Unbiased estimation of the number of segregating sites across unequal sample sizes

William Hemstrom[1 †], Mark R. Christie[1,2†]

[1]Department of Biological Sciences, Purdue University; 915 W. State St., West Lafayette, IN, USA

[2]Department of Forestry and Natural Resources, Purdue University; 715 W. State St., West Lafayette, IN, USA

† Corresponding authors:
William X. Hemstrom; Department of Biological Sciences, Purdue University; 915 W. State St., West Lafayette, Indiana 47907; whemstro@purdue.edu, 503-730-5529

Mark R. Christie; Department of Biological Sciences & Department of Forestry and Natural Resources, Purdue University; 915 W. State St., West Lafayette, Indiana 47907; christ99@purdue.edu, 765-494-2070

Abstract

Geneticists use a wide range of approaches to measure genetic diversity across the genome. While estimators which explicitly measure allele frequencies (such as expected heterozygosity, nucleotide diversity, and Tajima's D) are well developed for single nucleotide polymorphism (SNP) data, estimators of numbers of polymorphic loci or alleles, such as private allele counts or allelic richness, are not as common. The number of segregating sites (or SNPs) is one such estimator, calculated as the proportion of nucleotide sites that have more than one allele. This underutilized estimator can provide informative estimates of genetic diversity across multiple scales, from genes to chromosomes to entire genomes, and is particularly informative when used in conjunction with frequency estimators such as expected heterozygosity. Unlike estimates of allelic richness or private alleles, however, segregating site counts are rarely adjusted to correct for unequal sample sizes or differences in missing data among populations or sample groups. Here, we introduce an estimator for the number of segregating sites expected in a sample group following rarefaction, which allows for the unbiased comparison of the number of segregating sites among multiple sample groups.

Introduction

Estimating the degree of genetic diversity has long been a critically important and widespread practice in population, evolutionary, and conservation genetics (Allendorf, 1986; Chapman et al., 2009; David, 1998; DeWoody et al., 2021; Hedrick & Kalinowski, 2000; Moritz, 2002). Genetic diversity, or the diversity alleles, genotypes, and segregating loci within a population (Frankham et al., 2002), is the both the basic foundation upon which natural selection acts, and thus affects the speed and degree of genetic adaptation (Kardos et al., 2021), and is strongly correlated with fitness both at the individual and population level even when no obvious functional ties are known (reviewed in DeWoody et al., 2021). Accurate estimates of genetic diversity are therefore critical for monitoring the health of populations and for predicting their capacity to respond to changing environmental conditions (Lai et al., 2019; Reid et al., 2016; Visser, 2008). As a result, genetic diversity is considered a key component of biodiversity (Hvilsom et al., 2022; Schmidt et al., 2023).

Estimators of genetic diversity vary widely, but essentially fall into two categories: those that explicitly measure variation in allele frequencies and those that do not. Allele frequency dependent statistics, such as expected heterozygosity (H_e), F-statistics (F_{IS}, F_{ST}, and F_{IT}; (Weir & Cockerham, 1984), nucleotide diversity (π), and Tajima's θ (Tajima, 1989) all rely principally on and seek to measure allele frequencies across surveyed loci; whereas other statistics, such as Watterson's θ (Watterson, 1975), counts of alleles per locus, the number of observed segregating sites (from which Watterson's θ is derived), and private allele counts do not. These different measures therefore focus on different aspects of genetic diversity, which is not trivial: evolutionary forces do not act equally on the number of loci/alleles in a population and the frequency of those alleles (Fu, 2022). Tajima's D, for example, is a powerful and broadly used

statistic fundamentally based on the difference between Tajima's θ and Watterson's θ that can be used to both detect selection and population demographic changes (Tajima, 1989) due to the way that those forces act on the balance of allele frequencies within populations. For example, a population which has many polymorphic loci and a high expected heterozygosity may have undergone different historical demographic processes than a population with the same number of polymorphic loci but a far lower heterozygosity. Both types of measures therefore provide important ecological and evolutionary insight.

