COMPUTER PROGRAMS

CLONALITY V.0.4: a randomization-based program to test for heterozygosity-genet size relationships in clonal organisms

FRANCK PRUGNOLLE, MARC CHOISY and THIERRY DE MEEÛS

Laboratoire Génétique et Evolution des Maladies Infectieuses, UMR 2724 CNRS-IRD-UMI, Centre IRD, 911 Av. Agropolis, BP 64501, 34394 Montpellier Cedex 5, France

Abstract

CLONALITY **V.0.4 is a program for testing heterozygosity-genet size relationships in clonal organisms using a randomization procedure. The software has been developed under the Borland Delphi developing environment and a Windows-executable version is freely downloadable from http://gemi.mpl.ird.fr/SiteSGASS/Prugnolle/ClonalityPage.html. The program compares the observed F_{IS} of the population with the F_{IS} expected if genets (multilocus genotypes present in multiple copies within the population) were chosen randomly from the set of different multilocus genotypes. The randomization procedure is performed with the same number of genets and the same number of repetitions per genet as what is observed in the original data set.**

Although inbreeding depression is a common phenomenon in living organisms (e.g. Kelly 2005), it is difficult to measure in natural populations. A way to approach this difficulty is to correlate individual heterozgyosity – measured using polymorphic genetic markers – with fitness-associated traits (David 1998). Clonal organisms are expected to exhibit wide variations in fitness among individuals in natural populations (Gardner & Mangel 1999). As such, they are particularly suitable as model species to examine the relationship between inbreeding depression and fitness in natural populations.

One fitness-correlated trait of clonal organisms, easily measurable in the field, is the genet size (e.g. Hammerli & Reusch 2003; Prugnolle *et al.* 2004; Alberto *et al.* 2005). A genet corresponds to the collection of individuals produced clonally and thus sharing the same multilocus genotype at all loci in the genome (e.g. De Meeûs *et al.* 2007).
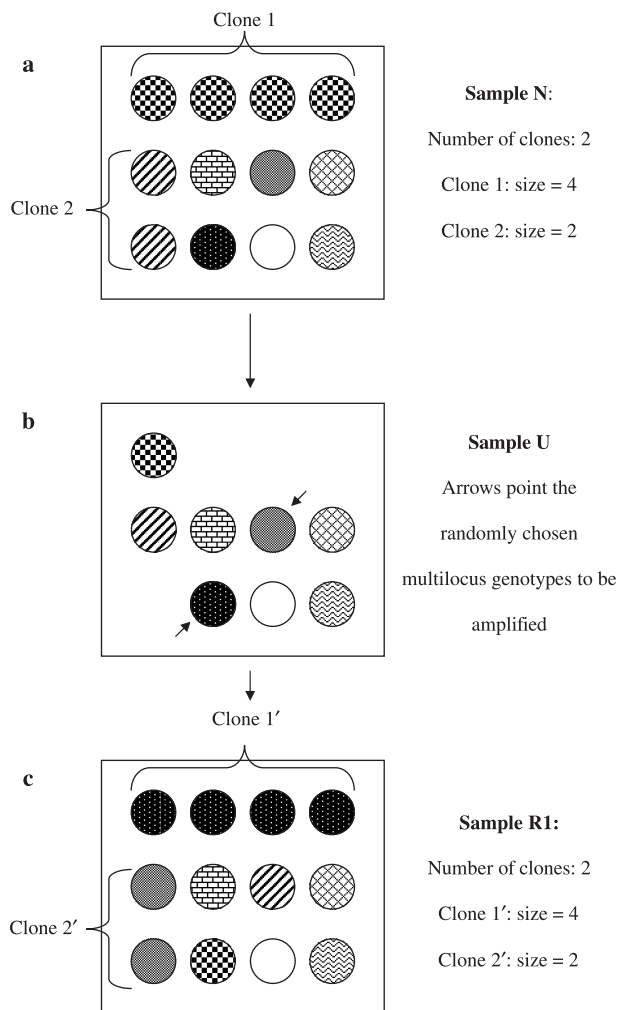
Is genet size positively correlated with individual heterozygosity in natural populations? One way to test the correlation is obviously to analyse the relationship between genet size and individual heterozygosity using parametric

