APPENDIX

Probability, Statistics, and Coding

> He has a lot of extremely abstruse, in fact almost esoteric mathematics. Mathematics, incidentally, of a kind which I certainly do not claim to understand. I am not a mathematician at all. My way of reading Sewall Wright's papers, which I think is perfectly defensible, is to examine the biological assumptions the man is making, and to read the conclusions he arrives at, and hope to goodness that what comes in between is correct.
>
> **(Theodosius Dobzhansky 1962, p. 346)**

> Current research in population genetics employs advanced mathematical methods that are beyond the reach of most biology students.
>
> **(James F. Crow 1986, p. xi)**

The gulf between mathematical population genetics and the understanding of most biologists has greatly increased over the past 25 years because of the introduction of a variety of new theoretical and computational approaches (Guest Box 5). This can make it difficult for conservation geneticists to analyze data with recent computational approaches, and to publish their results in peer-reviewed conservation journals (Box A1).

The main purpose of this Appendix is to provide a basic understanding of the mathematical and statistical approaches used in this book. We also discuss the importance of **bioinformatics**, **coding,** and computational approaches in conservation genomics, while providing references to appropriate publications and training opportunities for those who want to learn why bioinformatics is so empowering and not that difficult! We have modeled this Appendix after the appendices in Crow & Kimura (1970) and Crow (1986), with substantial use of Dytham (2011). We recommend Whitlock & Schluter (2020) as an excellent general text for statistical analysis of biological data.

We have not tried to provide mathematical rigor, but rather intend to clarify the general nature and limitations of the mathematical and statistical approaches used in this book. We aim to provide a conceptual understanding of different statistical approaches and show how to interpret results, rather than to teach details about how to actually conduct a certain statistical test or likelihood estimation.

The basic idea of statistical inference is simple: a sample is taken from a population, and you want to extrapolate from that sample to make general conclusions about the population from which the sample was taken. To do this, you must understand the relationship between a population **parameter** and a sample **statistic**. A sample statistic can be the mean, median, mode, or some other quantity that describes a characteristic of that particular sample. Statistics are computed from samples because all the items or individuals that constitute the entire population can seldom be collected and measured.

If you do measure the entire population and calculate a mean, though, this value is a parameter of the population,

Box A1 Problems understanding sophisticated computational approaches

The first author of this book was an Associate Editor in the early days of the journal *Conservation Biology*. He handled a manuscript that applied some fairly sophisticated mathematical population genetics theory to a problem in conservation. He received the following review comments from a well-known population geneticist: "According to the Instructions to Reviewers for this journal, manuscripts should be understandable to conservation managers and government officials. It is not reasonable to expect either of these groups to understand stochastic theory of population genetics." This problem is much worse now than in the early 1990s because of the increased sophistication of computational approaches as presented in this Appendix. It is becoming increasingly important to analyze empirical data with complex statistical approaches. However, it is also becoming increasingly difficult to evaluate the reliability of these analyses.

not a statistic. In most situations, the parameter represents the truth; that is, the true population value that we want to estimate from the sample statistic. When a sample statistic is used to estimate the corresponding population parameter, it is often written with a "hat" (e.g., $\hat{F}_{ST} = 0.010$) to distinguish it from the population parameter (e.g., $F_{ST} = 0.015$). However, the "hat" is not necessary if the context clearly describes a value (F_{ST}) as a statistic or an estimate. In this Appendix, as in many texts, we do not use a hat to indicate a statistic, but rather we make it clear from context when we are discussing a statistic or estimate.

In this Appendix, we first give a brief historical perspective and explain the major differences between the three approaches to statistical inference. Then, we present concepts of probability and basic statistics, including hypothesis testing. Finally, we return to discuss in more detail likelihood and Bayesian approaches, along with the coalescent and **Markov chain Monte Carlo (MCMC)** methods and their importance in conservation genetics. We finish by discussing why bioinformatics is not so difficult, why **filtering** of next-generation sequencing (NGS) data is extremely important but challenging, and why **simulations** are so useful in conservation genetics.

A1 Paradigms

There are three main approaches, or paradigms, to statistical inference: frequentist, likelihood, and **Bayesian** approaches. Likelihood methods are sometimes classified within the frequentist approach (Section A5).

The Bayesian philosophy and statistical approach to data analysis were developed in the 18th century by the Reverend Thomas Bayes. The classic frequentist approach was formalized later, during the early 1900s, by K. Pearson and R.A. Fisher (both from England), as well as J. Neyman (from Poland); it quickly became dominant in science. Modern likelihood analysis was developed almost singlehandedly by R.A. Fisher between 1912 and 1922. A revival of the Bayesian approach has occurred during the past 20 years or so, thanks to advances in computer speed and simulation-based algorithms such as MCMC (Section A6) that allow the analysis of complex probabilistic models containing multiple interdependent parameters such as genotypes, population allele frequencies, population size, migration rates, and variable mutation rates among loci.

The **frequentist** approach to statistical inference generally involves four steps: stating a hypothesis, collecting data, computing a summary statistic (e.g., $F_{ST} = 0.01$), and then inferring how frequently we would observe our statistic (0.01) by chance alone, if our null hypothesis (H_0) is true (e.g., H_0: $F_{ST} = 0.00$). If our statistic is so large (e.g., $F_{ST} > 0.09$) that we expect to observe it very infrequently by chance alone (e.g., only once per 100 independent experiments), we would reject the null hypothesis. The frequentist approach determines the expected long-term frequency of an observation or a summary statistic, if we were to repeat the experiment or observation many times. Frequentist approaches typically use the **moments** of the distribution (a summary statistic) and thus are called "methods of moments." The moments are the mean and variance, as well as skewness and kurtosis. These concepts are discussed in the text that follows.

Likelihood approaches typically involve four steps: collecting data, developing a mathematical model with parameters (e.g., F_{ST}), plugging the raw data into the model (not a summary statistic), and computing the likelihood of the data for each of all possible parameter values; for example, $F_{ST} = 0.00, 0.01, 0.02$, up to 1.00. This requires many computations or iterations. We then identify the parameter value that maximizes the likelihood of obtaining our actual data under the model. The main advantage of likelihood over frequentist (moments) approaches is that likelihood uses the raw data (e.g., allele counts at each locus separately) instead of a summary of it (e.g., F_{ST} averaged across loci; Section A5). Thus, more

information is used from the data (e.g., interlocus variation in F_{ST}), and therefore the estimates of parameters (and inference in general) often should be more accurate and precise (e.g., Williamson & Slatkin 1999; Wang et al. 2016).

The Bayesian approach is distinct in that (1) it can incorporate prior information (e.g., data from previous studies) to compute a probability estimate (i.e., a **posterior probability**), and (2) it directly yields the probability that the hypothesis of interest is true (e.g., H_A: $F_{ST} > 0.00$). (Note: H_A in frequentist statistics is the alternative hypothesis, which is the hypothesis of interest.) Thus, Bayesian statistics more directly tests a hypothesis than frequentist methods that assess how frequently we expect to observe a summary statistic (e.g., $F_{ST} = 0.10$) if the null hypothesis is true. Recall that the null hypothesis is not the direct hypothesis of interest in the frequentist approach, but rather is the hypothesis we test and may reject (Section A4).

Bayesian methods combine a likelihood calculation with prior information to obtain a modified likelihood estimate called the posterior probability (Section A6). Further, Bayesian approaches compute the probability (posterior probability) of the parameter given the data, whereas likelihood computes the probability of the data for a given parameter value in order to find the most likely value (maximum likelihood value). For example, when estimating N_e, the Bayesian approach outputs the probability (posterior) for different N_e values (e.g., for $N_e = 1 - 500$) given the data (Section A6), whereas likelihood finds the parameter values that maximize the probability of the data (Section A5).

Bayesian and likelihood approaches are model-based. It is important to understand model-based approaches because they "open doors for population geneticists and phylogeographers to the repertoire of likelihood-based analyses, including maximum likelihood estimation of model parameters and likelihood-ratio hypothesis tests" (Beaumont et al. 2010, p. 437). Model-based approaches explicitly employ demographic models that include parameters such as population size and migration rates (e.g., Beaumont 1999). Model-based approaches have a goal of computing a likelihood function; that is, the probability of the data as a function of the parameters within a given model. An advantage is that a parameter estimated from two models (stable population versus declining population) can be compared to infer which model best fits your empirical data to determine whether your population is declining.

We will return to model-based Bayesian and likelihood methods again, after considering the important concepts of probability, statistical distributions, and hypothesis testing. Such concepts will help explain the different methods of statistical inference and modeling.

A2 Probability

Probability was defined in 1812 by a French mathematician, Pierre Simon Laplace, as a value between 0 and 1 that measures the certainty of some event. A probability of 1.0 means the event is 100% certain to occur. An example is the probability of the A_1 allele being transmitted into a gamete by an A_1A_2 heterozygote; this probability is 0.5 (Table 6.1). Probability concepts, including probability distributions such as posterior distributions, are important in statistics for using samples from a population to make inferences about the population, based on the sample characteristics (Section A3).

Two important probability rules that we often use in genetics are the addition and the product rule. The **addition rule** is illustrated in Box 5.1. The addition rule (also known as the **sum rule**) is the probability of two or more mutually exclusive events occurring, which equals the sum of the separate probabilities of each event (Box 5.1). In conservation genetics, we often study the probability of mutually exclusive events, such as an individual originating from either population *X* or population *Y*. The sum of mutually exclusive events adds to one (1.0). For example, using Bayesian assignment tests (e.g., Sections 9.9.4 and 22.4), the estimated probability of an individual (multilocus genotype) originating from population *X*, versus *Y* or *Z*, might be 0.00, 0.01, or 0.99, respectively, all of which sum to a total probability of 1.0.

The **product rule** says that the probability of two independent events occurring simultaneously is equal to the product of the probabilities of the two events. The product rule is illustrated as follows: the probability that a heterozygous parent will transmit both the *A* allele at a locus (*Aa*) and the *B* allele at another locus (*Bb*) is (0.50)(0.50) = 0.25, assuming independent loci. For an example application, consider a wildlife forensics case where the four-locus genotype from a bloodstain is *Aa/Bb/CC/dd*. What is the probability of randomly sampling a second individual with an identical genotype from this population, if the genotype frequencies are as follows: *Aa* = 0.25, *Bb* = 0.50, *CC* = 0.10, and *dd* = 0.10? Using the product rule and assuming gametic equilibrium: $P(AaBbCCdd) = (0.25)(0.50)(0.10)(0.10) = 0.00125$.

The probability of an event can be estimated from a large number of observations (e.g., flipping a coin hundreds of times and computing the long-term frequency of heads versus tails). This is called an **empirical probability** because it is obtained through empirical observations. This conceptual framework involving repeated events and their long-run frequency is known as the frequentist approach to probability and statistics.

The above concepts of probability are "objective probabilities." That is, there is no subjectivity, best guess,

or intuition involved in computing the probability. For example, we know from Mendel's laws that each allele at a locus generally has an equal probability of being transmitted (e.g., a 50% chance). Furthermore, if we did not know the probability (0.50), we could empirically estimate the probability via repeated observations (e.g., repeated transmissions of alleles).

A disadvantage of this frequentist approach is that it generally cannot give probability estimates for rare or infrequent events. Further, frequentist probability estimates cannot incorporate common sense or prior knowledge because the estimates are based only on a sample. For example, if you flip a coin 10 times and obtain only three heads, your probability estimate will be 0.30. However, prior knowledge that unfair coins are rare would lead us to suspect that the estimate of 0.30 is too low (and should be close to 0.50). In this case, a more subjective approach to estimating probability could be used to incorporate all available information, and thereby obtain an estimate closer to 0.50.

Subjective probability is an important concept because it facilitates an alternative approach for describing probabilities. It can take into account previous knowledge, data, or best guesses. For example, when computing the probability of extinction for a certain population, we can use input parameters in a population viability model (e.g., VORTEX, Section 18.9), which include best guesses or intuitive predictions. When modeling population viability and the cost of inbreeding on population growth, we might use the average cost measured across mammals in captivity, if no data exist for our particular mammal species. The average cost of inbreeding is approximately a 30% reduction in juvenile survival of progeny by matings between full sibs (F = 0.25) for mammals in captivity (Section 17.4). This best guess of the cost is a somewhat subjective probability if we do not measure the cost in the actual species and population being studied.

Another example of a subjective probability is estimating the probability of a 1°C temperature increase due to global warming. This type of computation is often conducted using a somewhat subjective model and parameter values; for example, including uncertainties inherent in the feedback processes that must be included in climate models.

Subjective probabilities are used in the Bayesian statistical approach (Section A6), which uses Bayes' theorem to incorporate prior information. The Bayesian approach uses a modifiable (or relativist) view of probability by using **prior probability** estimates (from prior knowledge) and then updating them with new data (from new observations) to give an improved posterior probability estimate.

A2.1 Joint and conditional probabilities

We often must compute the probability of two events (E) occurring at the same time. This leads us to consider joint and conditional probabilities. For example, in order for inbreeding to increase the risk of population extinction, it is necessary that inbreeding reduces individual fitness ($E1$ = inbreeding depression) and that the reduced individual fitness leads to reduced population performance ($E2$ = reduced population growth rate). Here, $P(E1 \text{ and } E2)$ is the joint probability of $E1$ and $E2$. Joint probabilities are important in the modeling of complex processes (e.g., Bayesian inference of processes) that have multiple sources of variation; for example, allele frequency changes are influenced by multiple sources of variation such as drift, selection, and migration (Beaumont & Rannala 2004).

A conditional probability is the probability of an event given that another event has happened. Conditional probabilities are used whenever considering events that are not independent. For example, if the effect of inbreeding on fitness increases with environmental stress, then we could compute the probability of inbreeding depression conditional upon a certain stress such as temperature change (resulting from global warming or an unusually hot summer). A conditional probability, the probability of $E2$ given $E1$ (i.e., conditioned on $E1$), is defined as follows:

$$P(E2|E1) = \frac{P(E1\&E2)}{P(E1)} \tag{A1}$$

Note that conditioning on an independent event does not change the probability of the event: $P(E1 \mid E2) = P(E1)$. Note that the "&" symbol means that $E1$ and $E2$ both occur.

Bayes' theorem is used to obtain a posterior probability conditioned on the data available from a sample. The posterior probability $P(E1 \mid E2)$ uses the prior probability $P(E1)$ conditioned on the event $E2$ (the sample of data). Thus, the Bayesian approach computes revised (updated) estimates of the probability of event $E1$ by conditioning on new data ($E2$), as data become available. A prior probability can be rectangular (flat) and thus uninformative. For example, we could consider that microsatellite mutation rates range from 10^{-2} to 10^{-6}, with all values having an equal probability (a flat probability distribution). Alternatively, we could use a bell-shaped prior probability distribution with a higher probability for mutation rates between 10^{-3} and 10^{-4}, which is consistent with published observations suggesting that microsatellite mutation rates are often near 10^{-3} or 10^{-4} (Section 12.1.2).

A2.2 Odds ratios and LOD scores

Another probability concept important in conservation genetics is that of "odds." The probability of an event can be expressed as the odds of an event. The odds ratio for

an event E is computed as the probability that E will happen divided by the probability that E will not happen. For example, the probability of 0.01 has the odds of 1 to 99 (or 1/99). Odds ratios (also called likelihood odds ratios) are used, for example, in paternity analysis to decide whether one candidate father is more likely than another candidate to be the true father (Marshall et al. 1998).

Odds ratios are also used in assignment tests to decide whether population X is more likely than population Z to be the origin of an individual (Banks & Eichert 2000). For example, we can compute the probability (expected genotype frequency, e.g., $2pq$, for a heterozygote) of a multilocus genotype originating (occurring) in Pop X versus Pop Z. If the logarithm of the ratio of the probabilities is very large (e.g., $\log_{10}\{P(\text{Pop } X)/P(\text{Pop } Z)\}$), we can conclude that Pop X is the origin of the individual. For example, we might decide to assign individuals to Pop X if the log of the odds (LOD) ratio is at least 2.0. In this case, with LOD = 2.0, we expect only 1/100 erroneous assignments where an individual assigned to Pop X actually originates from Pop Z. If the LOD score is 3.0, we expect only one in 1,000 erroneous assignments (e.g., Banks & Eichert 2000).

A3 Statistical measures and distributions

A **statistic** is a single measure of an attribute of a sample (e.g., its arithmetic mean value). A statistic is computed from a sample because the entire population usually cannot be collected or measured, as mentioned above. Five categories or kinds of statistical tests can be delineated based on the questions they address: descriptive statistics, tests for differences, tests for a relationship, multivariate exploratory methods, and estimators of population parameters (Dytham 2011).

A3.1 Types of statistical descriptors or tests

Descriptive statistics are computed to describe and summarize sample data during the initial stages of data analysis, without fitting the data to a probability distribution or model (e.g., the normal distribution or model). Since no probability models are involved, descriptive statistics are not used to test hypotheses or to make testable predictions about the whole population. Nevertheless, computing descriptive statistics is an important part of data analysis that can reveal interesting features in the sample data. Examples of descriptive statistics are the mean and variance, which are described in Section A3.2. Descriptive statistics are often called summary statistics (e.g., H_e, F_{IS}), as in approximate Bayesian computations (e.g., Section A7; Tallmon et al. 2008).

