Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 966–970 doi: 10.1111/2041-210X.12550

APPLICATION

RClone: a package to identify MultiLocus Clonal Lineages and handle clonal data sets in R.

Diane Bailleul[1,2], Solenn Stoeckel[3] and Sophie Arnaud-Haond[1,2]*

[1]*IFREMER, UMR MARBEC, Station de Sète, Avenue Jean Monnet, CS 30171, 34203 Sète Cedex, France;* [2]*OREME – Station Marine, Université Montpellier, 2 rue des Chantiers - CC 99009, 34200 Sète, France; and* [3]*INRA, UMR1349 Institute for Genetics, Environment and Plant Protection, 35650 Le Rheu, France*

Summary

1. Partially, clonal species are common in the Tree of Life. And yet, population genetic models still mostly focus on the extremes: strictly sexual versus purely asexual reproduction. Here, we present an R package built upon GENCLONE software including new functions and several improvements.

2. The RClone package includes functions to handle clonal data sets, allowing (i) checking for data set reliability to discriminate multilocus genotypes (MLGs), (ii) ascertainment of MLG and semi-automatic determination of clonal lineages (MLL), (iii) genotypic richness and evenness indices calculation based on MLGs or MLLs and (iv) describing several spatial components of clonality. RClone allows the one-shot analysis of multipopulation data sets without size limitation, suitable for data sets now increasingly produced through next-generation sequencing.

3. A major improvement compared to existing software is the ability to determine the threshold to cluster similar MLGs into MLLs, based on implemented simulations of sexual events. Several functions allow data importation, conversion and exportation with adegenet, Genetix or Arlequin.

4. RClone is provided with two vignettes to handle analysis on one (*RClone_quickmanual*) or several populations (*RClone_qmsevpops*).

Key-words: clonal diversity, clonal population, clonality, multilocus genotypes, multilocus lineages, software, spatial autocorrelation

Clonality is a widespread trait across the Tree of Life (Halkett, Simon & Balloux 2005) allowing organisms to produce offspring without sexual reproduction. These offspring/clones are genetically identical to their relatives, at the exception of somatic mutations. A key step in genetic analysis of potentially clonal data set involves a genotypic analysis to discriminate multilocus genotypes (MLGs). The study of clonality includes its detection as well as the estimation of its quantitative and qualitative consequences on the demographic and evolutionary trajectories of populations. The study of clonality thus requires the ability to distinguish between two central components that are the demographic individual (ramet, i.e. demographic unit which could be a module in clonal plant, a colony in corals or an individual aphid in insects) and the genetic individual (genet, i.e. cluster of ramets that are all derived from a single event of sexual reproduction followed by clonal multiplication, *in a clonally propagating organism, the entity that persists and evolves* – Ayala 1998). For this purpose, molecular markers might be able to identify ramets from the same genet, as those are supposed to share the same MLG. However, slightly distinct MLGs may belong to the same clonal lineage, due to the occurrence of somatic mutations (Klekowski 2003), and scoring errors (Douhovnikoff & Dodd 2003; Meirmans & Van Tienderen 2004). Working with large data sets, which is now made possible by new-generation sequencing (NGS), increases the probability to have to manage uncommon somatic mutations and decreases the time allocated to resolve scoring errors ambiguities. To make ecological and evolutionary analyses possible and easier considering such issue, the concept of multilocus lineages (MLL) was thus introduced (Arnaud-Haond *et al.* 2007b) to define clusters of MLGs belonging to the same genet, therefore sharing the same original event of sexual reproduction, but appearing slightly distinct either due to somatic mutations or scoring errors.

