*Conservation Genetics* **5:** 539–543, 2004.



# Counting alleles with rarefaction: Private alleles and hierarchical sampling designs

Steven T. Kalinowski
*Department of Ecology, Montana State University, Bozeman, MT 59717, USA (fax: +1-406-994-3232; e-mail: skalinowski@montana.edu*

Received 5 January 2004; accepted 16 February 2004

*Key words:* allelic richness, estimation, genetic diversity, rarefaction, sample size, species diversity, species richness

**Abstract**

The number of alleles (allelic richness) in a population is a fundamental measure of genetic variation, and a useful statistic for identifying populations for conservation. Estimating allelic richness is complicated by the effects of sample size: large samples are expected to have more alleles. Rarefaction solves this problem. This communication extends the rarefaction procedure to count private alleles and to accommodate hierarchical sampling designs.

## Introduction

The amount of genetic diversity within populations is a fundamental parameter in evolutionary and conservation biology. High levels of genetic variation are expected to increase the potential of populations to respond to selection and to maintain the health of individuals. The simplest measure of genetic diversity at a locus is the number of alleles (allelic richness). A related statistic, the number of unique alleles in a population (private allelic richness) is a simple measure of genetic distinctiveness.

The primary disadvantage of using allelic richness as a measure of genetic diversity is that it is highly dependent on sample size: large samples are expected to contain more alleles than small samples. Similarly, more alleles are expected to be found in a region sampled many times than in a region sampled few times. Private allelic richness has the same problem: large samples are expected to have more private alleles than small ones. On the other hand, intensive sampling of genetically similar populations may reduce the number of alleles private to any population. Therefore, a region that has been sampled intensively may appear to have fewer private alleles than a region sampled less intensively.

These problems have a straightforward statistical solution: rarefaction can be used to compensate for differences in sample size and number. Rarefaction has a long history of use in the ecological literature for estimating species diversity[1] (see Hurlbert 1971 for its first use; see Simberloff 1979; Gotelli and Colwell 2001 for reviews) but is used sporadically by conservation geneticists (Leberg 2002). Here, I extend the rarefaction technique to accommodate hierarchical sampling designs and to count private alleles.

## Allelic richness

In this paper I use 'allelic richness' to indicate a measure of genetic diversity in either a sample or a population. The allelic richness of a sample, $a_g$, is the expected number of alleles that a sample would

have had if the sample size had been $g$ genes instead of $N (g \leq N)$. The allelic richness of a population, $\alpha_g$, is the expected number of alleles in a sample of $g$ genes taken from a population. In most conservation genetic applications, $a_g$ is of much less interest than $\alpha_g$.It turns out that $a_g$ is useful to estimate $\alpha_g$, but the distinction between the statistics is important – in realistic applications, the variance of $\alpha_g$ will be greater than $a_g$ (see below).

The parameter $g$ can be considered a standardized sample size. For example, if a study aims to compare allelic richness between several populations, then $g$ must be less than the smallest sample size. However, $g$, is more than a standardized sample size; it indicates how sensitive $\alpha_g$ is to the presence or absence of rare alleles. When $g$ is large, rare alleles will have a big effect upon $\alpha_g$. When $g$ is small, rare alleles have little effect upon $\alpha_g$. When $g$ takes its smallest useful value, $g = 2$, the allelic richness of a population is equal to the expected heterozygosity ($H$) of the population plus one, i.e. $H + 1 = \alpha_2$ (Smith and Grassle 1977). This last point is important, for it shows that expected heterozygosity is a special case of allelic richness.

I now present formulae for several measures of allelic richness, starting with the number of alleles expected in a single sample. Consider a survey of genetic variation in which the sample size varies across populations. Let $N_{ij}$ represent the number of copies of the $i$th allele in the sample from the $j$th population; let $N_j$ represent the total number of genes sampled from the $j$th population; and let $m$ represent the total number of distinct alleles observed at the locus ($N_j = \sum_{i=1}^{m} N_{ij}$).

Rarefaction was invented to calculate $a_g$, in order to compare the allelic richness of samples having different sizes (Hurlbert 1971). Hurlbert showed that the expected number of alleles in a sample of $g$ genes selected at random (without replacement) from a sample of $N_j$ genes is equal to

$$a_g^{(j)} = \sum_{i=1}^{m} P_{ijg}, \tag{1}$$

where

$$Q_{ijg} = \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}} = \prod_{u=0}^{g-1} \frac{N_j - N_{ij} - u}{N_j - u} \tag{2a}$$

and

$$P_{ijg} = 1 - Q_{ijg}. \tag{2b}$$

Smith and Grassle (1977) showed that $a_g^{(j)}$ is a minimum variance, unbiased estimate of the allelic richness of the $j$th population, $\alpha_g^{(j)}$,

$$\hat{\alpha}_g^{(j)} = a_g^{(j)} = \sum_{i=1}^{m} P_{ijg}, \tag{3}$$

where $\hat{}$ indicates that $\hat{\alpha}_g^{(j)}$ is an estimate of $\alpha_g^{(j)}$.