The average allele count per locus (which is usually corrected to allelic richness; see Kalinowski (2004) has historically been one of the most prevalent allele frequency independent measures of genetic diversity. While this is a particularly useful measure in datasets of microsatellite and other heavily polyallelic markers, it is less so in those composed of single-nucleotide polymorphisms (SNPs). SNP datasets are often biallelic by design, and thus allele counts per locus vary little among sample groups and are correspondingly less informative. In contrast, the location and number of segregating sites per population is a useful alternative in SNP datasets (Hartl et al., 1997), especially given the speed of calculation and ease with which comparisons can be made across samples and genomic locations.

While the number of segregating sites is well-defined mathematically (Fu, 2022), it is not without problems. In particular, estimates of the number of segregating sites per sample group will be biased whenever sample sizes are not equal across the sample groups under comparison at all loci. Specifically, sample groups with large samples will tend to have a higher number of segregating sites, since low-frequency variants will be observed on average much more frequently in sample groups with more sequenced gene copies than in those with fewer. This bias can occur either because of unequal numbers of individuals among sample groups or due to

unequal proportions of missing data. This problem also affects estimates of allele counts per locus and private alleles, but corrections for both of those estimators which use rarefaction to estimate those parameters under a common sample size are well developed (Kalinowski, 2004).

Here we present rarefaction-corrected estimators for 1) the probability that any given locus will segregate with a reduced sample size and 2) the expected total number of segregating sites across all loci within a sample group. We show that these estimators are highly accurate via comparison to re-sampling using simulated data. These estimators are currently implemented in the snpR R package via the function calc_seg_sites (Hemstrom & Jones, 2023).

Methods

Probability of observing segregating loci via rarefaction

Rarefaction can be used to estimate the probability of observing a segregating site at a specific locus using much of the same framework used for calculating allelic richness. In brief, allelic richness (or the expected number of distinct alleles expected at a given locus under a common sample size g across sample groups) can be estimated for a given sample group j by summing the probability of observing each i of m unique alleles using the counts of those alleles in sample group N_{ij} and the total sample size in that sample group N_j.Allelic richness is calculated by comparing the number of possible ways to draw g gene copies without sampling allele i ($\binom{N_j - N_{ij}}{g}$) to the total number of possible combinations of gene copies that can be drawn ($\binom{N_j}{g}$); the inverse of this ($1 - \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}}$) is therefore the probability of observing allele i in

sample group N_j, and the sum of this value across all m alleles gives the expected number of alleles observed at a locus in sample group j, α_g^j (Hurlbert, 1971; Kalinowski, 2004):

$$\alpha_g^j = \sum_{i=1}^{m} 1 - \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}} \quad (1)$$

The expected number of segregating loci in a sample group for a draw of g gene copies can be derived similarly. For a locus i to be segregating in sample group j, all alleles drawn across all gene copies must be identical. If alleles are independent within each locus (the loci are at Hardy-Weinburg Equlibrium, HWE) and initial sample size (N) is infinite, the probability ($P(S_j)$) of observing a segregating site at a locus is the inverse of the probability of drawing only one allele in g draws with replacement:

$$P(S_j) = 1 - \sum_{i=1}^{m} f_{ij}^g \quad (2)$$

where f_{ij} is the allele frequency of allele i in sample group j. When N is finite, however, draws are conducted with replacement and thus binomial coefficients must be used instead to determine the probability of drawing a specific allele:

$$P(S_j) = 1 - \sum_{i=1}^{m} \frac{\binom{N_{ij}}{g}}{\binom{N_j}{g}} \quad (3)$$

However, HWE is often not a desirable assumption to make. Even if filtering is employed to remove loci which do not conform to HWE, the degree of conformity, and thus the degree of statistical bias in estimating $P(S_j)$, typically varies somewhat among sample groups. For example, in a sample of 100 genotypes with a minor allele frequency of 0.05, only five minor

alleles are expected and two out of three possible combinations of minor homozygotes and heterozygotes that produce that frequency will not deviate from HWE at $\alpha = 0.05$ according to an exact test (Wigginton et al., 2005). However, re-sampling these genotypes to a sample size of, say, ten genotypes will produce quite different $P(S_j)$ (roughly 0.7, 0.6, and 0.5, for purely heterozygotes, one homozygote and three heterozygotes, and two homozygotes and one heterozygote, respectively)..

