PROGRAM NOTE

HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness

STEVEN T. KALINOWSKI

Department of Ecology, Montana State University, Bozeman, MT 59717, U.S.A.

Abstract

The number of alleles in a sample (allelic richness) is a fundamental measure of genetic diversity. However, this diversity measure has been difficult to use because large samples are expected to contain more alleles than small samples. The statistical technique of rarefaction compensates for this sampling disparity. Here I introduce a computer program that performs rarefaction on private alleles and hierarchical sampling designs.

Keywords: alleles, allelic richness, estimation, genetic diversity, rarefaction

Received 14 June 2004; revision received 3 September 2004; accepted 6 October 2004

Two statistics frequently used to measure genetic diversity include gene diversity (expected heterozygosity) and allelic richness (number of alleles). Of the two, gene diversity is probably used more frequently. This may be because the allelic richness of a sample is affected by the size of a sample — large samples are expected to have more alleles than small samples. Rarefaction is an infrequently used statistical method that accounts for this effect to produce unbiased estimates of allelic richness (Hurlbert 1971; Smith & Grassle 1977; Leberg 2002). Kalinowski (2004) recently extended the rarefaction method to count private alleles and to accommodate hierarchical sampling designs that had two levels. In this note, I describe algorithms implemented by the computer program, HP-RARE, to estimate allelic richness and private allelic richness for hierarchies with an arbitrary number of levels. I will concentrate on the algorithm used to perform the calculations. See Kalinowski (2004) for a more general discussion of rarefaction.

Rarefaction is a statistical method for estimating how many alleles are expected in a sample of specified size taken from a taxon. By 'taxon', I mean either a population or a set of populations that have been placed in the same category. By 'sample', I mean either a sample of genes taken from a population, or a set of samples taken from populations within a taxon. For our purposes, sample size refers not only to the number of genes sampled from a population, but also the number of samples within a taxon.

Let the vector N represent the actual sample size taken from a taxon, and let the vector S represent the size of a balanced sample for which an estimate of the allelic richness of the taxon is desired (see succeeding disucussion for an example). A balanced sample is one that has the same number of samples at each level in the hierarchy and the same number of genes within the populations sampled. The allelic richness of a taxon, A^{S}_{taxon}, is the number of alleles expected in a sample of size S taken from the taxon. In order to estimate A^{S}_{taxon}, we must calculate the probability, $P^{\mathrm{S}}_{i,taxon}$, that each allele ($i = 1, 2 \ldots m$) would be present in a subsample of size S taken without replacement from the actual sample of size N collected from a taxon

$$A^{\mathrm{S}}_{taxon} \triangleq \sum_{i=1}^{m} P^{\mathrm{S}}_{i,taxon} \qquad (1)$$

The previous points deserve repeating. A^{S}_{taxon} is the number of alleles expected in sample of size S taken from actual populations. A^{S}_{taxon} is calculated from the probabilities of alleles being present in subsamples of size S taken from the samples collected from the populations. Succeeding formulae will desribe how to calculate $P^{\mathrm{S}}_{i,taxon}$.

The private allelic richness of a taxon, Π^{S}_{taxon}, is the expected number of private alleles in a sample of size S taken from the taxon. It is estimated by

$$\Pi^{\mathrm{S}}_{taxon} \triangleq \sum_{i=1}^{m} U^{\mathrm{S}}_{i,taxon} \qquad (2)$$

where $U^{\mathrm{S}}_{i,taxon}$ is the probability that the i^{th} allele will only be found in the subsample from the taxon indicated (formulae for calculating $U^{\mathrm{S}}_{i,taxon}$ are given below). Both of these estimators (Equations 1 and 2) are unbiased and have minimum variance (see Smith & Grassle 1977 for a sketch of a proof).

Correspondence: Steven T. Kalinowski. E-mail: skalinowski@montana.edu

The probability, $P^S_{i,taxon}$ in Equation 1 is most easily calculated by

$$P^S_{i,taxon} = 1 - Q^S_{i,taxon} \tag{3}$$

where $Q^S_{i,taxon}$ is the probability that a sample of size S from a taxon does not contain the i^{th} allele (formulae for calculating $Q^S_{i,taxon}$ are given below). The probability $U^S_{i,taxon}$ is equal to the probability of the i^{th} allele being found in a taxon multiplied by the probability, $Q^S_{i,taxon'}$, that the allele is not present in any other subsample

$$U^S_{i,taxon} = P^S_{i,taxon}Q^S_{i,taxon'} \tag{4}$$

Calculating these probabilities is not difficult, but their formulae are lengthy and use set notation relatively difficult to interpret. Therefore, I will describe algorithms to estimate allelic richness. This is most easily done with a heuristic example. Consider a survey of the genetic variation in a widespread North American species (Table 1). In this survey, samples were collected from three countries: the United States, Canada, and Mexico. Multiple regions were sampled in each country. Multiple states/provinces were sampled from each region, and multiple sites were sampled from each state. Assume that the sampling was not balanced. Rarefaction will be used to estimate how many alleles are expected in samples that contain 20 genes per site, three sites per state, two states per region, and three regions per country, i.e., S = {20, 3, 2, 3}.