Correspondence: Franck Prugnolle, Fax: +33 (0)467 41 62 99;
E-mail: prugnoll@mpl.ird.fr
Authors Franck Prugnolle and Marc Choisy contributed equally to this work.

methods. By definition, these methods unfortunately rely on particular distributions of residuals (e.g. normal, Poisson or binomial distributions), which are not always easily identifiable (Manly 1997). Permutation or randomization tests appear as powerful alternatives to standard parametric tests for analysing such relationships. They have indeed the advantage of making no hypothesis on the distribution of the data (such as normality) (Manly 1997). The increase in computing resources makes such computer-intensive methods more and more attractive for researchers.

The CLONALITY V.0.4 program uses a randomization approach to test for heterozygosity–genet size relationship in natural populations of clonal organisms.

Principle of the test. If genet size is positively correlated with individual heterozygosity (i.e. genets of higher size display a higher heterozygosity than genets of lower size), the statistics F_{IS}, which measures the departure of a population from Hardy–Weinberg expectations, should be lower than expected under the null hypothesis of no relationship (H_O). The method developed in the program relies upon randomization to create a large number of data sets (replicates) that could have arisen under the null hypothesis of no relationship between genet size and heterozygosity (H_O). These random data sets thus give an approximation of the distribution of the statistics F_{IS} under H_O. The more

Fig. 1 Schematic representation of the randomization procedure applied by the program CLONALITY V.0.4. In this scheme, a square represents one population. Circles within the square represent the different individuals. Circles displaying different draws display different multilocus genotypes. (a) Original sample (Sample N); (b) sample where copies of clones were deleted (Sample U); (c) random sample (Sample Ri). The expected F_{IS} is computed from this later sample.

data sets are generated, the better this distribution is approximated.

Randomised data sets are generated as follows (see also Fig. 1): (i) within the observed data set (sample N), the program detects the number of different genets within each subpopulation (i.e. the number of multilocus genotypes having copies) and measures their size (i.e. the number of copies; Fig. 1a); (ii) only a single copy of each different multilocus genotype is retained (sample U; Fig. 1b); (iii) a new sample is generated (sample Ri) by amplifying randomly chosen multilocus genotypes from sample U so that the sample size and the distribution of genet size are kept identical to those observed in sample N, in each subpopu-

lation (Fig. 1c); (iv) the procedure is repeated a large number of times (samples $R1$ to Rn, N being the total number of randomizations).

F_{IS} is then computed for the observed data set (f_o) and for each of the randomly generated ones (f_{Ri}) using the Weir & Cockerham's (1984) estimate f. The distribution of f_{Ri} allows to compute the P value of the test, which corresponds to the proportion of times $f_{Ri} \leq f_o$, i.e. P value $= [\Sigma(f_{Ri} | f_{Ri} \leq f_o) + f_o]/\Sigma(f_{Ri} + f_o)$ (eqn 1). The program was tested on 1000 simulated data set generated under the null hypothesis (using under EASYPOP version 1.8 updated from Balloux 2001). Parameters of the simulations were, for each population, full clonality, 100 individuals, 10 loci with 100 alleles each and displaying a K-allele model of mutation and a mutation rate of 10^{-5}. The K-allele model treats all alleles as equivalent, with mutation from any allele to any other allele equally probable. This model is frequently used to simulate the evolution of microsatellite-like loci. The simulations were run for 20 generations. The analysis of these 1000 data sets with CLONALITY V.0.4 provided 1000 P values, the distribution of which did not differ from a uniform distribution (Kolmogorov–Smirnov test, P value $= 0.36$), with a mean at 0.51 and 43 values equal or lower than 0.05, which is in agreement with the 5% expected under the null hypothesis (binomial test, P value $= 0.35$).

In CLONALITY V.0.4, a graphic window allows to follow the evolution of the P value as the number of randomizations increases. This provides the possibility to check for the stability of the P value when randomizations stop, which is a fundamental check when using Monte Carlo methods (Manly 1997).

The program. The software executable for Microsoft Windows can be downloaded from http://gemi.mpl.ird.fr/SiteSGASS/Prugnolle/ClonalityPage.html. The source code has been written under the Borland Delphi developing environment and is available upon request to *F*. Prugnolle. Previous uses of an early version of this program can be found in Prugnolle *et al*. (2004) and Alberto *et al*. (2005).

