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

William Hemstrom1†, Mark R. Christie1,2†

**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 Hemstrom; Department of Biological Sciences, Purdue University; 915 W. State St., West Lafayette, Indiana 47907; [hemstrow@gmail.com](mailto:hemstrow@gmail.com), 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](mailto:christ99@purdue.edu), 765-494-2070

# Abstract

Unlike the suite of genetic estimators which rely on direct measurements of allele frequencies (such as expected heterozygosity, nucleotide diversity, and Tajima’s θ), estimators for the numbers of polymorphic loci or alleles, such as private allele counts or allelic richness, are not as commonly used or well developed for single nucleotide polymorphism (SNP) data. One such estimator which does not rely on estimates of allele frequencies is the number of segregating sites or SNPs (*S*), calculated as the number of nucleotide sites that have more than one allele across the genome. *S* 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 allele frequency dependent estimators such as expected heterozygosity. However, segregating site counts are rarely adjusted to correct for unequal sample sizes or differences in missing data among populations or sample groups, and when they are, typically fail to account for deviations from Hardy-Weinberg Proportions (HWP). Here, we introduce an unbiased estimator for the number of segregating sites expected in a sample group following rarefaction (*S’*) and we use simulated data sets to illustrate that *S’* allows for accurate comparisons of the number of segregating sites among multiple sample groups with varied sample sizes and deviations from HWP. We also demonstrate that *S’* can be incorporated into estimates of Watterson’s θ to reduce bias in samples not in Hardy-Weinberg Proportions.

**Keywords:** Watterson’s θ, rarefaction, segregating sites, genomics, genetic diversity

# Introduction

Estimating 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). Functional genetic diversity, or the diversity of alleles, genotypes, and segregating sites that contribute to phenotypic variation within a population (Frankham et al., 2002), is the basic foundation upon which natural selection acts, and thus affects the speed and degree of genetic adaptation (Kardos et al., 2021). Genetic diversity is thus strongly correlated with fitness both at the individual and population level, a trend that holds even in neutral sites with no obvious functional ties (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 require measurements of allele frequencies and those that do not. Allele frequency dependent statistics, such as expected heterozygosity (), F-statistics (, , and ; (Weir & Cockerham, 1984), nucleotide diversity (), and Tajima’s (Tajima, 1989) all rely principally on and seek to measure allele frequencies across surveyed loci; whereas other statistics, such as Watterson’s (Watterson, 1975), counts of alleles per locus, the number of observed segregating sites (from which Watterson’s is derived), and private allele counts do not. These different measures therefore focus on different aspects of genetic diversity, which can be informative given that 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 (also commonly referred to as Tajima’s π) and Watterson’s that can be used to both detect selection and population demographic changes due to the way that those forces act on the balance of allele frequencies within populations (Tajima, 1989). For example, a population which has many polymorphic loci and a moderate expected heterozygosity may have undergone different historical demographic processes than a population with the same heterozygosity but many more polymorphic loci (Gattepaille et al., 2013). Alternatively, regions of the genome which have recently undergone an incomplete selective sweep often contain many loci with rare alleles whereas regions which are subject to balancing selection may have similar numbers of polymorphic loci but far higher average minor allele frequencies (Pennings & Hermisson, 2006). 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) was historically one of the most prevalent allele frequency independent measures of genetic diversity, used primarily for datasets of microsatellite and other heavily polyallelic markers. It is less useful in data composed of single-nucleotide polymorphisms (SNPs) because SNP datasets are often filtered to remove non-biallelic loci (Hemstrom et al., 2024), and thus allele counts per locus vary little among sample groups and are correspondingly less informative. By contrast, the location and number of segregating sites per population, *S*, 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 *S* 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 numbers of samples will tend to have a higher number of segregating sites, since low-frequency variants will be observed on average 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).

The bias in *S* that can occur due to unequal sample sizes or varying amounts of missing data is problematic given that one of the core uses of the parameter *S* is in estimating θ, the mutation-rate scaled effective population (*Ne*) size of a population in mutation-drift equilibrium (Wang, 2005), equal to in diploid populations (where *μ* is the population mutation rate). There are several measures of θ—all of which measure genetic diversity—such as the average number of pairwise differences between haplotypes, as in the above mentioned Tajima’s θ (Tajima, 1989), the number of different haplotypes, and *S*. All θ estimators (and *S*) are essentially descriptors of and can be calculated from the site frequency spectra (SFS), and are thus particularly useful for developing tests of neutrality and detecting evolutionary and demographic processes which shape the SFS (Achaz, 2009). More detailed explorations of θ estimators, and the tests derived from them, can be found in Achaz (2009) and Wang (2005).

