Molecular Ecology (2007) **16**, 5115–5139 doi: 10.1111/j.1365-294X.2007.03535.x

INVITED REVIEW

Standardizing methods to address clonality in population studies

S. ARNAUD-HAOND,* C. M. DUARTE,† F. ALBERTO* and E. A. SERRÃO*
**CCMAR — CIMAR Laboratório Associado, Univ. Algarve, Gambelas, 8005-139, Faro, Portugal, †IMEDEA, CSIC-Univ. Illes Balears, C/Miquel Marques 21, 07190 Esporles, Mallorca, Spain*

Abstract

Although clonal species are dominant in many habitats, from unicellular organisms to plants and animals, ecological and particularly evolutionary studies on clonal species have been strongly limited by the difficulty in assessing the number, size and longevity of genetic individuals within a population. The development of molecular markers has allowed progress in this area, and although allozymes remain of limited use due to their typically low level of polymorphism, more polymorphic markers have been discovered during the last decades, supplying powerful tools to overcome the problem of clonality assessment. However, population genetics studies on clonal organisms lack a standardized framework to assess clonality, and to adapt conventional data analyses to account for the potential bias due to the possible replication of the same individuals in the sampling. Moreover, existing studies used a variety of indices to describe clonal diversity and structure such that comparison among studies is difficult at best. We emphasize the need for standardizing studies on clonal organisms, and particularly on clonal plants, in order to clarify the way clonality is taken into account in sampling designs and data analysis, and to allow further comparison of results reported in distinct studies. In order to provide a first step towards a standardized framework to address clonality in population studies, we review, on the basis of a thorough revision of the literature on population structure of clonal plants and of a complementary revision on other clonal organisms, the indices and statistics used so far to estimate genotypic or clonal diversity and to describe clonal structure in plants. We examine their advantages and weaknesses as well as various conceptual issues associated with statistical analyses of population genetics data on clonal organisms. We do so by testing them on results from simulations, as well as on two empirical data sets of microsatellites of the seagrasses *Posidonia oceanica* and *Cymodocea nodosa*. Finally, we also propose a selection of new indices and methods to estimate clonal diversity and describe clonal structure in a way that should facilitate comparison between future studies on clonal plants, most of which may be of interest for clonal organisms in general.

Keywords: clonal diversity, clonal size, clonal subrange, clonality, methods, molecular markers, power law, sampling design, spatial autocorrelation, species richness

Received 15 May 2007; revision 27 July 2007

Introduction

Clonality is a life-history strategy, particularly widespread in plants, allowing organisms to produce offspring without sexual reproduction, hence typically genetically identical — at the exception of possible somatic mutations — to themselves. Despite the large number of clonal species present across a wide variety of taxa and habitats, evolutionary theory and models are mostly based on singular genetic individuals. A specific consideration of clonality is largely lacking, probably because ecological and particularly evolutionary studies of clonal plants have long been deterred by the difficulty in discriminating between genetically distinct individuals and clonal replicates [i.e. to discriminate

Correspondence: S. Arnaud-Haond, E-mail: sarnaud@ifremer.fr; eserrao@ualg.pt
Present address: Ifremer, Centre de Brest BP70, Department DEEP, 29280 Plouzané, France

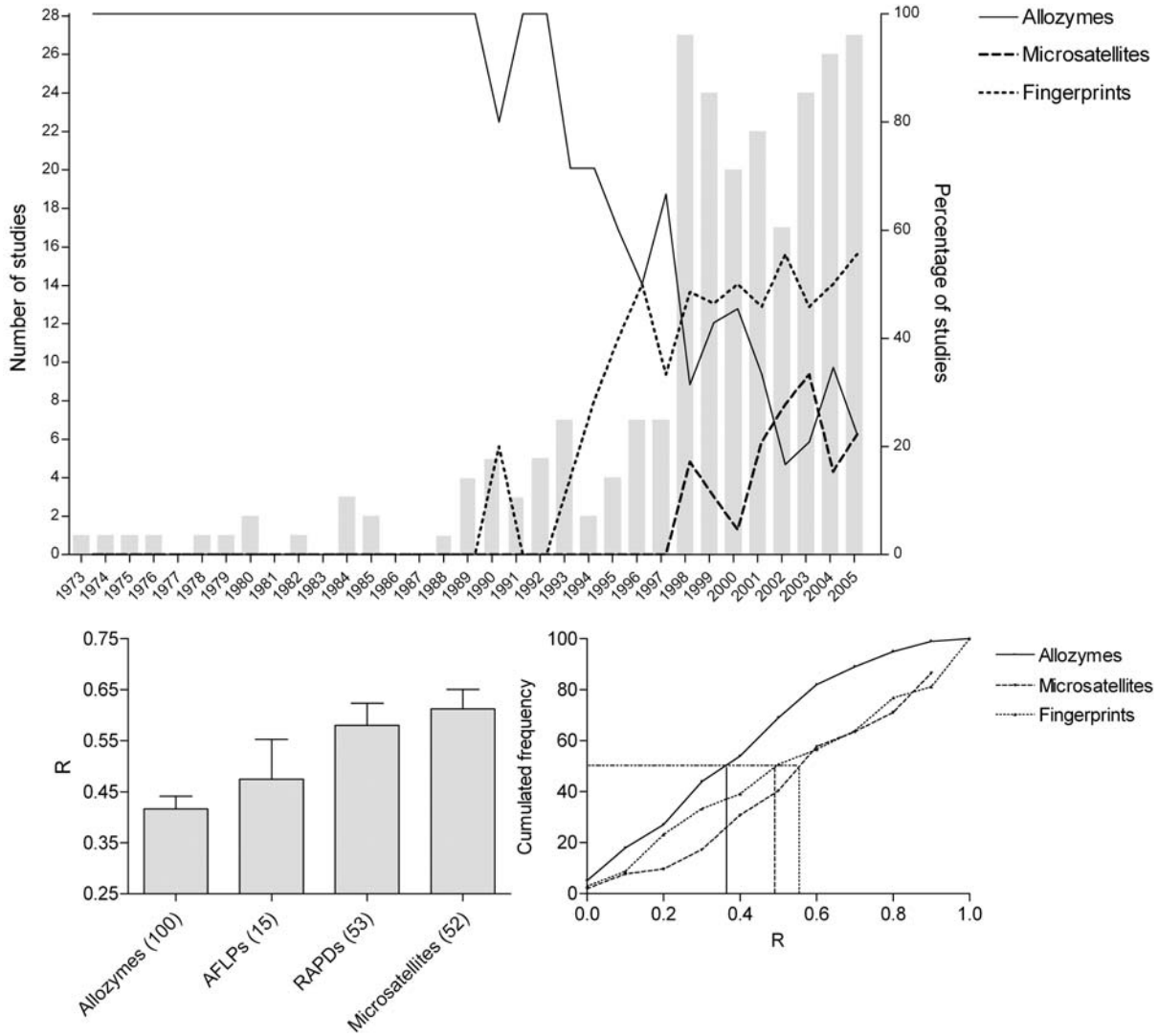

Fig. 1 (a) Time course of the number of studies on clonal plants using molecular markers per year, among the 247 published studies on clonal plants reviewed (bars), and the temporal evolution of the percentage of studies using allozymes (–), multibanding (RAPDs, AFLP and fingerprints; ···) and microsatellites (---). (b) The distribution of clonal diversity (R) estimated with Allozymes, Fingerprints (RAPDs, AFLP), and Microsatellites markers over the 297 studies reviewed, presented as the average (± SE) for the studies using different marker types on the left panel, and as the cumulated frequency of increasing *R*-values on the right panel with lines pointing at the median values of R for different marker types.

between distinct *genets* and distinct *ramets*; *sensu* Harper (1977)]. The advent and subsequent development of markers powerful enough to resolve genotypic identity has now bypassed that bottleneck, stimulating research efforts towards the examination of the genetic structure of clonal plant populations. This is indicated by the fact that 83% of the articles on clonal plants published in that area over the past three decades, as revealed by a literature search on the ISI Web of Knowledge, were produced after 1995 (Fig. 1a). The bulk of these articles characterized the genetic structure of clonal populations through the computation of general indices of genetic structure, such as heterozygosity, *F* estimators or spatial autocorrelation analysis, all methods developed for nonclonal organisms and therefore not explicitly addressing the issue of clonality. Yet the clonal nature of the populations poses specific challenges that impinge on their genetic structure, and this introduces some uncertainties in the interpretation of results derived in the past. Moreover, the implications of the clonal nature of the organisms studied are so pervasive that clonality affects the study of population genetics even at the sampling stage. This aspect has not been specifically addressed as yet, possibly leading to errors in the use and interpretation of the indices applied.

A substantial fraction of the research effort has attempted to characterize the extent of clonality in populations through the use of diversity indices, borrowed from the species' diversity literature. These include the ratio of the number of genotypes (or clonal lineages) over the number of samples (Ellstrand & Roose 1987), the Shannon-Wiener index (Pielou 1966; Peet 1974), the complement of Simpson's index (Gini 1912; Simpson 1949) and the corresponding evenness indices. However, the use of different indices across studies precludes an efficient and useful comparison of their results in terms of clonal diversity. In general, none of the available software for general population genetics analyses includes routines and options for clonal organisms, signalling a lack of sufficient awareness of the specificities of clonality and the need for a standardized set of indices and methods. Some specific software have been developed in the last few years, allowing the analysis of some clonal components at the intrapopulation levels (Stenberg *et al*. 2003; Meirmans & Van Tienderen 2004; Peakall & Smouse 2006; Arnaud-Haond & Belkhir 2007). Also, none of the calculations used so far specifically consider how different clones are distributed in space, which is a fundamental trait of the genetic structure of clonal populations (van Groenendael & de Kroon 1990; Reusch 2001), and it was only very recently that a software was released allowing those features to be specifically analysed for clonal organisms (Arnaud-Haond & Belkhir 2007). Hence, there is a need to standardize the methods used to characterize the genetic structure of clonal organisms both in order to facilitate the gathering and integration of future data and their comparison among studies.

