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

# Abstract

Population geneticists use a wide range of approaches to measure genetic diversity across the genome. While measures which explicitly measure allele frequencies (such as expected heterozygosity, nucleotide diversity, and Tajima’s D) are well developed for single nucleotide polymorphism (SNP) data, alternative estimators which provide measures that reflect the number of polymorphic loci, 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 missing data among populations or sample groups. Here, we introduce an unbiased estimator for the number of segregating sites expected in a population following rarefaction, which allows for an unbiased approach for comparing the number of segregating sites among multiple groups of samples.

# 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 in part controls population adaptability (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 thus critical for monitoring the health of populations and for predicting their future ability to respond to changing environmental conditions (Lai et al., 2019; Reid et al., 2016; Visser, 2008); as such genetic diversity is a key component of biodiversity (Hvilsom et al., 2022; Schmidt et al., 2023).

Estimators of population genetic diversity are diverse, 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 (), F-statistics (, , and ; Weir & Cockerham, 1984), nucleotide diversity (), and Tajima’s (Tajima, 1989) all rely principally on and seek to measure the frequency of the alleles across surveyed loci; whereas other statistics, such as Watterson’s (Watterson, 1975), counts of alleles per loci, the number of observed segregating sites (from which Watterson’s is derived), and private allele counts do not. This is not a trivial difference: evolutionary forces do not act equally on the number of loci/alleles in a population and the frequency of those alleles. Tajima’s D, for example, is a powerful and broadly used statistic fundamentally based on the difference between Tajima’s and Watterson’s 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, for example, may have undergone different historical demographic processes than a population with the same number of polymorphic loci but a far lower heterozygosity. As such, both types of measures 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 microsatellite and other datasets utilizing heavily polyallelic markers, it is less so in single-nucleotide polymorphism (SNP) datasets. SNP datasets are usually biallelic by design, and thus allele counts per locus vary little populations and are therefore less informative. The location and number of segregating sites per population is a useful alternative in SNP datasets, especially given the speed of calculation and ease with which comparisons can be made across samples and genomic locations [@?CITE].

However, 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 higher number of segregating sites, since low-frequencies variants will be observed on average much more frequently in sample groups with more sequenced gene copies than in those with few. This 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 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 loci would be observed to be segregating under a reduced sample size and 2) the expected total number of segregating sites across all loci within a population. We show that these estimators are highly accurate via comparison to simulated re-sampling using simulated data. These estimators are currently implemented and available in the snpR R package via the function calc\_seg\_sites (W. Hemstrom & Jones, 2023), where they can be automatically run for all populations with either automatic, per-locus optimized control of target sample size or globally via manual control.

# 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 across populations) can be estimated for a given population by summing the probability of observing each of unique alleles using the counts of those alleles in the population and the total sample size in that population .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 population , and the sum of this value across all alleles gives the expected number of alleles observed at a locus in population , (Hurlbert, 1971; Kalinowski, 2004):

The expected number of segregating loci in a population for a draw of gene copies can be derived similarly. For a locus to be segregating in population , 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 ( 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 population . 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, 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 , 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 according to an exact test (Wigginton et al., 2005). However, re-sampling these to, 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 population given independent sampled genotypes(not gene copies) from the pool of observed genotypes. Here, is the number of observed homozygote genotypes of type in population 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. Note that and will be half the value of their equivalents and for diploid species, one third for triploids, and so on.

Interestingly, this method, like the richness method and related private allele rarefaction approaches, can smoothly account for varying amounts of missing data at specific loci in different populations by varying across loci. Specifically, setting equal to the smallest observed across all populations 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 populations 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 populations. Given that the expected number of segregating sites at locus in population , , is equal to , 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 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 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 .

## 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 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 and and their variances given the number of sampled genotypes was between 10 and 100. (*i.e.*, , 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 by summing across all 100 loci for each set of draws. We likewise calculated the variance of 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 parameter we observed in each simulation.

We implemented equations 4-7 in the R package snpR (W. 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, although that behavior is adjustable. To test this implementation, we compared the number of segregating sites present using both the default snpR implementation 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 methods described here for calculating and performed well. Individual values for each locus and from each population were within the 95% confidence intervals calculated from their respective simulations ~95% of the time, with no substantial bias across minor allele frequency or the number of simulations (Figures 1-2), although confidence intervals calculated from simulations with higher values tended to contain slightly more often (Figure S1). Note that variation along the generally correlated minor allele frequency/ axes visible in Figure 1 are due to variations in genotype frequencies in the simulated data for a given minor allele frequency. Given that the values estimated using equation 4 track the confidence intervals calculated from simulations, this is accounted for adequately.

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, 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 values are slightly non-normal across samples. Specifically, the medians, but not the means, of the observed samples are therefore slightly above (Figure S2). Both point estimates of the number of segregating sites and 95% prediction intervals are similarly accurate across a wide range of both values and simulation counts (Figure 4). The snpR implementation using automatically defined values for each locus also performed well, with values closely aligned with mean simulated values (Figure S3).

# Discussion

We present here an 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 population sample at any reduced size, and therefore provides a way to standardize that metric across samples from different populations. 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 an excellent counterpoint to 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., 2024).

In cases like 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 populations, 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 method we present here should therefore be useful for future studies of genetic diversity, and is currently available for use via the function “calc\_seg\_sites()” in the R package “snpR” (W. Hemstrom & Jones, 2023). An example demonstrating its use can be found in the Supplementary Example 1.

**Funding**

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

# 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

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.

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

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

# 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 either *N* = 10 or *N* = 100, corresponding to = 10 or = 100. Points vary from the trend lines due to the addition of missing data: each locus has a random, independent percentage of missing data which varies between 0 and 30. Points are colored depending on if they were within a 95% confidence interval based on 10,000 simulated rarefaction trials of each population to each *N*, which is marked for each point with an error bar.

**Figure 2:** Trends in difference 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. Loci are colored depending on if the expected probabilities were within the observed 95% confidence intervals.

**Figure 3:** The distribution of the total number of segregating sites () observed for 10,000 replicate simulated rarefaction trials to either *N* = 10 or *N* = 100 (corresponding to either = 10 or = 100, respectively) for starting population sizes of either *N* = 250 or *N* = 2500. The mathematically expected number of segregating sites and 95% prediction intervals on that number 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 () 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 note 95% quantile limits. The mathematically expected number of segregating sites () and 95% prediction intervals on that number are shown in yellow error bars to the right of each distribution.