CLONALITY V.0.4 HELP

By Franck Prugnolle, Marc Choisy and Thierry De Meeûs

What does CLONALITY V.0.4 do?

CLONALITY V.0.4 uses a randomisation approach to test for a positive heterozygosity-genet size relationship in clonal organisms.

Principle of the test:

A genet corresponds to the collection of individuals produced clonally and thus sharing the same multilocus genotype at all loci in the genome (de Meeus et al., 2007). If genet size is positively correlated with individual heterozygosity (i.e. genet of higher size display a higher heterozygosity than genet 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_0). The method developed in the program rely upon randomisation to create a large number of datasets (replicates) that could have arisen under the null hypothesis of no relationship between genet size and heterozygosity (H_0). These random-datasets thus give an approximation of the distribution of the statistics F_{IS} under H_0. The more datasets are generated, the better this distribution is approximated.
Randomised datasets are generated as follows: i) within the observed dataset (sample N), the program detects the number of different genets within each sub-population (i.e. the number of multilocus genotypes having copies) and measures their size (i.e. the number of copies); ii) only a single copy of each different multilocus genotype is retained (sample U); 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 sub-population; iv) the procedure is repeated a large number of times (samples R_1 to R_n, n being the total number of randomisations).

F_{IS} is then computed for the observed dataset (f_o) and for each of the randomly generated ones (f_{Ri}) using the Weir and 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. $p-value = \frac{\sum (f_{Ri} / f_{Ri} \leq f_{\grave{a}}) + f_0}{\sum f_{Ri} + f_0}$ eq (1). Under the null hypothesis H_0, the distribution of the p-values follows a uniform distribution (tested on 1000 simulations, *p-value* = 0.36 for the conformity of the distribution to the uniform one).

In the program, a graphic window allows to follow the evolution of the *p-value* as the number of randomisations increases. This provides the possibility to check for the stability of the p-value when randomisations stop, which is a fundamental check when using Monte Carlo methods (Manly, 1997).

Input file:

The input file is a simple text file. It must present at least two populations (se below how to deal with single population data) and alleles must be obligatorily coded by three digits. The two alleles are entered successively without space between them. Missing data are not recognised 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).

Example given for three loci and two populations.

Line1: Number of populations (<200) [space]Number of loci (<100) [space] Maximum number used to code an allele (<999) [space]Number of digit used to code alleles (=3) [tabulation]
Line 2: Name of locus 1 [Tabulation]
Line 3: Name of locus 2 [Tabultation]
Line 4: Name of locus 3 [Tabultation]
Line 5: Population number 1 [Tabulation] alleles of locus 1 [Tabulation] alleles of locus 2 [Tabulation]alleles of locus 3[return]
Line 6: Population number 1 [Tabulation] alleles of locus 1 [Tabulation] alleles of locus 2 [Tabulation]alleles of locus 3[return]
Line 7: Population number 1 [Tabulation] alleles of locus 1 [Tabulation] alleles of locus 2 [Tabulation]alleles of locus 3[return]
Line…
etc
Line …
Line …Population number 2 [Tabulation] alleles of locus 1 [Tabulation] alleles of locus 2 [Tabulation] alleles of locus 3[**No Return at the end of the file**]
[End of file]

A real example. Note that for locus 1, individual 1's alleles are: 248 and 248. (see also the file "Example.txt" provided with the program for another example).

```
2 3 312 3
Locus1
Locus2
Locus3
1       248248  155155  231231
1       248248  155155  237276
1       245248  155155  231231
1       248248  155155  237276
1       248248  155155  231276
1       248248  153155  231276
1       245248  155155  231231
1       245248  155155  231231
1       248248  153155  231231
1       248248  153153  231276
2       245248  155155  231276
2       233248  153155  231255
2       233248  153155  231255
2       245248  155155  231237
2       248248  155155  246276
2       248248  155155  246276
2       248248  155155  243273
2       233248  155155  231231
2       233248  153155  255273
2       248248  155155  231231
2       245245  153155  231246
2       248248  155155  273276
2       248248  155155  246276
```

Loading the data file:

To load the data file, use the program interface (button "Data file").

User options:

Once the input file has been loaded using the program interface, several options are offered. The user has first to enter the number of randomisations to be performed (default 1000) and the name of the output file (default "results.txt"). The user may also choose to perform the test (i) overall loci overall populations (default option), (ii) for each locus independently (check box "per locus analysis") or (iii) for each population independently overall loci (check box "per population analysis"). The user may finally choose to create the files that contain the value of each p-value after each iteration by checking the check box "create files iterations / p-values".

Output files:

The program CLONALITY V. 0.4 provides several output files.

1) The first output file is called upon user's request (hereafter referred as "name of the output file". This file contains the results of the randomisation 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 dataset and (iii) the *p*-value of the test (i.e. the proportion of times f_{Ri} were inferior to f_o).
2) 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").
3) The third output file is a file containing the original dataset but without the repetitions (i.e. a file where only one copy of each multilocus genotype within each population is conserved) (named "Withoutrepeats_ name of the output file"). This output file is formatted to be used with FSTAT (Goudet, 1995).
4) 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) overall loci overall populations (named "overall p-values_ name of the output file"), (ii) for each locus (named LocusX p-values_ name of the output file, with X $\in$ [1, total number of loci])or (iii) for each population (PopY p-values_name of the output file" with Y $\in$ [1, total number of populations].

Temporary files: Two temporary files are created during the running process (Tempmicrosat.txt and Temp.txt) that are only for computing purposes. You can delete these files safely after each run.

References:

De Meeus, T., Prugnolle, F. & Agnew, P. (2007) Asexual reproduction: Genetics and evolutionary aspects. Cellular and Molecular Life Sciences, 64**,** 1355-1372.

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

Manly, B. F. J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall.

Weir, B. S. & Cockerham, C. C. (1984) Estimating F-statistics for the analysis of population structure. Evolution, 38**,** 1358-1370.