The definition of MLLs relies on the computation of genetic distances between pairs of MLG and the determination of a threshold distance distinguishing pairs of ramets belonging to the same genets and pairs of ramets from the same genets. Some studies used *a priori* fixed genetic distances (number of different alleles between genotypes) and *a posteriori* methods to ascertain the occurrence of MLLs (incompatibility and deletion procedures, Van Der Hulst *et al.* 2003; use of *Psex* calculation: probability that the same genotype – nonidentical loci apart – appears through independent sexual reproduction

*Correspondence author. E-mail: sarnaud@ifremer.fr

events, among clustered MLG, Arnaud-Haond *et al*. 2007b). The threshold could be computed by comparing the tail of the clonal distribution with that of the sibling distribution (Douhovnikoff & Dodd 2003). Cluster algorithms such as UPGMA combined to genetic distances were also used to define lineages (Kamvar, Brooks & Grünwald 2015). A more adequate concept would be similar to gap analysis (ABGD, Automatic Barcode Gap Discovery, Puillandre *et al*. 2012) to determine threshold distance allowing to discriminate species based on sequence data. The use of a graphical visualization of the threshold, defined as a valley into the frequency distribution of pairwise distances of the data set, was proposed (Meirmans & Van Tienderen 2004) and used alone (Clark & Jasieniuk 2011) or in combination with other methods, such as *Psex* (Arnaud-Haond *et al*. 2007a). However, the choice of the threshold leading to the clustering of MLG into MLLs was still somehow arbitrarily user defined.

The choice of the threshold should be based on the knowledge of lowest distances susceptible to derive from sexual reproduction *versus* the largest distances compatible with clonal reproduction, that is comparison of Genetic Distance Spectrum (GDS). Arnaud-Haond *et al*. (2007a) suggested that a more accurate option would be to compare the GDS of seeds or seedling from that of ramets, to determine specifically such a threshold for a given species and a given set of markers. Although this strategy is more accurate, information about the GDS among seeds or seedling is seldom available. RClone thus proposes to simulate a pseudo-observed generation of pure sexual reproduction from the genotyped population and to compare the genetic distances of the genotyped population and the pseudo-observed generation. This ideally leads to the identification of a threshold (the minimum genetic distance among sexually produced MLG) allowing the discrimination of clonal lineages (corresponding to MLLs). It can be noted that the ability to determine MLLs delineation without ambiguity, and thus to use an automatic algorithm to define MLLs, will strongly depend on the discriminative power of the data set analysed (see Example section).

The discrimination of clonal lineages and the assemblage of similar MLG into MLLs is based on the examination of the GDS based on genetic distances (i) for each pair of unique MLG in the sample initially characterized molecularly, compared (ii) to each pair of unique MLG sexually produced from simulations. Sexual events could be simulated either through pure outcrossing or through a combination of outcrossing and selfing (selfing occurring at a rate determined by the amount of clonal replicates present in the sample). Genetic distances computed can be based on the number of different alleles (Chakraborty & Jin 1993) or on the difference in length between alleles (Rozenfeld *et al*. 2007) which relies on a similar rational as Bruvo distance (Bruvo *et al*. 2004). The occurrence of MLLs can therefore be extracted directly from the graphical output of GDS. Indeed, rather narrow peaks at the very beginning of the distribution, followed by a 'gap', preceding the peak at larger distances are a signature of the possible occurrence of MLLs (Meirmans & Van Tienderen 2004). The superposition of this GDS with one obtained after a sexual reproduction event can thus be used to ascertain the threshold below which independent events of sexual reproduction are unlikely to have produced slightly distinct MLGs. The threshold can be defined either in percentage of the distribution or directly at a distance value and is kept at user discretion. When the threshold is defined, the function extracts potential clusters of MLGs belonging to the same MLLs in a table and computes the resulting MLLs list automatically. This function is designed mainly for data including a large number of loci, as data set including too few loci might erroneously cluster MLG into MLLs. For such data set, it is thus strongly advisable to opt for a manual examination of GDS to decide whether or not MLLs can be defined, an option included in RClone.