Tests for differences address questions such as: Do populations A and B have different heterozygosities? Here, the null hypothesis is that A and B have the same heterozygosity. Tests for differences can also be used to compare distributions. For example, we might ask if the shape of the distribution of allele frequencies is different in populations A and B, or if the proportion of alleles with a frequency below some threshold is less in population A than is expected in a large, stable (nonbottlenecked) population (Example 6.2; Luikart et al. 1998). There are many statistical tests for differences, including **parametric** and **nonparametric** tests (e.g., *t*-tests and signed-rank tests).

Tests for relationship ask questions like: Is fitness related to heterozygosity? A null hypothesis might be: Heterozygosity is not associated with juvenile survival. Two classes of tests for relationships are correlation and regression. Correlation assesses the degree of association without implying a cause and effect. Regression fits a relationship (e.g., linear or curvilinear) between two variables so that one can be predicted from the other, implying a cause and effect relationship. That said, we must recall that correlation does not prove causation. The effect of inbreeding on fitness traits can be predicted via regression (Section 17.4 and Figure 17.14).

We could imagine a scenario where inbreeding is associated with reduced fitness, but inbreeding is not the direct cause. For example, if individuals from population A are more inbred, but also have poorer nutrition than individuals from population B, a correlation (between populations) for individual growth rate versus inbreeding could be caused by the environment, not genetics. Interactions can complicate the assessment of relationships (e.g., genotype-by-environment interactions; Section 11.1.4). There are many ways to test for correlations, compute regressions, and account for interactions. One use of regressions is to compare many regression models to test for effects of environmental factors on genetic structure, as implemented in the computer program *GESTE* by Gaggiotti et al. (2009). See Underwood et al. (2018) for another application of *GESTE*.

Multivariate exploratory techniques ask questions such as: (1) What patterns exist in the data? (2) Can we assign individuals to groups based on multilocus genotypes? (3) Which factor (e.g., locus) is most informative when assigning individuals to groups? Multivariate exploratory techniques can help identify hypotheses to test. In large datasets with multiple factors (e.g., many loci, morphological or environmental measurements), we might not initially test a specific hypothesis because so many potential hypotheses exist. Exploratory techniques are more appropriate for generating hypotheses than for formally testing them (i.e., they do not yield *P*-values, likelihoods, or probability values). A wide range of statistical approaches exists, such as principal component analysis (PCA), frequency correspondence analysis (FCA), multidimensional scaling (MDS), or cluster analysis (20.4.2).

Informative uses of PCA and potential misinterpretations of PCA outputs are described in Novembre & Stephens (2008).

Statistical estimators infer a population parameter by using data that are related to that parameter. For example, we could infer the effective population size (N_e) from data on the temporal change in allele frequencies between two generations. Change in allele frequencies is influenced by N_e, but might also be influenced by **sampling error**, population structure, demographic status (expanding/declining), and selection or mutation rates. There are different approaches to statistical estimation (using the method of moments, maximum likelihood, Bayesian, and approximate Bayesian methods).

Statistical tests can be divided into two classes: parametric and nonparametric. Parametric statistics assume that the data follow a known distribution—usually the normal distribution. Parametric distributions can be defined completely using very few parameters, such as the mean and variance, in a function or mathematical expression. Parametric statistical tests are generally more powerful than nonparametric tests, and thus are often preferred. An example of a parametric test is the *t*-test, which assumes a normal distribution; it can be used to compare mean heterozygosity from two population samples if the distribution of heterozygosity among loci is similar to the normal distribution (Archie 1985).

Nonparametric statistics require few or no assumptions about the distribution of the data or test statistic. Therefore, nonparametric statistics are called distribution-free tests. Some of them are also called "ranking tests" because they often involve ranking observations to generate an empirical cumulative distribution. These tests are generally less powerful, but more appropriate than parametric tests if the data do not follow a parametric distribution. An important example is the Wilcoxon signed-rank test (a nonparametric version of the *t*-test), which is often used to test for lower mean heterozygosity in one population compared with another population (Luikart et al. 1999), or to test for a population bottleneck (Luikart & Cornuet 1998).

Permutation tests (also called **randomization tests**) can be extremely useful using computer-based randomization that makes no assumptions about distributions. We can use this approach to test for **Hardy–Weinberg (HW) proportions** in Example 5.3, where low expected numbers make a chi-square test problematic. A computer program would randomize genotypes by sampling, or creating, 40 diploid individuals from a pool of 61 copies of the 100 allele and 19 copies of the 80 allele. A chi-square value is then calculated for 1,000 or more of these randomized datasets, and its value compared with the statistic obtained from the observed dataset. The proportion of chi-square values from the randomized datasets that give a value as large as or larger than the observed value provides an unbiased estimate of the probability that the null hypothesis is true.

A3.2 Measures of location and dispersion

In statistics, the population is defined as the totality of the individuals with some characteristic we are studying. The sample is the subset of our observations. We compute sample statistics to infer the population value of a parameter (e.g., the mean). For any trait X, the general formulae for the population mean (μ) and sample mean ($\bar{x}$) are as follows:

$$\mu = \frac{\sum x_i}{N} \tag{A2}$$

$$\bar{x} = \frac{\sum x_i}{n} \tag{A3}$$

Where i is the individual number, and N and n are the population size and sample size, respectively.

The mean is a statistical measure of "central value" or the central location (of a distribution). The **arithmetic mean** is given by Equations A2 and A3. Another kind of mean important in population genetics is the **harmonic mean** (Equation A4), which gives more weight to observations with small values. This equation is used in Chapter 7 (Equation 7.8) to estimate the effective population size (N_e) over many generations (t):

$$\frac{1}{N_e} = \frac{1}{t}\left(\frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \ldots + \frac{1}{N_t}\right) \tag{A4}$$

The harmonic mean is used for computing the multigenerational effective population size from successive N_e estimates from each of several separate generations. An interesting controversy in conservation genetics results from, in part, confusing the arithmetic and harmonic mean when computing the ratio of N_e to N_C. The N_e, averaged across generations, is always computed as a harmonic mean, whereas N_C averaged across generations is often computed as an arithmetic mean (Frankham 1995). The harmonic mean is strongly influenced by low values, causing the (harmonic) mean N_e estimates to be lower than the (arithmetic) mean N_C. Thus, the estimates of N_e/N_C ratios (averaged across generations) can be biased low due to a statistical artifact of using the harmonic mean of N_e but the arithmetic mean of N_C (Section 7.10; Kalinowski & Waples 2002).

The **geometric mean** is also used in population genetics. The geometric mean is an average or mean value that informs us of the central tendency of a set of numbers by using the product of their values with the root of the number of values to be averaged. Thus, the geometric mean of 12 and 3 is the square root of (12×3), which is 6. The

geometric mean is used to quantify individual fitness in variable environments. Haaland et al. (2019) tested how selection for **bet-hedging** can adaptively reduce fitness variation. This study used evolutionary simulation models to test if variance-prone strategies are outperformed by bet-hedging strategies. Bet-hedging is a long-term strategy maximizing the geometric mean fitness of a genotype in variable environments by reducing an individual's fitness variance despite lowering arithmetic mean fitness. In bet-hedging, reproduction across generations is an inherently multiplicative process, so the success of a lineage over time is best estimated by geometric mean fitness across generations (rather than the arithmetic mean). The geometric mean of two (or more) numbers is typically larger than the harmonic mean but smaller than the arithmetic mean.

Other familiar measures of central location are the median and mode. An advantage of the median is that it is less influenced than the mean by the skewness of the distribution of the statistic, so the median is resistant to extremely high or low outlier values. Thus, the median is said to be a relatively robust or resistant measure of central location.

A statistical measure of variability (or dispersion) of points around the mean is the variance. If all points have the same value, there is no dispersion and the variance is zero. If points have only very high and very low values, the variance would be high. The variance is the average of the squared deviations from the mean, and is computed as follows: the mean is subtracted from each observation point, this difference is squared, and finally, the average of the squares is computed. The population variance (σ_x^2) and sample variance (s_x^2) are computed as follows:

$$\sigma_x^2 = \frac{\sum (x-\mu)^2}{N} \tag{A5}$$

$$s_x^2 = \frac{\sum (x-\bar{x})^2}{n-1} \tag{A6}$$

Where n (and N) is the number of sample (and population) observations, as above.

The standard deviation is another important measure of dispersion. It is computed as the square root of the variance ($s_x = \sqrt{s_x^2}$). We take the square root of the variance to avoid having to think in terms of squared measures, which are less interpretable (e.g., it is easier to interpret the height of individuals than the "height squared"). Furthermore, one standard deviation under the normal (bell-shaped) distribution encompasses 68% of the central area, while two standard deviations encompass 95%, and finally three standard deviations contain 99% (99% fall between $\mu \pm 3\sigma$; Section A3.3). Probability distributions such as the normal distribution, and their use for describing dispersion, are discussed more in the next section.

The **standard error** is a measure of the dispersion of a statistic (e.g., the sample mean, $\bar{x}$) computed from a sample. The standard error of the mean (SEM) is the standard deviation of a distribution of means for repeated samples from a population. The SEM should not be confused with the standard deviation, which describes the probability distribution of the underlying population parameter *(μ)* for the dataset. For example, the standard error describes the distribution of the SEM of heterozygosity in a dataset, whereas the standard deviation describes the probability distribution of the **population parametric** heterozygosity in the dataset (Section A3.3 and Example A1). Unfortunately, standard error and standard deviation are often confused or not clearly differentiated in publications.

A3.3 Probability distributions

Probability distributions are crucial to understand because statistical tests and estimators require the use of a probability distribution. Different types of variables (height, temperature, F_{ST}) have different probability distributions (Figure A1).

Probability distributions are generally illustrated graphically as a curve or frequency histogram. The total area under a probability curve is 1.0. The probability of a rare or unusual observation is represented as a small area (e.g., 0.05) in the tail(s) of the distribution. We can obtain an empirical estimate of a probability distribution by plotting the relative frequency (histogram) of occurrence of each observation, for example, the height of each individual, in a sample.

An example probability distribution is that from a common likelihood-based estimator of the parameter N_e (Figure A2b). A widely used Bayesian probability distribution in population genetics is that for the discrete variable k (number of demes or "genetic clusters"), which is computed by the program *STRUCTURE* (Pritchard et al. 2000). This distribution gives the probability of each of the possible values of k ($k = 1, 2, 3 \ldots$) and thus helps us to determine how many populations are represented by our sample of individuals from across a landscape, for example.

The probability distribution of any continuously distributed variable is defined as the probability of a random variable being less than or equal to a particular value $P(X \leq x) = P(x)$. Here, $P(x)$ is called the **probability distribution function** (or cumulative distribution function). The derivative of the probability distribution is called the **probability density function** (PDF). The area under any segment of a PDF curve is the probability of X being in a certain interval. Note that a PDF is the output of Bayesian analyses (posterior distribution), for example, used to estimate the probability of some parameter such as N_e, F_{IS}, or mutation rate. Note that a maximum likelihood estimate

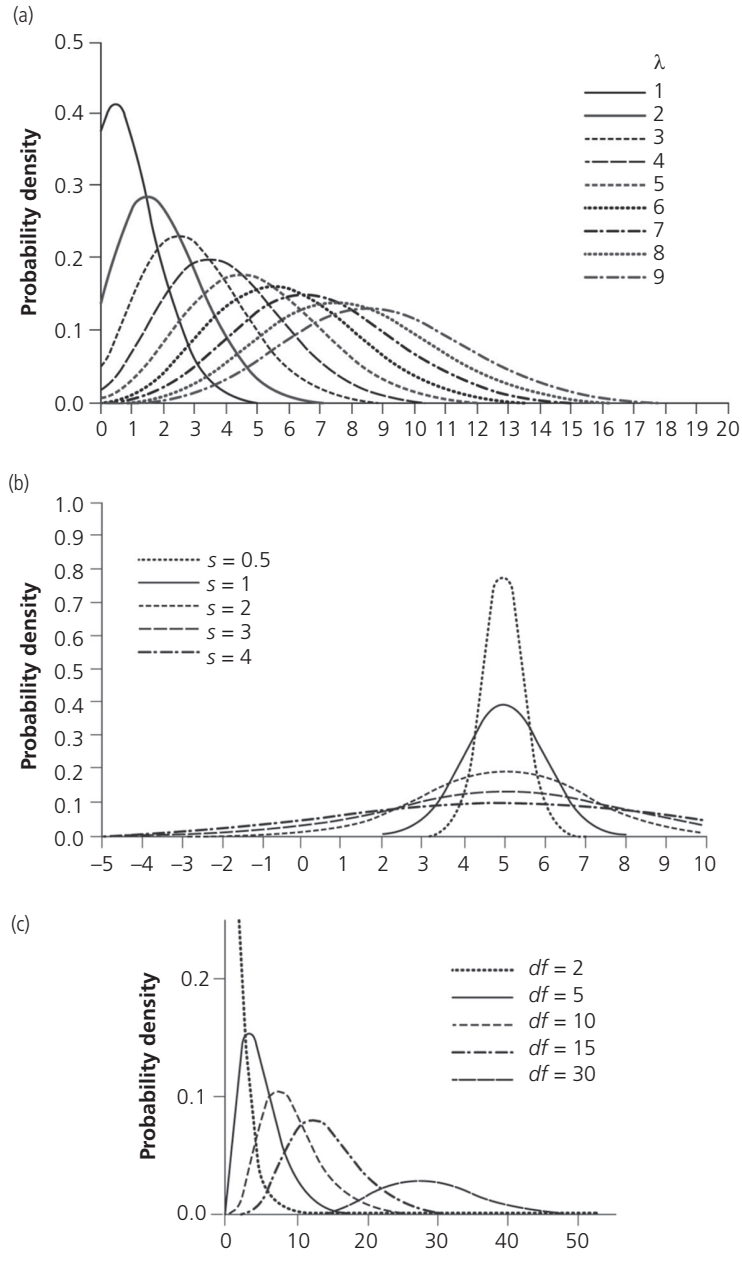

Figure A1 Probability density distributions often used in conservation genetics. (a) The Poisson distribution with a mean (λ) from 1 to 9. (b) Normal (Gaussian) distribution with standard deviation (*s*) from 0.5 to 4, and mean 5. (c) The chi-square distribution (*df* is the degrees of freedom). Modified from P. Bourke (personal communication).

(MLE) is the parameter value that maximizes the probability (density) of the data (Section A.5). The likelihood curve is the density for different parameter values, and thus is not itself a PDF (and does not typically have an area of 1.0).

The population probability distribution can be estimated empirically by computing the cumulative frequencies of observations in a sample; for example, by plotting a histogram of cumulative frequencies of

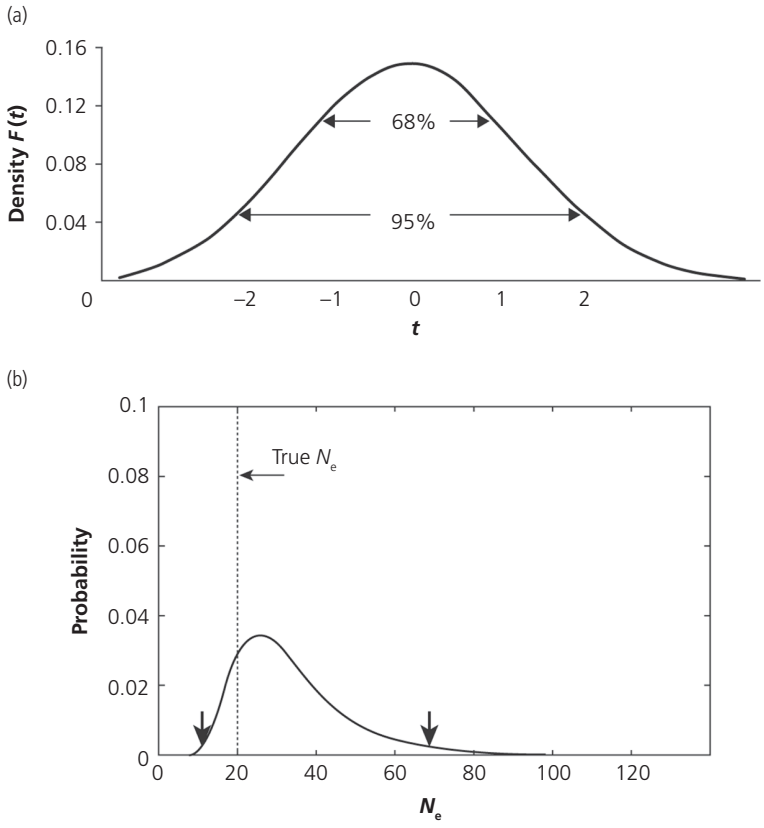

Figure A2 (a) A normal sampling distribution for the statistic *t* showing the upper and lower 68% and 95% confidence limits [tα/2, t1−α/2], where the α (critical/threshold *P*-value) equals 0.32 and 0.05, respectively (see text). (b) Probability distribution (likelihood curve) output from a likelihood-based estimation of N_e, and 95% support limits identified by vertical arrows placing 2.5% of the area in each tail of the distribution. Modified from Berthier et al. (2002).

observations having values less than *x*. The accuracy of the empirical distribution—as an estimate of the population probability distribution—increases with large sample sizes.

