# manuscript

31 October 2023

## Abstract

## Introduction

## 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 $g$ across populations) can be estimated for a given population $j$ by summing the probability of observing each $i$ of $m$ unique alleles using the counts of those alleles in the population $N_{ij}$ and the total sample size in that population $N_j$. This is done 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 population $N_j$, and the sum of this value across all $m$ alleles gives the expected number of alleles observed at a locus in population $j$, $\alpha_g^j$ (**hurlbertNonconceptSpeciesDiversity1971?**; **kalinowskiCountingAllelesRarefaction2004?**):

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

The expected number of segregating loci in a population for a draw of $g$ gene copies can be derived similarly. For a locus $i$ to be segregating in population $j$, 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 $N$ 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:

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

where $f_{ij}$ is the allele frequency of allele $i$ in population $j$. 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:

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

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 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 $\alpha = 0.05$ according to an exact test (Wigginton et al., 2005). However, re-sampling these to, say, ten genoytpes should 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) as we will see below.

To remedy these problems, I propose the following estimator of $P(S_j)$:

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

where the $P(S_j)$ is given by the probability of exclusively drawing any $k$ of $h$ possible homozygote genotypes in population $j$ given $\gamma$ 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 population $j$ and $n_j$ is the total number of observed genotypes of all types. Note that $n_j$ and $\gamma$ will be half the value of their equivalents $N_j$ and $g$ 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 $g$ or $\gamma$ across loci. Both can be set to one less than the smallest observed $N_j$ or $n_j$, 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 $N_j$ or $n_j$ will instead return the observed allele diversity or segregating site status, respectively.

This is particularly useful given that $E(N_S)$ 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 $q$ in population $j$, $E(N_{S_{jq}})$, is equal to $P(S_{jq})$, $E(N_S)$ can be calculated by summing $P(S_{jq})$ across all $Q$ loci:

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

with $\gamma$ set accordingly for each locus. In this case, $0 \leq E(N_{S_{jq}}) \leq 1$ 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 $P(S_{jq})$. As such, the variance of $P(S_{jq})$ for each locus is given by

$$\sigma^2_{P(S_{jq})} = P(S_{jq}) \times (1 - P(S_{jq}))$$

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

$$\sigma^2_{E(n_S)} = \sum_{q=1}^{Q} \sigma^2_{P(S_{jq})}$$

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

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

and

$$PI_{N_S} = E(N_S) \pm Z\sqrt{\sigma^2_{E(n_S)}(1 + (1/Q))}$$

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 $\alpha$.

### *Emperical Validation*

To validate equations 4 and 5, I 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. I added missing data to each population assigning each locus a missing data rate $R_{mq}$ from a uniform distribution such that $R_{mq} \sim U(0, .3)$, ensuring that overall allele frequencies in each population were maintained. I then used the methods described above to estimate $P(S_{jq})$ and $E(N_S)$ and their variances given $\gamma = 30$. For comparison, I also conducted 10,000 random draws of size $\gamma$ from each loci in each population, then calculated $P(S_{jq})$ and its variance for each locus empirically and $E(N_S)$ by summing across all 100 loci for each set of draws. I likewise calculated the variance of $N_S$ directly across all sets of random draws.

To compare my calculated estimates to the empirical simulations, I used the “exact” method from the R package `binom` (**rajBinomBinomialConfidence2022?**) to calculate 95% confidence intervals for each empirical $P(S_{jq})$. I also used normal approximation to calculate 95% confidence and prediction intervals for $E(N_S)$ and $N_S$ using $\sigma^2_{E(n_S)}$ and the variance directly observed from the simulations, respectively.

## Results

The methods described here for calculating $P(S_j)$ and $E(N_S)$ performed well. Individual $P(S_{jq})$ values for each locus were within the confidence intervals derived from the simulated values for 97 and 96% of loci from the $n = 100$ and $n = 1000$ populations, respectively (Figure 1). Note that variation along the generally correlated minor allele frequency/$P(S_{jq})$ axes is due to variations in genotype frequencies in the simulated data for a given minor allele frequency. $P(S_{jq})$ values estimated using equation X account for this adequately.

$E(N_S)$ values estimated with equation X were similarly accurate and were within the 95% confidence intervals produced using the simulated $N_S$ for both population sizes. Likewise, the 95% prediction intervals calculated using $\sigma^2_{E(n_S)}$ via equation X contained 96.1 and 95.3% of the simulated $N_S$ values for $n = 100$ and $n = 1000$, respectively.

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