doi: 10.1111/j.1471-8286.2006.01522.x

PROGRAM NOTE

GENCLONE: a computer program to analyse genotypic data, test for clonality and describe spatial clonal organization

SOPHIE ARNAUD-HAOND* and KHALID BELKHIR†

*Laboratory of 'Ecology and Evolution of Marine Organisms', CCMAR, F.C.M.A.-Universidade do Algarve, Faro, Portugal, †Laboratoire Génome, Populations, Interactions, Adaptation, Université de Montpellier II, France

Abstract

GENCLONE 1.0 is designed for studying clonality and its spatial components using genotype data with molecular markers from haploid or diploid organisms. GENCLONE 1.0 performs the following tasks. (i) discriminates distinct multilocus genotypes (MLGs), and uses permutation and resampling approaches to test for the reliability of sets of loci and sampling units for estimating genotypic and genetic diversity (a procedure also useful for nonclonal organisms); (ii) computes statistics to test for clonal propagation or clonal identity of replicates; (iii) computes various indices describing genotypic diversity; and (iv) summarizes the spatial organization of MLGs with adapted spatial autocorrelation methods and clonal subrange estimates.

Keywords: clonality, individuals resampling, locus combination, multilocus genotypes, software, spatial distribution

Received 15 June 2006; revision accepted 10 July 2006

Clonal species, from unicellular organisms to marine invertebrates, are dominant in many habitats. In ecological studies on clonal organisms, particularly in clonal plants with cryptic rhyzomatic connections, the discrimination of individuals issued from sexual or clonal reproduction, or the estimation of sexual vs. clonal reproduction, are among the most challenging technical issues. Molecular markers are particularly useful since genetically identical individuals issued from clonal reproduction can theoretically be recognized on the basis of their multilocus genotypes. However, reliable recognition of clonal identity, on the basis of molecular markers, requires specific statistical tests and procedures. Two software packages have been recently developed, MLGSIM (Stenberg *et al*. 2003) and GENOTYPE and GENODIVE (Meirmans & Van Tienderen 2004), that provide some of those required tests. The software developed here, GENCLONE 1.0 implements new and improved statistical features such as accounting for deviation from Hardy–Weinberg equilibrium while testing for clonality, and specially adapted analyses for studying the spatial components of clonality.

GENCLONE requires the following information for each individual: (i) a name; (ii) one to two spatial coordinates (when available); and (iii) genotype at each locus for codominant markers. The options available to users can be divided in three sets of analysis, corresponding to the three 'upper panels'. (i) 'Test' — for checking for locus and 'sampling unit' reliability for optimal multilocus genotypes (MLGs) and genetic individuals recognition; (ii) 'MLG' — for computing various genotypic richness and diversity descriptors; and (iii) 'Spatial components' — for describing various spatial aspects of clonality.

Tests

These procedures use permutation approaches to test for data quality. (In other words, the power of the analysed sample and loci set to obtain an accurate estimate of the maximum number of multilocus genotypes present in the dataset and in the sampling area, respectively). All possible datasets corresponding to all possible combinations of loci (with L the number of analysed loci) and sampling units (with N the total number of 'sampling units') are generated,

Correspondence: Sophie Arnaud-Haond, Fax: +61 7 4725-1570; E-mail: s-arnaud@ualg.pt or belkhir@univ-montp2.fr
http://www.ualg.pt/ccmar/maree/software.php?soft=genclon