Here we provide an overview, on the basis of a review of the published literature, of current methods to assess the genetic structure of clonal plant populations and formulate new methods where appropriate. We specifically focus on indices and statistics to (i) relate genotypic and clonal identity, (ii) describe clonal diversity, and (iii) describe the spatial pattern of clonal distribution. We examine the properties of the statistics most commonly encountered in the literature, on the basis of simulated and empirical microsatellite data sets of populations of the clonal seagrasses *Posidonia oceanica* and *Cymodocea nodosa* (Alberto *et al*. 2003a, b, 2005) used as test cases. These simulated and empirical data sets are also used to examine and discuss the implications of clonality for sampling design.

Literature survey

We searched the published literature for studies using molecular markers to assess population genetic structure of clonal plants published between 1973 and 2003. We did so by searching the ISI Web of Knowledge for entries of published studies including the terms 'plants' and ('clonality' or 'clonal' or 'clone' or 'asexual') and a variety of molecular markers [e.g. allozymes, microsatellites, random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeats), and screening the references obtained for molecular analysis of clonal plants. A first screening of the literature delivered about 450 studies, of which further scrutiny revealed only about 280 to be relevant, 246 of which could be retrieved and analysed. Additionally, searches on genetic structure of nonplant clonal organisms were also conducted, for articles published between 2000 and 2005, of which 51 were analysed. The list of those references can be found in Table S1, Supplementary material, summarizing the information extracted from each article. For each article, the methods used to estimate and describe clonal diversity and spatial clonal distribution, as well as the spatial design of the sampling were extracted (Table S1, Table 1).

The examination of the publication trends shows a major growth in the number of published studies on population structure of clonal plants using molecular markers (Fig. 1a), as well as a shift in the relative use of different markers. The publication effort on population structure of clonal plants increased abruptly in 1998 coinciding with the advent of the use of microsatellite markers (Fig. 1a). All published studies used allozymes until the early 1980s, when the introduction of fingerprinting approaches in the literature led to a shift in methods followed by an uprise in the use of microsatellites as the most powerful markers to assess clonal membership yet available (Fig. 1a).

Genotypic vs. clonal membership, estimating sexual input

The genotyping of sampling units, or ramets, with multiple independent markers will allow their assignment to several groups of multilocus genotypes (MLGs). Two additional steps are necessary before being able to reasonably assume that (i) all replicates of the same MLG are part of the same clone, or genet; and (ii) each distinct MLG belongs to a distinct clone, or genet (Halkett *et al*. 2005b). The first part requires estimating the probability of finding identical MLGs resulting from distinct zygotes, and the second requires a careful analysis of the pairwise differences among MLGs in order to detect possible somatic mutations or scoring errors that may result in distinct MLGs characterizing sampling units actually belonging to the same clone. Procedures to accomplish both these steps are detailed below and illustrated in Box 1.

The analysis of clonal populations requires the capacity to assess the likelihood that two individuals with the same multilocus genotype, within the power of the markers used, are indeed part of the same clone and therefore unlikely to be derived from distinct sexual reproductive events. These tests have been used in about 30% of the reviewed studies. For the calculation of this probability, the population allelic

Box 1 Genotypic vs. clonal membership

a) Assessing whether all replicates of the same MLG are part of the same clone

The probability of a given genotype i under the assumption of Hardy–Weinberg equilibrium can be estimated as:

$$p_{\text{gen}} = \sum_{i=1}^{l} (f_i)2^h \qquad \text{(eqn 1)}$$

where l is the number of loci, f_i the frequency of each allele at the i^{th} locus (estimated using the round-robin method, see text), and h the number of heterozygous loci in the sample.

When taking into account departures from Hardy–Weinberg equilibrium (using F_{IS}), this equation becomes:

$$p_{\text{gen}}(F_{\text{IS}}) = \prod_{i=1}^{l} [(f_i g_i) \times (1 + (z_i \times (F_{\text{IS}(i)})))]2^h \qquad \text{(eqn 2)}$$

where l is the number of loci, h is the number of heterozygote loci, and f and g are the allelic frequencies of the alleles f and g at the i^{th} locus (with f and g identical for homozygotes), $F_{\text{IS}(i)}$ is the F_{IS} estimated for the i^{th} locus (using allelic frequencies estimated with the round-robin method), and $z_i = 1$ if the i^{th} locus is homozygous (for $f_i = g_i$) and $z_i = -1$ if the i^{th} locus is heterozygous.

When the same genotype is detected n times in a sample of N sampling units, the probability that the repeated genotypes originate from distinct sexual reproductive events (i.e. from different zygotes, thus being different genets), derived from the binomial expression, is:

$$p_{sex} = \sum_{i=n}^{N} \frac{N!}{i!(N-i)!} [p_{\text{gen}}]^i [1 - p_{\text{gen}}]^{N-i} \qquad \text{(eqn 3)}$$

In this calculation, the probability of the genotype p_{gen} can be replaced by $p_{\text{gen}}(F_{\text{IS}})$ to consider possible departures to Hardy–Weinberg equilibrium, in order to obtain a more conservative estimate of p_{sex}.

A Monte Carlo procedure can be applied to ensure that the set of loci used provides enough power to discriminate all MLGs present in the sample:

Fig. B1.1: Box plot describing the genotypic resolution of microsatellites in a data set of the seagrass *Cymodocea nodosa* containing 220 sampling units genotyped using nine microsatellites, analysed for of all possible combinations C_l^K of K loci ($K = 1, \ldots, l$; l is the number of loci available). the edges of the boxes show the minimum and maximum number of genotypes and the central line shows the average number of genoptypes identified in the sample using X microsatellites (Alberto *et al*. 2005). The example illustrated here shows that a set of seven loci allows an accurate determination of the number of genotypes in the sample.

b) Ascertaining that each distinct MLG belongs to a distinct clone, or genet (Halkett *et al*. 2005a); *defining clonal lineages (MLL)*

This procedure can be used if the distribution of genetic distances among sampling units does not follow a strict unimodal distribution but shows high peaks toward low distances, susceptible to reveal the existence of somatic mutations or scoring errors in the data set resulting in low distances among slightly distinct MLG actually deriving from a single reproductive event. The use of the frequency distribution of distances to detect such events