A3.3.1 Sampling distributions

Importantly, in frequentist statistics, we use a sampling (probability) distribution of our sample statistic to obtain an estimate of the population parameter. The **sampling distribution** is defined as the distribution of an infinite number of samples of the same size as the sample in your study (Figure A2a). The point here is to realize that our sample is just one of a theoretically infinite number of samples we could have taken. Keeping the sampling distribution in mind, we understand that the statistic we computed from our sample is likely near the center of the sampling distribution, as most of the samples would likely have the statistic near the center.

Readers should not confuse probability distributions of sample statistics (e.g., the mean heterozygosity of a sample) with probability distributions of the underlying population parameter (e.g., heterozygosity computed from the entire population; see the standard error and SEM in Section A3.2). Two important characteristics of sampling distributions are (1) they have lower variance than parameter distributions, simply because each sample mean includes multiple observations (and thus the probability distribution for a mean value is narrower than for individual observations); and (2) they approach the normal distribution when sample sizes are large, which is a surprising principle of the **central limit theorem** as mentioned in Section A3.3.4. The low variance and central limit theorem explain why we often see the normal distribution used for conducting a statistical test and computing confidence intervals.

A3.3.2 Binomial distribution

An important probability distribution in genetics is the binomial distribution. The binomial is one of several theoretical probability distributions used for modeling (approximating) the distribution of observed data that occur in discrete classes, such as genotypes at a locus, as opposed to a continuous distribution of observations, such as height. The binomial is useful for modeling the

proportion of binary events (male versus female births; transmission of allele *A* versus *a*; or survival versus death) that occur in a population sample of size *n*. Note that when more than two events are possible, we can use the multinomial distribution—a simple extension of the binomial.

The binomial distribution contains information on the number of times, *x*, an event with probability π occurs in a fixed number of observations *n*. The binomial distribution is defined as:

$$P(x = m) = \frac{n!}{(n-m)!m!}\pi^m(1-\pi)^{n-m} \qquad \text{(A7)}$$

The factorials in the fraction give the number of ways, *m*, that positive outcomes (transmission of *A*) can occur out of *n* events (offspring). The binomial has a variance of:

$$V(x) = n\pi(1-\pi) \qquad \text{(A8)}$$

For example, if the probability of transmitting the *A* allele is $\pi = 0.50$, then out of 100 transmissions (offspring), we expect a mean of $100 \times 0.50 = 50$ transmissions of the *A* allele, with a variance of $100 \times 0.50 \times 0.50 = 25$ (standard deviation = 5.0). When the number of observations (*n*) becomes large, the binomial approaches the normal distribution.

A3.3.3 Poisson distribution

The Poisson distribution is another discrete distribution that is widely used in conservation genetics and ecology (Figure A1a). The Poisson distribution assumes an event is rare and the events are independent. The Poisson distribution often is used to model events that occur in a spatial or temporal sample. For example, the Poisson distribution is used to model the probability of mutations through time (e.g., under the coalescent, Section A10) because mutations are rare events that arise randomly.

The Poisson distribution is also used to model variance in family size (reproductive success), as in Section 7.3 (Figure 7.4). Ecologists use the Poisson to test whether the distribution of organisms over space is uniform versus random. For example, if the observed variance in distance between individuals is less than the mean distance, then the spacing is more uniform than random because the mean equals the variance in a Poisson distribution.

The Poisson distribution is widely used to model a stable-sized population, with the mean family size (number of offspring per mating pair) equal to two and the variance equal to two (Figure A1a, second curve from the left). This is called the "ideal" **Wright–Fisher population model**, which is the most widely used and among the simplest models for simulating data for conservation genetics applications (e.g., Waples & Faulkner 2009; Hoban et al. 2012). In such an ideal model, the effective population size (N_e; Section 7.1) equals the census size (N_C). However, in natural populations N_e is generally less than N_C because the variance in family size is often high (>2.0; Figure 7.5). Thus, the Poisson distribution is not always the most appropriate distribution for modeling N_e or variance in reproductive success in natural populations (Waples 1989).

Under the Poisson distribution, the probability of any number *x* occurrences is:

$$P(x) = \frac{e^{-\mu}\mu^x}{x!} \qquad \text{(A9)}$$

where μ is the mean number of occurrences, e is the base of the natural log (~2.718), and *x*! is *x* factorial.

A3.3.4 Normal distribution

The normal (Gaussian) distribution is the most widely used continuous distribution: it is the famous symmetric bell-shaped curve (Gauss 1809; Figure A1b). The binomial distribution approaches the normal distribution as sample sizes increase. Thus, for example, the shape of the distribution of the mean heterozygosity (H_e) approaches a smooth bell shape when the sample size of loci approaches 30–50 loci.

The normal distribution is useful for modeling many observed variables (e.g., heterozygosity) because of the central limit theorem, which states that the distribution of the sample mean will approach the normal distribution as the sample size of observations increases, even if the observed variable itself is not normally distributed!

A3.3.5 Chi-square distribution

The chi-square distribution is another continuous distribution widely used in statistics and in conservation genetics. It is asymmetric, unlike the normal distribution (Figure A1), and ranges from zero to infinity. The chi-square distribution is used to model and conduct tests comparing variance measures; thus, the chi-square probability distribution is used when studying, for example, the spatial variance in allele frequencies (F_{ST}) or the temporal variance in allele frequencies (F_k; Antao et al. 2011). The chi-square distribution can be used to compute confidence intervals around F_{ST} or around N_e estimates that are computed from temporal variance in allele frequencies. Chi-square tests using the chi-square probability distribution are discussed in Section 5.3 (Examples 5.1 and 5.2).

Chi-square tests use numbers (not proportions), and if the "expected number" in any class (e.g., genotypic class) is less than one, we should consider using an exact multinomial test based on the multinomial probability distribution (Section 5.3.1). **Exact tests** (Example 5.3) are performed by determining the exact probabilities of all

possible sample outcomes, and then summing the probabilities of all equal and less probable sample outcomes to obtain the exact probability of the observed outcome.

Interval estimates are usually more useful than point estimates. In fact, without an interval estimate, a point estimate (e.g., mean H_e, F_{ST}, or N_e) is generally of little value. Two kinds of interval estimates often used in conservation genetics are confidence intervals (for frequentist approaches) and support limits or credible intervals (for likelihood-based and Bayesian approaches).

Confidence intervals (CIs) give the range of values within which the true population parameter (e.g., population mean) is likely to occur, with some chosen probability (usually 95% or 99%). Thus, CIs are measures of spread. Publications often report 95% CIs, which should span all but 5% of outcomes from repeated, independent sampling events. Note that error bars (e.g., on histograms) often report ±1 standard errors (±1 SE, or standard deviations of the mean), which represent 68% CIs for normal/Gaussian distributed statistics (Example A1). Note also that 95% CIs are nearly twice as wide as 68% CIs; that is, a 95% CI represents approximately ±2 SE (Figure A2a).

To compute a 95% CI, we choose an **alpha value** (α), typically 0.05. Alpha is the critical threshold *P*-value used for rejecting the null hypothesis (e.g., if $P < 0.05$). For a sample statistic $t(x)$, we can compute a $[(1 - \alpha)\ 100\%]$ confidence interval as CI$[t\alpha/2, t1-\alpha/2]$, with lower and upper confidence limits of $t\alpha$ and $t1-\alpha/2$, respectively (where t_n is the nth quantile of the sampling distribution of the population parameter, T).

Support limits are used in likelihood and Bayesian approaches instead of CIs. Support limits can be computed, like confidence limits, such that the estimated sampling distribution (likelihood or posterior distribution) has cutoff points placing 2.5% of the probability density area in each tail. For an illustration, see Figure A2b. Support limits are generally reported with, and plotted on, a probability curve (likelihood or posterior distribution), which allows visualization of the probability of different outcomes just by "eyeballing" the curve (Figure A2b). This makes interpretation of probability estimates (from probability curves) more straightforward than frequentist CIs (Example A1).

A4 Frequentist hypothesis testing, statistical errors, and power

Hypothesis testing is widely used in many scientific disciplines. It requires a formal statement called the null hypothesis (H_0), followed by a statistical test of the null hypothesis, which assesses the probability of the null being true, by computing a *P*-value or a likelihood (probability) distribution. The null hypothesis is a negative statement that mirrors the alternative hypothesis. For example, a null hypothesis might be: Population *X* is stable or growing. The alternative hypothesis is: Population *X* is declining. Many funding agencies suggest that a well-written hypothesis should include a "because phrase," such as "population *X* is declining because of inbreeding depression," which helps researchers to develop predictions and focus on the testing of alternative hypotheses.

Errors in rejecting the null hypothesis can arise because we usually have only a small sample from an entire population, and because statistical tests only relate to the probability that the null hypothesis is false. Two kinds of errors, false positive and false negative, are possible when conducting a statistical test. A false positive test result consists of rejecting the null when it is true. A false negative result is failing to reject the null when it is false (Table A2). The choice of the level α thus inevitably involves a compromise between significance and power, and consequently between the false positive and false negative error rates. False positive errors are also called Type I errors. False negatives are called Type II errors.

The lower the *P*-value, the more confident you are that the null hypothesis (H_0) is false. For example, if $P < 0.001$, you expect that in less than one of 1,000 independent experiments you would observe an outcome as unusual as the one observed. A *P*-value of 0.05 is often used in hypothesis testing as the threshold (α value) for rejecting the null hypothesis. When $P = 0.05$, we have five chances in 100 of rejecting the null when it is true (a false positive). The use of 0.05 is arbitrary and other α values can be used (0.10 or 0.01) depending on the importance of avoiding a false positive test result (Type I error).

A low false positive rate (choosing a low critical α value) will increase the false negative (Type II) error rate. Therefore choosing the appropriate α depends on the relative importance of avoiding a false positive versus false negative error. For example, consider the following null hypothesis: Population *X* is stable or growing. An important question is: Would it be more risky to erroneously reject the null (wrongly accept or conclude that population *X* is declining) or to erroneously fail to reject the null (wrongly conclude that population *X* is stable or growing)? If we wrongly conclude the population is stable but it is actually declining (a false negative), it could lead to extinction of the population or species.

In conservation biology, it often is more risky to make a false negative error than to make a false positive error. False positive errors can be the more risky kind of error in other sciences, such as human medicine, where we must not reject the null when it is true. For example, we would not want to reject the following null hypothesis that "medication *X* has side-effects," unless we are highly certain (e.g., $P < 0.001$) the null hypothesis is false and there are no side-effects.

Example A1 Comparison of different types of error bars

Consider a hypothetical study where you discover a brain protein (LDE, language development enzyme) that causes people to speak articulately (Streiner 1996). You think LDE is in higher concentrations in administrators than in other people. You sample 25 administrators and 25 other people (as a control group) and compute the mean and standard deviation (Table A1). You present the data in a bar graph to make them more visually interpretable (Figure A3).

Table A1 Levels of language development enzyme (LDE) in the cerebrospinal fluid of administrators and controls (Streiner 1996).

Group	Number	Mean	SD
Administrators	25	25.83	5.72
Controls	25	17.25	4.36

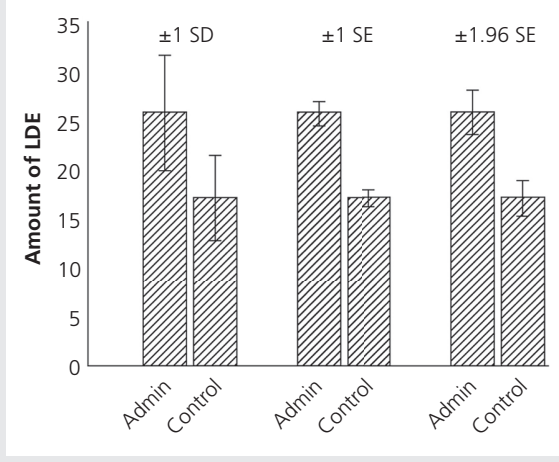

Figure A3 Computing error bars using SDs, SEs (i.e., SDs of the mean), and 95% CI (assuming the normal distribution). Note that ±1.96 SEs represent 95% CIs. Thus, we are 95% confident that the value of language development enzyme (LDE) in the population of administrators is somewhere within the interval shown. This suggests that the amount of LDE is significantly higher in administrators than in the control group. A *t*-test would support this conclusion ($P < 0.05$). From Streiner (1996).

But how do you compute the error bars to extend above and below each histogram bar? In all studies, it is important to report the standard deviation because this shows the dispersion of the actual raw data points. However, the reader generally also wants to know the sample-to-sample variation. For example, if we repeat this study 100 times, how much variation between the means of each study would we expect? Stated another way, how much confidence do we have in the estimation of the population mean from our sample mean? For this, we must compute a standard error (i.e., a standard deviation of the mean).

Should we report one or two standard errors? We are generally interested in a range of values in which we are 95% certain. Thus we could report 2 SEs, which should contain approximately 95% of the study means (Figure A3). Furthermore, 2 SEs are used to compute exact 95% CIs (assuming a normal distribution) when testing for statistically significant differences between population means.

For example, using our table of the normal distribution, we find that 95% of the area falls between −1.96 and +1.96 SEs (SDs of the means, for this example). We compute 95% CIs as follows:

$$95\%\,\mathrm{CI} = \bar{x} \pm (1.96 \times \mathrm{SE})$$

where $\bar{x}$ is the mean.

Of course, ±1.96 SDs of the mean nearly equals ±2 SDs of the mean. CIs show the range in which statistically significant differences exist between means. Showing 95% CIs (or ±2 SEs) supports statistical testing (Section A4) and allows for an "eyeball test" of significance. Note that this eyeball approach does not work accurately when more than two groups are compared because of issues of multiple tests.

Continued

Example A1 *Continued*

How do we interpret the error bar results? If the top of the lower bar (controls) and the bottom of the upper bar (administrators) do not overlap, then the difference between the groups is significant at the 5% level. We could then conclude that administrators have higher concentrations of LDE. Modified from Streiner (1996).

A4.1 One- versus two-tailed tests

Statistical tests can either be one- or two-tailed. In a one-tailed test, the alternative hypothesis (H_A) is a deviation in only one direction (Figure A4), for example, H_A: population *X* is declining. However in a two-tailed test, the alternative hypothesis would be H_A: population *X* is declining or growing (i.e., changing in size; Figure A4a). Thus a two-tailed test checks for deviations in either of two directions. A one-tailed test is appropriate when (1) biological evidence suggests a deviation in one direction (e.g., a population has declined), so we conduct a one-tailed test for reduced allelic diversity; or (2) we only care about a deviation in one direction. For example, we might use a one-tailed test to detect reduced heterozygosity in a population that recently became isolated, if we care only about detecting a reduction in heterozygosity.

One-tailed tests generally have more power than two-tailed tests. Thus it is important to understand the difference between one- and two-tailed tests, and to use one-tailed tests when possible and appropriate. A one-tailed test (e.g., *t*-test) is more powerful because more of the "rejection region" (all 5%, not just 2.5%) is located in the one tail that we are interested in, making it easier to reject the null hypothesis (Figure A4 panel (a) versus panel (b)).

A4.2 Statistical power

An important consideration when choosing a statistical approach is its statistical power (i.e., the probability of detecting an effect when it occurs). For example, the power of a statistical test for detecting a population decline is important in conservation genetics (e.g., Tallmon et al. 2012; Luikart et al. 2021). Power is also defined as the probability of rejecting the null hypothesis (H_0) when it is false.

Power is related to the false negative error rate as follows: Power = 1 – β, where β is the false negative error rate). Thus, the power of a test depends on the choice of β, such that choosing a small β leads to more power. Other factors that influence power, besides β, are the effect size (strength of the effect, e.g., severity of population decline) and the sample size (e.g., number of individuals and loci sampled).

Power is also influenced by the chosen statistical test itself. For example, we mentioned that parametric tests (*t*-test for loss of heterozygosity) are expected to be more powerful than nonparametric tests. A relevant example for conservation genetics is that the most powerful test for detecting a decline in heterozygosity is not the standard *t*-test, but rather a paired *t*-test. The paired test is more powerful because it treats each locus individually and thereby reduces the influence of interlocus variation that often is high. For example, different loci in a sample might have heterozygosity (H_e) ranging from 0.2–0.8, but the between-sample difference in mean H_e that we are testing might be only 0.6 versus 0.5 (e.g., in a large versus a small population). Interestingly, Wilcoxon's nonparametric test often does not have less power than the parametric (paired) *t*-test when monitoring for loss of heterozygosity using two temporally spaced samples (Luikart et al. 1998).