To remedy these problems, we propose the following estimator of $\hat{P}(S_j)$:

$$\hat{P}(S_j) = 1 - \sum_{k=1}^{h} \frac{\binom{n_{kj}}{\gamma}}{\binom{n_j}{\gamma}} \quad (4)$$

where $\hat{P}(S_j)$ is given by the probability of exclusively drawing any k of h possible homozygote genotypes in sample group j given γ independent sampled genotypes (not gene copies) from the pool of observed genotypes. Here, n_{kj} is the number of observed homozygote genotypes of type k in sample group j and n_j is the total number of observed genotypes of all types, including heterozygotes. For example, in a sample group containing sequence data for ten genotypes at a single bi-allelic loci in which one genotype is *AA*, two are *AG*, and four are *GG*; h is two, n_{kj} will be one and four for the *AA* and *GG* genotypes, respectively, and n_j will be ten. A subsample of three individuals (γ = 3) will therefore yield $\hat{P}(S_j) \approx 0.76$, implying that a segregating site would be observed roughly 76% of the time if three samples were to be drawn from this sample group at random.

Interestingly, this method, like the richness method and related private allele rarefaction approaches, can account for varying amounts of missing data at specific loci in different sample

groups by varying γ across loci. Specifically, setting γ equal to the smallest observed n_j across all sample groups after accounting for missing data at each locus retains the highest amount of information possible at each locus while standardizing sample sizes. Note that setting $\gamma = n_j$ will result in either $\hat{P}(S_j) = 1$ or 0 depending on if the locus is segregating or not in the observed data without rarefaction. Thus, all sample groups other than that with the smallest n_j will be sampled to the size of the smallest group.

Applying the optimum γ at each locus across sample groups is particularly useful given that $E(N_S)$, or the expected total number of segregating sites across all loci, is often of specific interest as a measure of genetic diversity when comparing sample groups. Given that the expected number of segregating sites at locus q in population j, $E\left(N_{S_{jq}}\right)$, is equal to $\hat{P}(S_{jq})$, $E(N_S)$ can be calculated by summing $\hat{P}(S_{jq})$ across all Q loci:

$$E(N_S) = \sum_{q=1}^{Q} \hat{P}(S_{jq}) \tag{5}$$

with γ set accordingly for each locus. In this case, $0 \leq E\left(N_{S_{jq}}\right) \leq 1$ for all loci (and thus $0 \leq E(N_S) \leq Q$).

Usefully, under this framework each locus represents a single Bernoulli trial in which it can be observed to be segregating or not with probability $\hat{P}(S_{jq})$. As such, the variance of $\hat{P}(S_{jq})$ for each locus is given by

$$\sigma^2_{\hat{P}(S_{jq})} = \hat{P}(S_{jq}) \times \left(1 - \hat{P}(S_{jq})\right) \tag{6}$$

and, if each locus is independent, the variance of $E(N_S)$ is equal to the sum of $\sigma^2_{\hat{P}(S_{jq})}$ across all loci:

$$\sigma^2_{E(N_S)} = \sum_{q=1}^{Q} \sigma^2_{\hat{P}(S_{jq})} \quad (7)$$

Confidence and prediction intervals can then be derived using standard approaches for the sum of random, independent Bernoulli trials. When Q is large, for example, the distribution of $E(N_S)$ should be approach normal and confidence and prediction intervals can be derived using standard normal approximation using the equations:

$$CI_{N_S} = E(N_S) \pm Z\sqrt{\frac{\sigma^2_{E(n_S)}}{Q}} \quad (8)$$

$$PI_{N_S} = E(N_S) \pm Z\sqrt{\sigma^2_{E(n_S)}\left(1 + \left(\frac{1}{Q}\right)\right)} \quad (9)$$

where Z is given by the normal quantile function $Z = Q_X(1-\alpha)$ with $\mu = 0$ and $\sigma = 1$ for a desired confidence level α.