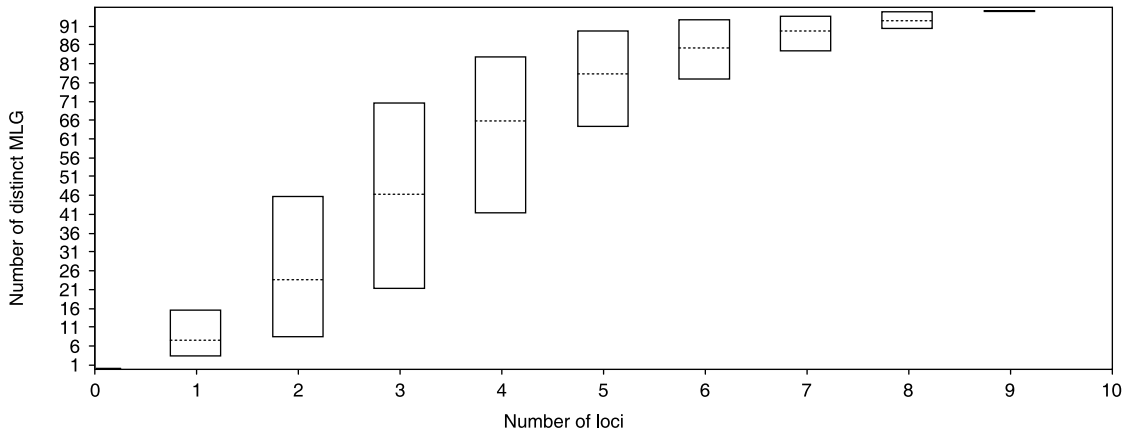

Fig. B1.1

Box 1 *Continued*

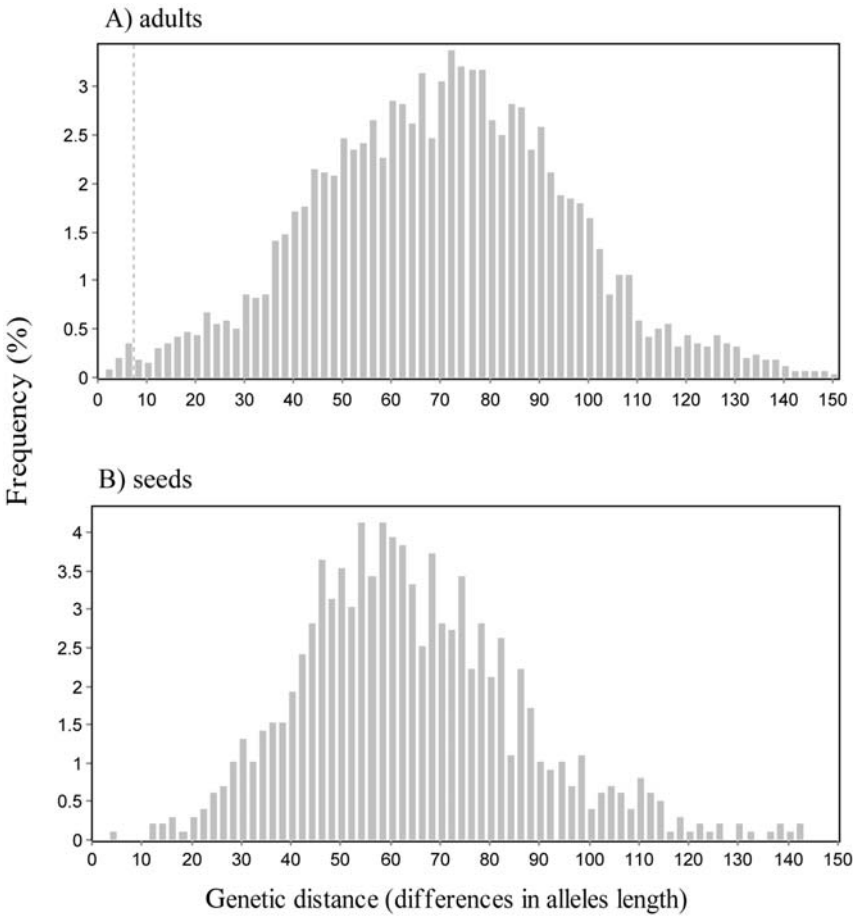

Fig. B1.2

has been proposed four times so far, to our knowledge (Douhovnikoff & Dodd 2003; Meirmans & Van Tienderen 2004; Arnaud-Haond *et al*. 2005; Rozenfeld *et al*. 2007). In a recent work on *Posidonia* (Arnaud-Haond *et al*. 2007) we introduced the concept of MLL to design genets represented by slightly distinct MLG, due to mutation or scoring errors. We propose a two step approach, consisting in (i) screening each MLG pair presenting extremely low distance, and originating a primary small peak in the frequency distribution of distances, making it bimodal rather than unimodal (see the dashed line in Fig. B1.2). Then we propose (ii) using p_{sex} on the set of identical loci in order to estimate the likelihood that those slightly distinct MLG would actually be derived from distinct reproductive events. When such likelihood was lower than a chosen threshold (in that case 0.01), then the slightly distinct MLG may be considered as being derived from the same genet and being slightly distinct representatives of the same MLL. Numerous distance metrics can be chosen, such as the number of distinct alleles, Jaccard similarity in particular for multibanding patterns (Douhovnikoff & Dodd 2003) or the number of microsatellite motifs (Arnaud-Haond *et al*. 2007) under the hypothesis of a stepwise mutation model for somatic mutations.

Fig. B1.2: (A) Frequency distribution of the pairwise number of alleles differences between MLGs for the same sample of *C. nodosa* (Alberto *et al*. 2005), compared with (B) the frequency distribution of the pairwise distances in a set of seeds from the same location (Cadiz, Spain) in which neither identical MLG nor somatic mutation are expected. The *x*-axis represents the number of allele differences and the *y*-axis is the frequency distribution for each x rank. The dashed line in the adult distribution represents the threshold below which identical MLG have a p_{sex}, estimated after excluding the slightly different loci, that supports the slightly distinct MLG as having originated from the same MLL (i.e. from the same zygote).

frequencies can be estimated using a 'round-robin' method (Parks & Werth 1993; Arnaud-Haond *et al*. 2005). This subsampling approach avoids the overestimation of the rare allele frequencies, by estimating the allelic frequencies for each locus on the basis of a sample pool composed of all the MLGs distinguished on the basis of all the loci, except that for which allelic frequencies are estimated. This procedure is repeated for all loci, and the unique genotype probability (p_{gen}) is then estimated under the assumption of Hardy–Weinberg equilibrium (Box 1, equation 1).

A constraint on this procedure is the possible occurrence of departures from panmixia in the population studied, as may occur due to selfing and biparental inbreeding, or high linkage disequilibrium. In these cases, the estimated probability p_{gen} may be significantly lower than the real probability of occurrence of a given repeated MLG originated from different zygotes. The corresponding p_{sex} may in those cases represent an underestimation of the likelihood of encountering this particular MLG twice or more. It has been proposed that the genetic composition of the population could be taken into account to improve the estimate of p_{gen}, by using samples collected at the zygote stages (for example seeds) in order to assess the level of linkage disequilibrium and departure from Hardy–Weinberg in the population of sexual individuals (Gregorius 2005). Yet, for those species, numerous among clonal plants, that experience large variance in reproductive success or variable selection regimes in space and time, this approach may not be realistic, or may even lead to more biased results than the classical estimates of p_{gen}. We therefore recommend the use of F_{IS} values obtained using allelic frequencies estimated with of the round-robin method, to improve estimates of p_{gen} by taking into account departures from Hardy–Weinberg equilibrium, as first suggested by Young *et al*. (2002: Box 1, equation 2).

These estimates of p_{gen}, or of the upper bound of its confidence interval, are often (about 13% of studies) used to ascertain whether replicated MLGs result from clonal reproduction. This is not appropriate, as the p_{gen} is the probability of finding a given MLG_i when analysing only one sampling unit, not the probability of finding that MLG_i in the N sampling units collected and analysed. A similar problem occurs with other methods, used in 5% of the articles reviewed, estimating the probability for a given MLG_i to occur n times due to sexual reproduction as p_{gen}^n. This calculation actually delivers the probability of finding n times the MLG_i when analysing exactly n sampling units, instead of the probability of MLG_i occurring n times in a sample of N sampling units. Therefore, these calculations do not address the question 'are one or more of the n replicates of a given MLG_i encountered in a sample of N sampling units likely to be issued from independent events of sexual reproduction?'. To address this question when the same genotype i is detected more than once (n) in a sample composed of N sampling units, the probability that the sampling units with the same genotype actually originate from distinct sexual reproductive events (i.e. from separate genets) is best derived from the binomial expression describing p_{sex} (Tibayrenc *et al*. 1990: Box 1, equation 3, Parks & Werth 1993), which has only been used in 6% of the articles reviewed.

In very particular cases of high clonal dominance and very low clonal diversity, a limitation exists to this method. First, it will not be known whether the estimates of allelic frequencies on the basis of very few sampled chromosome will accurately represent population allelic frequencies (if all existing genets have been included in the sample, as in a monoclonal population) or if there are many more genets in the population but which the sampling scheme was unable to detect. Second, and above all, the low statistical power in such a data set is likely to lead to nonsignificant probabilities p_{sex}, thus not allowing exclusion of the possibility that the most common MLGs would have occurred independently several times in the studied population as a result of distinct events of sexual reproduction. Such situation is paradoxical as this implies that in the cases where the dominance of clonality would be more obvious, it may not be possible to demonstrate its occurrence statistically. One recommendation in such cases may be the increase in sample size, or the extension of the sampling area, to attempt collecting more distinct and rare MLG, if they exist in the population . The increase in the number of distinct MLGs sampled would indeed increase the reliability of allelic frequency estimates and the statistical power to ascertain the clonal identity of the numerous identical MLGs. If however, a population contains only one or only a few genotypes, even with very high sampling effort no further MLGs are detected, and although the allelic frequencies of the population are exhaustively sampled, statistical power associated with p_{sex} may be low. The recommendation in those cases is to increase the number of variable loci in the analysis, towards levels at which the probability of finding the exact same MLG but originated from distinct zygotes, would be very low.

It may be wise to proceed with these tests of clonal identity for identical multilocus genotypes before engaging in analyses that assume these to derive indeed from the same clone. A further test for the likelihood of clonal identity between two samples with the same multilocus genotype may be to sample, using a Monte Carlo procedure, subsets of loci and examine the robustness of the inferred clonal membership to changes in the power of the analysis. Indeed, this procedure allows testing whether or not the power to discriminate the maximum number of distinct genotypes is satisfactorily reached with the number of markers used, thereby allowing the accurate estimation of the clonal diversity (Arnaud-Haond *et al*. 2005, see Box 1, Fig. B1.1).

Once the set of loci has been assessed to be powerful enough to resolve all distinct clones in a set of samples (i.e. each MLG corresponds to a single clone), the second step is to ascertain the clonal membership of each MLG (i.e. each clone corresponds to a single MLG). Indeed, the assignment of genetic identity of clones has recently been questioned (Klekowski 2003). Multiple MLGs belonging to the same clone may be found either due to the existence of somatic mutation or scoring errors (Douhovnikoff & Dodd 2003), which would lead to the overestimation of the number of clones in the sample analysed. This potential bias can be tested for by inspecting the frequency distribution of genetic distances among pairs of MLGs (Douhovnikoff & Dodd 2003; Meirmans & Van Tienderen 2004). The occurrence of somatic mutation or scoring errors at a significant rate is expected to be reflected in the existence of a peak in the frequency distribution of genetic distances at very low, non-null, genetic distances (Douhovnikoff & Dodd 2003; Van der Hulst *et al*. 2003: see Box 1, Fig. B1.2A and B). A threshold of genetic distance can in those cases be defined, below which the hypothesis that distinct MLGs belong to the same clone cannot be rejected (Douhovnikoff & Dodd 2003; Meirmans & Van Tienderen 2004). These MLGs will then be assembled into groups of distinct 'multilocus lineages' (MLLs) corresponding to the best possible identification of distinct clonal lineages (Arnaud-Haond *et al*. 2007; Diaz-Almela *et al*. in press).

