ARES: software to compare allelic richness between uneven samples

E E van Loon, D F R Cleary, Cécile Fauvelot

Molecular Ecology Notes (2007) 7, 579–582 doi: 10.1111/j.1471-8286.2007.01705.x

PROGRAM NOTE

ARES: software to compare allelic richness between uneven samples

E. E. VAN LOON,* D. F. R. CLEARY*† and C. FAUVELOT*‡
*Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, PO Box 94062, 1098 SM Amsterdam, The Netherlands, †National Museum of Natural History, 'Naturalis', PO Box 9517, 2300 RA Leiden, The Netherlands

Abstract

Allelic richness is one of the most basic measures of genetic diversity. Its calculation is, however, still problematic because estimates depend on sample size. This paper describes an R library that calculates mean allelic richness with confidence bounds for a range of sample sizes. It takes a file in GENEPOP format as input, or alternatively a binary data matrix with columns representing different individuals and rows representing different alleles. The output is tabular as well as graphical. Unlike existing tools, ARES does extrapolate beyond the sample size and provides confidence bounds for these predictions.

Keywords: allelic richness, extrapolation, rarefaction, uneven sample size

Received 27 September 2006; revision accepted 8 January 2007

The preservation of genetic diversity within populations appears to be critical to species survival under changing environmental conditions (Allendorf 1986; Pease *et al*. 1989; Petit *et al*. 1998; Saccheri *et al*. 1998; Spielman *et al*. 2004; Reusch *et al*. 2005; Vellend & Geber 2005). Allelic richness (the number of alleles) is increasingly used as a means to estimate levels of genetic diversity in evolutionary and conservation biology, in order to identify endangered breeding units, populations or species. The primary disadvantage of using allelic richness for this purpose is that its value is highly dependent on sample size, making comparisons between samples of different sizes difficult (e.g. Stout & Vandermeer 1975). One way to avoid this problem is to use rarefaction (El Mousadik & Petit 1996; Kalinowski 2004). With rarefaction, the sample size at which allelic richness among samples is compared is that of the smallest sample. This, however, can lead to relatively inaccurate results and leaves valuable information unused (Gotelli & Colwell 2001). In the context of species richness estimation, it has been suggested to extrapolate small samples rather than rarefying larger samples in order to avoid this problem (Colwell & Coddington 1994; Melo *et al*. 2003). Nevertheless, present user-friendly software for estimating allelic richness that extrapolates and provides confidence intervals around estimates is lacking. This is an important hiatus because obtaining reliable estimates of allelic richness may be both time-consuming and expensive (especially at large scales and for rare species), and it may not always be feasible to provide well-replicated samples that can be used for subsequent hypothesis testing with, for example, *t*-tests or ANOVAS. Moreover, confidence intervals are more informative than the simple result of hypothesis tests (reject or do not reject H_0) since they provide a range of plausible values for the unknown parameter.

Recently, Colwell *et al*. (2004) and Mao *et al*. (2005) have introduced a new model to extrapolate an incidence-based accumulation curve: a combination of a moment-based richness estimator for calculating richness up to the number of individuals present in the sample and a truncated binomial mixture model for extrapolating beyond that. Using this model, allelic richness estimates can be obtained with confidence bounds for any sample size, and tests can be confidently performed to investigate whether two samples of differing size differ with respect to allelic richness. We call our implementation of the aforementioned model ARES (Allelic Richness EStimation), a model that closely follows the notation and definitions given by Colwell *et al*. (2004). Input for ARES core functions is a binary data matrix with the individuals columnwise, and the alleles row-wise. Briefly, the data for h individuals

Correspondence: E. E. van Loon, CBPG/IBED, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands. Fax: +31 20 5257431; E-mail: vanloon@uva.nl
‡Present address: Environmental Science, University of Bologna at Ravenna, Via S. Alberto 163, I-48100 Ravenna, Italy.

are expressed as an *S*-by-*h* allele–individual incidence matrix consisting of the presence indicators Z_{ij}, which is 1 if the *i*th allele is present in the *j*th individual; and 0 if the *i*th allele is absent in the *j*th individual. The model is based on two statistical assumptions: (i) an allele has the same probability of being present in each individual (ϕ_i), and (ii) the Z_{ij} are independent, given the probability of encountering allele *i*, over all *i* and *j*.

The allelic accumulation function, which gives the expected number of alleles observed in *h* individuals, is the sum of the probabilities, across the alleles, that each allele is not absent from all *h* individuals:

$$\tau(h) = \sum_{i=1}^{S}(1 - (1 - \phi_i)^h) \tag{1}$$

where ϕ_i is the probability that an allele is present in an individual *i*. Alleles with identical presence probabilities can be grouped. For *G* of these groups eqn 1 becomes

$$\tau(h) = S\sum_{k=1}^{G}\pi_k(1 - (1 - \psi_k)^h) \tag{2}$$

where ψ_k is the presence probability of the *k*th group of alleles, and π_k is the number of individuals in the *k*th group divided by the total number of individuals.