A statistical power of 0.80 is often considered by statisticians as reasonably high power for detecting the event of interest (e.g., population decline, migration, fragmentation) making it worth conducting the study of interest. A problem in science, and particularly in conservation biology, is the failure of researchers to compute the power of statistical tests. Fortunately, power analyses are becoming easier to conduct, thanks to the increasing availability of computer simulation programs that allow simulation of different population scenarios (e.g., population declines) and marker numbers and types.

Ryman and his colleagues have provided a valuable discussion of statistical power when asking the straightforward question "Are all these samples drawn

Table A2 False positive (Type I) and false negative (Type II) errors that can result when testing a null hypothesis (H_0).

	Do not reject H_0	**Reject H_0**
H_0 True	Correct	False positive error
H_0 False	False negative error	Correct

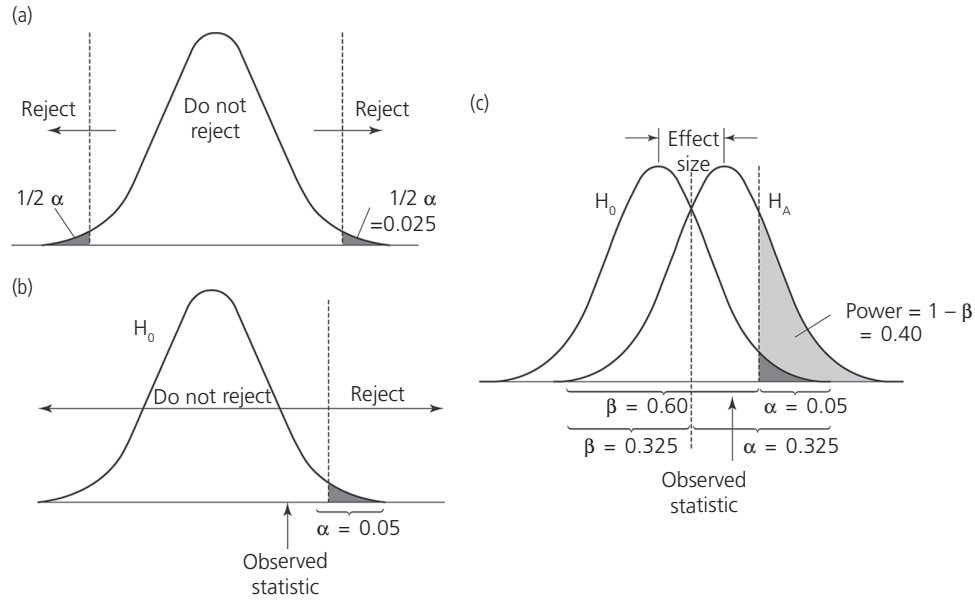

Figure A4 Two-tailed test (a) in contrast to a one-tailed test (b). A two-tailed test is appropriate when we do not know the expected direction of deviation. Panel (b) shows the conventional $P < 0.05$ (alpha = $\alpha = 0.05$) as a threshold to reject the null hypothesis, whereas panel (c) shows a more balanced approach of choosing an alpha value leading to similar risk of false positive (Type I) versus false negative (Type II) errors. Note that the risk of a false negative error (beta, β) is 0.60 when alpha is 0.05. However, if we choose an alpha of 0.325, beta will also be 0.325. Further, note that the "observed statistic" does not fall in the tail (right side of the vertical dotted line for $\alpha = 0.5$) in panel (b) ($P > 0.05$), so we would not reject the null hypothesis. However, in panel (c), we would reject the null because the statistic is smaller than the threshold of rejection, α ($P < 0.325$). Modified from Taylor & Dizon (1999).

from the same population?" (Ryman & Palm 2006; Ryman et al. 2006). They discuss the assessment of power, the use of simulations, and comparison of the performance of different statistical approaches using genotypic data.

A4.3 Problems with *P*-values

A problem with *P*-values and hypothesis testing via the frequentist approach is that *P*-values can be more difficult to interpret than Bayesian posterior probabilities (Section A6). "*P*-values do not measure the probability that the studied hypothesis is true" (Wasserstein & Lazar 2016, p. 131). A *P*-value should be interpreted as the chance that if the null hypothesis is true, you will get a similar or more extreme result if you repeat an experiment many times. A value of $P < 0.05$ is sometimes misinterpreted to mean that there is a 95% probability that the alternative hypothesis is true. This is different from the actual definition, given in the previous sentence.

P-values can overstate the strength of evidence, compared with Bayesian approaches. For example, Malakoff (1999) found that a statistically significant increase in acid rain pollutants detected in some lakes by frequentist analyses disappeared upon a Bayesian reexamination. As explained in Gaggiotti (2010), the reason for this is that hypothesis testing based on *P*-values is appropriate only if the effects of all the various factors that influence the final result are minimized by randomization. This is impossible in typical ecological studies that rely on observational data. Thus, the rejection of the null hypothesis can be highly significant but does not necessarily mean that the alternative hypothesis can be considered as plausible. Bayesian approaches, on the other hand, can incorporate the uncertainty due to various factors, which can strongly reduce the support for the rejection of the null hypothesis.

Another problem with *P*-values often arises when the *P*-value is low, but not significant. If $P = 0.06$, researchers might not reject the null, and subsequently conclude there is no effect, for example, no evidence the population is declining. However, as mentioned earlier, the choice of $\alpha = 0.05$ is often arbitrary, and in fact, $P = 0.06$ is suggestive of an effect (especially if the power of the test is low). Recall that if the effect size is small, we are unlikely to obtain a significant *P*-value (e.g., $P < 0.05$), unless sample sizes are very large and power is high (Section A4.2).

Another problem with *P*-values is that negative results ($P > 0.05$) are sometimes difficult to publish, and can lead

to a bias in the scientific literature, with an underrepresentation of studies that find no significant effect. For example, out of all the studies done on the correlation between heterozygosity and fitness, a greater proportion of those finding a significant correlation may be published than those not finding a correlation. This potential lack of publication of negative results has been called the "file drawer effect," because negative results often end up in a file drawer, unpublished.

Solutions to *P*-value problems were described by Halsey (2019). First, we can supplement or replace *P*-values with the **Bayes factor** to quantify strength of evidence for the null and alternative hypotheses. This approach is especially useful when you keep collecting data to recompute evidence for or against your hypothesis. Second, we can supplement *P*-values with estimates of confidence in the *P*-value and the probability that a result is not a false positive. Third, we can supplement the *P*-value by reporting effect sizes and confidence in the effect size estimate. Finally, The American Statistical Association provided a "Statement on Statistical Significance and *P*-values" with six principles underlying the proper use and interpretation of *P*-values (Wasserstein & Lazar 2016). This section has discussed some of these six principles.

A5 Maximum likelihood

Likelihood is the probability of observing the data given some parameter value (e.g., $mN = 50$), under a certain statistical model (e.g., island model of migration). Maximum likelihood (ML) methods estimate the parameter value that maximizes the probability of obtaining the observed data under a given model. For example, we might compute the likelihood of each of many effective population sizes (N_e = 10, 11, 12 . . . up to 500), and then choose the best point estimate of N_e as the value that has the highest (maximum) likelihood (e.g., $N_e = 25$ in Figure A2b).

An advantage of likelihood analysis is that it is model-based and thus allows easy comparison of different models (even complex models), thereby improving inference about complex processes (e.g., different dispersal patterns, mutation models, stable versus declining population size) that might explain the data. Likelihood analysis is often used to test the fit of two different models by using the ratio of the MLE for one versus the other model. For example, if the first model is far more likely than a second one (e.g., $\log_{10}(\text{MLE1}/\text{MLE2}) > 3$), we might reject the second model MLE2 (Section A2.2). The two models might be, for example, a stable versus declining population, or alternatively, the existence of two versus three subpopulations. Note that when $\log_{10}(\text{MLE1}/\text{MLE2}) > 3$, the probability of MLE1 is generally 1,000 times more likely that MLE2 (e.g., $P < 0.001$); when >2 the probability of MLE1 is considered to be 100 times more likely that MLE2 ($P < 0.01$).

Likelihood methods are sometimes classified as frequentist methods. For example, when we compute the expected frequency in a large number of trials of a likelihood ratio (or a likelihood value); for example, as part of a statistical test, this is a frequentist approach.

The main advantage of ML approaches is that they use all the data in their raw form, and not some summary statistic (e.g., $\hat{H}_e$ or $\hat{F}_{IS}$). Because likelihood methods use a maximum of information from the data, they should, in theory, be more accurate and precise than moments-based methods (Luikart & England 1999). For example, likelihood-based methods use the raw data (number and genealogical divergence of each allele) to estimate N_e (or mN), and not a single summary statistic (e.g., H_e), as in classical moments-based estimators of N_e (or mN; see Miller & Waits 2003).

Different datasets can give the same summary statistic (e.g., F_{ST}), whereas different datasets are less likely to yield the same ML estimates. For example, two independent sets of temporally spaced samples can have the same temporal F_{ST} even though they have different numbers of alleles. When using the summary statistic temporal F_{ST} to estimate N_e, we would not be using the information about rare alleles, and thus might not achieve the most accurate or precise estimate of N_e. In another example, two independent metapopulations could have the same F_{ST}, but have different proportions of rare alleles. Information about the proportions of rare alleles can help to infer whether a metapopulation is stable, fragmenting, or growing in size (e.g., Ciofi et al. 1999).

In practice, ML methods are often more accurate and precise than moment-based methods. For example, estimators of N_e based on likelihood provide tighter CIs and less biased point estimates (Williamson & Slatkin 1999; Berthier et al. 2002). A likelihood version of the temporal N_e-estimation method outperforms a method-of-moments estimator by reducing the bias caused by rare alleles, by effectively down-weighting contributions of rare alleles. A problem with rare alleles is they can be lost between sampling periods and thereby bias the estimator (Luikart et al. 1999). A possible disadvantage of likelihood-based estimators is they generally require large sample sizes and can be biased and less precise than simpler summary statistics (moments-based methods), if sample sizes consist of fewer than 30 to 50 individuals (e.g., see Lynch & Ritland 1999).

A6 Bayesian approaches and Markov chain Monte Carlo

Bayesian inference differs from classical frequentist statistics in two main ways. First, probabilities are defined and

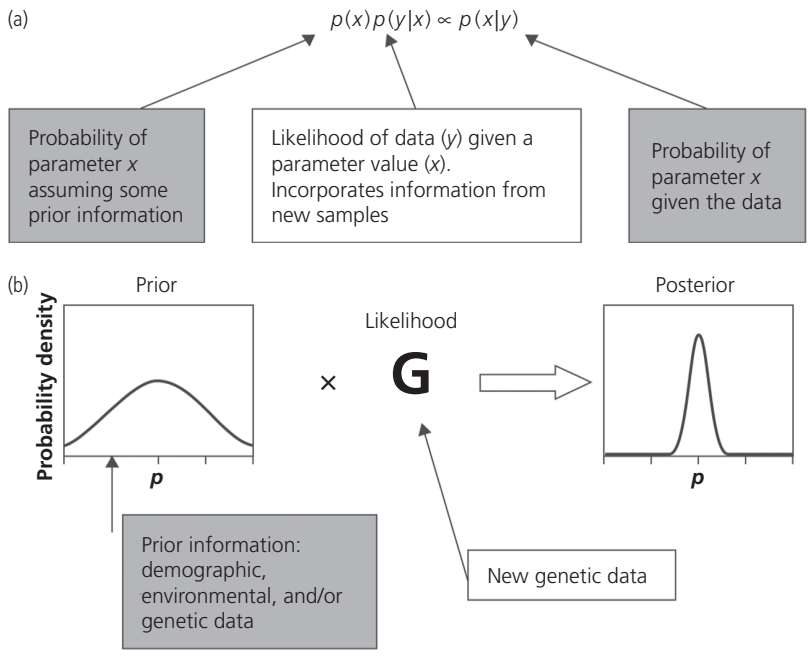

Figure A5 (a) Simplified Bayesian mathematical expression showing how the Bayesian approach allows us to combine the information from the data with prior information about the parameters of the model in order to obtain their posterior distribution (to estimate a parameter). (b) Graphical illustration of how prior information (in the prior probability distribution) is modified by the multiplication of it by the likelihood function (from the standard likelihood-based approach) to obtain a posterior probability distribution. Modified from O. Gaggiotti (personal communication).

interpreted differently. In frequentist statistics, *P*-values are interpreted as the average outcome of a repeated experiment in a large number of trials. *P*-values are interpreted as the probability of the test statistic being as extreme (or more extreme) than observed, if the null hypothesis is true. A frequentist test might yield $P = 0.05$, meaning there is a 5% chance of observing the test statistic simply by chance alone.

Bayesian computations yield a more straightforward and direct probability answer that is easier to interpret than a *P*-value. For example, a Bayesian (posterior) probability might yield a probability of $P = 0.95$, meaning there is 95% probability that your hypothesis is true (e.g., that N_e is less than 100, for example). Recall that in the more complicated and less direct frequentist approach, we would construct a null hypothesis (e.g., H_0: N_e is greater than or equal to 100), and then reject the null if the *P*-value is low (e.g., $P < 0.05$); thereby finding support for the alternative hypothesis of interest—N_e is less than 100.

Furthermore, Bayesian posterior probability distributions (and support limits) are easier to interpret than CIs because probability distributions show the probability visually as the area under a curve (e.g., in the tails of a probability distribution). We immediately get a feel for the width and degree of skewness of the probability distribution by observing the posterior distribution, which we cannot get from reading confidence limits. Thus, a probability distribution (e.g., posterior probability distribution) carries more information than a classical CI and it gives a better feel for the relative probability of different parameter values (e.g., small versus large N_e, mN, or F_{IS}; Ayres & Balding 1998).

Second, perhaps the main advantage of the Bayesian approach is the ability to include prior data or information when estimating the posterior probability that a hypothesis is correct. Bayes' theorem was developed to allow easy updating of an existing estimation when presented with new data, such as observations from a new experiment. Classical frequentist statistics generally require each experiment to be totally independent and without reference to previous experiments. Prior information (previous data or even a hunch) can be incorporated into the computation of a probability (posterior probability) by multiplying the likelihood function by the prior information (Figure A5; Example A2).

Example A2 The effect of an informative prior in a coin-flipping experiment

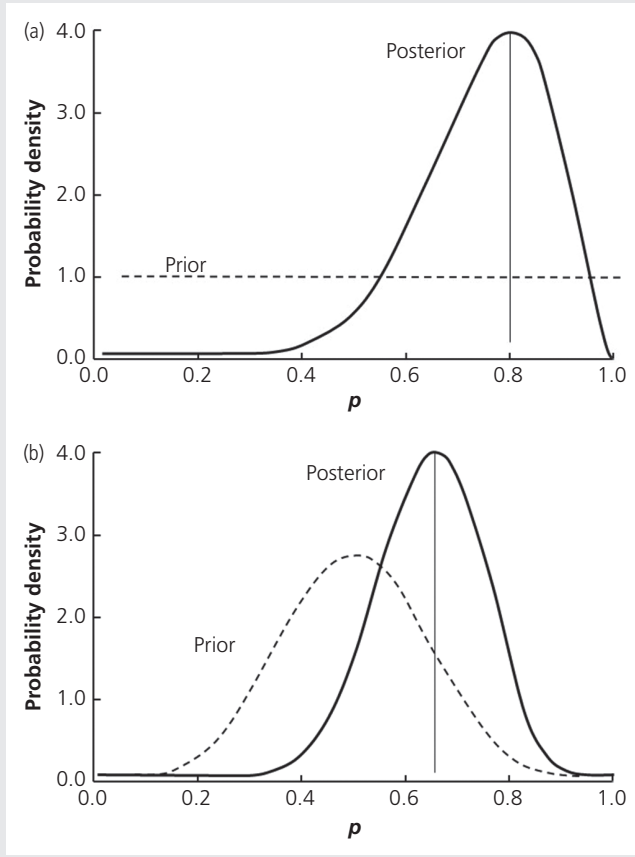

Figure A6 Effect of using an informative prior on the posterior density curves for p (probability of flipping a head) in a Bayesian analysis of a coin-flipping experiment in which 8 heads occur in 10 flips. A noninformative (flat) prior is used in (a), while an informative prior with a mode at $p = 0.50$ is used in (b). The mode for the posterior distribution with the flat prior is at 0.80, whereas the mode for the posterior distribution using the informative prior is 0.65. Thus, the posterior distribution was shifted because of the use of the informative prior with a peak at $p = 0.50$. The probability density value on the y-axis shows the relative probability of observations on the x-axis. Modified from Lewis (2001).