The tendency for studies to use a growing number of increasingly polymorphic markers will likely lead to an increase in the number of apparent MLGs relative to the number of MLLs in the sample, as more somatic mutations and scoring errors are expected as marker number and resolution increase. This suggests that the procedure described above should be routinely used to avoid bias in clonal diversity estimates (Loxdale & Lushai 2003). Although the concept of clone was first introduced by the ancient Greeks to design entities issued from asexual reproduction, and did not necessarily imply exact genetic identity (unlike the term genet, defined much later by Harper in 1977), it has been traditionally used in biology to refer both to biological units derived from asexual reproduction and those sharing genetic identity. Indeed, the capacity to ascertain genetic identity is a recent achievement, and the consequences of the ambiguity of the traditional use of the term 'clone' are only now becoming apparent (Tibayrenc & Ayala 2002). At this stage, the concept of 'clonal lineages', defined as 'the asexual descendants of a given genotype differing from the originator only via mutation and mitotic recombination' (Anderson & Kohn 1995) may therefore be more precise and operative than that of the more ambiguous term 'clones'.

Only once these tests have been conducted that the indices described below may be considered indices of clonal, and not genotypic, diversity, which is a requirement to assess the spatial distribution of the clonal lineages. It is indeed important to recognize that the terms 'clonal lineages' (or MLLs) and 'clonal' do not necessarily correspond to 'genotypes' (or MLGs) and 'genotypic', respectively. This step is also required to obtain reliable estimates of the rate of clonal vs. sexual reproduction. The successful assessment of the level of 'individual' or 'clonal lineage' (arising from a single zygote) through these two steps is also particularly important to further apply classical population genetic analyses such as F_{IS} or F_{ST}, or autocorrelation analysis (see below) in order to extract information on inbreeding, heterozygote selective values, dispersal and migration rate via sexual propagules vs. clonal spread.

One of the most common problems affecting the estimates reported in the literature is the lack of resolution due to the limited polymorphism of the markers used. This precludes the accurate discrimination of some distinct lineages that falsely appear identical, on the basis of the set of markers used, leading to the overestimation of clonal input (i.e. the underestimation of clonal diversity).The comparison of the average clonal diversity derived using five types of molecular markers across the studies reporting clonal richness suggest that microsatellites and RAPD are more efficient in distinguishing among clones on the basis of their multilocus genotypes than AFLP or allozymes are (Fig. 1b). Indeed studies with microsatellites or RAPD tend to report higher clonal diversity than studies using AFLP, with the mean clonal diversity across the studies reviewed here increasing from allozymes to fingerprints and to microsatellites (Fig. 1b). There has been a shift in the use of these markers, from a dominance of studies using allozymes to a rapid spread of the use of fingerprints and microsatellites (Fig. 1a). However, it is also important to note that whatever kind of marker can led to erroneous estimates if the polymorphism is insufficient, as was observed comparing two sets of distinct microsatellites revealing very contrasting results for the seagrass *Posidonia oceanica* (Alberto *et al*. 2003a, Arnaud-Haond *et al*. 2005).

Description of the components of clonal diversity

As in studies addressing species biodiversity (e.g. Peet 1974), several components can be used to estimate clonal diversity in a particular population: clonal richness, representing either the absolute number or the proportion of distinct entities (clonal lineages or genets) present in the sample relative to the number of sampling units; clonal heterogeneity, which is influenced both by the richness and the relative abundance of the entities in the sample; and clonal evenness, describing the equitability of the distribution of the sampling units (or ramets) among these entities.

Clonal richness

The simplest and most widely used (about 72% of the studies) index of clonal richness is the number of genotypes of

Box 2 Clonal richness estimates

The index of clonal diversity proposed by Ellstrand & Roose (1987) for a sample of size N in which G genotypes are discriminated is estimated as:

$$P_d = \frac{G}{N} \quad \text{(eqn 4)}$$

This modification was proposed by Dorken & Eckert (2001):

$$R = \frac{(G-1)}{(N-1)} \quad \text{(eqn 5)}$$

such that the smallest possible value in a monoclonal stand is always 0, independently of sample size, and the maximum value is still 1, when all the different samples analysed correspond to distinct clonal lineages.

These indices provide an estimate of the clonal (vs. sexual) input, once the set of loci allowed assessing the clonal membership, as previously detailed. Else, this index may overestimate clonal input, as it will ignore the reproduction of the same multilocus genotype through sexual reproduction (Stoddart 1983; Uthike *et al*. 1998). To estimate the extent of this possible bias in estimating sexual input, one method was developed (Stoddart 1983; Stoddart & Taylor 1988) involving two of those components. The first is the estimate of genotypic diversity in the sample:

$$G_o = \frac{1}{\sum_{i=1}^{G} p_i^2} \quad \text{(eqn 6)}$$

where p_i is the observed frequency of the i^{th} of G genotypes, as described in Stoddart (1983). This first component happens to be also the inverse of the Simpson index of genotypic heterogeneity commonly used to describe clonal diversity (equation 20). It is used in a ratio with the second component, the expected genotypic diversity under Hardy–Weinberg and random assortment between all pairs of loci:

$$G_e^* = \frac{1}{\left(D + \frac{P}{N}\right)} \quad \text{(eqn 7)}$$

where D is the sum of all p_i^2 for all p_i where $(p_i \times N) > 1$, and P the sum of p_i for all $(p_i \times N) < 1$. The clonal input is then estimated as:

$$\frac{G_o}{G_e^*} \quad \text{(eqn 8)}$$

When the data set used is made of markers exhibiting high polymorphism and allowing an optimal discriminating power, a very high number of genotypes may be expected and P will be negligible. The estimator (equation 19) will approximate estimator (15) as the number of multilocus lineages is more accurately estimated, and when reaching full resolution of MLLs P_d (or R) provides then a reliable estimate of the clonal input.

the population estimated by G, the number of multilocus genotypes or lineages detected in a sample. This index is obviously dependent on the sample size. As proposed for species richness S, the rarefaction method used to compare allelic richness estimates (Petit *et al*. 1998) or a permutation approach should be used (Leberg 2002) to compare two samples differing in sample size, n and $N > n$. These methods allow the estimation of expected G in the second population if only n units would have been sampled. A bootstrap approach can be used to subsample n individuals from the total sample universe available (N), and reiterate this process to estimate the average G, along with confidence intervals (Arnaud-Haond & Belkhir 2007).

After G, the most commonly (about 38% of the studies) used index of clonal richness is the 'clonal diversity' index P_d as proposed by Ellstrand & Roose (1987), the fraction of distinct clonal lineages in the population relative to the number of sampling units (Box 2, equation 15). The expected confidence limits of P_d can be derived from tables of confidence limits of percentages depending on sample size (Sokal & Rohlf 1995, Table P). Examination of these tables reveals that P_d estimates are very sensitive to sample size for low percentage values (i.e. strongly clonal populations). Indeed, this estimator can be seriously biased when analysing data from population with an extreme composition, such as monoclonal stands (richness will be overestimated), particularly when sample sizes are small. As an example, the finding of a single MLG among 20 individuals (i.e. a monoclonal set) would still lead to an estimated P_d of 0.05, the same as encountering five distinct clonal lineages among 100 sampling units. To attenuate this flaw for the extreme cases of monoclonal or low richness stands with small sample size, a slight modification has been proposed by Dorken & Eckert (2001) as R (Box 2, equation 16). Clonal diversity ranges across all possible values (from monoclonal $R = 0$ to absence of clonality R or $P_d = 1$) across studies (Fig. 1b, Table 2), reflecting the variable extent of clonality of populations. Moreover, studies including comparative analyses of R or P_d across populations typically display broad differences among populations of individual species (Table S1). Numerous examples can be observed in all kinds of organisms, where the same species can occur

Table 1 Sampling geometries, strategies, and statistics used for clonal plants in 246 reviewed articles. The symbols are linking this information to the text and to the raw data available in Table S1, Supplementary material, detailing the findings of the literature review. The percentage of studies using various sampling geometries (shape of the area sampled) and sampling strategies (choice of sampling units) is detailed; the frequency of the statistics used to describe clonal richness and diversity, as well as to ascertain clonal identity of the replicates of the same MLG are also detailed. Finally, recommendations are suggested as to the use of sampling geometries, strategies and the choice of statistics (labelled * and ** corresponds to recommended and highly recommended methods, respectively)