Input file. The input file is a simple text file. It is very similar to the one used for FSTAT (Goudet 1995) for diploid organisms. However, it must present at least two populations (see below how to deal with single-population data) and alleles must be obligatorily coded with three digits. The two alleles are entered successively without space between them. Missing data are not recognized and individuals that present missing genotypes must be excluded from the file. If only one population is available, the user has to create a second fake population of only homozygous individuals, all displaying the same homozygous genotype at all loci (for instance, genotype 100100). An example file (called Example.txt), from Prugnolle *et al*. (2004) is also available on the CLONALITY V.0.4 web page.

User options. Once the input file has been loaded using the program interface, several options are offered to the user. The user has to enter the number of randomizations to be performed (n) and the name of the output file. The user may also choose to perform the test (i) for over all loci over all populations (default option), (ii) for each locus independently (check box named 'per locus analyses') or (iii) for each population independently over all loci (check box named 'per population analysis'). The user may finally choose to create the files that contain the P values obtained after each iteration by checking the check box named 'create files iterations/P values'.

Output files. The program CLONALITY V.0.4 provides several output files. The first output file is named following the user's request (hereafter referred to as 'name of the output file'). This file contains the results of the randomization procedure overall populations and loci, per locus and/or per population (if requested). In this file, several statistics are given: (i) the observed F_{IS} estimate (f_o), (ii) the average F_{IS} (f_{Ri}) computed overall randomised data sets, and (iii) the P value of the test (i.e. the proportion of times $f_{Ri} \leq f_o$; cf. equation 1).

The second output file is named 'Clones_name of the output file'. This text file gives basic information regarding the sampled populations: the sample size ('Sample size'), the total number of different multilocus genotypes observed ('NDGenotypes'), the number of clones ('NbClones', i.e. the number of multilocus genotypes present at more than one copy) and the number of copy per clone that is the genet size ('Repeats/Clone').

The third output file (named 'Withoutrepeats_name of the output file') is a file containing the original data set but without the repetitions (i.e. a file where only one copy of each multilocus genotype within each population is conserved). This output file is formatted to be used with the program FSTAT (updated from Goudet 1995).

The last files provided by the program (if requested by the user) are those that contain the value of each P value obtained after each iteration. Files are provided either (i) for over all loci over all populations (named 'over all P values_name of the output file'), (ii) for each locus [named 'LocusXp-values_name of the output file', with X ∈ (1, L), if there are L loci], or (iii) for each population ('PopYp-values_name of the output file' with Y ∈ [1, Pops], if Pops is the total number of populations).

References

Alberto F, Gouveia L, Arnaud-Haond S, Perez-Llorens JL, Duarte CM, Serrao EA (2005) Within-population spatial genetic structure, neighbourhood size and clonal subrange in the seagrass *Cymodocea nodosa*. *Molecular Ecology*, **14**, 2669–2681.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

David P (1998) Heterozygosity-fitness correlations: new perspectives on old problems. *Heredity*, **80**, 531–537.

De Meeûs T, Prugnolle F, Agnew P (2007) Asexual reproduction: genetics and evolutionary aspects. *Cellular and Molecular Life Sciences*, **64**, 1355–1372.

Gardner SN, Mangel M (1999) Modeling investments in seeds, clonal offspring, and translocation in a clonal plant. *Ecology*, **80**, 1202–1220.

Goudet J (1995) FSTAT (version 1.2): a computer program to calculate *F*-statistics. *Journal of Heredity*, **86**, 485–486.

Hammerli A, Reusch TBH (2003) Inbreeding depression influences genet size distribution in a marine angiosperm. *Molecular Ecology*, **12**, 619–629.

Kelly JK (2005) Family level inbreeding depression and the evolution of plant mating systems. *New Phytologist*, **165**, 55–62.

Manly BFJ (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Prugnolle F, Choisy M, Theron A, Durand P, De Meeûs T (2004) Sex-specific correlation between heterozygosity and clone size in the trematode *Schistosma mansoni*. *Molecular Ecology*, **13**, 2859–2864.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.