An example of using the Bayesian approach to incorporate prior information is estimating the effective population size (N_e) when the population census size is known (e.g., $N_C = 250$). Here, we can use the prior knowledge of N_C, and knowledge that N_e cannot be more than twice the census size ($N_e \leq 2N_C$; Chapter 7). Thus, the prior probability of N_e being greater than 500 equals zero ($P[N_e > 500] = 0.0$; as in Berthier et al. 2002 or Tallmon et al. 2008). Furthermore, we know that N_e is often less than 50% of N_C (Frankham 1995; Kalinowski & Waples 2002). This information can be used to give more weight to N_e estimates near or below 50% N_C (e.g., using a prior probability distribution).

Another example use of prior information is in models that incorporate mutation dynamics. Published data suggest that most microsatellites have mutation rates between 10^{-2} and 10^{-5}. So, we might use a flat (rectangular) prior ranging between 10^{-2} and 10^{-5}, when modeling humans or other mammals. We also know that the average mutation rate is near 5×10^{-4}. Thus we might use a more informative prior (e.g., bell-shaped rather than flat) with a high probability peak near 5×10^{-4}. For example, Beaumont (1999) used a prior mutation rate greater than zero for monomorphic loci, thereby allowing the use of monomorphic loci when testing for population bottlenecks. Other bottleneck inference tests do not use monomorphic loci (Luikart & Cornuet 1998). See Lewis (2001) for a simple example of Bayesian computation.

The Bayesian approach to incorporating prior information can be especially useful in conservation biology because it facilitates decision-making when data are few and we want to integrate all available knowledge. In conservation biology, we often must make decisions based on limited data. For example, wildlife managers often must decide if the size of a population is large enough to allow harvest, or alternatively, if the population needs protection, monitoring, or supplementation. A US National Academy of Sciences panel recommended that fisheries scientists use Bayesian methods to help estimate fish population status and guide management policies (Malakoff 1999). Harvest quotas could be more appropriate and flexible if the risks of population decline were calculated directly via Bayesian statistics incorporating prior information such as the probability that harvest actions might endanger a stock.

The main criticism of Bayesian approaches is they can be strongly influenced by prior information, and thus be less objective than classical approaches. For example, two different people could use different prior information and obtain different results. A counterargument is that we can quantify the effects of different priors (e.g., via sensitivity analysis using different priors); thus we can (and should) consider the magnitude of influence of the prior when making management decisions. Often prior information has little influence on the posterior, especially if data are extensive. Unfortunately, such sensitivity analyses are not always conducted. It is reasonable to use both Bayesian and classical frequentist methods in certain applications, such as the estimation of F_{ST}, N_e, or mN, especially when the performance of one or both methods has been poorly evaluated (Section A9).

An important general contribution of Bayesian approaches is that they allow for computations using complex models that that could not be achieved using other statistical approaches (Beaumont & Rannala 2004). Bayesian computation using complex models has been greatly facilitated by MCMC computational methods.

A6.1 Markov chain Monte Carlo (MCMC)

MCMC is a method for simulating random samples from a probability distribution. Importantly, it is very useful because it can be used in cases where there is no straightforward equation that gives the distribution. This is generally the case, for example, in Bayesian inference where there is typically no analytical formula for the posterior distribution. However, by simulating random samples from this distribution we can then obtain an estimate of the posterior mean and credible interval for a parameter.

MCMC can generate probability distributions difficult to obtain from analytical equations, including likelihood equations (Storz 2002; Wu & Drummond 2011). Analytical equations often cannot be developed to describe complex processes with many variables, such as population size, allele frequencies, and mutation rates. MCMC allows simulation of a special kind of stochastic process known as a **Markov chain**. A Markov chain generates a series of random variables whose future state depends only on the current state at any point in the chain (Beaumont & Rannala 2004).

MCMC allows us to obtain random samples from the sample space, even when the sample space is enormous (e.g., billions of phylogenies or genealogies). MCMC combines: (1) a Markov chain model, that is, a model involving a random walk (chain of random steps) in which the next step is determined by the characteristics of the current or previous step; and (2) the **Monte Carlo process** of drawing a random number that is necessary at each step of the random walk (Monte Carlo is a city famous for gambling, which also uses random events like the rolling of a dice).

MCMC is illustrated by an analogy of a robot taking a random walk in a square field (Figure A7). Each step of the robot varies in length and direction, randomly. Eventually, the robot visits every space within the field. However, the robot spends more time in spaces that are on hilltops at higher elevation (i.e., having higher probability). This

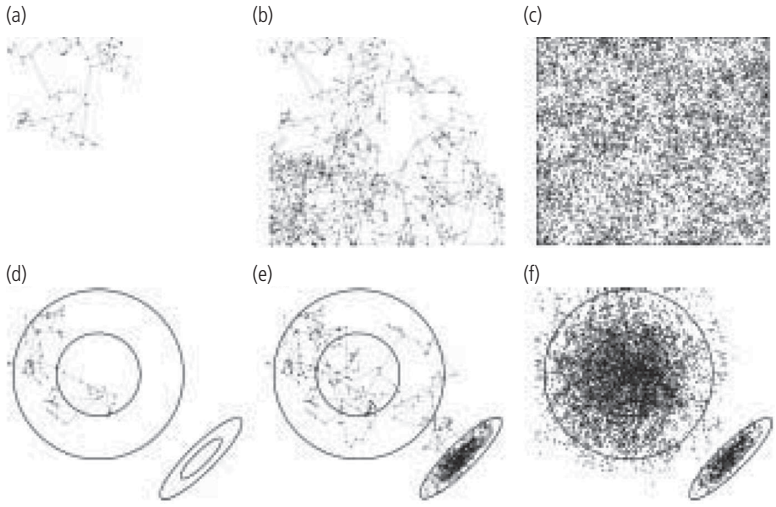

Figure A7 Principles behind MCMC methods using a simple analogy of a "random walk" in a square field (top panels) by a robot. The robot begins its walk in the upper left corner and continues for (a) 100 steps, (b) 1,000 steps, and (c) 10,000 steps until nearly every portion of the field has been covered. Now, supposing that two hills are present, represented by the concentric circles and smaller concentric ovals (d, e, and f). The robot will take steps to points in proportion to their elevation. Thus higher points will be visited more often than lower ones. The proportion of time spent in any place approximates the probability of that location. From Lewis (2001).

is achieved by using a model with the following two main rules: (1) if a step takes the robot uphill, the robot will automatically take it; and (2) if a step would take the robot downhill, the robot only takes the step with a probability depending on the elevation reduction (this probability can be computed several ways, e.g., via Metropolis or **Metropolis–Hastings methods**).

The first few steps or thousands of steps are called the **burn-in**, and are discarded to reduce influence of the starting point (bias). Once burn-in is achieved, the simulation has **converged**; that is, become independent of the starting point (Figure A8). The remaining steps, after convergence, give a good approximation of the landscape (probability space). This simulation of a random walk allows for estimation of the parts of the sample space with the highest probability (e.g., maximizing the probability of the data, given the model), as in ML estimation (Section A5). Under the Bayesian approach, MCMC simulation is often used to sample from the posterior distribution of a parameter in order to generate the posterior probability estimate of the parameter.

A problem with MCMC approaches is we sometimes are not sure we have conducted a long-enough burn-in to achieve convergence and thus avoid bias (Figure A8). To make sure an MCMC-based program has converged and provides consistent results, you can conduct multiple independent runs of the program and test for similarity of results among replicates. Also, MCMC simulation programs are generally difficult to code, and thus errors are relatively likely and can be difficult to detect. In addition, MCMC approaches are computationally slow, making evaluations of performance difficult using numerous simulations for a wide range of scenarios (N_e, mN, etc.; Section A9).

MCMC is primarily used within Bayesian approaches, but can also be used in ML estimation. For example, some available software programs can use flat priors (or no priors) and give as output a likelihood (probability) curve or a posterior distribution if prior information is used (e.g., Beaumont 1999; Beaumont & Wang 2019).

A7 Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation (ABC) employs a Bayesian framework, incorporating prior information, to output an approximate posterior probability distribution (Bertorelle et al. 2010). This posterior distribution is only an estimation of the full posterior because all the raw data (e.g., full allelic distributions) are not used to compute the posterior. Instead, the posterior is usually approximated by summarizing the data using multiple different summary statistics (Tallmon et al. 2008; Sousa et al. 2009). For example, a full (exact) Bayesian approach might conduct MCMC simulations to obtain the exact posterior probability of the raw sample (allele number and frequency distribution), using each of thousands of simulated datasets (e.g., genealogies, for a population model under consideration, such as a stable population). Here, for example, we might consider population models (and simulations) with $N_e = 10, 20, 30$, and so on, if we were estimating the N_e for our observed data.

Unlike exact Bayesian computation, ABC would (1) replace (summarize) the raw observed data with multiple summary statistics of the data such as F_{ST}, H_e, and H_o, and then (2) compute the same summary statistics for

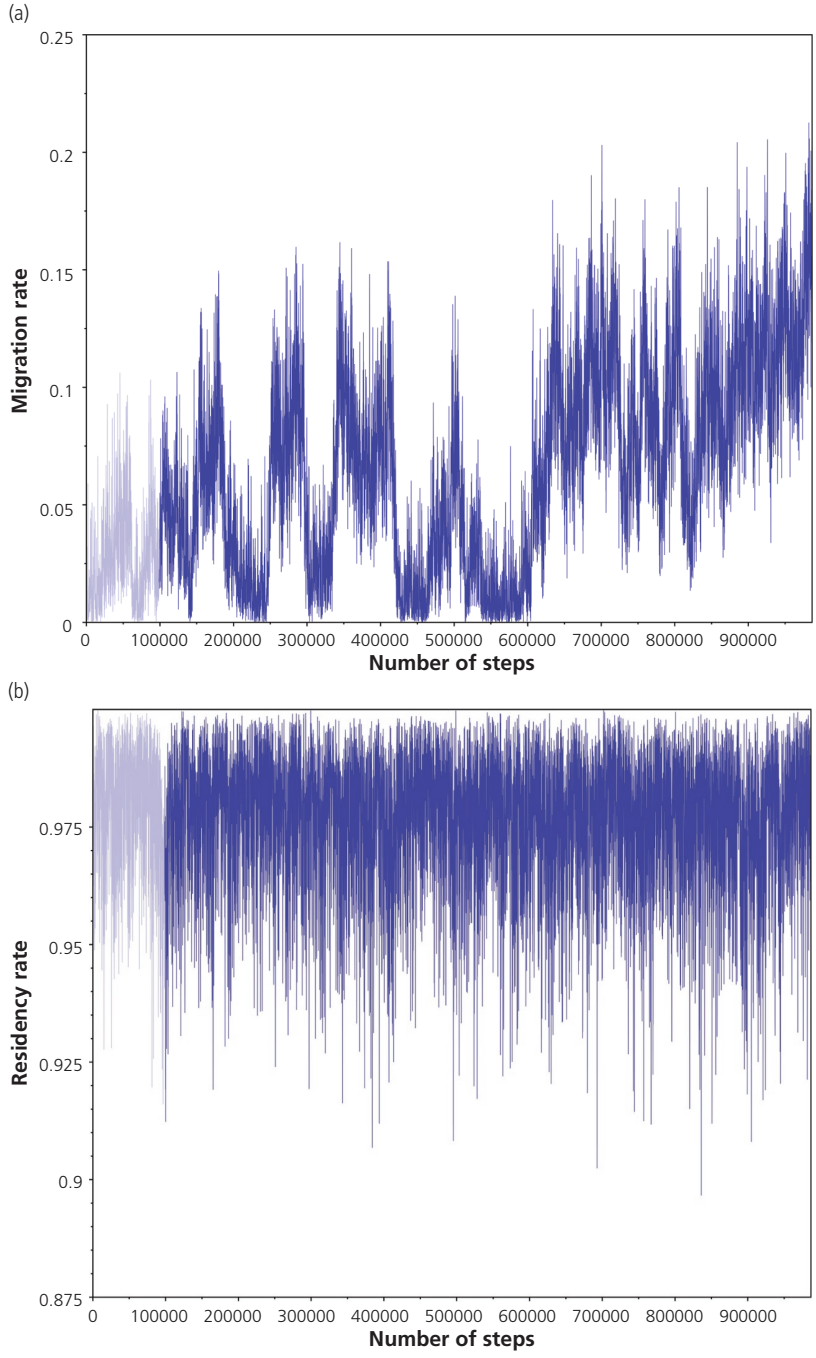

Figure A8 Trace files showing (a) lack of convergence, and (b) convergence of MCMC random walks to estimate the amount of exchange among a set of three subpopulations during runs of program *BayesAss.* (a) The *y*-axis shows the proportion of individuals in subpopulation 2 that migrated in from subpopulation 1. (b) The *y*-axis shows the residency rate of subpopulation 3 (i.e., the proportion of individuals that did not migrate out to other subpopulations). The burn-in period is the first 100,000 steps (*x*-axis) on the left side of each panel in lighter-colored blue-gray trace. To achieve convergence, we can often run the burn-in period longer. Figure made in program *Tracer* using output from *BayesAss* v3.0 trace files and microsatellite data from bluehead sucker. Courtesy of Steven Mussmann.

each of thousands of simulated population datasets (for the population models under consideration), and finally (3) match the observed data summary statistics to those from simulated populations to choose the population parameter (N_e) estimate that best fits our data. This ABC approach is also called "summary statistic matching" because we match our empirically observed summary statistics (F_{ST}, H_e, and H_o) to those from simulated population datasets (each with a different effective size, e.g., N_e = 10, 20, 30) to find the parameter estimate (e.g., N_e) that is most probable for our population, according to the matching of our empirical and simulated summary statistics.

ABC methods have become popular because they use nearly all the information from the data (Beaumont et al. 2002), yet they are usually far less computationally demanding than fully Bayesian (MCMC) approaches. Thus, their performance can be evaluated thoroughly (Section A9), and they can be used with large datasets containing many loci or, when conducting complex analyses, with numerous parameters such as the population size, migration rate, and sex ratio. Finally, an experienced modeler can construct an ABC model in hours or days (e.g., Cornuet et al. 2008; Wegmann et al. 2010), whereas it can take weeks to construct a fully Bayesian MCMC model and the risks of programming errors can be far higher (M. Beaumont, personal communication).

Advantages of ABC are that they allow (1) comparisons of most demographic scenarios that can be simulated, and (2) estimation of key parameters of the model such as the bottleneck minimum effective size (Stoffel et al. 2018). ABC also allows incorporation of uncertainty in model specification by defining priors. These attributes have made the coalescent-based ABC method a preferred approach for inferring population bottlenecks and demographic histories. A possible disadvantage of using ABC is that "translating genome-wide observed data into a set of summary statistic values that are readily useable by ABC programs can be a challenge" (Elleouet & Aitken 2018, p. 537). Another disadvantage is we often cannot estimate many demographic parameters with accuracy in a model.

ABC methods have been replaced in some situations for NGS data analysis because it can be slow to simulate large sequence datasets even with efficient coalescent methods (Elleouet & Aitken 2018). Other methods based on **site frequency spectra** (SFS) such as *fastsimcoal* are flexible and relatively quick to use for demographic modeling with large genomic datasets (Excoffier et al. 2013). Nonetheless, ABC remains useful for investigating simple demographic scenarios with small to moderate N_e.

A8 Parameter estimation, accuracy, and precision

Here we consider statistical frameworks (moments, likelihood, Bayesian) for inferring population parameters. To estimate a population parameter, such as the mean (μ), we usually compute a sample statistic ($\bar{x}$) from a sample of individuals. We can estimate a population parameter using different sample statistics such as the arithmetic mean, harmonic mean, median, or mode. To further complicate things, to compute an estimator, such as the mode, we can use different approaches, including moment methods, ML estimation, or Bayesian estimation.

The sample moments, (e.g., $\bar{x}$, $\bar{x}^2$, and $\bar{x}^3$) are used to obtain estimates of location, variance (scale), and shape of the population distribution, respectively. Moment-based estimators are widely used (e.g., in classic frequentist statistics), but can yield biased estimates when the underlying population distribution is nonnormal, especially when the higher moments ($\bar{x}^2$, $\bar{x}^3$) are not considered. An example of such bias is the classical F_{ST}-based estimator of N_e, which is often a biased estimator because: (1) the underlying probability distribution of F_{ST} is often skewed with a long tail (unlike the normal distribution); and (2) the moment estimator (F_{ST}) incorporates information only from the first two moments, and does not contain information on skewness (i.e., shape) of the sampling distribution.

MLE infers a parameter by finding the parameter value that maximizes the likelihood of obtaining the sample data (assuming some model such as Mendelian inheritance or a Wright–Fisher equilibrium population). MLE is increasingly used in population genetics for several reasons:

1. It yields probability distributions that are easy to interpret (Section A5 and Figure A2b, A5b), rather than just a point estimate and confidence interval, as in moments methods.
2. MLE can help to evaluate and choose the best estimators (including moment-based estimators of the mean or variance, when data are normally distributed).
3. Faster computers and software programs increasingly allow the computation of MLE estimates (*LAMARK*, *MIGRATE*, *MSVAR*, *fastsimcoal*, *dadi*, and other computer programs).