Empirical Validation

To validate equations 4 and 5, we simulated genotypic data for two populations with sizes 100 and 1000, each with 100 bi-allelic loci with minor allele frequencies spaced equally between 0.01 and 0.1. We added missing data to each population by assigning each locus a missing data rate R_{mq} from a uniform distribution such that $R_{mq} \sim U(0, 0.3)$, ensuring that overall allele frequencies in each population were maintained. We then used the methods described above to estimate $\hat{P}(S_{jq})$ and $E(N_S)$ and their variances given the number of sampled genotypes was

between 10 and 100. (*i.e.*, $\gamma = 10, 20, 30, ..., 100$). For evaluation, we also conducted between 1,000 and 10,000 random draws (i.e 1,000, 2,000, …, 10,000) for each γ from each locus in each population, then calculated the statistic $\hat{P}(S_{jq})$ and its variance for each locus empirically and $E(N_S)$ by summing across all 100 loci for each set of draws. We likewise calculated the variance of N_S directly across all sets of random draws.

To compare our calculated estimates to the empirical simulations, we used the implementation of the Agresti-Coull (Agresti & Coull, 1998) method from the R package binom (Dorai-Raj, 2022) to calculate 95% confidence intervals for the $P(S_{jq})$ parameter we observed in each simulation.

We implemented equations 4-7 in the R package snpR (Hemstrom & Jones, 2023). The "calc_seg_sites" function is set to automatically determine γ for each locus based on the sample group with the smallest sample size after accounting for missing data by default. To test this implementation, we compared the number of segregating sites present using both the mathematical implementation (Equations 4-5) and 100 simulated rarefaction draws using 5,000 randomly sampled SNPs from five populations from a previously published dataset of monarch butterflies (Hemstrom et al., 2022).

Results

The equations described here for calculating $\hat{P}(S_{jq})$ and $E(N_S)$ performed well. Individual $\hat{P}(S_{jq})$ values for each locus and rarefaction size (γ) from each population were within the 95% confidence intervals calculated from their respective simulations 95-100% (98.5% of the time in total across all γ, iteration counts, and initial population sizes), with no substantial bias across

minor allele frequency or the number of simulations (Figures 1-2), although confidence intervals calculated from simulations with higher γ values tended to contain $\hat{P}(S_{jq})$ slightly more often (Figure S1).

Calculations of the total number of segregating sites after rarefaction were likewise very similar to those observed via simulating random draws for any γ or iteration counts. Specifically, $E(N_S)$ values calculated via Equation 5 were very close to the mean number of segregating sites in each population after rarefaction (Figure 3, Figure S2). 95% prediction intervals were generally very close to observed 95% quantiles from simulations across γ and simulation counts but were consistently slightly overestimated on both ends. This is expected given that Equation 9 assumes normality; the actual distribution of N_S values are slightly non-normal across samples. Specifically, the medians, but not the means, of the observed samples are therefore slightly above $E(N_S)$ (Figure S2). Both point estimates of the number of segregating sites and 95% prediction intervals are similarly accurate across a wide range of both γ values and simulation counts (Figure 4). Setting the γ equal to the lowest observed sample size for each locus also performed well, with $E(N_S)$ values closely aligned with mean simulated N_S values (Figure S3).

Discussion

We present here a method to correct for the probability that a given locus would be observed as segregating following rarefaction to a given sample size. This method provides for the straightforward estimation of the total number of segregating sites which would be observed in a sample group at any reduced size, and therefore provides a way to standardize that metric across

samples from different sample groups and studies. We also show that our approach is unbiased by allele frequency or missing data variation across individual samples or loci.

Our estimator should be useful given that the number of segregating sites can provide additional information that can complement or enhance measures of genetic diversity that are based on allele frequencies (such as observed and expected heterozygosity). For example, populations that have experienced a recent population expansion will often carry an excess of low frequency variants (Gattepaille et al., 2013) caused by the recent increase in the overall rate at which mutations are produced in the population and the relative lack of time for any such new variants to drift to higher frequencies. The average expected heterozygosity across segregating sites may be lower (or at least lower than expected) in such cases than populations which have been demographically static, but they will carry far more segregating sites. A recent study in yellow perch provides an excellent example of such: several recently expanded populations show a relatively slight difference in heterozygosity in comparison to other, more demographically static populations but segregate at far more loci (Yin et al., n.d.)