Table 1 Sampling locations for a heuristic example of a hierarchal survey of genetic variation

Country	Region	State/province	Site
United States	Northwest	Montana	Kalispell Missoula Butte GreatFalls
		Idaho	—
		Washington	—
		Oregon	—
	Southwest	California	—
		Arizona	—
		New Mexico	—
	Northeast	—	—
	Southeast	—	—
Canada	Atlantic	—	—
	Pacific	—	—
	Arctic	—	—
Mexico	—	—	—

I will begin by describing how to estimate allelic richness and then deal with private allelic richness. As Equations 1 and 3 show, estimating allelic richness requires calculating the probability that each allele is not found in a subsample taken without replacement from the sample collected from a taxon, $Q^S_{i,taxon}$. For taxa at the lowest level of the hierarchy, e.g. Kalispell, the formula of Hurlbert (1971) is used

$$Q^S_{i,(Kalispell)} = \frac{\binom{N_{(Kalispell)} - N_{i,(Kalispell)}}{S_0}}{\binom{N_{(Kalispell)}}{S_0}} \tag{5}$$

where $N_{(Kalispell)}$ is the number of genes sampled at the site, $N_{i,(Kalispell)}$ is the number of times the i^{th} allele was observed in the sample, and S_0 is the number of genes for which rarefaction will be done. Calculating Q for higher levels in the hierarchy, e.g. $Q_{(Montana)}$, begins by enumerating all the possible sets of sites that can be drawn from Montana. Let $X^S_{Montana}$ be the set of all these combinations

$$X^S_{Montana} = \begin{Bmatrix} \text{(Kalispell, Missoula, Butte),} \\ \text{(Kalispell, Missoula, Great Falls),} \\ \text{(Kalispell, Butte, Great Falls),} \\ \text{(Missoula, Butte, Great Falls)} \end{Bmatrix} \tag{6}$$

Each set has three sampling sites because we are rarefacting to three sites per state. Each of the four combinations of three sites is equally likely. The probability that a specific combination of subsamples (e.g. Kalispell, Missoula, and Butte) does not contain allele i, $Q^S_{i,(Kalispell,Missoula,Butte)}$ is equal to

$$Q_{i,(Kalispell,Missoula,Butte)} = Q^S_{i,(Kalispell)}Q^S_{i,(Missoula)}Q^S_{i,(Butte)} \tag{7}$$

$Q^S_{i,(Montana)}$ is then equal to

$$\begin{aligned} Q^S_{i,(Montana)} = {} & \frac{1}{4}Q^S_{i,(Kalispell)}Q^S_{i,(Missoula)}Q^S_{i,(Butte)} \\ & + \frac{1}{4}Q^S_{i,(Kalispell)}Q^S_{i,(Missoula)}Q^S_{i,(Great\ Falls)} \\ & + \frac{1}{4}Q^S_{i,(Kalispell)}Q^S_{i,(Butte)}Q^S_{i,(Great\ Falls)} \\ & + \frac{1}{4}Q^S_{i,(Missoula)}Q^S_{i,(Butte)}Q^S_{i,(Great\ Falls)} \end{aligned} \tag{8}$$

Now that we have calculated $Q^S_{i,(Montana)}$, the allelic richness for Montana is calculated from Equations 3 and 1. Q is calculated the same way for the higher levels in the hierarchy. For example, $Q^S_{i,(Northwest)}$ is calculated by from $Q^S_{i,(Montana)}$, $Q^S_{i,(Idaho)}$, $Q^S_{i,(Washington)}$, and $Q^S_{i,(Oregon)}$.

Calculating the probabilities needed to estimate private allelic richness is a little trickier. Kalinowski (2004) presented an approach based on complete enumeration of all possible combinations of samples. This is difficult for large studies, because the number of combinations becomes prohibitively large. Here I present a more efficient approach, starting with the private allelic richness for Kalispell, one of the taxa on the lowest level of the sampling hierarchy in the heuristic example.