Which estimator and approach perform best? This is a critical question in conservation genetics that is often ignored or underappreciated. It is especially important in light of the many new methods and computer programs published in recent years. The performance of an estimator (accuracy, precision, and robustness; Section A9) depends on the question, sample size, sample characteristics,

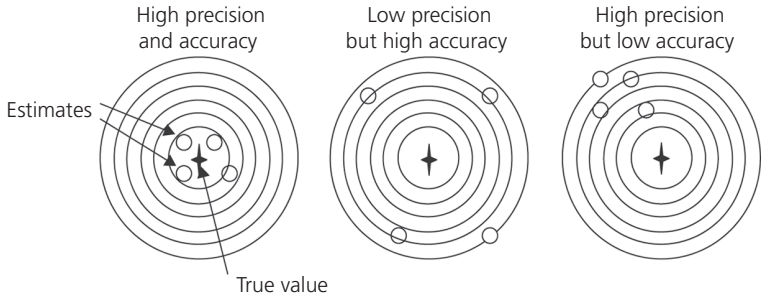

Figure A9 Illustration of the difference between accuracy and precision. Imagine these are archery targets with the bullseye (true value) in the middle.

the parameter being estimated, and the effect size. For example, MLE approaches are generally most efficient (Section A5) with large samples, but can be less efficient than moment methods when using small sample sizes, for example, fewer than 40 individuals. Efficiency refers to the ability to extract information from the data and to achieve high accuracy and precision in estimating the true population parameter.

Identifying the best estimator generally requires a **performance evaluation** comparing estimators. For examples of performance evaluations, see Section A9 and Tallmon et al. (2004a) and Wang (2002).

Accuracy and precision are crucial concepts related to estimators of central tendency and dispersion, respectively. Accuracy of an estimator is its tendency to yield estimates near the true population parameter value. For example, if we compute an estimate of the mean heterozygosity (H_e) for each of 10 independent samples, the accuracy is good if 50% of estimates are high and 50% are low. If most of the independent estimates were low (or high) the estimator is considered to be biased.

If an estimator has poor precision, the 10 estimates will scatter widely from each other—often on both sides (e.g., high and low) of the true value. A precise statistical estimator will have relatively narrow CIs, and the point estimates from independent estimations will cluster tightly together. An estimator can have low precision but high accuracy, or vice versa (Figure A9).

Genomics can improve precision and accuracy of estimates of population genetic parameters. For example, thousands of single nucleotide polymorphisms (SNPs) improve estimation of individual inbreeding, allowing detection of inbreeding depression in natural populations when 20–30 microsatellites do not (e.g., Excoffier et al. 2014). Similarly, Kardos et al. (2015a) showed that ~5,000 random SNPs provide more precise estimates of individual inbreeding than 10 generations of complete pedigree data, and that tens of thousands of mapped SNPs further improve precision. These empirical and theoretical studies suggest most of the published literature underestimates inbreeding effects on fitness, and that natural populations likely suffer more from inbreeding depression than previously thought.

Genomics and thousands of loci closely distributed on chromosomes can lead to enormous overestimation of precision. Averaging among loci while assuming independence creates pseudoreplication, thereby reducing the effective degrees of freedom (*df*) compared with the actual *df*. Waples et al. (2021) estimated the effective and actual *df* by quantifying variance of mean F_{ST} and mean r^2 as more loci were used. For r^2, the effective *df* plateaued after a few thousand loci. Pseudoreplication was less extreme for F_{ST}, but occurs when using tens of thousands of loci. Commonly used **jackknife** methods consistently underestimated variance in F_{ST}, producing highly conservative (narrow) CIs.

Several different estimators should often be used whenever assessing a given question. For example, it is useful to estimate both the mean and median because if they are different we can infer that the distribution might be skewed. Simply plotting data and visually inspecting them to identify errors or issues is advisable; for example, if the mean and median are very different or the expected and observed values are different (e.g., Figure 6.12). It is also useful to compute both moment-based and likelihood-based estimators, as we sometimes do not know which is most reliable or accurate. In general, when estimating parameters, it can help to use multiple methods and software programs, to avoid errors and to increase confidence in results.

Random and representative sampling is critical and often assumed without testing or evaluating the implications of the assumption. If sampling is not random or not representative, the statistical estimate can be biased. For an extreme example, imagine that we sample only 10 individuals from within only one family from a population containing hundreds of family groups. The sample is not random or representative of the population. The allelic richness statistic we estimate often will be low compared with the true population value, simply because the

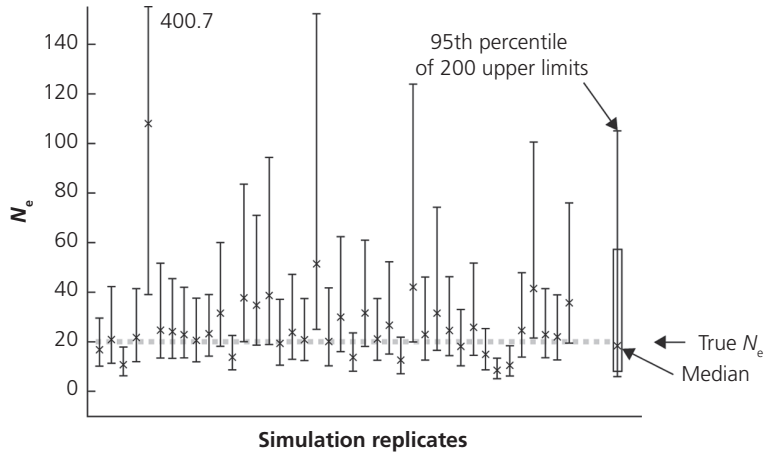

Figure A10 Example of power analysis showing 40 N_e estimates for each of 40 independent population simulation replicates based on allele frequency change over 5 generations at five loci. Point estimates ("X") of N_e, with CIs (vertical lines), for each of 40 independent simulated populations are shown. The box plot graph on the far right summarizes the accuracy of point estimates by comparing the median of the many point estimates with the true N_e (dashed horizontal line). The median is slightly biased low (arrow). The box plot upper limit is the upper 95th percentile of the upper confidence interval limits (over 200 simulations). Modified from Berthier et al. (2002).

individuals we sampled are closely related compared with individuals from a true population-wide sample (with random representation of all family groups).

A9 Performance evaluation

Performance testing is the quantification of the accuracy, precision, power, and robustness of a statistical estimator or test. This includes quantifying the bias caused by violating assumptions (random sampling, no selection, no population subdivision, independent loci, etc.); such violations often occur in real datasets for natural populations. Estimators and tests are often robust to certain violations of assumptions, but we should quantify robustness before using an estimator, test, or computational method for conservation.

Performance testing involves four main steps: (1) generate a test dataset (simulated or real) with a known parameter value for the parameter of interest (N_e, mN, etc.); (2) estimate the parameter (e.g., with a confidence interval); (3) repeat both steps 1 and 2, 1,000 times; and (4) compute the proportion of the 1,000 estimates that give the true parameter most accurately and precisely (Figure A10). This testing can include simulation of datasets for scenarios where assumptions of a method are violated (e.g., selection, or linkage between some loci).

Performance testing is important to allow conservation biologists to use statistical methods on real populations with minimal risk of making erroneous management decisions. Unfortunately, performance testing is rarely conducted thoroughly, but the growing availability of computer simulation programs (e.g., *SLiM* and *msprime)* makes performance testing increasingly feasible, even for relatively inexperienced investigators or students (Hoban 2014; Section A13).

Without performance evaluations, statistical methods are often used and then later found to be biased or produce erroneous results. For example, some assignment tests were shown to produce erroneous false positive error rates (wrongly identifying immigrants that were not immigrants, using simulation evaluations; Paetkau et al. 2004). Similarly, some N_e estimators were shown to produce misleading results and underestimate N_e (England et al. 2006). Also, problems with precision (erroneously narrow CIs) for N_e estimators have been reported when using genomic datasets with 100s to 1,000s of loci (Waples et al. 2021). The problems with these methods were not discovered or quantified until long after they were being used in natural populations.

A10 The coalescent and genealogical information

To coalesce means to fuse, unite, or come together. This refers to the process of tracing backward through time and the joining of (coalescence of) homologous gene copies (haplotypes) from different individuals in the same parent or ancestor (Figure A11). The word coalescent is used in several ways in the genetics literature. The coalescent is a model of a genealogical tree of DNA sequences sampled from a population to infer population genetic parameters. The coalescent theory was developed (Kingman 1982) to model genealogies so that allele frequency and genealogical patterns (trees) could be used to infer parameters such as population size, population growth, gene flow, time of divergence, and selection.

Why should we learn coalescent theory and coalescent model testing? It empowers us to estimate population genetic parameters and infer population demographic status—including testing for bottlenecks and identifying

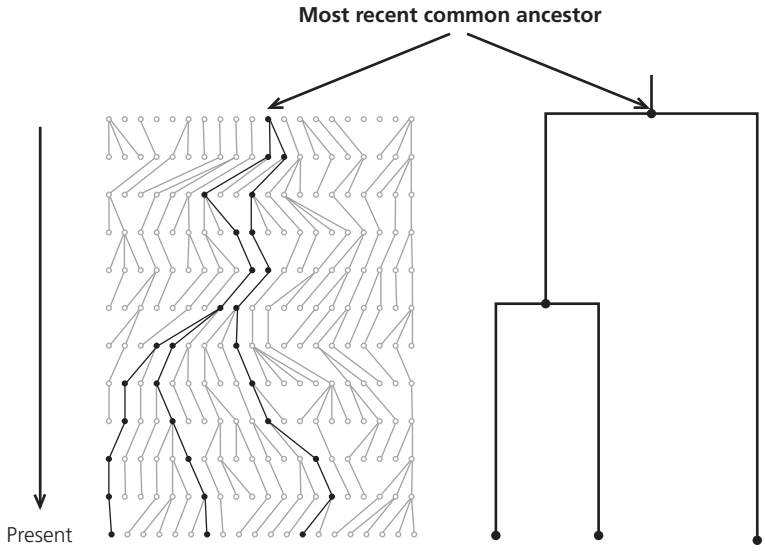

Figure A11 The coalescent approach for modeling the genealogy of individuals in a random mating population of 10 individuals for 11 generations. The complete genealogy is on the left; dark lines trace back through time (from bottom to top) the ancestries of three gene copies sampled in the present population. The "subgenealogy" of the three sampled gene copies is on the right. The two gene copies on the left coalesce most recently. The coalescence times are proportional to branch lengths. The average (and distribution) of coalescence times provide information about the tree shape, which is used to make inferences about demographic history. Modified from Rosenberg & Nordborg (2002) and Felsenstein (2004).

adaptive loci (e.g., Figures A14 & A15). Coalescent models (and simulations) can produce many theoretical genealogies that we can compare with observed data to test hypotheses about demographic history and parameters such as gene flow and effective population size. Knowing coalescent theory and concepts allows us to understand the growing number of publications using coalescent-based inferences for conservation and management of populations.

The coalescent approach yields a distribution of times to the most recent common ancestor between gene copies in a genealogy (Box A2; Kuhner 2009). Coalescences are the points of common ancestry, or internal nodes, on a genealogy (Figure A11). The coalescent can be used with frequentist, ML, or Bayesian statistical approaches; for example, to generate the expected distribution of allele frequencies (i.e., site frequency spectrum) to test hypotheses and estimate parameters (Box A3).

The coalescent is a powerful modeling approach for analyzing population genetic data (Rosenberg & Nordborg 2002). It involves a different way of thinking about population genetics compared with classic approaches. Classic approaches for modeling populations typically follow the inheritance of genes in a forward direction from parents to offspring to grand offspring. That is, individual parents are randomly mated to produce offspring, which are eventually mated to produce grand offspring, as is conducted in individual-based simulation modeling (Balloux 2001). In contrast, the coalescent approach looks backward in time and traces gene copies back from offspring to parents, to grandparents, and eventually to a single most recent common ancestor.

Coalescent analyses allow extraction of genealogical information from DNA sequences for alleles at a locus. Classic statistical estimators in population genetic (e.g., F_{ST}) do not use genealogical information; they use only allelic states, A1, A2, and their frequency. Since the advent of DNA sequencing (and restriction enzyme analysis; Chapter 3) most datasets contain information on allele relationships (i.e., mutational divergence between alleles or haplotypes). Microsatellites also contain genealogical information in the number of repeat unit differences between two alleles (assuming the stepwise mutation model discussed in Sections 3.3.2 and 12.1.2). Coalescent trees visualize relationships among alleles or haplotypes at a locus (Figure A11).

Genealogical methods, such as the coalescent, are generally not used to infer a phylogeny (but see Carstens et al. 2005); rather they estimate parameters of the stochastic evolutionary processes that give rise to genealogies, such as gene flow rates, population size, or population growth rates. For example, different population demographic histories yield differently shaped genealogies (Figure A12). Consequently, genealogical shape can be used to infer a population's demographic history (Emerson et al. 2001; Kuhner 2009)

Population growth yields star-like genealogies with an excess of long branches of similar length compared with a stable, stationary population size (Figure A12). Many long and similar-length branches are expected to arise during a long-term population expansion because new alleles (mutations) tend to persist a long time since genetic drift (lineage sorting) is negligible in fast-growing populations. If a population is large and has expanded from a small size

Box A2 Coalescent modeling

Coalescent modeling involves two main steps: first, we generate a random genealogy of individuals backward through time. Here it helps to envision clonal individuals (or haploid chromosomes such as mitochondrial DNA (mtDNA)). We start with a sample of clones and randomly connect them to parents, grandparents, great-grandparents, etc., until all clones coalesce into a single ancestor (the most recent common ancestor, MRCA; Figure A11). Going back in time, two lineages will coalesce whenever two clones are produced by the same parent. Going forward in time, lineages branch whenever a parent has two or more offspring, and branches end when no offspring are produced (i.e., **lineage sorting**, Section 7.8).

Second, we randomly place mutations on branches using Monte Carlo simulations and a random number generator while considering the mutation rate. We start by assigning some allelic state to the MRCA and then drop mutations along branches randomly moving forward. If a mutation is placed on a branch, then the allelic state (e.g., base pair state for a SNP or allele length for a microsatellite) must be determined by following rules of a model. For example, mutation models for SNPs or DNA sequences often include a higher probability of a transition than a transversion (e.g., an A to G transition is more probable than an A to T transversion). For microsatellite loci, the stepwise mutation model assumes that a mutation is equally likely to increase or decrease allele length by a single repeat unit (Section 12.1.2).

Coalescent modeling is computationally efficient because we only simulate the sampled lineages (subgenealogy in Figure A11), and not the entire population as is done for individual-based forward models. Simulating only the subgenealogy requires less record-keeping and saves computer time compared with the forward (individual-based) simulation modeling approach that requires record-keeping for all individuals including those not sampled.

In coalescent modeling, we often want to separate the two stochastic genealogical processes: (1) random neutral mutation; and (2) random genetic drift (caused by random reproduction and population demography). These two processes determine the genetic makeup of the population of lineages. Separation of the two is important because we are often interested in the biological phenomena of demography and reproduction, but not mutation processes (Rosenberg & Nordborg 2002). For example, we are often interested in testing for population expansion or population subdivision, but not mutation dynamics.

in the past, then most coalescent events are relatively old (i.e., branch lengths are relatively long) and originate near the time the expansion began, compared with the more randomly distributed coalescent event times in a stable population (Figure A12, A13, and A14).

Population bottlenecks lead to genealogies having most coalescence events during the historical period of small effective size (Figure A14). Weak bottlenecks can be differentiated from strong historical bottlenecks by the shape of the genealogy.

Coalescent trees also contain information on the allele frequency spectrum, e.g., the proportion of rare alleles or **singletons**. A singleton is a rare variant (allele) found as a single copy. All singletons are represented by mutations on branch tips of a coalescent tree, whereas common mutations (shared by multiple haplotypes) are near the base of coalescent trees (Figure A14).

To understand the value of coalescence modeling to detect changes in population size, we can compare the shape of coalescent trees (genealogies) to the associated SFS, which is also perturbed by bottlenecks (and population expansions) in a characteristic way. A recent weak bottleneck causes a deficit of rare variants (alleles), which can be seen in the SFS (Figure A14, bottom right). A strong historical bottleneck causes an excess of rare variants that accumulate during post-bottleneck population growth (Figure A14, bottom far right). Similarly, the magnitude of nucleotide diversity (π) and the number of segregating sites are different for a weak versus strong historical bottleneck (Figure A14).

We must sample many genes to obtain accurate and precise estimates of a population's demographic history. This is because random genealogical processes lead to many possible random genealogies for a given demographic history (Figure A13). To test if one particular history best fits our empirical dataset, we can simulate thousands of random genealogies for each population history; for example, a stable versus declining population. If one history (e.g., a decline) best fits our observed empirical data, then another history (stable population) can be rejected by comparing our observed data and the simulated genealogy data from the alternative population histories.

