manuscript

17 October 2023

# Abstract

# Introduction

# Methods

Outline:

* Background on methods for private alleles, richness
* Solution for segregating sites
  + Derivation
  + Can account for both different sample sizes and missingness across loci and populations.
* Simulations
  + Sampling approach, missingness, etc

## 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 , (**hurlbertNonconceptSpeciesDiversity1971?**; **kalinowskiCountingAllelesRarefaction2004?**):

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.

To remedy this, I 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 genotypes in population where the total number of observed genotypes of all types is . 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. In contrast, confidence intervals for individual values can be achieved via standard approaches for single Bernoulli samples of size .

## Emperical Simulations

To validate Eqns