Description	Symbols	Percentage of studies	Recommendation (if any)
Sampling			
Sampling geometry			
Undefined	G_u	46.7	Avoid
Linear	L	46.7	Avoid
Rectangles	Q	28.9	*,†
Square	S	10.7	**
Circle	C	1.5	**
Patches	p	2.5	*
Sampling strategy			
Undefined	nd	25.9	Avoid
Haphazard	h	26.4	Avoid
Regular	re	21.8	*
Random coordinates	ra	3.0	**
Minimum spacing	min	18.8	*,‡
Exhaustive	exh	6.6	*,§
Coordinates	coord.	33.7	
Statistics			
Richness			
No estimates	—	18.8	Avoid
Number of genotypes	G	68.0	*
Ratio (G/N)	P_d (or IC = $1 - P_d$)	37.1	*
Ratio $(G-1)/(N-1)$	R	1.0	**,¶
Resampling to standardize richness estimates to the minimum sample size	sub-sampling	0.7	**,††
Heterogeneity and evenness			
Simpson complement	D^*	31.0	**,‡‡
Simpson (or Fager) evenness	V	15.2	*
Shannon-Wiener	H'	5.0	*
Shannon-Wiener evenness	$V'H'$	1.0	*
P (getting the most common MLG by chance)	PG	1.5	*
Ascertain clonal identity (Studies not considering by default identical MLG=identical clones)		25.8	
Probability of a given MLG	p_{gen}	15.2	Avoid, §§
p(getting a given MLG n times by chance)	$p_{gen}{}^n$	6.3	Avoid, ¶¶
p(identical MLG to derive from distinct reproductive events)	p_{sex}	4.6	**
$1/G_{max\ simulated}$ or $(1 - p_{identity})$ (with the set of loci used)	$1-p_{identity}$	2.0	Avoid, †††

†If low perimeter/area ratio, note that squares and circles are inducing less edge effect.
‡If based on pilot studies or prior knowledge of average clonal size.
§If not detrimental to the population.
¶Minimize the bias when N is low (lower than 20).
††If necessary for comparison purposes.
‡‡The less redundant with classical richness estimates.
§§Is the probability of getting a given MLGi when analyzing only one sampling unit, without taking into account the number of sampling units, N, collected and analyzed.
¶¶Delivers the probability of getting n times a given MLG when analyzing exactly n and not N (sample size) sampling units.
†††An average value is not reliable as the probability may be extremely distinct among genotypes, besides, this method does not take into account the number of sampling units analyzed.

Table 2 Range of values reported for the main indices of clonal diversity and clonal size (linear) or surface area encompassed in different categories of clonal organisms (values for each study are detailed in Table S1)

Organisms	R or P_d	Simpson diversity	Simpson evenness	Clonal size (m)	Clonal area (m)
Terrestrial plants	[0.00, 1.00]	[0.00, 1.00]	[0.00, 1.00]	[0.25, 1000.00]	[1.00, 7000.00]
Aquatic plants	[0.00, 1.00]	[0.00, 0.99]	[0.00, 0.99]	30.00	—
Marine plants	[0.00, 1.00]	[0.00, 1.00]	[0.00, 1.00]	[8.00, 80.00]	[31.00, 6400.00]
Marine invertebrates	[0.03, 1.00]	—	—	—	—

both in monoclonal stands and in stands where the clonal diversity reaches, or almost, its maximum (Piquot *et al*. 1996; Ayre & Hughes 2000; Freeland *et al*. 2000; Kapralov 2004; Olsen *et al*. 2004; Halkett *et al*. 2005a) These observations show that the extent of clonality is highly flexible not just among but also within clonal species, suggesting considerable plasticity in the apportioning of reproductive effort between clonality and sexual reproduction.

Finally, several methods have been developed and mostly used for clonal invertebrates (Stoddart 1983; Stoddart & Taylor 1988; Uthike *et al*. 1998), to estimate the sexual vs. clonal input with a limited set of markers (see Box 2, equations 17–19).

Clonal heterogeneity

Clonal richness indices only describe the proportion of the sample that is variable and do not describe the distribution of the sampling units among MLLs (i.e. evenness). Indeed for the same amount of clonal richness, the sample could comprise either very few highly represented clonal lineages with several rare ones, or evenly distributed ones. Discriminating between these contrasting clonal compositions is essential, since clonal heterogeneity is a fundamental feature determining the ecology and evolution of the populations. This issue parallels the old debate in ecology, when the need to combine richness with evenness was proposed to describe species heterogeneity in communities (Simpson 1949; Peet 1974). Indeed species heterogeneity indices have been borrowed to describe clonal diversity (Parker 1979; Ellstrand & Roose 1987).

The most widely used index of clonal heterogeneity (28% of the studies reviewed) is the Simpson index (Simpson 1949), which was developed originally to calculate the probability that two individuals selected at random from the sample will belong to the same species. When applied to clonal diversity, this can be interpreted as estimating the probability that two sample units chosen at random from the sample universe would belong to the same clonal lineage (Box 3, equations 20–22). The reciprocal index (Hurlbert 1971; Hill 1973), reflects the 'apparent number of *clonal lineages* in the sample' (Box 3, equation 23).

The Shannon-Wiener's index is the best known and most used diversity index in ecology, although it has only been used in about 6% of the articles on clonal diversity. It was derived independently by Shannon and Wiener (Wiener 1948, Shannon & Weaver 1949 both in Washington 1984; see also Washington 1984 for clarification on the incorrect use of the designation Shannon-Weaver). It should be noted that this last index is prone to a large sampling variance (Pielou 1966). For a given clonal richness, the Shannon-Wiener index is not expected to be very sensitive to the variation in the dominance of a particular MLL, whereas for a constant dominance it is more sensitive than the Simpson's index to the increase in the number of rare MLLs (Peet 1974).

The choice of index depends on the question posed. If the goal is the estimation of genotypic diversity or the amount of sexual vs. asexual reproduction in different populations, then the Shannon-Wiener's estimators may be most adequate. On the other hand, if the study addresses historical processes, such as the way colonization occurred in different populations, or ecological processes such as intraspecific competition under different environmental conditions, the Simpson's index may be more informative. However, the interpretation of spatial or temporal variability with either of these indices is often difficult given that they vary with both clonal richness and evenness, making it often necessary to assess these two components independently of each other. In all of the distinct types of organisms studied, widely diverse Simpson clonal heterogeneity values were reported ranging between 0 and 1 (Table 2), consistent with the similarly broad ranges of R.

Clonal evenness

As the indices of heterogeneity do not reflect equitability, the indices of evenness used in ecology have also been adapted to estimate the equitability in the distribution of clonal membership among samples. The indices of heterogeneity of Simpson and Shannon both have a corresponding index of evenness (Box 3, equations 26 and 27). Both of these most commonly used evenness indices (respectively in 12% and 1% studies) vary from 0 to 1 when all MLLs have equal abundance. The performance of equitability indices is

Box 3 Clonal heterogeneity and evenness estimates

Clonal heterogeneity

Simpson index: $\lambda = \sum_{i=1}^{G_{pop}} p_i^2$ (eqn 9)

where p_i is the frequency of the MLL*i* in the population, and G_{pop} the number of distinct MLLs in the population. An unbiased estimator of λ for a sample of size N is:

$$L = \sum_{i=1}^{G} \left[\frac{n_i(n_i - 1)}{N(N-1)} \right] \quad \text{(eqn 10)}$$

where G is the number of MLLs detected in the sample, and n_i is the number of sampled units with the MLL*i*.

The Simpson index can be modified to vary positively with heterogeneity (Pielou 1969), as an index first proposed in economical sciences (Gini 1912; Peet 1974), and the resulting *complement of Simpson index* then describes the probability of encountering distinct MLLs when randomly taking two units in the sample:

Simpson's complement: $D_{pop} = 1 - \sum_{i=1}^{G_{pop}} p_i^2$ (eqn 11)

for which the unbiased estimator from a sample of size N is $D^* = 1 - L$ that ranges from 0 to almost $1 - (1/G)$.

As proposed for species heterogeneity indices, the *reciprocal of Simpson index* is:

Simpson's reciprocal: $\frac{1}{\lambda}$ (eqn 12)

for which the unbiased estimator for a sample of size N is $1/L$.

Simpson's reciprocal ranges from 1 to G, and it can be interpreted as the number of equally represented MLLs required to obtain the same heterogeneity as observed in the sample (Hurlbert 1971; Hill 1973), or as the 'apparent number of *clonal lineages* in the sample'.

The Shannon-Wiener's index describes clonal diversity as:

$$H' = -\sum_{i=1}^{G_{pop}} p_i \log p_i \quad \text{(eqn 13)}$$

using the estimator:

$$H'' = -\sum_{i=1}^{G} \frac{n_i}{N} \log \frac{n_i}{N} \quad \text{(eqn 14)}$$

This index quantifies the level of uncertainty regarding the MLL of a sample unit taken at random (Pielou 1966). This index of clonal diversity increases with the number of MLLs and the evenness in the assignment of individuals (ramets) to the MLLs, since this leads to a greater uncertainty in predicting the MLL of a randomly drawn sample unit.

Clonal evenness

A way of describing clonal equitability, which is independent of clonal richness but not explicitly described by any diversity index (see above), is to use an evenness index. So far the most widely used evenness index in clonal plant studies is the Simpson's complement index (Hurlbert 1971; Fager 1972):

$$V = \frac{(D - D_{min})}{(D_{max} - D_{min})} \quad \text{(eqn 15)}$$

with D_{min} and D_{max} being the approximate minimum and maximum values of Simpson's complement index given the sample size N and the sample clonal richness G, estimated as:

$$D_{min} = \left[\frac{(2N - G) \times (G - 1)}{N^2} \right] \times \frac{N}{(N-1)} \text{ and}$$

$$D_{max} = \frac{(G-1)}{G} \times \frac{N}{(N-1)}$$

This evenness formulation can also be used with the Shannon-Wiener index (e.g. Hurlbert 1971), or alternatively evenness can also be estimated as V', the ratio of observed to maximal diversity (using either heterogeneity index). In this case, when using the Shannon-Wiener index, the corresponding evenness index, sometimes called Pielou's evenness (J', Pielou 1975) and hereafter referred to as such, can be estimated as:

$$J' = V'H'' = \frac{H''}{H''_{max}} \quad \text{(eqn 16)}$$

where $H'_{max} = \log G$.