Equations (2a) and (2b) form the foundation for future formulae, so deserve explanation. The denominator of the middle term in Equation (2a), $\binom{N_j}{g}$, is the total number of combination of $g$ genes can be made from $N_j$ genes (if sampling is without replacement. The numerator of the middle term in Equation (2a), $\binom{N_j - N_{ij}}{g}$, is the number of combinations of $g$ genes that do not include allele $i$ (if sampling is without replacement). $Q_{ijg}$, therefore, is the probability that a sample of $g$ genes taken from a sample of $N_j$ genes will not contain allele $i$. The right-hand term in Equation (2a) is a convenient expression for calculating $Q_{ijg}$ (Comps et al. 2001). $P_{ijg}$ (Equation (2b)) is the probability that such a sample will contain allele $i$.

Private allelic richness, $\pi$, is a convenient measure of how distinct a population is from other populations. (In the ecological literature, private alleles correspond to endemic species). Let $\pi_g^{(j)}$ represent the number of private alleles expected in a sample from the $j$th population if $g$ genes are sampled from each of $J$ populations. A minimum variance, unbiased estimator of $\pi_g^{(j)}$ is obtained using the approach of Smith and Grassle (1977)

$$\hat{\pi}_g^{(i)} = \sum_{i=1}^{m} \left[ P_{ijg} \left( \prod_{\substack{j'=1 \\ j' \neq j}}^{J} Q_{ij'g} \right) \right]. \tag{4}$$

Equation (4) is easily deconstructed. $P_{ijg}$ is the probability that a sample of $g$ genes taken from the $j$th sample contains at least one copy of allele $i$; $\prod_{\substack{j'=1 \\ j' \neq j}}^{J} Q_{ij'g}$ is the probability that samples of $g$ genes taken from all of the other samples do not contain allele $i$.

Now, I extend the rarefaction technique to accommodate hierarchical sampling. When

sampling is hierarchical, the number of populations sampled per region must be standardized, as well as the number of genes per population. I retain the notation above, and add the following. Let $S_k$ represent the number of populations that have been sampled in the $k$th region; let $R$ represent the number of regions in the study, $J = \sum_{k=1}^{R} S_k$; and let $r$ represent the standardized number of populations per region. Let $C_r$ represent the total number of ways that $r$ samples can be sampled from the $R$ regions, $C_r = \prod_{k=1}^{R} \binom{S_k}{r}$. Let $X_k$ represent the set of populations from region $k$. Let $Y_{kcr}$ represent the $c$th set among the $\binom{S_k}{r}$ cardinality-$r$ subsets of $X_k$.

If $r$ populations are sampled per region, and $g$ genes are sampled per population, the expected number of alleles in region $k$ is estimated by

$$\hat{\alpha}_{r,g}^{(k)} = \sum_{i=1}^{m} \left[ \frac{1}{\binom{S_k}{r}} \sum_{c=1}^{\binom{S_k}{r}} \left( 1 - \prod_{j \in Y_{kcr}}^{S_k} Q_{ijg} \right) \right]. \quad (5)$$

This is also a minimum variance, unbiased estimator (Smith and Grassle 1977). El Mousadik and Petit (1996) presented an alternative to Equation (5), but round off errors in their method produce a modest amount of bias (S. Kalinowski unpublished), so Equation (5) is recommended.

Counting private alleles in hierarchical sampling designs requires defining private. I use 'private alleles' to describe alleles that are observed in only one population, and 'regionally private alleles' to describe alleles that are observed in only one region. The expected number of private alleles in region $k$, when $g$ genes are sampled from $r$ populations per region, is estimated by

$$\widehat{\prod}_{r,g}^{(k)} = \sum_{i=1}^{m} \left\{ \frac{1}{C_r} \sum_{c_1=1}^{S_1} \sum_{c_2=1}^{S_2} \cdots \sum_{c_R=1}^{S_R} \left[ \left( \sum_{j \in Y_{kc_k}r} \left( P_{ijg} \prod_{\substack{j' \in Y_{kc_k}r \\ j' \neq j}} Q_{ij'g} \right) \right) \prod_{\substack{k'=1 \\ k' \neq k}}^{R} \prod_{j \in Y_{kc_{k'}r}} Q_{ijg} \right] \right\} \quad (6)$$

The expected number of regionally private alleles in region $k$, when $g$ genes are sampled from $r$ populations per region, is estimated by

$$\hat{\rho}_{g,r}^{(k)} = \sum_{i=1}^{m} \left\{ \frac{1}{C_r} \sum_{c_1=1}^{S_1} \sum_{c_2=1}^{S_2} \cdots \sum_{c_R=1}^{S_R} \left[ \left( 1 - \prod_{j \in Y_{kc_k}r} Q_{ijg} \right) \prod_{\substack{k'=1 \\ k' \neq k}}^{R} \prod_{j' \in X_{k'c_{k'}r}} Q_{ij'g} \right] \right\}. \quad (7)$$

Again, these are minimum variance, unbiased estimates (Smith and Grassle 1977).