In any realistic situation, one takes a random sample of *H* individuals from the population (hence an *S*-by-*H* allele-individual incidence matrix). The observed data matrix consists of alleles that have at least one $Z_{ij} > 0$. Let s_j be defined as the number of alleles found in exactly *j* individuals: s_0 is the number of alleles present in the population but not in the observed sample, s_1 is the number of alleles found in only one individual, etc. s_j are called empirical counts. The observed allelic richness in the sample set is therefore

$$S_{obs} = \sum_{j=1}^{H} s_j \tag{3}$$

and the total number of alleles in the population is

$$S = s_0 + S_{obs} \tag{4}$$

We use the following form of the *Chao*2 richness estimator (Chao 1989) to estimate *S*

$$\tilde{S} = \frac{(H-1)s_1^2}{2Hs_2} + S_{obs} \tag{5}$$

where s_1 is the number of alleles that occur in a single individual and s_2 is the number of alleles that occur in exactly two individuals.

Assuming that *H* individuals have been observed, the expected number of alleles for a sample of size *h* (i.e. $\tilde{\tau}(h)$ for $h < H$) is given by

$$\tilde{\tau}(h) = S_{obs} - \sum_{j=1}^{H}\alpha_{jh}s_j \quad \text{with} \quad h = 1, 2, \ldots, H$$
$$\alpha_{jh} = \begin{cases} \dfrac{(H-h)!(H-j)!}{(H-h-j)!H!} & \text{for} \quad (j+h) \le H \\ 0 & \text{for} \quad (j+h) > H \end{cases} \tag{6}$$

The variance of this estimator is given by

$$\tilde{\sigma}^2(h) = \sum_{j=1}^{H}(1 - \alpha_{jh})^2 s_j - \frac{\tilde{\tau}^2(h)}{\tilde{S}} \tag{7}$$

Eqns 6 and 7 allow calculating approximately 95% confidence intervals for $\tau(h)$ by using $\tilde{\tau}(h) \pm 1.96\tilde{\sigma}(h)$.

To estimate the number of alleles within samples of the population, larger than the observed sample (i.e. $\tilde{\tau}(h)$ for $h > H$), a formulation based on eqn 2 is used

$$\tilde{\tau}(h) = S_{obs} + S_{obs}\sum_{k=1}^{G}\omega_k\frac{(1-\psi_k)^H(1-\psi_k)^h}{1-(1-\psi_k)^H}$$
$$\omega_k = \frac{\pi_k(1-(1-\psi_k)^H)}{\sum_{m=1}^{G}\pi_m(1-(1-\psi_m)^H)} \quad \text{with} \quad k = 1, 2, \ldots, G \tag{8}$$

where ψ_k is the presence probability of the *k*th group, and π_k is the number of individuals in the *k*th group. In eqn 8, the parameters ψ_k, π_k and *G* need to be estimated. This is achieved by maximizing the log conditional likelihood of the observed empirical counts $s_1, s_2, \ldots, s_H$ (given the observed allelic richness S_{obs}), using an expectation maximization algorithm (Mao *et al*. 2005). Confidence intervals for $\tau(h)$ are estimated by taking bootstrap resamples (typically between 100 and 1000) from the likelihood of the empirical counts ($s_1, s_2, \ldots, s_H$) given a random S_{obs} generated as a binomial variable with size $\tilde{\tau}(\infty)$ and probability

$$p(\infty) = \left(2 + \sum_{k=1}^{G}\omega_k\frac{(1-\psi_k)^H}{1-(1-\psi_k)^H}\right)^{-1} \tag{9}$$

This algorithm allows computing reliable allelic richness estimates for sample sizes ranging from one up to any number of individuals (Colwell *et al*. 2004). The software provides estimates of mean allelic richness, along with confidence bounds in order to test whether two populations differ with respect to allelic richness.