Box 4 Power law (Pareto) distribution of clonal membership

The distribution of elements into size classes has been shown to follow a power law for a very broad diversity of systems and phenomena, all of which (from distributions in social sciences to astrophysics and the commonality of gene expression) conform to a particular probability density distribution referred to as the Pareto distribution (e.g. Pareto 1897 in Vidondo *et al*. 1997; Ueda *et al*. 2004). A power law distribution applies to systems where the distribution of elements into classes is highly skewed, with much fewer large classes than small ones. The use of a power distribution allows the efficient and parsimonious description of the distribution of the studied elements into classes. We therefore propose here the use of the Pareto distribution as a continuous approximation to describe the discrete distribution of sample units, or ramets (elements) into groups of clonal sizes (classes), where clonal sizes are defined by the number of sampling units belonging to that clone (MLL). This relationship is described by the equation:

$$N_{\geq X} = aX^{-\beta} \qquad \text{(eqn 17)}$$

where $N_{\geq X}$ is the number of sampled ramets belonging to lineages (MLLs) containing X, or more, ramets in the sample of the population studied, and the parameters a and β are fitted by regression analysis. In practice, the power slope (–β) is derived as the slope of the fitted log-log regression equation describing the rate of decline in the relative frequency of ramets that belong to MLLs of size equal to or larger than a given number of ramets X (when both are in log scale; Fig. B4.1). The parameter β (–slope) therefore indicates the scaling of the partitioning of the ramets among MLL size classes (Fig. B4.1).

Fig. B4.1: (a) Distribution of replicates among MLLs in *Cymodocea nodosa* from Alfacs Bay (Alberto *et al*. 2005), showing the steep decline in number of MLLs with increasing clonal membership typical of power law distributions; (b) transformed into a log-log reverse cumulative distribution.

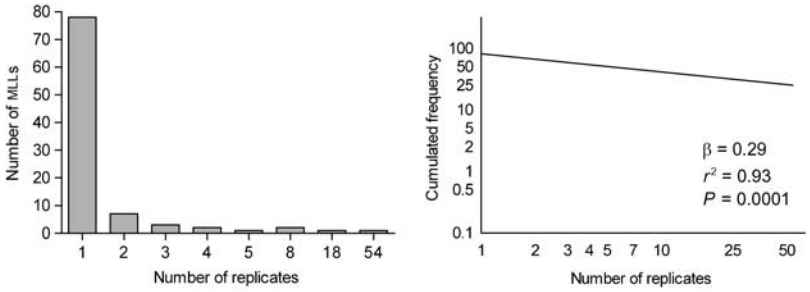

Fig. B4.1

dependent on that of the heterogeneity indices they are based upon: if based on the Shannon-Wiener index, they will give more weight to the rarer components (species or genotypes) than when based on the Simpson index. In addition, a review of these and other evenness indices (Smith & Wilson 1996) reports that *V′H″*(= J′, equation 27) remains sensitive to changes in richness (also shown here below) despite intended to be independent of richness. As for richness and diversity, Simpson evenness values encompass the maximum, or almost the maximum, range (Table 2).

Clonal distribution

In fact, the problem on hand amounts to the description of the distribution of elements (ramets) into classes (clonal lineages, or genets), so that the use of a density distribution may be more appropriate than the calculation of a compound index. An overview of the literature shows that the distribution of replicates among lineages, when detailed, is always left skewed (all of the 45 studies reporting this information) with an exponential decay (Table S1). Transformed in a reverse cumulative frequency distribution, this empirical distribution can be approximated by a power law distribution, appropriately described by the Pareto distribution (e.g. Pareto 1897 in Vidondo *et al*. 1997; Box 4). All of the distributions of clonal membership found in the literature review conformed to the Pareto. This distribution indeed applied to a range of clonal organisms encompassing herbaceous plants and trees (Parks & Werth 1993; Hangelbroek *et al*. 2002; Chung *et al*. 2004; Nagamitsu *et al*. 2004), corals (Bastidas *et al*. 2001; Le Goff-Vitry *et al*. 2004), bivalves (Taylor & Foighil 2000), and ostracods (Cywinska & Hebert 2002). The model was shown to appropriately fit all of the distributions, with all

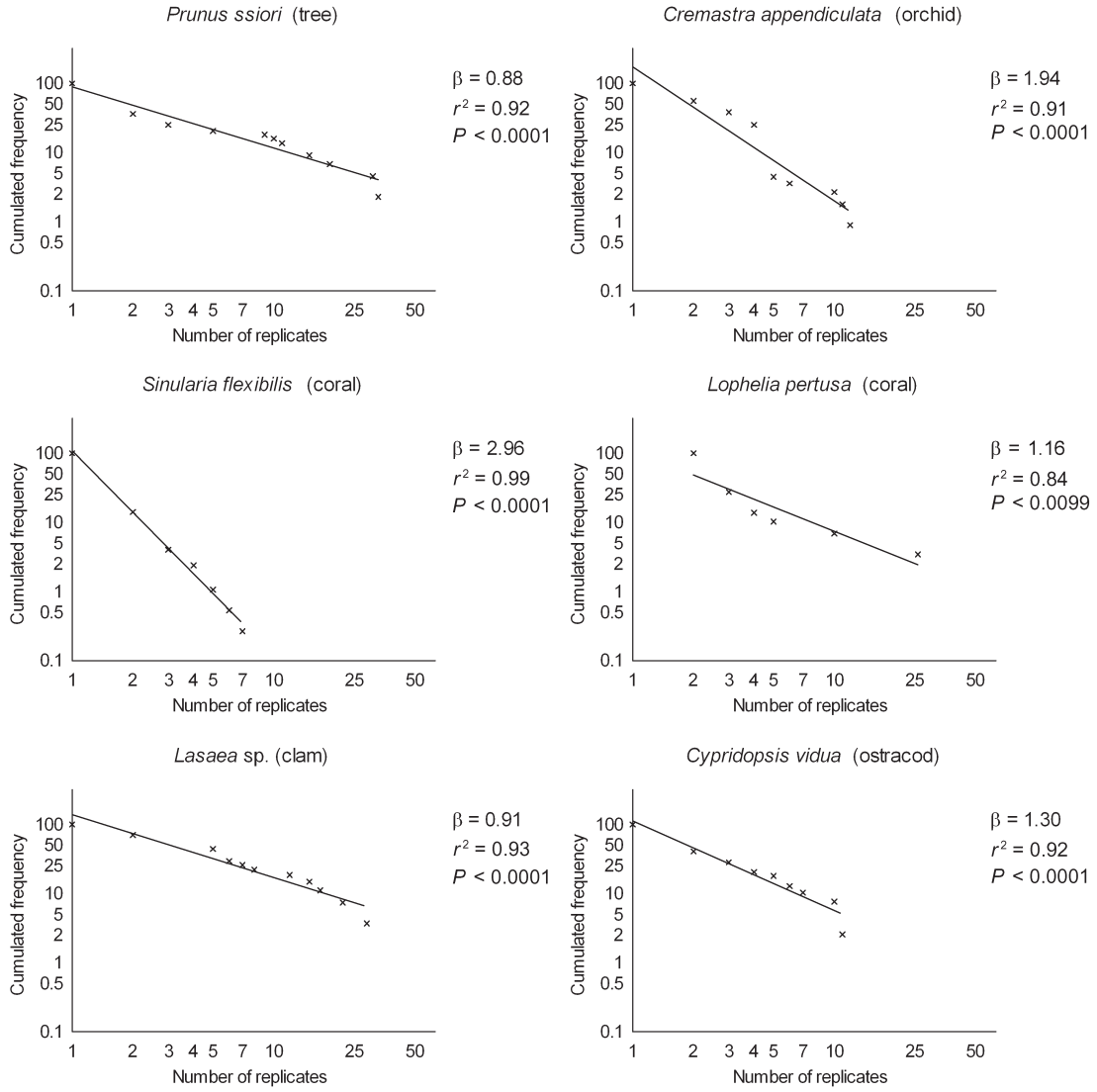

Fig. 2 Pareto plots showing the distribution of clonal membership across a range of species of terrestrial plants (Chung *et al*. 2004; Nagamitsu *et al*. 2004) and marine invertebrates: corals (Bastidas *et al*. 2001; Le Goff-Vitry *et al*. 2004), clams (Taylor & Foighil 2000) and ostracods (Cywinska & Hebert 2002). Pareto plots represent the fraction of sampling units belonging to clones representing by ≥ X units as a function of X on a double logarithmic scale (Y = proportion of sampling units belonging to clonal lineages represented in the samples by X or more sampling units, and X = observed clonal sizes quantified as the numbers of sampling units found for every clonal lineage). This plot should display a straight line if the distribution of clonal membership conforms to a Pareto distribution, and the Pareto parameters can be estimated from the least squares regression line. The coefficient β, describing the Pareto distribution (–1 × regression slope), the correlation coefficient (r^2) and the significance of the regression (*P* value) are given for each panel.

regressions showing high significance and high r^2 values spanning from 0.84 to 0.99 (Fig. 2). Hence, the Pareto model adequately describes the frequency distribution of clonal membership for populations of clonal organisms.