Monte Carlo simulation can be also be used to estimate the five measures of allelic richness described above (e.g., King et al. 2001). For example, $\alpha_g^{(j)}$ can be estimated by randomly drawing $g$ genes (without replacement) from a sample taken from the $j$th population. If this is done many times, the average number of alleles in the simulated samples is an estimate of $a_g^{(j)}$ and thus of $\alpha_g^{(j)}$. Monte Carlo estimates of allelic richness for hierarchical sampling designs are obtained by randomly sampling $r$ populations per region as well as $g$ genes per sample. This approach will be useful when $\binom{S_k}{r}$ or $C_r$ is large. Monte Carlo estimation could also be used to accommodate sampling hierarchies that include more than two levels. Equations (3)–(5) could be extended to accommodate hierarchies with more than two levels, but the Monte Carlo approach would probably be easier to implement.

## An example: A reanalysis of human microsatellite data

An example shows that rarefaction can substantially change estimates of allelic richness and private allelic richness. I present the microsatellite data of Jin et al. (2000) as an example. I selected this data as an example because: (1) the data set is especially large (64 loci), (2) the species is well-studied (humans), and (3) the methods used by the authors are typical. Jin et al. (2000) genotyped 64 microsatellite loci in 11 populations among five regions: Africa, Asia, Europe, North and South America, and Oceania. Both the number of populations sampled

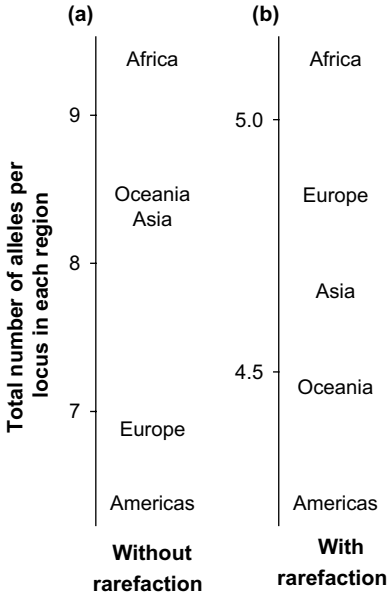


*Figure 1.* The allelic richness in each region in the microsatellite data of Jin et al. (2000): with (b) and without (a) rarefaction (Equation (5)). Rarefaction included 10 genes per sample, and one sample per region.

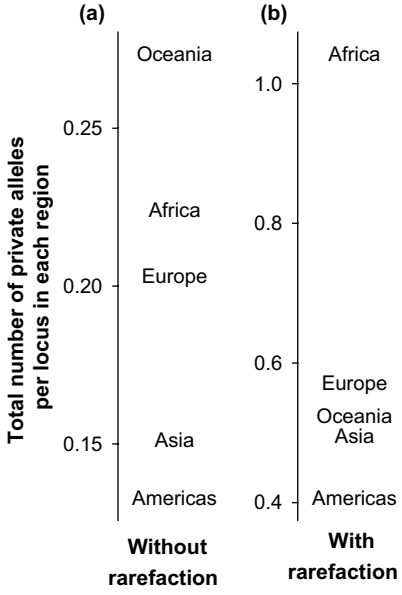


*Figure 2.* The total number of private alleles in each region in the microsatellite data of Jin et al. (2000): with (b) and without (a) rarefaction. Rarefaction included 10 genes per sample, and one sample per region (Equation (7)).

per region varied (1–3) as well as the number of genes sampled per population (10–26). Two of the goals of the study were to identify which continent had the most genetic variation and the most unique alleles. Jin et al. did not use rarefaction to compensate for variation in sampling effort.

In the raw data of Jin et al., the samples from Oceania (New Guinea and Australia) had the second highest total number of alleles (Figure 1a) and the highest number of private alleles (Figure 2a). However, these samples were also the two largest in the study. When rarefaction was used to standardize samples to 10 genes per population and one population per region, the samples from Oceania had relatively fewer alleles (Figures 1b and 2b). These results agree very well with a human study with larger sample sizes (Rosenberg et al. 2002).