In cases where biologically important conclusions can be drawn from the difference between heterozygosity or other allele frequency-based estimators of diversity and the number of segregating sites, it is particularly important that the latter is properly calculated across sample groups, since failing to correct for differences in sample size or data missingness could mask biologically interesting signals of demographic history and obscure a critical facet of overall genomic diversity. The $E(N_S)$ method we present here should therefore be useful for future studies of genetic diversity across disciplines. It is currently implemented and available for use via the function “calc_seg_sites()” in the R package “snpR” (Hemstrom & Jones, 2023).

Example code demonstrating the use of the use of this function using a “vcf” file (Danecek et al., 2011) can be found in Supplementary Example 1.

Funding

This work was supported in part by the National Science Foundation (grant numbers DEB-1856710 and OCE-1924505).

Data Availability

The scripts used to generate data, run the simulations, and produce the plots presented in this paper are available at: https://github.com/ChristieLab/seg_sites_rarefaction. The R package “snpR”, which implements the equations described here, is available from https://github.com/hemstrow/snpR.

References

Agresti, A., & Coull, B. A. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Allendorf, F. W. (1986). Heterozygosity and fitness in natural populations of animals. *Conservation Biology: The Science of Scarcity and Diversity*, 57–76.

Chapman, J. R., Nakagawa, S., Coltman, D. W., Slate, J., & Sheldon, B. C. (2009). A quantitative review of heterozygosity–fitness correlations in animal populations. *Molecular Ecology*, *18*(13), 2746–2765. https://doi.org/10.1111/j.1365-294X.2009.04247.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

David. (1998). Heterozygosity–fitness correlations: New perspectives on old problems. *Heredity*, *80*(5), 531–537. https://doi.org/10.1046/j.1365-2540.1998.00393.x

DeWoody, J. A., Harder, A. M., Mathur, S., & Willoughby, J. R. (2021). The long-standing significance of genetic diversity in conservation. *Molecular Ecology*, *30*(17), 4147–4154. https://doi.org/10.1111/mec.16051

Dorai-Raj, S. (2022). *binom: Binomial confidence intervals for several parameterizations* [Manual]. https://CRAN.R-project.org/package=binom

Frankham, R., Briscoe, D. A., & Ballou, J. D. (2002). *Introduction to conservation genetics*. Cambridge university press.

Fu, Y.-X. (2022). Variances and covariances of linear summary statistics of segregating sites. *Theoretical Population Biology*, *145*, 95–108. https://doi.org/10.1016/j.tpb.2022.03.005

Gattepaille, L. M., Jakobsson, M., & Blum, M. G. (2013). Inferring population size changes with sequence and SNP data: Lessons from human bottlenecks. *Heredity*, *110*(5), 409–419. https://doi.org/10.1038/hdy.2012.120

Hartl, D. L., Clark, A. G., & Clark, A. G. (1997). *Principles of population genetics* (Vol. 116). Sinauer associates Sunderland, MA.

Hedrick, P. W., & Kalinowski, S. T. (2000). Inbreeding depression in conservation biology. In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 31, Issue Volume 31, 2000, pp. 139–162). Annual Reviews. https://doi.org/10.1146/annurev.ecolsys.31.1.139

Hemstrom, W. B., Freedman, M. G., Zalucki, M. P., Ramírez, S. R., & Miller, M. R. (2022). Population genetics of a recent range expansion and subsequent loss of migration in monarch butterflies. *Molecular Ecology*, *31*(17), 4544–4557. https://doi.org/10.1111/mec.16592

Hemstrom, W., & Jones, M. (2023). snpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources*, *23*(4), 962–973. https://doi.org/10.1111/1755-0998.13721

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology*, *52*(4), 577–586. https://doi.org/10.2307/1934145

Hvilsom, C., Segelbacher, G., Ekblom, R., Fischer, M. C., Laikre, L., Leus, K., O'Brien, D., Shaw, R., & Sork, V. (2022). Selecting species and populations for monitoring of genetic diversity. *IUCN Publication*.