Example

We generated two *in silico* pseudo-observed populations (here called *synthetic* populations) using a simulation program written with Python 2.7 (Foundation Python Software, available on: http://wwz.ifremer.fr/clonix/Logiciels/SimulatorClonix). The simulation was conducted among 10 000 generations, with meiotic mutation rate of 10^{-3} and somatic mutation rate of 10^{-6}. One population was of 10 000 individuals with a clonality rate (c) of 0·9999, and the other was of 1000 individuals with $c = 0{\cdot}4$. Each population was of 100 loci, 10 alleles possible per loci. The *synthetic* populations obtained were analysed with RClone package (v1.0.1, GPL license) in R (R Core Team 2015) with a method similar to vignette *RClone_quickmanual*. The function *genet_dist* computed a matrix of pairwise genetic distances between pairs of unique MLGs of the *synthetic* population, based on number of different alleles. The function *genet_dist_sim* allowed to simulate a sexual reproduction event between pairs of unique MLGs (outcrossing) or pairs of MLG (partial selfing) of the *synthetic* population. *genet_dist_sim* computed matrices of pairwise genetic distances within the resulting population, here called *simusex* generation. The three matrices were visualized as superimposed histograms to identify the occurrence of a valley between them, through which we could identify a diagnostic threshold. This threshold is defined by the interval between the lowest boundaries of distances among sexually produced MLGs compared to the possible narrow peak at small distances which maximal value should not exceed this lower boundary. Only, somatic mutations could lead to the first peaks of the *synthetic* distribution, as scoring errors would not occur *in silico* neither artefact such as mutations induced by PCR (polymerase chain reaction).

From the first *synthetic* population ($c = 0{\cdot}9999$), 29 MLG were identified, encompassing 1–4322 ramets (median: 58, first quartile: 12, third quartile: 327). Among the 406 pairs of these genotypes, the 41 first pairs of genets were distinct for only 1–4 alleles (Fig. 1). The following pairs of genets were distinct from 74 alleles. The distribution of genetic distances obtained in the *simusex* generation (including or not selfing) matched mainly with the second peaks. The mean genetic distance between pairs after a sexual event was of 108·9 alleles (minimum: 24

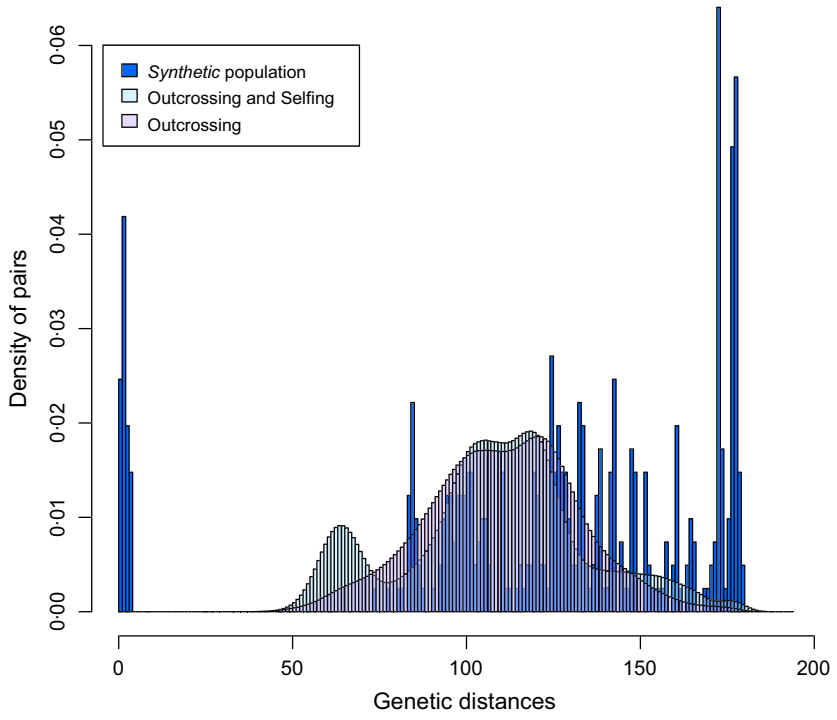

Fig. 1. Frequency distribution of the pairwise number of alleles differences between MLG for the *synthetic* population (population *in silico*: $c = 0{\cdot}9999$, $N = 10\ 000$), compared with the frequency distribution of the pairwise distances after 10 000 sexual events (one generation each, outcrossing and selfing) in which neither identical MLG nor somatic mutations are expected, and with the frequency distribution of the pairwise distances after 10 000 sexual events (one generation each, outcrossing), without selfing.

alleles) with selfing and 111·1 alleles (minimum: 14 alleles) without selfing. The gap between the two main peaks defined without a doubt lineages that encompassed MLGs distinct from four or fewer alleles. The clustering of those MLGs resulted in the definition of 17 MLLs, obtained with function *MLL_generator*, encompassing 1–5797 ramets (median: 37, first quartile: 3, third quartile: 468).

