Analysis of Allelic Diversity in Subdivided Populations

S. T. Rodríguez-Ramilo, A. Caballero

Introduction

The analysis of genetic diversity is usually one of the first steps followed in almost all population genetics and evolutionary studies, as well as in the conservation of genetic resources. In subdivided populations, gene diversity (or expected heterozygosity) can be partitioned into within and between-subpopulation components (Nei 1973), with important applications for priorisation of populations in conservation (see review by Toro et al. 2009). The statistic most frequently used to quantify the degree of gene frequency differentiation between subpopulations is Wright´s (1969) F_{ST}. However, allelic richness (the number of different alleles segregating in the population) is an alternative criterion to evaluate genetic diversity and differentiation between subpopulations (El Mousadik and Petit 1996, Foulley and Ollivier 2006, Toro et al. 2009). Here we present a partition of allelic diversity into within and between-subpopulation components in analogy to the corresponding partition for gene diversity. A parameter to measure allelic differentiation between subpopulations (A_{ST}) arises from this partition and is compared with F_{ST} by means of computer simulations for a finite island model under a range of migration rates.

Methods

Gene frequency differentiation. In a structured population with n subpopulations, the total gene diversity or expected heterozygosity (H_T) can be partitioned into a component within subpopulations (H_S) and another ($H_T - H_S$) between subpopulations,

$$H_S = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} p_{i,k}^2\right), \qquad H_T = 1 - \sum_{k=1}^{K}\left(\sum_{i=1}^{n}\frac{p_{i,k}}{n}\right)^2$$

(Nei 1973), where $p_{i,k}$ is the frequency of allele k for a given locus in subpopulation i and K is the total number of alleles in the population. The between-subpopulation component of gene diversity ($H_T - H_S$) is also the average Nei´s minimum distance between populations

$$D_G = \frac{1}{n^2}\left[\sum_{i,j=1}^{n} d_{G,ij}\right],$$

where $d_{G,ij} = \frac{1}{2}\sum_{k=1}^{K}\left(p_{ik} - p_{jk}\right)^2$ is the gene frequency distance between subpopulations i and j. Consequently, $H_T = H_S + D_G$, and Wright´s (1969) F_{ST} is defined as $F_{ST} = \left(H_T - H_S\right)/H_T = D_G/H_T$.

Allelic differentiation. The rarefaction methodology (Hurlbert 1971) is used to estimate the number of expected alleles in samples of a specified size. In this approach, the smallest

Dpto. de Bioquímica, Genética e Inmunología. Facultad de Biología. Universidad de Vigo. 36310 Vigo, Spain

sample size is chosen as a reference to examine the number of alleles present in all samples. In the context of a subdivided population, if N_{ik} represents the number of copies of the kth allele from the sample of a given subpopulation i and N_i represents the total number of genes present in that subpopulation, the allelic richness at one locus is denoted as the expected number of different alleles that a sample had if the sample size had been g genes (usually the smallest sample size) instead of N_i ($\geq g$). The expected number of different alleles in a sample of genes taken at random is then equal to

$a_i = \sum_{k=1}^{K} \left(1 - P_{ik}\right)$, where

$$P_{ik} = \binom{N_i - N_{ik}}{g} \Big/ \binom{N_i}{g}$$

is the probability that allele k does not occur in a sample of g genes chosen at random (El Mousadik and Petit 1996).

Following a derivation analogous to that of F_{ST} it is possible to define a partition of allelic diversity into within and between-subpopulation components. The within-subpopulation component of allelic diversity is

$$A_S = \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) - 1 \cdot$$

Now, an allelic dissimilarity or distance between two subpopulations may be defined as the number of alleles present in a subpopulation and absent in the other. Thus, the average allelic distance between subpopulations i and j can be obtained as

$$d_{A,ij} = \frac{1}{2}\sum_{k=1}^{K}\left[\left(1 - P_{ik}\right)P_{jk} + P_{ik}\left(1 - P_{jk}\right)\right]$$

(see Foulley and Ollivier 2006). The average distance between all subpopulations is

$$D_A = \frac{1}{n^2}\left[\sum_{i,j=1}^{n} d_{A,ij}\right] \cdot$$

Hence, the total allelic diversity (A_T) is the sum of both components,

$$A_T = A_S + D_A = \left[\frac{1}{n}\sum_{i=1}^{n}\left(a_i + \frac{1}{n}\sum_{j=1}^{n} d_{ij}\right)\right] - 1 = \left[\frac{1}{n^2}\sum_{k=1}^{K}\sum_{i,j=1}^{n}\left(1 - P_{ik}P_{jk}\right)\right] - 1 \,,$$

Resulting in a definition of the coefficient of allelic differentiation,

$A_{ST} = \left(A_T - A_S\right)/A_T = D_A/A_T$.

Simulations. In order to investigate the behaviour of A_{ST} in a subdivided population and its comparison with F_{ST}, computer simulations were run assuming a finite island model with a range of migration rates. First, a single population of N_T = 10,000 diploid individuals was initially set where all individuals originally carried the same allele at a given neutral locus. The population was run for 100,000 generations assuming random mating. In each generation mutation to new allelic variants (infinite alleles model) occurred with Poisson probability and rate u = 0.00001. Over this period of time mutation-drift equilibrium was reached for the expected heterozygosity and the number of alleles.