Watterson’s θ however, is particularly relevant in discussions of *S.* It is an estimator calculated by determining the expected value of *S* given a sample of *K* independent gene copies with *S* segregating sites at a locus with free recombination between loci at mutation-drift equilibrium (Watterson, 1975). As such, Watterson’s θ already corrects for differences in sample sizes; however, the assumption of gene-copy independence upon which Watterson’s θ depends is a potential source of bias given that it is not met whenever genotypes do not follow Hardy-Weinberg Proportions (HWP). This is a common scenario, particularly in managed populations or those of conservation concern where effective populations sizes may be very small and/or inbreeding common (Charlesworth & Willis, 2009; Hedrick & Kalinowski, 2000; Keller & Waller, 2002; Willi et al., 2022). Specifically, Watterson’s θ is an overestimate of θ in inbred populations, given that expected probability of a site segregating for a given *K* should be lower when there are less heterozygotes, and an underestimate in outbred populations where the inverse is true. Other established methods for correcting for differences in sample sizes when estimating *S*, such as taking the sum of an SFS projected down to a smaller sample size according to Gutenkunst et al. (2009) also assume allelic independence and should therefore have the same issue. Bias in Watterson’s θ could therefore lead to incorrectly assuming over-optimistic *Ne* in inbred populations or generate skewed signals of population growth or directional selection via Tajima’s D, among other issues.

To help address this problem, and because *S* is itself an inherently useful estimator of genetic diversity (both genome-wide and for identifying particular regions of interest - *e.g.*, mutation hot spots), we here 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. Importantly, these estimators do not assume Hardy-Weinberg proportions and can be integrated into downstream estimates of θ. We show that these estimators are highly accurate via comparison to re-sampling using simulated data. We have implemented both estimators 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 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 across sample groups) can be estimated for a given sample group by summing the probability of observing each of unique alleles using the counts of those alleles in sample group and the total sample size in that sample group . Allelic richness is calculated by comparing the number of possible ways to draw gene copies without sampling allele () to the total number of possible combinations of gene copies that can be drawn (); the inverse of this () is therefore the probability of observing allele in sample group , and the sum of this value across all alleles gives the expected number of alleles observed at a locus in sample group , (Hurlbert, 1971; Kalinowski, 2004):

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

where is the allele frequency of allele in sample group . 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:

However, HWP is often not a desirable assumption to make. Even if filtering is employed to remove loci which do not conform to HWP, the degree of conformity, and thus the degree of statistical bias in estimating , typically varies somewhat among sample groups. For example, in a sample of 50 genotypes with a minor allele frequency of 0.05, only five minor alleles are present and two out of three possible combinations of minor homozygotes and heterozygotes that produce that frequency will not deviate from HWP at 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 (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 :

where is given by the probability of exclusively drawing any of possible homozygote genotypes in sample group given independent sampled individuals/genotypes(not gene copies) from the pool of observed individuals/genotypes. Here, is the number of observed homozygote genotypes of type in sample group and 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* istwo, will be one and four for the *AA* and *GG* genotypes, respectively, and will be ten. A subsample of three individuals ( = 3) will therefore yield ≈ 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 across loci. Specifically, setting equal to the smallest observed 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 = will result in either = 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 will be sampled to the size of the smallest group.

Applying the optimum at each locus across sample groups is particularly useful given that , 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 in population , , is equal to , , the expected total *S* in a population, hereafter referred to as *S’*, can be calculated by summing across all loci:

with set accordingly for each locus. In this case, for all loci (and thus ).

Usefully, under this framework each locus represents a single Bernoulli trial in which it can be observed to be segregating or not with probability . As such, the variance of for each locus is given by

and, if each locus is independent, the variance of *S’* is equal to the sum of across all loci:

Confidence and prediction intervals can then be derived using standard approaches for the sum of random, independent Bernoulli trials. When is large, for example, the distribution of *S’* should be approach normal and confidence and prediction intervals can be derived using standard normal approximation using the equations:

where is given by the normal quantile function with and for a desired confidence level .