Calculating the probability that an allele is unique to the Kalispell sample, $U^{S}_{i,(Kalispell)}$ is done in steps, starting with the probability that an allele is found in a Kalispell subsample but none of the other subsamples from Montana. I represent this probability as $U^{S}_{i,(Kalispell/Montana)}$. Let $X^{S}_{(Montana/Kalispell)}$ be the set of all combinations of three sites from Montana that contain the Kalispell site

$$X^{S}_{(Montana/Kalispell)} = \left\{ \begin{matrix} \text{(Kalispell, Missoula, Butte),} \\ \text{(Kalispell, Missoula, Great Falls),} \\ \text{(Kalispell, Butte, Great Falls)} \end{matrix} \right\} \quad (9)$$

$U^{S}_{i,(Kalispell/Montana)}$ is then calculated

$$\begin{aligned} U^{S}_{i,(Kalispell/Montana)} = & \frac{1}{3} P^{S}_{i,(Kalispell)} Q^{S}_{i,(Missoula)} Q^{S}_{i,(Butte)} \\ & + \frac{1}{3} P^{S}_{i,(Kalispell)} Q^{S}_{i,(Missoula)} Q^{S}_{i,(Great\ Falls)} \quad (10) \\ & + \frac{1}{3} P^{S}_{i,(Kalispell)} Q^{S}_{i,(Butte)} Q^{S}_{i,(Great\ Falls)} \end{aligned}$$

We continue our calculation of $U^{S}_{i,(Kalispell)}$ by finding the probability that the i^{th} allele is unique to the subsamples taken from the northwestern United States. There are now three different possible sets of subsamples that could be taken from the Northwest Region that include Montana,

$$\begin{aligned} X^{S}_{(NW/Montana)} = & [\text{(Montana, Washington) (Montana, Oregon)} \\ & \text{(Washington, Idaho)}] \end{aligned} \quad (11)$$

The probability that the i^{th} allele is unique to the Kalispell among the subsamples taken from the Northwest region is

$$\begin{aligned} U^{S}_{i,(Kalispell/NW)} = & \frac{1}{3} U^{S}_{i,(Kalispell/Montana)} Q^{S}_{i,(Washington)} \\ & + \frac{1}{3} U^{S}_{i,(Kalispell/Montana)} Q^{S}_{i,(Oregon)} \quad (12) \\ & + \frac{1}{3} U^{S}_{i,(Kalispell/Montana)} Q^{S}_{i,(Idaho)} \end{aligned}$$

Similarly, the probability that the i^{th} allele is unique to the Kalispell among the subsamples taken from the United States is

$$\begin{aligned} U^{S}_{i,(Kalispell/US)} = & \frac{1}{3} U^{S}_{i,(Kalispell/NW)} Q^{S}_{i,(NE)} \\ & + \frac{1}{3} U^{S}_{i,(Kalispell/NW)} Q^{S}_{i,(SE)} \quad (13) \\ & + \frac{1}{3} U^{S}_{i,(Kalispell/NW)} Q^{S}_{i,(SW)} \end{aligned}$$

And last, the probability that the i^{th} allele is unique to the Kalispell subsample among all the subsamples in the study is

$$U^{S}_{i,(Kalispell)} = U^{S}_{i,(Kalispell/US)} Q^{S}_{i,(Mexico)} Q^{S}_{i,(Canada)} \quad (14)$$

Estimating the private allelic richness for other levels in the hierarchy is done in a similar method. For example, calculating the private allelic richness of the Montana taxon begins by calculating the probability that an allele is present in Montana, but not in any of the other states in the Northwest.

HP-RARE is a computer program that uses the above approaches to estimate allelic richness and private allelic richness for hierarchical study designs. In addition, HP-RARE will output the size of each sample, the number of alleles observed in each sample, and the expected heterozygosity for each sample.

HP-RARE was written using Microsoft Visual Basic.Net and runs on the Microsoft Windows operating system. The accuracy of the calculations was checked by using simulation to draw subsamples from samples. HP-RARE uses GENEPOP files to read genotypic data (Raymond & Rousset 1995; http://wbiomed.curtin.edu.au/genepop/). Hierarchies are read from text files. Output is to text files. HP-RARE is available for download from www.montana.edu/kalinowski.

References

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52**, 577–586.

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics*, in press.

Leberg PL (2002) Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology*, **11**, 2445–2449.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Smith W, Grassle JF (1977) Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283–292.