Two examples of uses for ARES are illustrated in this paper: a spatial and a temporal comparison of allelic richness in different populations. Our data consist of microsatellite data from the tropical butterfly *Drupadia theda* (Felder), a rainforest lycaenid species typical of pristine and moderately disturbed forest. The data were originally collected to investigate the recently introduced species-genetic diversity correlation (Vellend & Geber 2005) in disturbed landscapes. Butterflies were collected in four

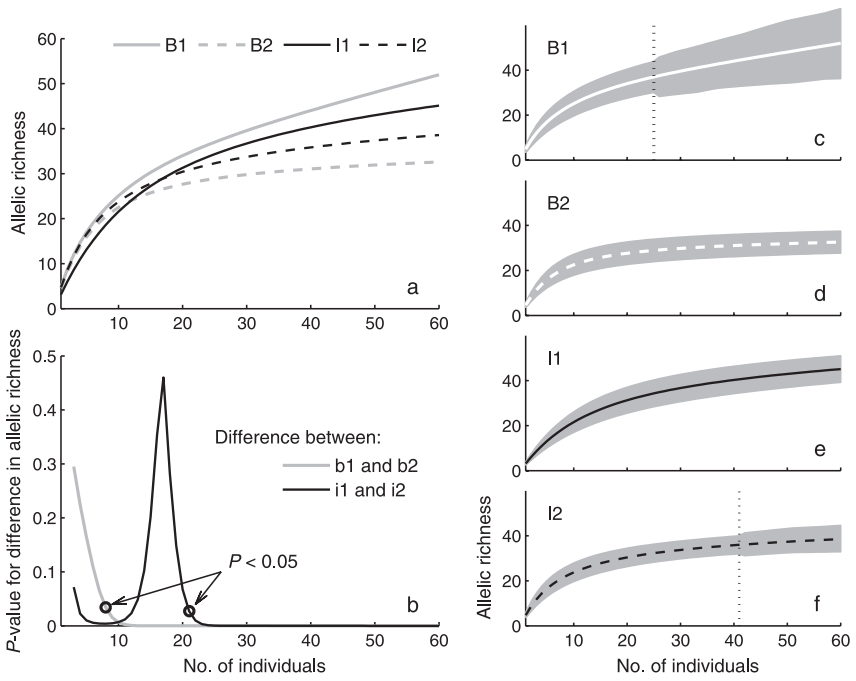

Fig. 1 Allelic richness at four sites sampled in the year 2000: a large unburned isolate (I1), a small unburned isolate (I2), once-burned forest (B1) and twice-burned forest (B2). (a) Comparison of mean allelic richness among all four landscapes. (b) *P* values when testing with a *t*-test the differences in mean allelic richness between I1 and I2 as well as B1 and B2; the circles indicate the minimum sample size where the the null hypothesis of equal allelic richness for I1 and I2 (B1 and B2) is rejected with a 0.05 confidence level. (c–f) mean allelic richness curves with 95% confidence bounds as shaded area. Vertical dotted line denotes the sample size (for I2 and B1 at 25 and 41 individuals, respectively, for I1 and B2 the samples were larger than 60 individuals). To the left of these vertical lines the mean and 95% confidence bounds are calculated analytically and to the right of the line they are calculated with a bootstrap analysis.

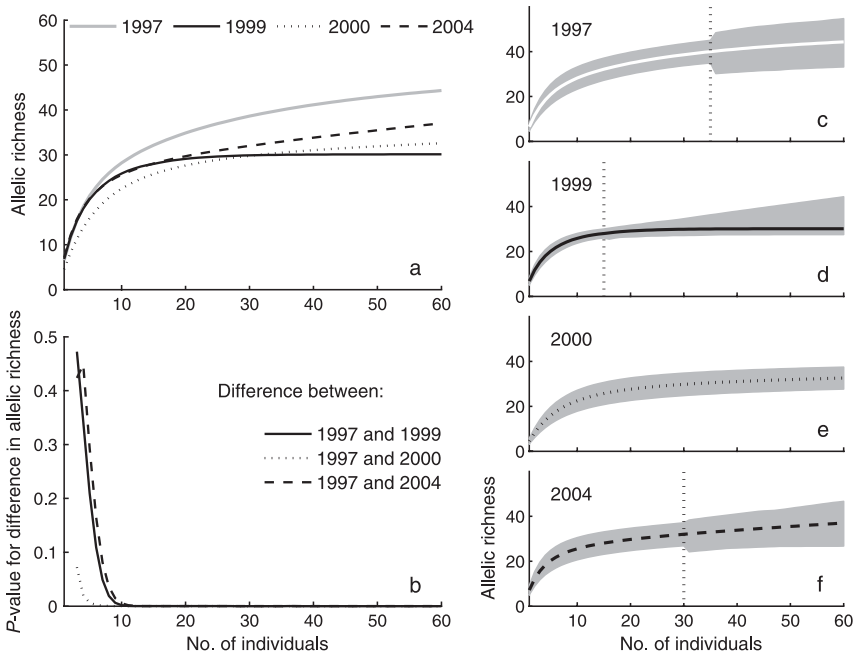

Fig. 2 Allelic richness at B2 (a twice-burned forest) in the years 1997, 1999, 2000 and 2004. For an explanation of the various graphs, see Fig. 1.

differentially disturbed landscapes (I1, I2, B1 and B2) in East Kalimantan, Indonesia Borneo. Also a time series of samples were taken in one landscape (B2) before and after the 1997–1998 El Niño Southern Oscillation (ENSO) event. The species was sufficiently abundant in all environments to enable reliable sample size for robust estimates of allelic richness. However, sample sizes differed among landscapes (sample sizes were much lower in burned than in unburned forests). A more elaborate description of landscapes and data is available in Cleary *et al*. (2006).