*S’* can then be incorporated into calculations of Watterson’s θ () if desired using a slight modification of equation 1.4 from Watterson (1975):

where and is the number of individuals/genotypes sampled when calculating *S’*. If using a variable per locus to account for variable missing data rates, the mean of should be used instead.

## Simulations and Empirical Validation

To validate equations 4 and 5, we simulated genotypic data for two populations with N = 250 and 2500, each with 100 bi-allelic loci with minor allele frequencies spaced equally between 0.01 and 0.5. We added missing data to each population by assigning each locus a missing data rate from a uniform distribution such that , ensuring that overall allele frequencies in each population were maintained. We then used the methods described above to estimate , *S’*,and their variances given a number of sampled genotypes between 10 and 100. (*i.e.*, γ , 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 from each locus in each population, then calculated the statistic and its variance for each locus empirically and *S* by summing across all 100 loci for each set of draws. We likewise calculated the variance of *S* directly across all sets of random draws. We also repeated this procedure with data derived from coalescent simulations to test the validity of and *S’* estimates with data with a more realistic SFS (see Supplementary Methods).

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 parameter observed in each simulation.

To determine the potential bias of our *S’* estimates to Watterson’s θ (Watterson, 1975) and *S* estimates derived from the sum of a projected SFS (Gutenkunst et al., 2009), we calculated each using inbred, neutral, and outbred data derived from or simulated based on data from a previously published study on monarch butterflies (*Danaus plexippus*; Hemstrom et al., 2022; see Supplementary Methods). We calculated F*IS* (Weir & Cockerham, 1984), Watterson’s θ, *S* via a projected SFS, and *S’* using γ = 25 for each population, then subsampled each to 25 individuals 1000 times and determined the empirical *S* for each subsample. To plot Watterson’s θ and *S* on a common scale and to determine the degree of bias in each statistic depending on deviation from HWP, we divided *S,* θ, and *S’* by their mean values across populations in each case.

We implemented equations 4-7 in the R package snpR ( Hemstrom & Jones, 2023). The “calc\_seg\_sites” function is set to automatically determine 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 and resulted in high accuracy and precision. Individual values for each locus and rarefaction size (γ) from each population were within the 95% confidence intervals calculated from their respective simulations 95-100% (97.8% 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 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 or iteration counts. Specifically, *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 and simulation counts but were consistently slightly overestimated on both ends. This is expected given that Equation 9 assumes normality; the actual distribution of *S* values were slightly non-normal across samples and as a consequence, the median, but not the mean, *S* values of the observed samples were slightly above *S’* (Figure S2). Both point estimates of the number of segregating sites and 95% prediction intervals were similarly accurate across a wide range of both values and simulation counts (Figure 4). Results were very similar when using data derived from coalescent simulations (Figure S3-4). Setting the γ equal to the lowest observed sample size for each locus also resulted *S'* values closely aligned with mean simulated *S* values (Figure S5).

Watterson’s θ estimates were biased in both inbred and outbred populations in comparison to *S* values obtained via direct rarefaction whereas *S’* or adjusted Watterson’s θ estimates calculated using *S’* were not (Figure 5). Specifically, Watterson’s θ was overestimated in inbred populations (those with a positive *F*IS) and underestimated in outbred populations. Surprisingly, *S* values derived from projected SFS were not consistently biased by HWP divergence but were generally less accurate that *S’* estimates: raw (un-normalized for statistic/statistic comparison) *S’* estimates were closer to subsampled *S* values than those produced via a projected SFS across all populations (Figure S6).

# Discussion

We present here an unbiased method to correct for the probability that a given locus will segregate following rarefaction to a given sample size. This 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 and, unlike other approaches, is robust to deviations from Hardy-Weinberg genotypic proportions.

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 populations than in those 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., 2024). Our estimator of total *S* is particularly useful in populations where most loci are skewed away from HWP, such as in inbred or outbred populations. Calculating Watterson’s θ from *S’* for use in such cases is straightforward.