For the second *synthetic* population ($c = 0{\cdot}4$), 944 MLG were identified encompassing 1–5 ramets (mean: 1·06, SD: 0·30; median, first quartile and third quartile: 1). Among those, the 30 first pairs of genet were distinct for 1–4 alleles (Fig. 2). The following pairs of genets were distinct from at least 58 alleles. The mean genetic distance after a sexual event in the *simusex* generation was of 107·1 alleles (minimum: 65 alleles) with selfing and 107·3 alleles (minimum: 51 alleles) without selfing. With a threshold of 4 different alleles, the clustering results in the recognition of 915 MLLs, encompassing from 1 to 5 ramets (mean: 1·09, SD: 0·36).

Either MLGs or MLLs lists obtained can furthermore be used with others functions of the package RClone. Briefly, the most used clonal-related indices in the literature can be computed with *clonal_index*: *R*, the richness index; *S*, the Simpson index applied to the clonality; *Hill*, the inverse index; *H''*, the Shannon–Wiener index; and corresponding evenness indices: *V*, the Simpson evenness index and *J'* Shannon–Wiener index. The description of the distribution of size (in terms of number of sampling units) among MLGs or MLLs can also be addressed through the Pareto distribution (Pareto 1887 in Vidondo *et al.* 1997) with *Pareto_index*. When sampling coordinates are available, functions allowing to study the spatial components of clonality are also included: *edge_effect* computed edge effect by comparison of the average distance between lineages of single unit and the centre of the sampling area with the average distance of all units and the centre. *agg_index* computed aggregation of clones by comparison of the probability of clonal identity between pairs of closest spatially units to that of all randomly chosen units pairs. Spatial autocorrelation analyses are used to determinate the scale of spatial dependence of clonal and genetic diversities, with autocorrelation. *clonal_sub* computed clonal subrange which corresponds to the spatial scale beyond which the clonality no longer affects the genetic structure. Finally, a recapitulative function *genclone* gathered 17 statistics describing clonal data and, when relevant, significance (*P*-values).

RClone, the natural and legitimate R version of the popular software dedicated to the analysis of clonal population, GENCLONE (Arnaud-Haond & Belkhir 2007), proposes all functions implemented in this mother software including the study of spatial components of clonality. RClone also implements a semi-automatic procedure to define MLLs and compatibility of MLL with others functions of the package. RClone also relaxes previous limitations of GENCLONE software in terms of number of samples and loci, and multi-population handling, and enables the analyses of large data sets derived from the broader access to NGS. RClone also includes *Psex* ascertainment by critical probabilities (*P*-values) based on populations simulations, a method derived from MLGsim (Stenberg, Lundmark & Saura 2003) and MLGSIM 2.0 (Ivens, van de Sanden & Bakker 2012) implemented with the authors' permission. RClone thus provides, considering also the spatial components included, different functions than the only other R package dedicated to clonality (poppr, Kamvar, Tabima & Grünwald 2014) which is more focused on the preparation of clonal data set for their analysis with classical population genetics software and adegenet (Jombart 2008; Jombart &