Now, the next step would be to be able to interpret the Pareto distribution in terms of diversity and evenness. Simulations, described in detail in the Supplementary material, were performed to explore cases where evenness would vary when diversity would be fixed, and conversely, in order to relate the properties of diversity and evenness in the populations studied and the shape and parameters of the Pareto distribution. The results (Fig. 3) show that the slope of the Pareto distribution, β, increases exponentially with increasing evenness of the distribution of sampling units into MLLs (with r^2 ranging from 0.62 to 0.93 depending on the richness level). A high evenness with clonal lineages all having approximately comparable sizes, will therefore result in a steep slope (high β value), whereas the outcome of a skewed distribution with very few, large clonal lineages and many small ones will be a shallow slope (low β

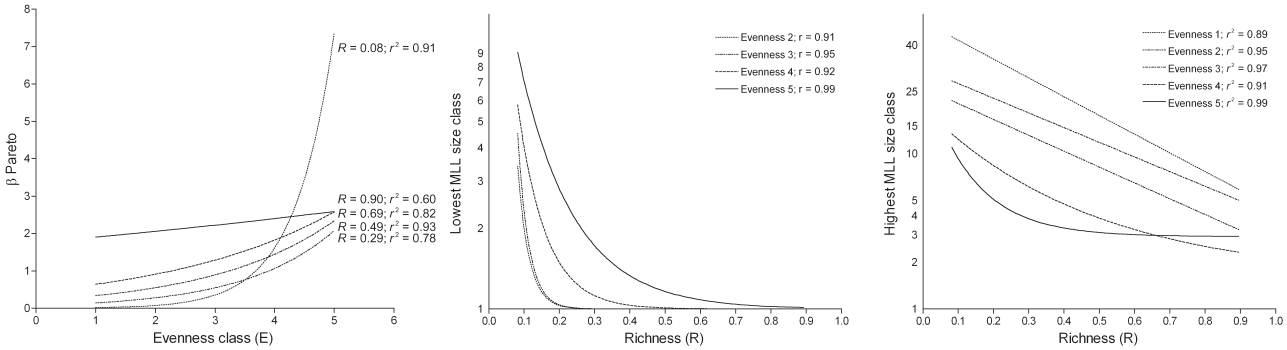

Fig. 3 The relationship between the parameters describing the Pareto distribution, and the richness and evenness level in the samples analysed. The data were obtained on the basis of simulations by distributing replicates (N = 50) among groups of genotypes (G = 5, 15, 25, 30, 45) across an increasing level of evenness (E = 1–5). The right panel illustrates the exponential increase of the Pareto parameter β with the level of evenness (r^2 between 0.60 and 0.93) for the five levels of richness explored. The left panel shows the exponential decrease in the size of the smallest size of genotype groups (in terms of number of replicates) with the increase in richness (r^2 between 0.91 and 0.99) for the five levels of evenness used.

value). Also, the results of the simulation showed that the sizes of the smallest and highest MLLs classes decrease exponentially with increasing richness (with r^2 ranging, respectively, from 0.88 to 0.99 and from 0.89 to 0.99 depending on the evenness level), so that as the lowest and highest size classes tend to get larger, the richness is expected to decrease. The Pareto distribution is therefore influenced both by richness and evenness, and provides an intuitive, graphical depiction of the heterogeneity in the distribution of replicates among lineages, which appears to be of universal application to populations of clonal organisms.

The representation of the Pareto distribution synthesizes the information in graphical form, rather than simply as a compound numerical estimate as the other indices reviewed here do, providing a clear depiction of the size distribution of clonal lineages in the population (Fig. 2). The β values obtained by compiling these data from the literature were spanning between 0.88, indicating a skewed distribution with dominance of some big clonal lineages and 2.96 indicating much higher evenness, although the minimum (i.e. most skewed) we observed, with β = 0.06, is a meadow of *Posidonia oceanica* dominated by a very big genet surrounded by several marginally represented MLG; Fig. 4). In the highest evenness scenario where all lineages bear the same number of replicates, estimation of the Pareto distribution parameters by regression is likely not to be possible as only one or two points would be available, but in this particular case, the interpretation of this finding as revealing extreme evenness is sufficient, provided enough lineages have been sampled. Furthermore, the maximum clonal size reached in terms of number of sampling units, and the frequency of those relatively dominant clonal lineages can also be observed on the graph (Fig. 2). An additional property, is that the application of the Pareto distribution allows calculation of the fractal dimension of the process under study, here the distribution of clonal size in the population, which equals 1 + β (Schroeder 1991), allowing, among other applications, the simulation of populations with a genetic structure similar to observed ones. Finally, the use of the Pareto distribution to describe the distribution of ramets into clonal lineages has the additional advantage that it is based on linear regression, providing estimates of uncertainty, therefore allowing statistical comparisons, which is not readily possible for other indices of clonal diversity.

The use of the Pareto distribution to describe clonal diversity is exemplified here for the Mediterranean seagrass (*P. oceanica*) populations sampled. The fitted Pareto distributions yielded β values that ranged between 0.033 ± 0.015, for Sa Paret (Cabrera, Balearic islands; Fig. 4), a population which was dominated by a large clonal lineage that contained most of the shoots sampled (35 of 40 shoots), and 1.48 ± 0.52, for the Acqua Azzurra (Sicily, Italy) three populations where almost all (33) genotypes were observed once, except two represented three and four times (data not shown). Figure 4 shows contrasting Pareto distributions illustrating clonal structure in four populations with highly contrasting richness (R spanning from 0.10 to 0.77) and evenness (as estimated by Simpson evenness V ranging from 0.20 to 0.73). Both richness and evenness influence the shape of the Pareto distribution and its associated β value, as can be observed by comparing samples with similar R and distinct V (Campomanes and Playa Cavallets) or, conversely, with similar V and distinct R (Carboneras and Playa Cavallets), all pairwise comparisons revealing contrasting Pareto distribution and associated β values. Yet, consistent with the results of simulations (Fig. 2), increasing richness from Carboneras to Playa Cavallets (R increasing from 0.3 to 0.73) is reflected in decreasing maximum MLL size (from 15 to 5) that can be easily derived from the Pareto plot (provided comparable sample sizes, which is the case here). In the same way, the comparison of samples like Campomanes and Playa Cavallets

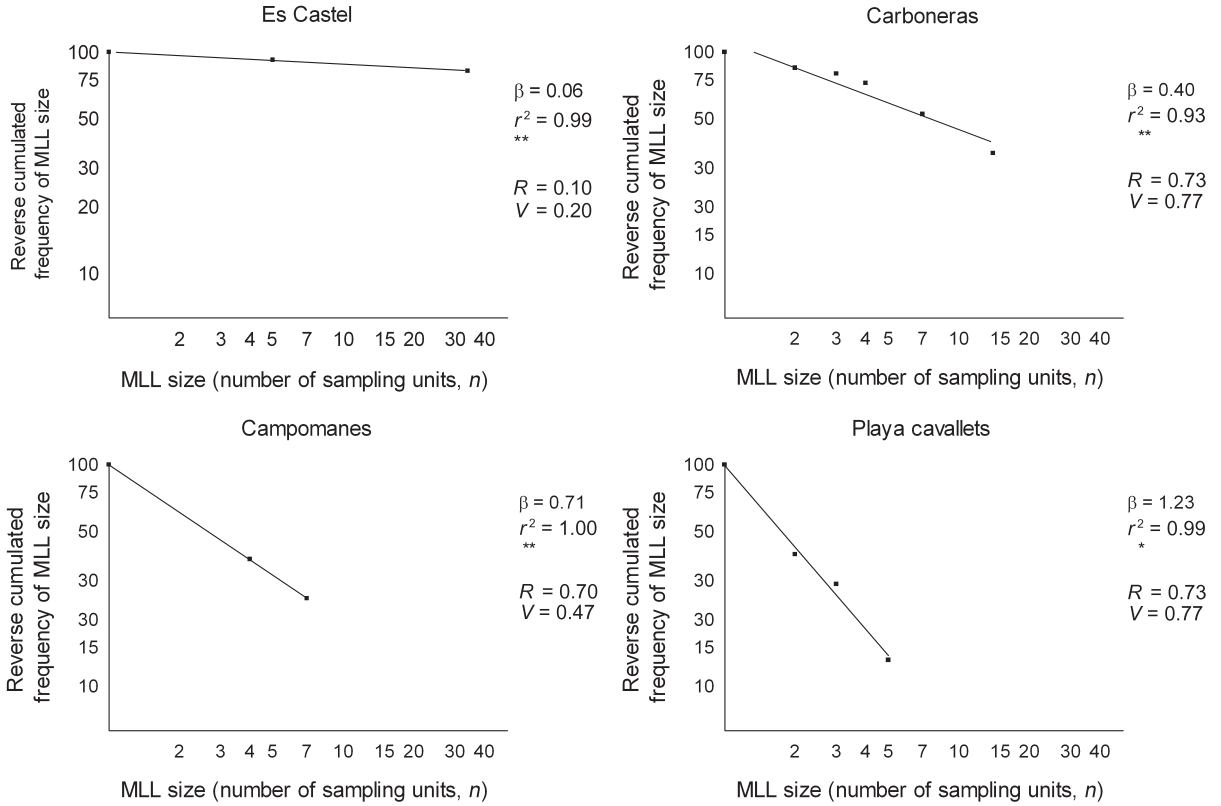

Fig. 4 Pareto plot of clonal membership distribution in four populations of the seagrass *Posidonia oceanica* (Es Castel, Porto Colom, Campomanes, Playa Cavallets). The coefficient β, describing the Pareto distribution, the correlation coefficient (r^2) and the significance of the regression (*P* value) are given for each panel.

shows how, *R* being equal, a higher evenness (as measured by Simpson index of evenness *V* increasing from 0.47 to 0.77) translate into a steeper Pareto slope (with the associated β parameter of Pareto increasing from 0.40 to 1.23).