An analysis of allelic richness with ARES yielded the results shown in Figs 1 and 2. Figure 1 describes the allelic richness of the four sites as observed in the year 2000. The four graphs at the right show the mean allelic richness curves with 95% confidence bounds shaded. In these graphs, a vertical dotted line denotes the sample size.

The mean and 95% confidence bounds at the left of these vertical lines are calculated analytically (namely eqns 6 and 7) and at the right of the lines, they are calculated with a bootstrap analysis (eqns 8 and 9). The divergences between confidence bounds are a direct consequence of the change from interpolation to extrapolation. Mean allelic richness accumulation curves at the four sites are shown in a single graph (upper left graph) to facilitate comparison. Here it can be seen that allelic richness was highest in populations sampled from areas in the large unburned isolate (I1) and once-burned forest (B1). For the comparison of I1 and I2, it can furthermore be seen that small sample sizes can lead to unreliable results, since for sample sizes of less than 18 individuals, allelic richness in I2 (small unburned isolate) is higher than in I1. The graph at the lower left shows, for a given sample size, the result of a one-sided *t*-test comparing allelic richness between I1 and I2, and between B1 and B2. The circles show the minimum sample size for which one can conclude that allelic richness of I1 > I2 and B1 > B2 (using a *P* level of 0.05 for rejecting H_0).

Figure 2 shows an analysis similar to that presented in Fig. 1. Here, the allelic richness is compared among years for site B2 (the twice-burned forest). Importantly, allelic richness was significantly higher pre-ENSO (1997) than post-ENSO (1999, 2000 and 2004). Moreover, the allelic richness seems to increase over time since burning (namely the increasing curves for 1999, 2000 and 2004.

ARES is available at CRAN (http://cran.r-project.org/). The archive contains the source code in R (www.r-project.org/), a brief text describing how the software can be used and the data that were used in this study.

Acknowledgements

Chang Xuan Mao is kindly acknowledged for providing help in applying his algorithm. This study was conducted within the Virtual Laboratory for e-Science project (http://www.vl-e.nl), supported by a BSIK grant from the Dutch ministry of education, culture and science and the ICT innovation program of the ministry of economic affairs.

References

Allendorf FW (1986) Genetic drift and the loss of alleles versus heterozygosity. *ZooBiology*, **5**, 181–190.

Chao A (1989) Estimating population size for sparse data in capture–recapture experiments. *Biometrics*, **45**, 427–438.

Cleary DFR, Fauvelot C, Genner MJ, Menken SBJ, Mooers AØ (2006) Parallel responses of species and genetic diversity to El Niño Southern Oscillation-induced environmental destruction. *Ecology Letters*, **9**, 304–310.

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **345**, 101–118.

Colwell RK, Mao CX, Chang J (2004) Interpolating, extrapolating and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717–2727.

El Mousadik A, Petit R (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic of Morocco. *Theoretical and Applied Genetics*, **92**, 832–839.

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

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling design. *Conservation Genetics*, **5**, 539–543.

Mao CX, Colwell RK, Chang J (2005) Estimating the species accumulation curve using mixtures. *Biometrics*, **61**, 433–441.

Melo AS, Pereira RAS, Santos AJ, Shepherd GJ, Machado G, Medeiros HF, Sawaya RJ (2003) Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? *Oikos*, **101**, 398–410.

Pease CM, Lande R, Bull JJ (1989) A model of population growth, dispersal and evolution in a changing environment. *Ecology*, **70**, 1657–1664.

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

Reusch TBH, Ehlers A, Hammerli A, Worm B (2005) Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proceedings of the National Academy of Sciences, USA*, **102**, 2826–2831.

Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature*, **392**, 491–494.

Spielman D, Brook BW, Frankham R (2004) Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences, USA*, **101**, 15261–15264.

Stout J, Vandermeer J (1975) Comparison of species richness for stream-inhabiting insects in tropical and midlatitude streams. *American Nature*, **109**, 263–280.

Vellend M, Geber MA (2005) Connections between species diversity and genetic diversity. *Ecology Letters*, **8**, 767–781.