## Tests of statistical significance

A few options are available to test the statistical significance of differences in the measures of allelic richness described above. A conservative approach for pairwise tests is to use a sign test across loci. For example, in the data of Jin et al. (2000), the African samples had more unique alleles (after rarefaction) than the European samples at 41 of 64 loci. Using the sign test, the one-tailed probability for a result this extreme under the null hypothesis that each region had the same number of unique alleles is 0.016.

Randomization of samples among regions can also be used to test the statistical significance of the observed difference between regions. This approach would not be useful for the data of Jin et al., because too few populations were sampled for each region. Parametric tests of statistical significance are available for simple (i.e. non-hierarchal study designs) comparisons of allelic richness and private allelic richness (Tipper 1979). These tests require estimating the sampling variance of $\hat{\alpha}_g$. This can be done only if the sample size is greater than $2g$ or if multiple samples have been taken from the same population (Smith and Grassle 1977; Tipper 1979).

The sampling variance of $\hat{\alpha}_g$, $\mathrm{Var}(\hat{\alpha}_g)$, is easily confused with a related variance – the variance in the number of alleles present in samples of $g$ genes subsampled from a larger sample. A heuristic example illustrates the difference. Assume that 22 genes (11 individuals) were sampled from a large population that has many alleles, and an estimate of $\alpha_{20}$ is desired. Equation (3) gives an unbiased

estimate of $\alpha_{20}$. As noted above, Smith and Grassle (1977) provide formulae for estimating $\mathrm{Var}(\hat{\alpha}_{20})$. Formulae have also been developed for calculating the sampling variance of the number of alleles in samples of $g$ genes taken from the complete sample (Heck et al. 1975), a quantity that Leberg (2002) estimated by simulation. This variance will be much smaller than $\mathrm{Var}(\hat{\alpha}_{20})$ – but only because samples of 20 genes will contain most (if not all) the alleles present in the complete sample of 22 genes. However, this sampling variance is seldom relevant. What is relevant is how different $\hat{\alpha}_{20}$ is likely to be if an independent sample was collected from the population (Simberloff 1979). Leberg's (2002) discussion of the effect of rarefaction upon the precision of estimates of allelic richness, therefore, is misleading.

## Discussion

The number of alleles or private alleles present in populations is useful for many conservation genetic applications. For example, allelic richness is useful for identifying populations that deserve special management. Petit et al. (1998) compared the allelic richness of populations of an endangered tree species to identify genetically diverse populations so that these populations could be protected. Alternatively, populations with low allelic richness might receive special management. Measures of allelic richness are also useful for inferring the evolutionary histories of populations (e.g., Castric and Bernatchez 2003). Expected heterozygosity can also be used as a measure of genetic diversity for these applications, but it is less sensitive to the presence of rare alleles. This is a disadvantage is many circumstances. For example, population bottlenecks reduce allelic richness faster than heterozygosity. This principle has been used to test for reductions in population size (e.g., Cornuet and Luikart 1996). The new measures of allelic richness presented here will allow conservation geneticists to use measures of allelic richness that fit their specific needs.

## Acknowledgement

I would like to thank N. Rosenberg for developing the notation used above.

## Notes

1. Rarefaction was invented by ecologists counting the number of species (species richness) in samples collected from different communities. In the equations and discussion below, I have translated this literature into generic terms for clarity.

## References

Castric V, Bernatchez L (2003) The rise and fall of isolation by distance in the anadromous brook charr (*Salvelinus fontinalis Mitchill*). *Genetics*, **163**, 983–996.

Comps B, Gömöry D, Letouzey J, Thiébaut B, Petit RJ (2001) Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics*, **157**, 389–397.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

El Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. *Theor. Appl. Genet.*, **92**, 832–839.

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, **4**, 379–391.

Heck KL, Van Belle G, Simberloff D (1975) Explicit calculation of the rarefaction diversity measurement and the determination of the sufficient sample size. *Ecology*, **56**, 1459–1461.

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

Jin L, Baskett ML, Cavalli-Sforza LL, Zhivotovsky LA, Feldman MW, Rosenberg NA (2000) Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. *Ann. Hum. Genet.*, **64**, 117–134.

King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA (2001) Population structure of Atlantic salmon (*Salmo salar L.*): a rangewide perspective from microsatellites. *Mol. Ecol.*, **10**, 807–821.

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

Petit R, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conserv. Biol.*, **12**, 844–855.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

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

Simberloff D (1979) Rarefaction as a distribution-free method of expressing and estimating diversity. In: *Ecological Diversity in Theory and Practice*. (eds. Grassle JF, Patil GP, Smith W, Taillie C), International Co-operative Publishing House, Fairland, Maryland.

Tipper JC (1979) Rarefaction and rarefiction – the use and abuse of a method in paleoecology. *Paleobiology*, **5**, 423–434.