Kalinowski, S. T. (2004). Counting Alleles with Rarefaction: Private Alleles and Hierarchical Sampling Designs. *Conservation Genetics*, *5*(4), 539–543. https://doi.org/10.1023/B:COGE.0000041021.91777.1a

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Funk, W. C. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences*, *118*(48), e2104642118. https://doi.org/10.1073/pnas.2104642118

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., Cao, H.-F., Zhang, B.-W., Yeh, C.-F., Hung, C.-M., Hung, H.-Y., Yang, M.-Y., Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences*, *116*(6), 2152–2157. https://doi.org/10.1073/pnas.1813597116

Moritz, C. (2002). Strategies to Protect Biological Diversity and the Evolutionary Processes That Sustain It. *Systematic Biology*, *51*(2), 238–254. https://doi.org/10.1080/10635150252899752

Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., Karchner, S. I., Hahn, M. E., Nacci, D., Oleksiak, M. F., Crawford, D. L., & Whitehead, A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*(6317), 1305–1308. https://doi.org/10.1126/science.aah4993

Schmidt, C., Hoban, S., Hunter, M., Paz-Vinas, I., & Garroway, C. J. (2023). Genetic diversity and IUCN Red List status. *Conservation Biology*, *37*(4), e14064. https://doi.org/10.1111/cobi.14064

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585 LP – 595.

Visser, M. E. (2008). Keeping up with a warming world; assessing the rate of adaptation to climate change. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1635), 649–659. https://doi.org/10.1098/rspb.2007.0997

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Yin, X., Schraidt, C., Sparks, M. M., Euclide, P. T., Ruetz III, C. R., Höök, T. O., & Christie, M. R. (submitted). Rapid, simultaneous increases in the effective sizes of adaptively divergent yellow perch (Perca flavescens) populations. *Proceedings of the Royal Society B: Biological Sciences*.

Figure Legends:

Figure 1: The expected probabilities of observing a segregating site at each locus ($\hat{P}(S_{j,q})$) for loci with different minor allele frequencies. Probabilities are derived from Equation 4 for each locus for population sizes of 250 and 2500, rarefacted to sample sizes of either 10 or 100 (corresponding to γ = 10 or 100). Since loci were not simulated in HWE, points vary to a small degree due to variation in genotype frequencies for a given minor allele frequency. Each locus had a random, independent percentage of missing data between 0% and 30%. Points are colored depending on estimates were contained within a 95% confidence intervals (marked with error bars) based on 10,000 simulated rarefaction trials for each minor allele frequency at each sample size.

Figure 2: Difference between the mathematically expected probability that a locus segregates after rarefaction ($\hat{P}(S_{j,q})$) and the observed probability of segregation ($P(S_{j,q})$) following simulated rarefaction across different rarefaction sizes (γ), minor allele frequencies, and number of simulated rarefaction events. Across all tested conditions, only 1.51% of parameter estimates fell outside the 95% CIs obtained via simulation.

Figure 3: The distribution of the total number of segregating sites (N_S) observed for 10,000 replicate simulated trials, rarefacted to either 10 or 100 samples (corresponding to either γ = 10 or 100, respectively) for starting population sizes of either N = 250 or N = 2500. The mathematically expected number of segregating sites $E(N_S)$ and 95% prediction are shown with solid yellow and dashed blue lines, respectively, for each distribution.

Figure 4: Trends in the distribution of the total number of segregating sites (N_S) observed following rarefaction across a range of rarefaction sizes (γ) and number of trials for starting population sizes of either N = 250 or N = 2500. Horizontal lines on each distribution indicate 95% quantile limits. The mathematically expected number of segregating sites ($E(N_S)$) and 95% prediction intervals are shown in yellow error bars to the right of each distribution.