**Unbiased estimation of the number of segregating sites across populations with unequal sample sizes via rarefaction**

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; [whemostro@purdue.edu](mailto:whemostro@purdue.edu), 765-496-0062

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 population genetic diversity. While measures which incorporate allele frequencies (such as observed and expected heterozygosity, nucleotide diversity, Tajima’s D, and so on) are well developed for single nucleotide polymorphism (SNP) data, there are only a few allele frequency independent estimators which are broadly in use. The number of segregating sites (or SNPs) is one of the more prevalent of these; however, unlike allelic richness or private allele frequencies, segregating site counts are rarely adjusted to correct for unequal sample sizes or missing data rates between populations. We here propose a corrected 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 between multiple populations.

# Introduction

Estimating the degree of genetic diversity has long been a critically important and widespread practice in conservation genetics (DeWoody et al., 2021). 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 selection acts and thus in part controls population adaptability (Kardos & Luikart, 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).

Estimators of population genetic diversity are themselves extremely diverse, but essentially fall into two categories: those that are depend on allele frequencies and those that do not. Allele frequency dependent statistics, such as expected heterozygosity (), the F-statistics (, , and )(Weir & Cockerham, 1984), nucleotide diversity (), and Tajima’s (Tajima, 1989) all rely principally on the frequency of the alleles across any surveyed loci; whereas allele frequency independent statistics, such as Watterson’s (Watterson, 1975), counts of alleles per loci, the number of observed segregating sites, 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.

The average allele count per locus (which is usually corrected to allelic richness (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 number of segregating sites per population is a useful alternative in SNP datasets given the large number of possible segregating sites that these datasets often query [@?CITE].

However, segregating sites counts per population are problematic whenever sample sizes are not equal across the populations under comparison at all loci. This can occur either directly due to unequal numbers of individuals or indirectly due to unqual proportions of missing data across populations. This problem also effects estimates of allele counts per locus and private allele numbers, 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 the probability that any given loci would be observed to be segregating under a reduced sample size and 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 (Hemstrom & Jones, 2023), where they can be automatically run for all populations either with 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 the same framework used to calculate 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 . This is done 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 at each locus (the locus is at Hardy-Weinburg Equlibrium, HWE) and 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 . However, in finite samples draws are conducted with replacement, and so binomial coefficients must instead be used to determine the probability of drawing only 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 between populations. 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 genoytpes should 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) as we will see below.

To remedy these problems, we propose the following estimator of :

where the 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. 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 or across loci. Both can be set to one less than the smallest observed or , the highest values at which rarefaction can be applied within a population, across all populations after accounting for missing data, and can thus vary across loci without bias. Setting either value to or will instead return the observed allele diversity or segregating site status, respectively.

This is particularly useful given that or the expected total number of segregating sites, is often of specific interest as a measure of genetic diversity when comparing populations. Given that the expected number of segregating sites at a specific 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 .

## Emperical 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 . For comparison, we also conducted 10,000 random draws of size from each loci in each population, then calculated 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 each empirical . We also used normal approximation to calculate 95% confidence and prediction intervals for and using and the variance directly observed from the simulations, respectively.

# Results

The methods described here for calculating and performed well. Individual values for each locus were within the confidence intervals derived from the simulated values for 97 and 96% of loci from the and populations, respectively (Figure 1). Note that variation along the generally correlated minor allele frequency/ axes is due to variations in genotype frequencies in the simulated data for a given minor allele frequency. values estimated using equation 4 account for this adequately.

values estimated with equation 5 were similarly accurate and were within the 95% confidence intervals produced using the simulated for both population sizes (Figure 2). Likewise, the 95% prediction intervals calculated using via equation 7 contained 96.1 and 95.3% of the simulated values for and , respectively (Figure 2).

# Discussion

We present here an method to correct for the probability that a given loci would be observed as segregating 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 population sample at any reduced size, and therefore provides a way to standardize that metric across samples from different populations. We also show that this approach is unbiased by allele frequency or missing data variation across individual samples or loci.

Our estimator should 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 actually 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 are segregating at far more loci (Schraidt et al., 2023).

In cases like this, it is particularly important that the number of segregating sites is being 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 a useful for future studies of genetic diversity, and is currently available for use via the function “calc\_seg\_sites()” in the R package “snpR” (Hemstrom & Jones, 2023).

# 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>

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>

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>

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., & Luikart, G. (2021). The Genetic Architecture of Fitness Drives Population Viability during Rapid Environmental Change. *The American Naturalist*, 0. <https://doi.org/10.1086/713469>

Schraidt, C. E., Ackiss, A. S., Larson, W. A., Rowe, M. D., Höök, T. O., & Christie, M. R. (2023). Dispersive currents explain patterns of population connectivity in an ecologically and economically important fish. *Evolutionary Applications*, *n/a*(n/a). <https://doi.org/10.1111/eva.13567>

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585 LP–595. <http://www.genetics.org/content/123/3/585.abstract>

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>

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, *76*(5), 887–893. https://doi.org/<https://doi.org/10.1086/429864>

# Figure Captions:

Figure 1: The expected probabilities of observing a segregating site at a locus for loci with different minor allele frequencies. Probabilities are derived from Equation 4 for each locus *a* for population sizes of 100 (left) and 1000 (right), each rarefacted to . Each locus has a random, independent percentage of missing data which varies between 0 and 30. Points are colored depending on if they fell within a 95% confidence level based on 10,000 random re-samplings of each population to 30 individuals.

Figure 2: The expected total number of segregating sites from populations of either 100 or 1,000 sampled to 30 individuals as calculated via Equation 5 vs that observed from 10,000 random re-samplings to 30 individuals. 95% Confidence and prediction intervals for are shown, calculated according to Equations 8 and 9. Observed values from each re-sample are colored depending on if they fell within the estimated 95% prediction interval. The mean across all re-samples is marked with a triangle, colored depending on if it fell within the estimated 95% confidence interval.