Natural selection also causes distinctively shaped genealogies at a locus. If the genealogy of one locus differs

Box A3 The coalescent used in frequentist, likelihood, and Bayesian approaches

The coalescent can be used for modeling or conducting statistical tests under different statistical frameworks including frequentist, likelihood, or Bayesian. For example, a frequentist coalescent approach might be used to test whether N_e is significantly smaller than 100. For this, we might (1) use the coalescent to simulate 1,000 independent datasets for a population with $N_e = 100$; (2) compute an estimate of N_e for each simulated population (to obtain a distribution of N_e estimates consistent with a true N_e of 100); and (3) calculate how frequently (out of the 1,000 datasets) we obtain a simulated N_e estimate as small as the empirical N_e estimate from our study population. If our population's N_e estimate is so small that it occurs only once in 1,000 simulated datasets, then we would conclude that our population's true (actual) N_e is significantly ($P < 0.001$) less than 100. This kind of approach was used in Funk et al. (1999) to test for small N_e in a salamander population.

In an ML approach for testing if N_e is significantly smaller than 100, the coalescent could be used to help compute the likelihood of $N_e = 1$, $N_e = 2$, $N_e = 3$ up to $N_e = 200$, given our raw data. Here, the coalescent could be used to simulate thousands of datasets for each N_e, and then compute the likelihood of each N_e ($N_e = 1$, $N_e = 2$, $N_e = 3$, etc.) given our real dataset. This would yield a probability (likelihood) distribution of N_e values (with $N_e = 1$, $N_e = 2$, $N_e = 3$ up to $N_e = 200$ on the *x*-axis). If all the area under the likelihood (probability) curve was less than 100 (i.e., did not include $N_e = 100$), we could conclude that our population's effective size is less than 100.

In a Bayesian approach, we would conduct the same computations as in the ML approach just described, using the coalescent. However, we then would modify the resulting likelihood distribution by multiplying it by a prior distribution to obtain a posterior distribution, as illustrated in Figure A5. This box illustrates how the coalescent can be used within different statistical frameworks to conduct statistical tests or estimate a population parameter.

significantly from most other loci in the genome, we might infer that selection has influenced the locus. Natural selection detection at a locus requires the study of many independent loci because demography and selection can generate similar genealogies (Figure A15).

A **selective sweep** (Section 10.3.1) will remove many alleles, similar to a bottleneck, and subsequently lead to a star-shaped genealogy as new mutations arise in the population during the generations following the selective sweep. This single-locus star genealogy resembles the star-shaped genealogies expected at loci genome-wide following population growth.

Balancing selection at a locus can mimic the effect of a recent bottleneck, increasing the proportion of alleles at intermediate frequencies at a single locus, similar to the way bottlenecks increase intermediate allele frequencies at loci genome-wide. To resolve between selection signatures and demographic signatures, we study many loci across the genome.

We can infer genealogies for loci at many points along chromosomes by using a sliding window approach. This approach of inferring and visualizing coalescent trees (or summary statistics) along chromosomes can localize genomic regions under selection, which is a "population genomics approach" for discovering adaptive loci (Chapter 4; Hohenlohe et al. 2010,).

A11 Bioinformatics, Linux, and coding

Bioinformatics is the application of code writing and informatics techniques used in statistical, mathematical, or computer science to understand and organize biological data (e.g., DNA sequence data). Acquiring the bioinformatic skills needed for analyzing NGS data is challenging. The ability to conduct Linux-based **scripting** to run software pipelines (workflows) to analyze large DNA sequence datasets is essential. Scripting has been defined as a way to manipulate large datasets and automate repetitive tasks (Mount 2004). A script is a file containing a series of statements interpreted by software (e.g., Linux **shell** software). Scripting and coding are sometimes used interchangeably.

Command-line scripting can be learned from online tutorials and workshops that teach the use of software and the Linux operating system (Hendricks et al. 2018; Stahlke et al. 2020). The Linux interface is the command shell, similar to DOS, which is widely used for genomic data analysis. Python is one of the most popular and

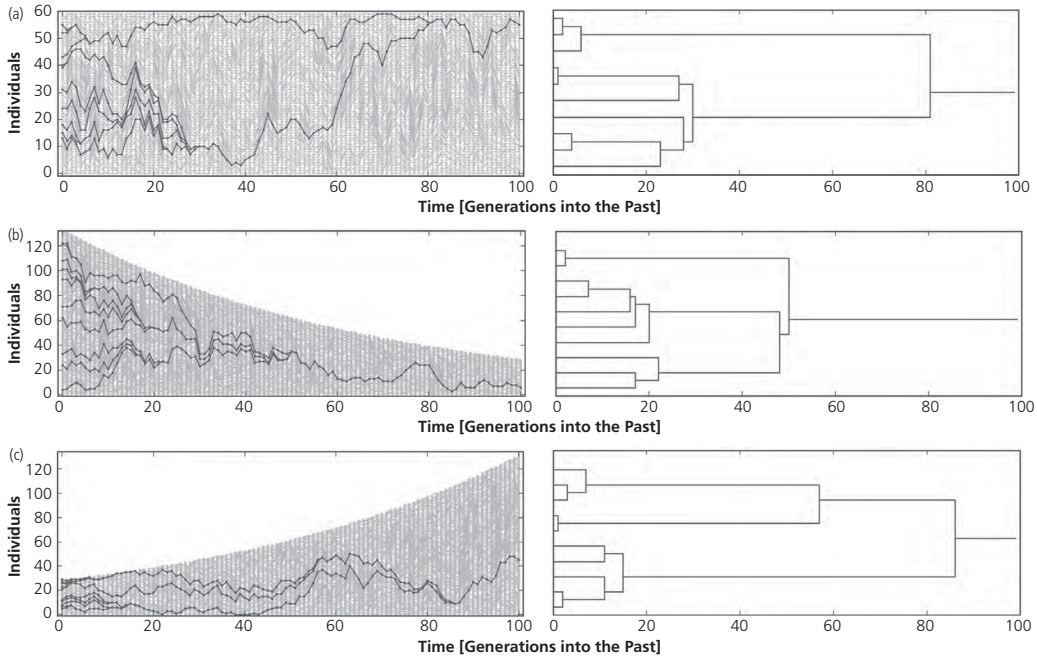

Figure A12 Three genealogies simulated with a coalescent model for a population with a mean N_e of 60. (a) Constant size population (random distribution of branch lengths). (b) Growing population (multiple relatively long branches, more variable distribution of branch lengths, and more coalescent events relatively near the time of initiation of population growth 100 generations in the past). (c) Declining population (fewer long branches, distribution of branch lengths less variable than random as in the stable population, and most coalescences are relatively recent). Similar to Figure A11, each complete genealogy is shown (top of each of the three panels) with dark zig-zag lines tracing back through time (from left to right) the ancestries of 10 gene copies sampled in the present population (time 0). The simplified "subgenealogy" of the 10 sampled gene copies is in the column on the right. Population size is indicated by the gray area on each top panel. The horizontal axis (time) is the same (100 generations) for all three trees. Figure courtesy of Peter Beerli.

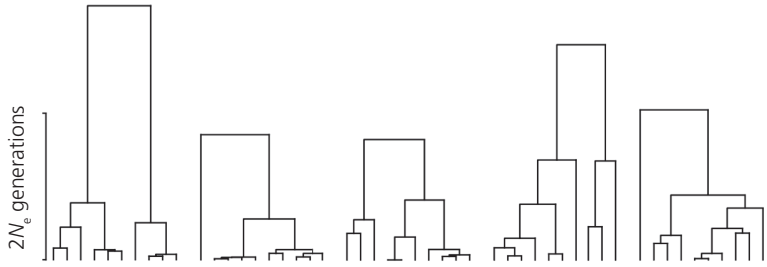

Figure A13 Genealogies for alleles at five loci with the same demographic history (stable with no selection). Note the wide range of coalescent branching times and time to most recent common ancestor (TMRCA) due to stochasticity typical of the coalescent. Thousands of different genealogies are possible for the same demographic history. Thus, we must study many independent loci to infer a population's demographic history with precision. Modified from Cutter (2019).

important programming languages for analyzing NGS data. Bioinformatic skills and scripting are needed to analyze DNA, RNA, epigenetic, protein sequence datasets, and also to conduct simulations, manage large databases, and for **data scraping** from databases (e.g., Garner et al. 2020) and the world wide web.

Code for Python or Linux scripts can be copied and pasted from user manuals and tutorials. The Seqanswers website is dedicated to NGS discussion (http://seqanswers.com). Also, the ConGen course (Andrews & Luikart 2014; Hendricks et al. 2018; Stahlke et al. 2020) provides video recordings of lectures that

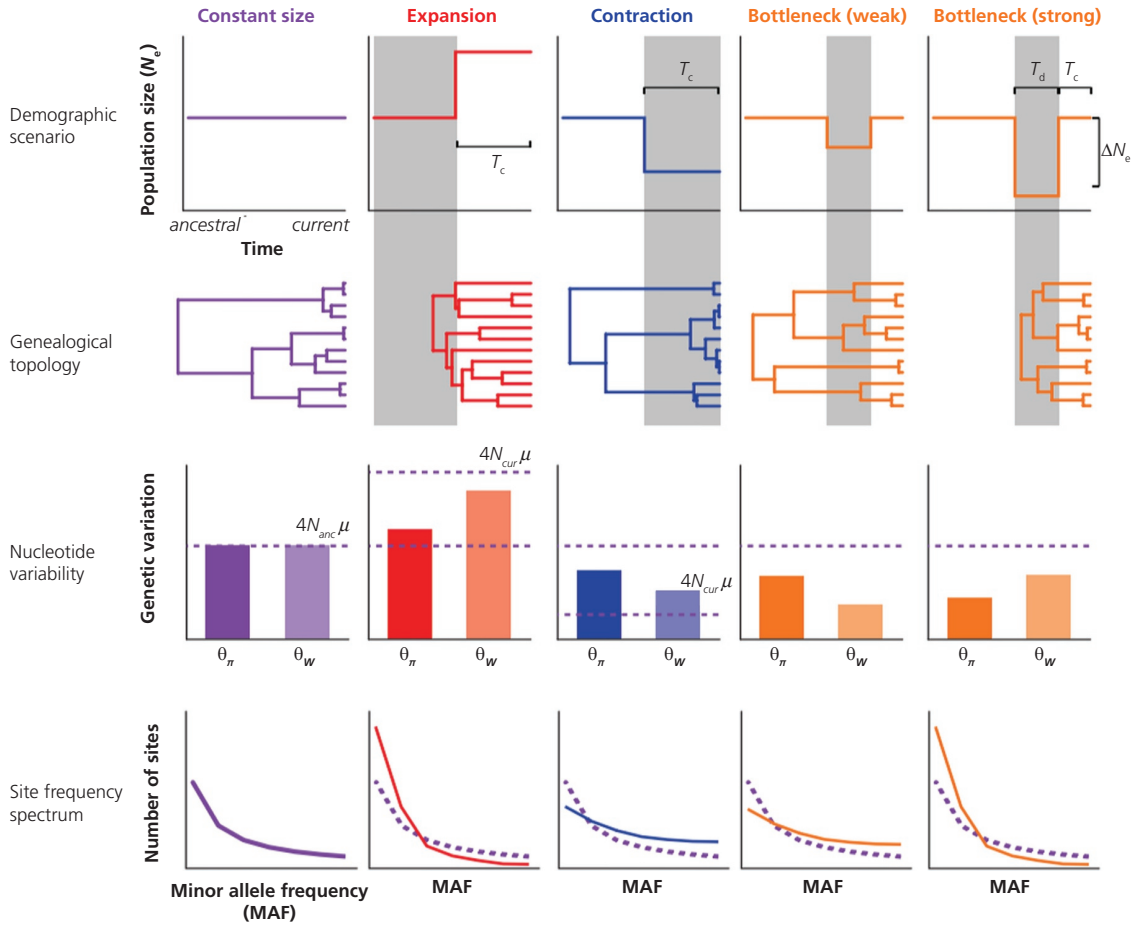

Figure A14 Population size changes perturb the shape of coalescent genealogies and patterns of molecular variation. Expansion or contraction of N_e is shown compared with a constant size (purple) for neutral loci. Coalescence is faster with small N_e (gray shaded regions), and small populations have less genetic variation (ancestral N_{anc}, current N_{cur}, where N equals N_e). Demographic perturbations shift the SFS and change the two metrics quantifying nucleotide variability: θ_π computed from nucleotide diversity ($4N\mu$) and Waterson's theta (θ_W) computed from segregating sites. These coalescent trees represent a single "representative" locus (sequence) for selectively neutral sites. We must study many loci to make general conclusions about demography because of stochastic differences among loci (Figure A13). The variance in coalescent times among loci in expanding and bottlenecked populations is lower than in a constant-size population. From Cutter (2019).

introduce Linux command line and scripting. Sharing of scripts among scientists is invaluable and people's scripts are often available on their web page, GitHub sites, and publications (e.g., Shafer et al. 2017, file named "Data S1. Scripts used…").

Point-and-click computer programs generally do not exist to conduct the quality control filtering, genotyping, and statistical analyses that are required to use NGS data for population management or wildlife research (Perkel 2017). Furthermore, most datasets are distinct and require novel analyses not available in existing computer programs, R, or bioinformatic pipelines.

Many computational approaches including MCMC, ABC, and likelihood-based methods are implemented in population genetic software programs to estimate parameters such as N_e, Nm, and detect selection and substructure. Some of these approaches require the use of command line as with Linux. The world's most widely used computer program to assess population structure and admixture (program *STRUCTURE*) is among the few with a point-and-click graphical user interface (GUI) in addition to a command-line version (Pritchard et al. 2000). A GUI (pronounced "gooey") allows computer users to communicate with the computer by moving a pointer around on a screen and clicking a button.

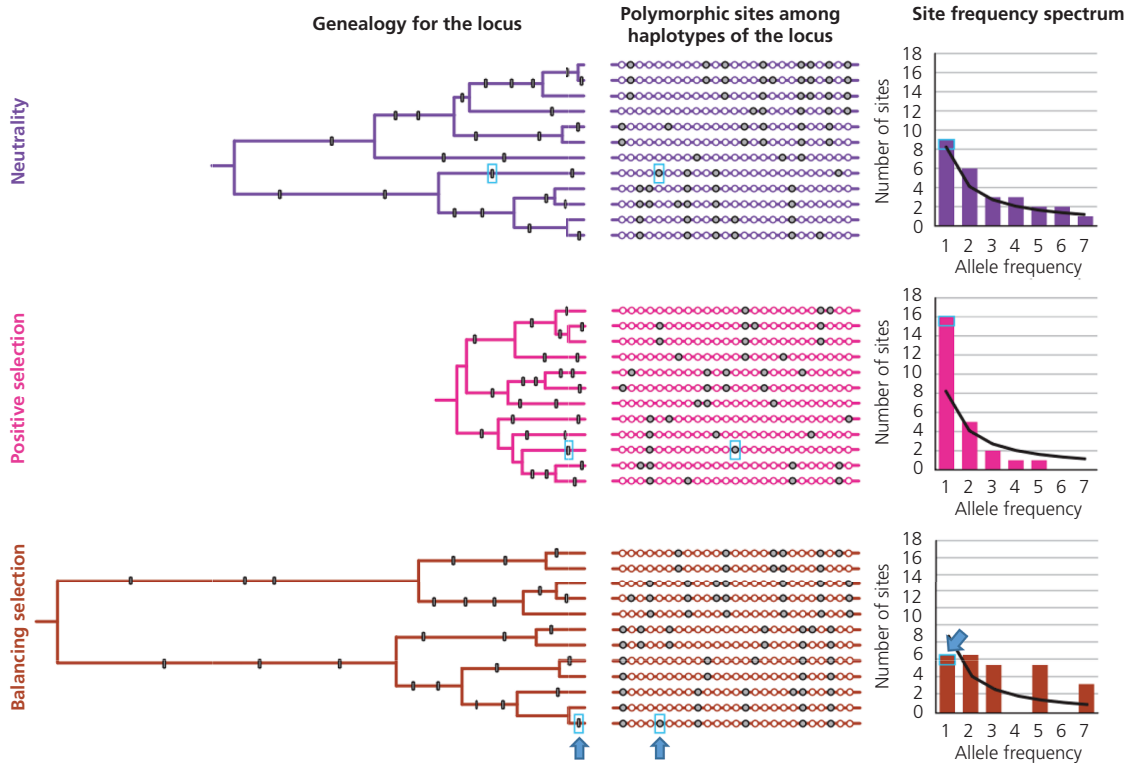

Figure A15 Three representations of polymorphisms across a locus (DNA sequence): a genealogy, a haplotype diagram of polymorphic sites, and allele frequency distribution (i.e., site frequency spectrum). This illustrates how a genealogy contains information about relationships among the alleles and also the allele frequency distribution (AFD). Singletons are on tree branch tips, and counted as a "1" on the AFDs. Natural selection modifies the shape of a genealogy relative to the shape of a neutral gene (purple). Positive selection causes a shorter time to the most recent common ancestor (pink). It also causes a greater proportion of new neutral mutations on terminal branches of the genealogy than expected under neutrality, along with an excess of singletons on the AFD, which develops after new mutations arise. The expected distribution curve for a neutral AFD is indicated with black lines. Long-term balancing selection causes a more ancient TMRCA (brown) and an excess of mutations on long internal branches of the genealogy, along with more shared variants among alleles, and a deficit of rare variants on the AFD. One example singleton variant is highlighted (see blue open rectangles and three arrows) to show its position on the genealogy, haplotype diagram, and AFD. From Cutter (2019).