At generation 100,000, the population was randomly split into n = 10 subpopulations of census size N = 1000. This subdivided population was maintained for a further 200,000 generation period assuming the same mutation rate as before and a migration rate per generation m between subpopulations (migration occurred with Poisson probability among randomly chosen subpopulations).

One hundred independent loci were assumed for each run. At different periods, the current allele frequencies for each locus were used to calculate the population genetic parameters (H_S, D_G, H_T, A_S, D_A and A_T) and averaged over loci to obtain estimates of F_{ST} and A_{ST}. Estimates of A_{ST} were obtained from different sample sizes (g) after rarefaction. All simulation scenarios were replicated three times and results were averaged over replicates.

Results and discussion

Figure 1 illustrates the main simulated parameters over the first 10,000 generations after subdivision. The equilibrium values (generations 185,000-200,000) are represented with symbols. After subdivision, A_S declined first and then increased slowly toward their equilibrium values and D_A increased accordingly. A_{ST} reached its equilibrium value rather faster than F_{ST}.

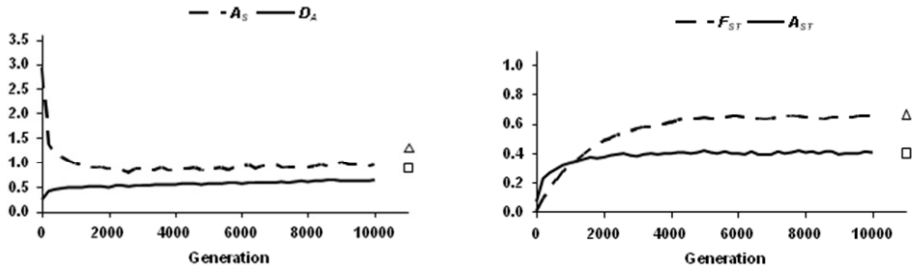

Figure 1: Average population genetic parameters over generations for a scenario of n = 10 subpopulations of N = 1000 individuals each, mutation rate u = 0.00001 per locus and generation and migration rate m = 0.0001 per generation (i.e., Nm = 0.1 migrants per subpopulation and generation). Right-hand side squares and triangles refer to the asymptotic values of the corresponding parameters. A_S: allelic diversity within subpopulations; D_A: allelic distance between subpopulations; F_{ST}: gene frequency differentiation coefficient. A_{ST}: allelic differentiation coefficient.

The equilibrium values of A_{ST} for different subpopulation sample sizes (g), using the rarefaction technique, are shown in Figure 2. It can be noted that the estimated values of A_{ST} are very close to their population values (symbols) even for very small subpopulation sample sizes. Therefore, A_{ST} is very robust against sample size variation.

Figure 3 compares the equilibrium values of F_{ST} and A_{ST} for a range of migration rates (m). For small values of migration, A_{ST} decays faster than F_{ST}, but for large migration rates, A_{ST} becomes almost insensitive to migration rate.

The parameter A_{ST} gives a measure proportional to the number of alleles in which two randomly chosen subpopulations differ. In the same way that a population with a larger number of alleles has a higher adaptive potential than another with a lower number of alleles, A_{ST} may indicate the degree of differential potentiality among subpopulations.

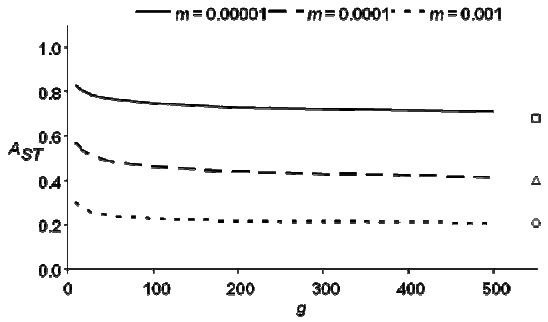

Figure 2: Asymptotic estimates of A_{ST} for different numbers of alleles (g) sampled within subpopulations and three different migration rates (m) per generation. The right-hand side symbols refer to the population genetic values of A_{ST}.

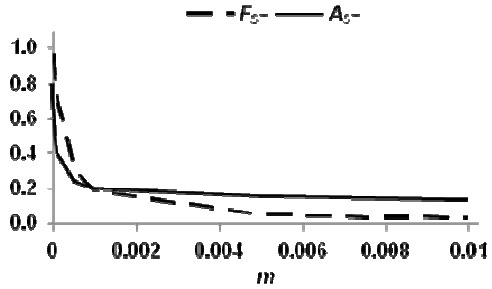

Figure 3: Asymptotic values of F_{ST} and A_{ST} for a range of migration rates (m) per generation. Other parameters as for Figure 1.

References

El Mousadik, A. and Petit, R. J. (1996). *Theor. Appl. Genet.,* 92: 832–839.

Foulley, J. L. and Ollivier, L. (2006). *Liv. Sci.*, 101: 150–158.

Hulbert, S. H. (1971). *Ecology*, 52: 577–586.

Nei, M. (1973). *Proc. Natl. Acad. Sci. USA,* 70: 3321–3323.

Toro, M. A., Fernández, J. and Caballero, A. (2009). *Livest. Sci.,* 120: 174–195.

Wright, S. (1969) *The Theory of Gene Frequencies*. Vol. 2. Univ. of Chicago Press, Chicago.