We note, however, that since our estimator relies on accurate genotype calls per locus, it is not suitable for use in low-coverage datasets where discrete genotype calls are unreliable and where genotype likelihoods may be preferable. In such cases, summing across an SFS calculated using appropriate low-coverage approaches may be a preferable avenue for calculating *S* (as well as the various derived θ estimators; see Korneliussen et al., 2013), although such methods generally assume allelic independence when calculating Watterson’s θ and require matched sample sizes to avoid bias when calculating *S*. Our approach does assume that loci segregate independently (i.e. are not linked) in that each locus is drawn independently during rarefaction, which may not always be true. While filtering out linked loci is an option in such cases (Hemstrom et al., 2024), new methods which retain linked sites without bias would increase power and are thus promising avenues for future work. That said, *S’* is still likely to be unbiased in such instances as long as many large haplotypic blocks are not out of HWP in a consistent way, as is supported by the lack of bias in *S’* results from our coalescent simulations (Figures S3 and S4). As a further filtering-related caution, we warn against a strong minor allele frequency filter when comparing *S’* values across populations given the different impacts minor allele frequency filters can have on populations with different site frequency spectra, although some filter permissive filtering is probably often justified to remove sequencing errors (Hemstrom et al., 2024).

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 *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

Achaz, G. (2009). Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics*, *183*(1), 249–258. https://doi.org/10.1534/genetics.109.104042

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

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*(11), 783–796. https://doi.org/10.1038/nrg2664

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

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*(10), 1–11. https://doi.org/10.1371/journal.pgen.1000695

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., Grummer, J. A., Luikart, G., & Christie, M. R. (2024). Next-generation data filtering in the genomics era. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-024-00738-6

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

Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, *17*(5), 230–241. https://doi.org/10.1016/S0169-5347(02)02489-8

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima’s D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*(1), 289. https://doi.org/10.1186/1471-2105-14-289

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

Pennings, P. S., & Hermisson, J. (2006). Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLOS Genetics*, *2*(12), e186. https://doi.org/10.1371/journal.pgen.0020186

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

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1395–1409. https://doi.org/10.1098/rstb.2005.1682

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

Willi, Y., Kristensen, T. N., Sgrò, C. M., Weeks, A. R., Ørsted, M., & Hoffmann, A. A. (2022). Conservation genetics as a management tool: The five best-supported paradigms to assist the management of threatened species. *Proceedings of the National Academy of Sciences*, *119*(1), e2105076119. https://doi.org/10.1073/pnas.2105076119

Yin, X., Schraidt, C., Sparks, M., Euclide, P., Hoyt, T., Ruetz, C., Höök, T., & Christie, M. (2024). Rapid, simultaneous increases in the effective sizes of adaptively divergent yellow perch (<em>Perca flavescens</em>) populations. *bioRxiv*, 2024.04.21.590447. https://doi.org/10.1101/2024.04.21.590447

# Figure Legends:

**Figure 1:** The expected probabilities of observing a segregating site at each locus () 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 = 10 or 100). Since loci were not simulated in Hardy-Weinberg proportions, points vary to a small degree due to variation in genotype frequencies for a given minor allele frequency. Each locus was simulated with a random, independent percentage of missing data between 0% and 30%. Points are colored depending on estimates were contained within 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:** Differences between the mathematically expected probability that a locus segregates after rarefaction () and the observed probability of segregation () following simulated rarefaction across different rarefaction sizes (), minor allele frequencies, and number of simulated rarefaction events. Across all tested conditions, only 2.2% of parameter estimates fell outside the 95% CIs obtained via simulation.

**Figure 3:** The distribution of the total number of segregating sites (*S*) observed for 10,000 replicate simulated trials, rarefacted to either 10 or 20 samples (corresponding to either = 10 or 20, respectively) for starting population sizes of either *N* = 250 or *N* = 2500. The mathematically expected number of segregating sites (*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 (*S*) observed following rarefaction across a range of rarefaction sizes () 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 (*S’*) and 95% prediction intervals are shown in yellow error bars to the right of each distribution.

**Figure 5:** Bias in θ or *S* values from Watterson’s θ, projected SFS summations, or *S’* calculations vs simulated rarefaction depending on F*IS*. Points depict estimated values from each method, distributions depict empirical estimates based on repeated direct rarefaction via sampling. Outbred (, left), neutral (, center), and inbred (, right) populations were simulated from phased monarch butterfly data, then randomly subsampled to 25 individuals 1000 times to calculate empirical *S* distributions. Points and distributions were standardized via division by the global mean across all populations to show the direction of bias depending on *F*IS and allow for direct comparison of *S* and θ values on a similar scale; values above the empirical distributions indicate overestimates and values below underestimates. Points (not distributions) are offset on the *x* axis for clarity—each value is plotted for the specific *F*IS of the corresponding population.