., et al. 2015. Evidence of inbreeding depression in the critically endangered parrot, the kakapo. *Animal Conservation, 18***,** 341-347.

Wickham, H. 2016. *ggplot2: elegant graphics for data analysis*, springer.

Wickham, H., et al. 2019. Welcome to the Tidyverse. *Journal of Open Source Software, 4***,** 1686.

Willis, S. C., et al. 2017. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources, 17***,** 955-965.

Willoughby, J. R., et al. 2015. The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology, 24***,** 98-110.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics, 16***,** 97.

Wright, S. 1932. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, n