Point-and-click pipelines exist for resequencing human and pathogen genomes using NGS data. For example, *EDGE* (Empowering the Development of Genomics Expertise), is a user-friendly bioinformatics pipeline to fully analyze common microbial genomes (including sequence assembly and species identification), using a computer interface to generate polished analyses (Perkel 2017).

User-friendly pipelines could help bridge the conservation genomics gap by making NGS data accessible to wildlife management and monitoring projects, which are often run by labs, technicians, and university students with limited bioinformatics expertise. However, such pipelines must be well validated before use. Validation requires repeat genotyping families or trios of individuals and running blind samples to ensure repeatability and low genotyping error rates. Conservation geneticists need validated pipelines and workflows to run scripts and programs linked together (pipelines) to filter and analyze NGS data.

A downside of a "point-and-click" pipeline is that users might not understand what assumptions underlie the analyses they are running. This can lead to erroneous results and management decisions, harming species and genetic diversity.

A12 Filtering and data quality

Among the most important aspects of population genomics data analysis is filtering of raw sequence reads (from NSG data) to identify and remove errors. For

example, filtering is required to identify sets of reads that belong to a single locus versus multiple loci (duplicated loci), to genotype individuals, and to resolve between outlier loci caused by selection versus errors in genotyping or data processing. Insufficient data filtering can fail to remove errors or add errors to a dataset (Section 4.6; Shafer et al. 2017; Larson et al. 2021).

Filtering can be difficult. We need to filter stringently to identify and genotype high-quality loci from NGS data. But we do not want to filter too stringently to remove high-quality loci or loci under selection that are of interest for studies of adaptation and fitness (Andrews & Luikart 2014).

Researchers have shown that stringency choices and bioinformatic processing (filtering) of NGS data can dramatically influence downstream population genetic inferences—with implications for management (Shafer et al. 2017; Larson et al. 2020). Shafer et al. (2017) observed major differences between *de novo* and reference-based approaches for genotype calling (for **restriction site-associated DNA sequencing (RADseq)** data) such that *de novo* discovered fewer SNP loci, and generally yielded lower F_{IS} values. They also found the SFS analysis was highly sensitive to the chosen filtering parameters or pipeline (Figure A16). This SFS sensitivity is problematic for demographic inference such as detecting population bottlenecks, which depends on the proportion of rare alleles (Figure A14).

Filtering NGS reads to identify new loci *de novo* (without a reference genome) requires a choice of the number of nucleotide site differences allowed between reads when combining reads together as a single (inferred) locus. If too many site differences are allowed among reads (e.g., >3 site differences per 150 bp read) it increases risk of erroneously combining reads from different (duplicated) loci. If not enough site differences are allowed among reads, some reads and alleles (e.g., divergent haplotypes) will be erroneously excluded from a locus, which leads to allele dropout and genotyping error.

Filtering loci using different minor allele frequency (MAF) thresholds provides an example of the effects of filtering on downstream data analyses. Students at a workshop used an MAF of 0.01, 0.05, 0.1, and 0.2 to filter a RADseq dataset and then computed F_{ST} using program *STACKS* (Catchen et al. 2013). Students discovered that as MAF increased, estimates of mean F_{ST} increased (Hendricks et al. 2018). This increase could result from the relationship between F_{ST} and expected heterozygosity such that low heterozygosity loci have a limited maximum F_{ST}-value for bi-allelic SNP loci (Allendorf & Seeb 2000). It suggests that stringent MAF filters, such as removing loci with MAF below 0.05 or 0.10, could skew metrics based on the SFS (or the expected heterozygosity) and inadvertently filter out useful loci including those under selection (Hendricks et al. 2018; Figure A17).

Filtering choices such as removing singletons can also influence F_{ST} and population structure analyses (Figure A17). To avoid such problems, researchers can test a range of filtering parameters to quantify filtering effects

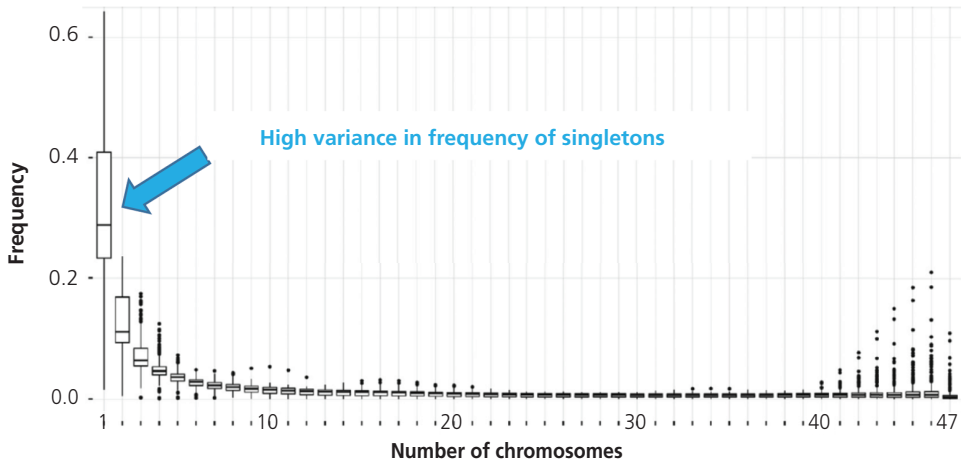

Figure A16 Distribution of the site-frequency spectrum (SFS) obtained from results of 312 different filtering-option combinations applied to a single RADseq dataset. Boxplots show the frequency of the minor allele (*y*-axis) at all segregating sites. Note the high variability in frequency of singletons (1 on the *x*-axis) among the 312 filtering options. Horizontal lines in box plots depict the median; box margins show interquartile range between 25% and 75%; whiskers extend to 1.5 times the interquartile range. Modified from Shafer et al. (2017).

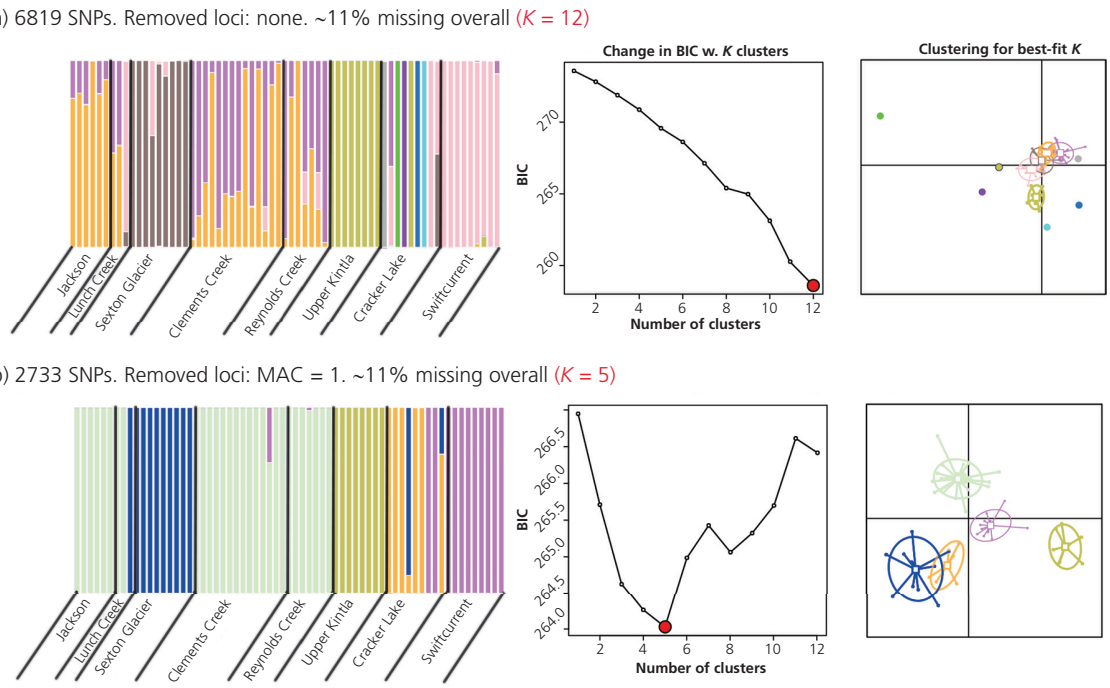

Figure A17 Comparisons of population structure in an endangered alpine stonefly (meltwater stonefly) inferred from discriminant analysis of principal components (DAPC) analyses after removing (and not removing) SNPs with a singleton (minor allele count (MAC) = 1). The best-supported *K* (number of populations) is in parentheses in red. From left to right: assignment plots where each vertical bar represents one individual, a plot of the Bayesian Information Criterion (BIC) for a range of *K* values (lower BIC indicates higher model support), and individual assignments to genetic clusters based upon two principal components. "Missing overall" refers to 11% of missing data, which were allowed to be missing during filtering to identify SNPs. The full 6,819 SNP dataset included singletons. The 2,733 SNP dataset has all singletons removed, which is roughly equivalent to a MAF cutoff of 0.015 for a sample size of 65 individuals. Modified from Hotaling et al. (2018).

and help avoid erroneous inferences caused by filtering (Shafer et al. 2017; Paris et al. 2017; Hotaling et al. 2018).

A13 Why simulations?

Computer simulations advance conservation genetics in multiple ways, including bridging the gap between computational method development and actual application of methods in natural populations (Section A9). Simulations also help bridge the gap as a tool for teaching population genetics, improving conceptual understanding of ecological and evolutionary processes and patterns, and most importantly by advancing the field generally, for many taxa, by allowing (simulation) studies of many taxa and scenarios (Epperson et al. 2010; Peng et al. 2012; Luikart et al. 2021).

Simulation studies are relatively quick, easy, and inexpensive to conduct, in part, because many computer simulation programs are available, as suggested by this quote: "Dozens of sophisticated, customizable software packages for simulation now makes simulation an accessible option for researchers in many fields" (Hoban et al. 2012, p. 110). Simulations are inexpensive compared with field and laboratory studies because you only need a computer and software for simulations.

Simulations can advance conservation by assuring biologists that genetic marker-based approaches produce reliable estimates of parameters they need such as population census size, effective size, inbreeding, inbreeding depression, F_{ST}-related divergence metrics, gene flow and dispersal (historical and contemporary), selection detection, etc. This "bridging the gap" to understand limits of marker-based population assessment helps make feasible the use of genetics for conserving natural populations.

Advancing conservation genetics broadly (for many species) via simulation studies is exemplified by Waples et al. (2013). These authors simulated >60 species with different age- and sex-specific demographic vital rates to establish our ability to reliably estimate N_e and N_b (effective number of breeders) in age-structured populations

using marker-based gametic disequilibrium statistics (r^2). This work was extended by Luikart et al. (2021) and others, who quantified bias and precision of the *LDNE*-based N_b estimator for detecting population declines in many species with different life histories and when using different DNA-marker types (Guest Box 10, Section 23.4.1). These simulation studies, along with new software (e.g., Antao et al. 2020), can help managers design statistically powerful N_b monitoring programs for early detection of population declines.

Simulation studies are often quick to complete and publish, and contribute to advancing knowledge by providing a theoretical framework for interpretation of field and lab data. Valuable simulation projects can be conducted in weeks or months, whereas field studies often require multiple field seasons. For example, Waples & Gaggiotti (2006) used the program *Easypop* to evaluate the relative usefulness of traditional contingency table tests versus population assignment tests to detect population subdivision. Also, Tallmon et al. (2012) used simulations to compare a traditional demographic and new genetic method to detect stable versus increasing population size trends. They compared the Lincoln–Peterson (LP) abundance (N) estimator and a one-sample effective population size (N_e) estimator (*LDNE;* Guest Box 10, Section 23.4.1). The genetic methods (*LDNE*) often outperformed LP when samples of 60–120 individuals were collected 5–10 generations apart, suggesting *LDNE* methods could supplement traditional demographic methods for detecting population growth.

Simulations are easily conducted for forward-time (individual-based) models (Waples et al. 2014) and also backward-looking coalescent models. For example, Berthier et al. (2002) and Wu et al. (2014) used coalescent simulations to evaluate the performance of estimators of the contemporary N_e, long-term historical N_e, and selection detection. Many simulation programs are available to produce datasets using forward-time and reverse-time simulation approaches (Hoban et al. 2014). Among the easiest simulators is *Easypop* (Balloux 2001; Waples & Gaggiotti 2006; Waples & Do 2008). Among the most flexible and power simulation packages for complex scenarios are *simuPOP* and *SLiM-3* (Peng & Amos 2008; Haller & Messer 2017). Other widely useful simulation programs include *CDMetaPOP* (Landguth et al. 2017) and *AgeStrucN*$_b$ (Antao et al. 2020).

Simulations allow study of relatively realistic and complex scenarios that cannot be studied using analytical or numerical equations. For example, simulations allow study of how demographic processes such as dispersal or reproduction interact with landscape features to affect probability of site occupancy, population size, and gene flow, which in turn determine spatial genetic structure (Epperson et al. 2010; Landguth et al. 2017). Simulations are crucial to advance **landscape genetics** and can help managers understand and manage landscape connectivity (Chapter 19). In another complex simulation, researchers determined how genetic architecture of a trait (i.e., the number and effect size of loci underlying a trait) influences population growth and persistence (Kardos & Luikart 2021). Simulations also are incorporated in software programs using Bayesian (MCMC) and ABC approaches to estimate population genetic parameters.

Conducting simulations can be fun and helpful for teaching population genetics. We can use programs like *PopG* (https://evolution.gs.washington.edu/popg/) to study how selection and drift interact to influence allele frequencies at a locus. Using simulations, students and managers can see the sobering result that even strong selection ($s = 0.1$, i.e., 10% survival difference) cannot overcome drift in small populations ($N_e < 50$), which can lead to loss of alleles that increase fitness.

Guest Box A A testable model-based perspective for conservation genetics

Mark A. Beaumont and Jo Howard-McCombe

Population genomic analysis is now central to many investigations in conservation biology, and the cost of NGS methods is increasingly affordable for many research budgets. Similarly, a range of different methods has been developed for analyzing and displaying genomic data.

Faced with a mass of data it can be tempting to display a range of statistics and diversity indices. However, these are not ends in themselves, and differences in diversity may reflect more the type of marker or mutation process. Visualization tools, particularly PCA and dendrograms, can be very useful, particularly for suggesting hypotheses. It is helpful to consider what underlying process may have given rise to the data, and then to uncover what this might be in terms of a population genetic model. A good example is provided by an investigation into the pattern of hybridization of wildcats with domestic cats in the Swiss Jura. Nussberger et al. (2018) estimated demographic parameters from genome data using a variety of approaches, and then Quilodrán et al. (2019) used simulation to test model-based explanations for the data.

This Appendix outlines several different methods that can help to structure and parameterize a model. It is, however, worth bearing in mind that whatever method is used to obtain the parameter estimates—whether based on ML, Bayesian theory, or least-squares—it only finds the best estimate given the structure of the model, and there is no guarantee that it provides a good explanation for your data. A useful quote to remember is from the statistician George Box: "all models are wrong, but some are useful" (Box & Draper 1987, p. 440). It is important to identify what level of complexity is useful in the context of specific research questions.

Simple models have the advantage that they are easier to interpret and often do not involve extensive computation (Wakeley 2004). However, it is also possible that limitations in genomic data or the type of analysis may mean that a simple model is also quite misleading. The connection between modern-day sequences and historical demography is a delicate genealogical thread that is easily disrupted by natural selection and population restructuring. A current narrative in human population genomics, for example, is that much of the earlier demographic history has been overwritten by recent events, and that a more accurate picture can be obtained from ancient DNA analysis of archeological and museum samples (Reich 2018).

Recently, efficient packages have become available for simulating genomic data under many scenarios. It can be helpful to use simulation just to check that the broad outlines of your model are consistent with what you consider important in your data (Section A13). Once you have a parameterized model, and you are happy that it explains your current data, it can then be used to make predictions that can be tested in future studies.

Even if, like most of us, you do not have the mathematical training or mindset to follow all the equations in any detail, it is important to understand the assumptions of the methods you are using well enough to know the possible pitfalls and limitations. It is useful to consider what cannot be captured in a model, and how this could potentially impact your results. Understanding and communicating complex population genetic theory is a key skill to translate research into effective conservation.

The genomics era is an exciting time in conservation biology. Unprecedented volumes of genetic data (increasingly from nonmodel species) allow population genomic tools to tackle previously intractable questions. Our advice is not to be overwhelmed by the quantity of data or statistical analyses, but to focus on specific research questions, develop testable hypotheses, and use the data to support evidence-based conservation.