and then the minimum, average and maximum number of discriminated MLGs for each class of number of locus (*l*) or sampling units (*n*) are obtained. When performed on the loci, this permutation procedure allows us to verify if an asymptote is reached when *l* tends towards *L*. It therefore allows ensuring that the set of loci used permit a good estimate of the real number of MLG present in the sample analysed. This procedure combined with the test for clonal identity detailed hereafter allows to ascertain the maximum efficiency of the chosen loci combination (Arnaud-Haond *et al*. 2005). When applied to individuals, it allows us to verify if the sampling density (in terms of the number of sampling units) is sufficient to reliably estimate the true number of MLGs present in the sampled area. When the number of individuals or loci is high, the computation time can be very long due to the huge number of possible combinations. We have therefore developed two complementary procedures based on resampling without replacement: resampling *x* times from the set of *L* loci; and resampling *x* times from the set of *N* individuals; (where *x* is chosen by the users) and then estimating the average number of MLGs which can be distinguished with this number *x* of loci or individuals. These resampling procedures also provide estimates of the maximum, minimum and average number of MLGs in the subset of data, as well as the maximum, minimum and the mean number of alleles and the heterozygosity (unbiased estimate, Nei 1978) for each subset of data generated. This is an alternative to the bootstrap, and to the rarefaction procedure (El Mousadik & Petit 1996), commonly used to compare the levels of diversity among sample sets of unequal size, and can also be used to compare allelic richness and heterozygosity in nonclonal organisms (Leberg 2002). The last test in this section is a test for clonal propagation based on the round robin method proposed by Parks & Werth (1993). This allows us to estimate, for each MLG, the probability P_{GEN} and the derived binomial P_{SEX}. These probabilities are used to test both for clonal identity and for clonal propagation (Arnaud-Haond *et al*. 2005; see also Tibayrenc *et al*. 1990; Gregorius 2005). A slightly more conservative test is also provided, which is based on estimates P_{GEN} (*f*) and P_{SEX} (*f*), of the same probabilities, but now taking into account the estimated F_{IS} in the population (Young *et al*. 2002). Finally, a genetic distance matrix can be computed (based on the number of different alleles among sampling units). The frequency distribution of genetic distances can, for example, help to screen for scoring errors or somatic mutations (Douhovnikoff *et al*. 2004; Meirmans & Van Tienderen 2004; Arnaud-Haond *et al*. 2005). With a high number of loci, or loci characterized by a high mutation rate, this frequency distribution can also help to define a threshold below which MLGs separated by low genetic distance and can be considered as belonging to the same 'clonal lineage', or of the same genetic individual.

MlG

This option allows us to compute the usual estimators of genotypic richness in a sample of *N* sampling units. These are: *G* = the number of distinct MLGs; *R* = the modified index of genotypic richness, as proposed by Dorken and Eckert (Dorken & Eckert 2001). Additionally, the commonly used indices of genotypic diversity, derived from species diversity indices, are computed. These are: the Simpson complement and, the Shannon-Wiener (Hurlbert 1971; Washington 1984) diversity and evenness indices, as well as Hill's Simpson reciprocal (Hurlbert 1971; Hill 1973) (which corresponds to the 'apparent number of genotypes in the sample').

Spatial components

These procedures allow us to summarize spatial aspects of clonal diversity when geographic coordinates of the sampling units are available. A map of MLGs can be drawn and exported as a bitmap file. The clonal subrange section plots the probability of clonal identity against distance (Harada & Iwasa 1996; Harada *et al*. 1997). Here we use a custom definition of distance classes (either as a number of distance classes, or as a list of predefined maximum distances for each class), and estimate of the 'clonal subrange' as the maximum spatial distance between two replicates of the same MLG (Alberto *et al*. 2005). Finally, autocorrelation procedures adapted to the existence of replicates are computed, using Loiselle *et al*. (1995) and Ritland (1996) kinship coefficients. Classical autocorrelation analysis are performed at the 'ramet level' (i.e. including all sampling units), and random permutations of the geographical coordinates are performed among sampling units in order to test for the significance of the observed spatial structure. Following Vekemans & Hardy (2004) F_{ij} (the average kinship for each distance class) and *b* (the slope of the regression) are estimated and tested for significance. At the 'genet level' (i.e. including only one copy of each MLG), autocorrelation is computed in three ways: (i) using central coordinates for each replicated MLG (Hämmerli & Reusch 2003; Alberto *et al*. 2005); (ii) using a weighted approach (Alberto *et al*. 2005; Wagner *et al*. 2005;) to remove the distances among pairs of identical genotype from the dataset; (iii) using a resampling approach in order to create and analyse subdatasets of size *g* (= the number of MLG identified), with each MLG being attributed randomly one of the spatial coordinates corresponding to one of the sampling units exhibiting this given MLG (Alberto *et al*. 2005). For this last procedure, confidence intervals are computed at 90% and 95% level, in order to test whether the 'observed' distribution obtained by resampling significantly depart from the 'random' distribution generated by randomly permuting spatial coordinates among MLGs.