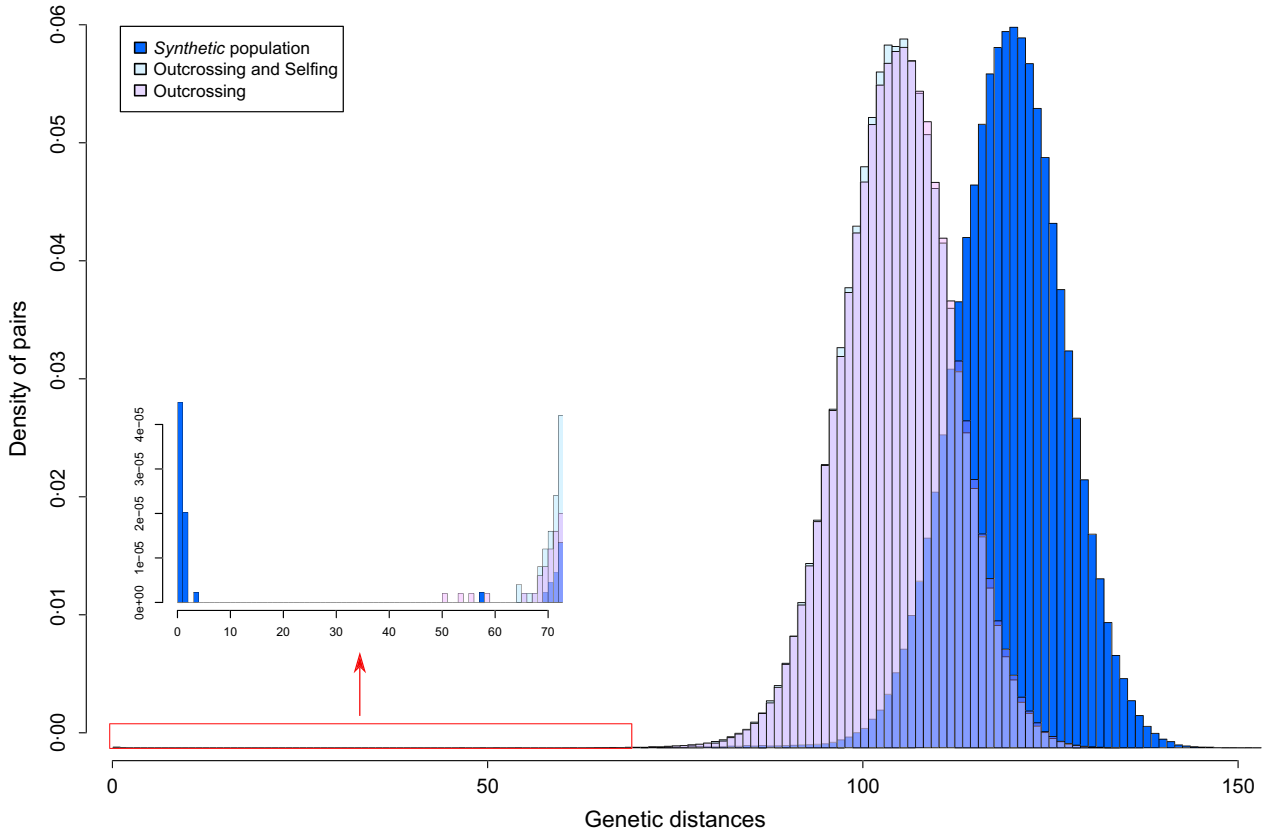

Fig. 2. Frequency distribution of the pairwise number of alleles differences between MLG for the *synthetic* population (c = 0·4, N = 1000), compared with the frequency distributions of the pairwise distances after 1000 sexual events accounting selfing and after 1000 sexual events without selfing. The small graph corresponds to a zoom of the area in red with a y-scale in terms of number of pairs instead density of pairs.

Ahmed 2011). Moreover, RClone on CRAN offers the benefits of an active and collaborative open-source platform: code availability, reproducible research and data transfer among packages.

The RClone functions require genotypes for co-dominant markers, indication for the haploid or diploid nature of the organism and x/y sampling coordinates for spatial analyses. Missing data are not supported yet as such and thus considered as new alleles if included. The functions available are distributed into four main themes: (i) tests checking for data set reliability to discriminate MLG, (ii) determination of clones among MLG and through clonal lineages with genetic distances, (iii) genotypic richness and evenness indices calculation with MLG or custom MLL and (iv) description of spatial aspects of clonality. Several functions allow data importation, conversion and exportation with adegenet, Genetix (Belkhir *et al.* 1996-2004) or Arlequin (Excoffier, Laval & Schneider 2005).

In conclusion, the RClone package provides adapted methods and statistics for clonal population study that are not yet implemented into R platform (aside clonal richness indices). Based on GENCLONE, RClone adds multiple populations handling, methods to test for clonal propagation from Stenberg, Lundmark & Saura (2003) and Ivens, van de Sanden & Bakker (2012), reinforcement of MLL determination methods, semi-automatic MLL definition and custom MLL handling in several tests. RClone is available on CRAN, provided with short manuals to handle quickly the functions: *RClone_quickmanual* vignette for one population and *RClone_qmsevpops* vignette for several populations. RClone will be kept active and implemented soon to handle missing data, other clonal indices, an IDE interface and possible suggestions included in users' feedbacks.

Acknowledgements

Diane Bailleul was funded by the French ANR project Clonix (ANR-11-BSV7-0007), and we wish to thank all the partners of the project for useful discussions and beta testing of the program, with a special thanks to Barbara Porro. We thank Per Stenberg for his help with MLGsim integration and acceptance, as well as Aniek B.F Ivens and Joke Bakker. We also thank the reviewers for their helpful comments and suggestions that improved the article and the package.

Data accessibility

Documentation and source code are freely available on CRAN (http://cran.r-project.org/web/packages/RClone) and GITHUB (https://github.com/dbailleul/RClone).

Genotypic data: Dryad (accession number: doi: 10.5061/dryad.9m8ff).

References

Arnaud-Haond, S. & Belkhir, K. (2007) GENCLONE: a computer program to analyse genotypic data, test for clonality and describe spatial clonal organization. *Molecular Ecology Notes*, **7**, 15–17.

Arnaud-Haond, S., Duarte, C.M., Alberto, F. & Serrão, E.A. (2007a) Standardizing methods to address clonality in population studies. *Molecular Ecology*, **16**, 5115–5139.

Arnaud-Haond, S., Migliaccio, M., Diaz-Almela, E., Teixeira, S., Van De Vliet, M.S., Alberto, F., Procaccini, G., Duarte, C.M. & Serrão, E.A. (2007b) Vicariance patterns in the Mediterranean Sea: east–west cleavage and low dispersal in the endemic seagrass *Posidonia oceanica*. *Journal of Biogeography*, **34**, 963–976.

Ayala, F.J. (1998) Is sex better? Parasites say "no". *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3346–3348.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. & Bonhomme, F. (1996-2004) *GENETIX 4.05, Logiciel Sous Windows TM Pour la Génétique des Populations*. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier, France.

Bruvo, R., Michiels, N.K., D'Souza, T.G. & Schulenburg, H. (2004) A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology*, **13**, 2101–2106.

Chakraborty, R. & Jin, L.I. (1993) Determination of relatedness between individuals using DNA fingerprinting. *Human Biology*, **65**, 875–895.

Clark, L.V. & Jasieniuk, M. (2011) polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, **11**, 562–566.

Douhovnikoff, V. & Dodd, R.S. (2003) Intra-clonal variation and a similarity threshold for identification of clones: application to Salix exigua using AFLP molecular markers. *Theoretical and Applied Genetics*, **106**, 1307–1315.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Halkett, F., Simon, J.-C. & Balloux, F. (2005) Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology & Evolution*, **20**, 194–201.

Ivens, A.B.F., van de Sanden, M. & Bakker, J. (2012) MLGsim 2.0: updated software for detecting clones from micro satellite data using a simulation approach. In: *Evolutionary Ecology of Mutualism*, pp. 107–111. PhD Thesis, University of Groningen.

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jombart, T. & Ahmed, I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Kamvar, Z.N., Brooks, J.C. & Grünwald, N.J. (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, **6**.

Kamvar, Z.N., Tabima, J.F. & Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Klekowski, E.J. (2003) Plant clonality, mutation, diplontic selection and mutational meltdown. *Biological Journal of the Linnean Society*, **79**, 61–67.

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

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rozenfeld, A.F., Arnaud-Haond, S., Hernández-García, E., Eguíluz, V.M., Matías, M.A., Serrão, E. & Duarte, C.M. (2007) Spectrum of genetic diversity and networks of clonal populations. *Journal of the Royal Society Interface*, **4**, 1093–1102.

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

Van Der Hulst, R.G.M., Mes, T.H.M., Falque, M., Stam, P., Den Nijs, J.C.M. & Bachmann, K. (2003) Genetic structure of a population sample of apomictic dandelions. *Heredity*, **90**, 326–335.

Vidondo, B., Prairie, Y.T., Blanco, J.M. & Duarte, C.M. (1997) Some aspects of the analysis of size spectra in aquatic ecology. *Limnology and Oceanography*, **42**, 184–192.

Received 20 January 2016; accepted 30 January 2016
Handling Editor: Timothée Poisot