Acknowledgements

We thank Kevin Dawson, Jim Coyer and Malia Chevolot for their help in improving the first versions of GENCLONE and correcting the English.

References

Alberto F, Gouveia L, Arnaud-Haond S *et al.* (2005) Spatial genetic structure, neighbourhood size and clonal subrange in seagrass (*Cymodocea nodosa*) populations. *Molecular Ecology*, **14**, 2669–2681.

Arnaud-Haond S, Alberto F, Procaccini G, Serrao EA, Duarte CM (2005) Assessing genetic diversity in clonal organisms: low diversity or low resolution? Combining power and cost-efficiency in selecting markers. *Journal of Heredity*, **96**, 1–8.

Dorken ME, Eckert CG (2001) Severely reduced sexual reproduction in northern populations of a clonal plant, *Decodon verticillatus* (Lythraceae). *Journal of Ecology*, **89**, 339–350.

Douhovnikoff V, Cheng AM, Dodd RS (2004) Incidences size and spatial structure of clones in second-growth stands of coast redwood *Sequoia sempervirens* (Cupressaceae). *American Journal of Botany*, **91**, 1140–1146.

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. *Theoretical and Applied Genetics*, **92**, 832–839.

Gregorius H-R (2005) Testing for clonal propagation. *Heredity*, **94**, 173–179.

Hämmerli A, Reusch TBH (2003) Genetic neighborhood of clone structures in eelgrass meadows quantified by spatial autocorrelation of microsatellite markers. *Heredity*, **91**, 448–455.

Harada K, Iwasa Y (1996) Analyses of spatial patterns and population processes of clonal plants. *Researches on Population Ecology*, **38**, 153–164.

Harada Y, Kawano S, Iwasa Y (1997) Probability of clonal identity: inferring the relative success of sexual versus clonal reproduction from spatial genetic patterns. *Journal of Ecology*, **85**, 591–600.

Hill MO (1973) Diversity and eveness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.

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

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

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understorey shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Nei M (1978) Estimation of heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**.

Parks JC, Werth CR (1993) A study of spatial features of clones in a population of bracken fern, *Pteridium aquilinum* (Dennstaedtiaceae). *American Journal of Botany*, **80**, 537–544.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, **67**, 175–185.

Stenberg P, Lundmark M, Saura A (2003) MLGSIM: a program for detecting clones using a simulation approach. *Molecular Ecology Notes*, **3**, 329–331.

Tibayrenc M, Kjellberg F, Ayala F (1990) A clonal theory of parasitic protozoa: the population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences. *Proceedings of the National Academy of Sciences, USA*, **87**, 2414–2418.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, **13**, 921–935.

Wagner HH, Holderegger R, Werth S, Gugerli F, Hoebee SE, Scheidegger C (2005) Variogram analysis of the spatial genetic structure of continuous populations using multilocus microsatellite data. *Genetics*, **169**, 1739–1752.

Washington HG (1984) Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. *Water Research*, **18**, 653–694.

Young AG, Hill JH, Murray BG, Peakall R (2002) Breeding system, genetic diversity and clonal structure in the sub-alpine forb *Rutidosis leiolepis* F. Muell. (Asteraceae). *Biological Conservation*, **106**, 71–78.