Relationship and possible redundancy between the different indices of clonal diversity

The various indices of clonal diversity discussed above are not independent of each other, as they are based on the same basic information, but differ in the weight each assigns to the basic clonal richness and to the equitability of the distribution of replicates among clonal lineages. Hence, the application of all these indices may be redundant, and a small subset may suffice to capture the crucial information in terms of richness and equitability.

The relationship between these different indices was assessed using correlation analysis on both empirical and simulated data sets. The empirical data set was obtained in the 34 populations of the seagrass *P. oceanica* used here as a test case. Monte Carlo simulations were also performed to explore the relationship between clonal diversity indices (richness, heterogeneity and evenness) and to test for the generality of the links between different indices of clonal diversity derived from the analysis of the *P. oceanica* data set. The details on the simulations conducted are provided in the Supplementary material.

All correlation estimates were transformed as '1-Pearson *r*' in order to perform a cluster analysis and draw a hierarchical tree using this index as an estimate of distance among indices. These analyses were performed using STATISTICA 6 software (StatSoft 2001).

The estimates of the indices derived from the simulated data set were indeed positively correlated to one another (Fig. 5). Qualitatively, the same correlation structure between indices was obtained on the basis of the *P. oceanica* data set, with *r*-values quantitatively similar to those obtained on the basis of simulations (data not shown). Examination of the correlation structure between the various indices showed that the genotypic richness *R*, the Simpson's complement *D*, the Shannon-Wiener *H*′, and Pielou's evenness *V*′*H*″ are very redundant (*r* = 0.82–0.95), whereas the Pareto β was the least redundant, followed — as expected — by the Simpson evenness *V*.

The redundancy between these indices is synthetically grasped upon examination of the cluster linking them (Fig. 5). A relationship with richness had been reported in the literature on species diversity for the Pielou evenness

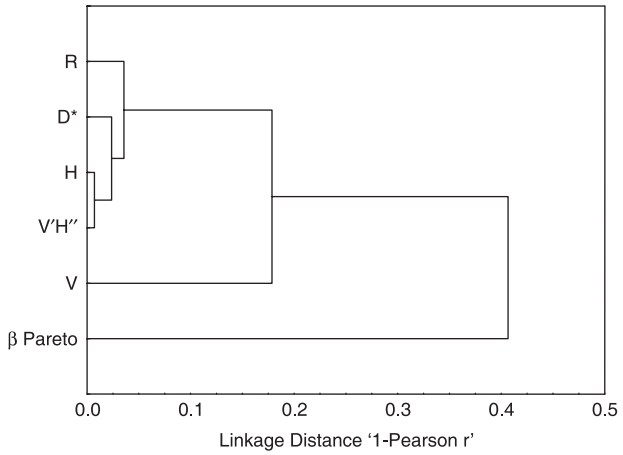

Fig. 5 Cluster analysis of indices describing clonal diversity obtained on the basis of simulated data, using '1-Pearson *r*' as clustering distance: clonal richness *R*, Shannon-Wiener's heterogeneity *H*' and Pielou evenness *V'H*', Simpson's complement heterogeneity *D* and evenness *VD*, and Pareto's β. The same correlation and cluster structure was obtained on the basis of the *Posidonia oceanica* data set, with *r* values quantitatively similar to those obtained on the basis of simulations.

index, when a small number of species (< 25) is observed, and can therefore apply to a large range of studies on clonal organisms, where the sample size per locality hardly encompasses 30–50, and the number of genotypes will therefore seldom be sufficient to avoid this bias (Smith & Wilson 1996). The Simpson evenness *V* (Hurlbert 1971) is least redundant with *R* (Fig. 1b; cf. Peet 1974), and appears therefore to be the most suitable index to estimate evenness in a given sample. Finally, the use of the Pareto distribution delivers the least redundant index (β), and can be useful to depict heterogeneity. Hence, the three main types of information required to fully describe diversity: richness, evenness and heterogeneity are adequately grasped by the combined use of *R*, *V* and the complement of the slope of the Pareto distribution (β), respectively. We therefore recommend use of these three metrics to describe clonal diversity.

Spatial analyses of clonal structure

Sampling geometry and strategy

In contrast with species consisting of unique genotypes, clonal populations have the capacity to spread and multiply common clonal lineages in space, so that inferences about the spatial genetic structure within clonal populations are unavoidably linked to the distribution of the clonal lineages in space. Sampling design choices can easily influence and bias estimators of clonal diversity. Definition of the sampling strategy must consider: (i) sample size, (ii) sampling area size, shape, and replication, (iii) sampling regime (random, haphazard, regular), and (iv) whether to impose any minimum spacing constraints in order to reduce clonal repetitions in the sample. Each of these choices critically affects the perceived genetic structure of the population and should be therefore adopted on the basis on an informed understanding of the consequences of alternative choices.

The choice of sampling design depends on the objectives of the study. If the main objective is comparison with previous studies, the best choice may be to use the same sampling methods and scheme, though possibly fraught with other problems, in order to avoid confounding the comparison with effects of differential sampling. When the objective is to estimate clonal diversity in a population, the ideal sampling design would be a random sampling along the distributional area of the entire target population so that every possible sampling unit would have equal probability of being included in the sample. In cases of patchy distribution of individuals, random coordinates can be generated and adjusted to the nearest possible sampling unit once on the field. Only this scheme would minimize bias in the estimation of diversity indices (Pielou 1966), and deliver the best approximation of the real population values (but see the next section on sampling density). This ideal sampling regime is, however, often practically difficult, particularly in cases of patchy distribution of populations, and bias derived from deviations from randomness and an uneven distribution across the whole population area are often introduced. An appropriate alternative may be a hierarchical (multistage sampling at randomly selected clusters) or a stratified random sampling design, particularly appropriate for heterogeneous populations where conspicuous subunits exist that can be defined as sampling strata (e.g. areas of different population densities). Each of the sampling cluster or stratum should be several times larger than the expected clonal size, and each sampled in sufficient numbers to yield representative estimates of the proportion of distinct clonal lineages within sampled areas. The replication of these areas in different zones of the population reduces the potential bias due to larger-scale heterogeneity within the target population, and in fact allows testing for such heterogeneity, an important advantage over simple random sampling.

Any sampling addressing the spatial structure of clonal lineages must be conducted along a two-dimensional area (i.e. avoiding linear transects, see below) recording the corresponding X–Y coordinates (absolute or relative). The location of the sampling units (i.e. the sampling coordinates) within this area can be selected randomly, haphazardly, or under a variety of regular schemes (e.g. simple lattice, hierarchical grid). Regular spacing of sampling coordinates should be based on information allowing best choice of relevant pore sizes (i.e. geographic distance between sampling units) in order to avoid bias. For instance, if

Table 3 Influence of the sampling area geometry on the estimation of the number of genotypes (G) and genotypic richness (R), on the importance of edge effect (E_e) and on its significance tested with a 1000 random resampling procedure (number of $p_{(\text{observe} > \text{random})} < 0.05$ in 10 tests; NS when none of the values was significant). An extrapolation procedure was used, based on the large rectangular sampled area of *Cymodocea nodosa* in Alfacs Bay (Alberto *et al*. 2005), to generate a high resolution virtual population with a similar clonal structure. Ten subsampling areas (about 50 m²) of each of the four following geometric shapes: circular (r = 4 m), squared (7 × 7 m), rectangular (14 × 3.5 m) or almost linear (38 × 1.3 m or 20 × 2.5 m), in which 30 sampling units were assigned random coordinates, were randomly set in the virtual population

	Circles	Squares	Rectangles	Lines
Perimeter/Area	0.50	0.56	0.70	1.44
G (± SE)	10.00 (± 0.42)	9.50 (± 0.69)	11.20 (± 0.31)	12.20 (± 0.70)
R (± SE)	0.31 (± 0.01)	0.29 (± 0.02)	0.34 (± 0.02)	0.39 (± 0.03)
E_e (± SE)	–0.10 (± 0.08)	0.11 (± 0.05)	0.35 (± 0.07)	0.15 (± 0.09)
P (Ee(obs) < Ee(random))*	NS	NS	6	4

*Number of E_e values significant ($P < 0.05$) after 10 resampling tests (1000 repetitions). NS when none of the 10 tests was significant.

clonal lineages are, on average, 2 m in diameter, regular sampling with a spacing of 4 m between neighbour nodes will grossly overestimate clonal diversity. Grid or regular sampling schemes also generate a discontinuous distribution of geographic distances among sample pairs that may, depending on the statistics used, result in difficulties to conduct spatially explicit analysis such as spatial autocorrelograms. Grid sampling designs, however, are best when the goal is to map the population for visual representation, as it provides a uniform sampling density over the studied domain. Haphazard or random designs can include any possible sampling distances, and are, in principle, free of any explicit or hidden assumptions concerning the extent of clonal lineages. However, haphazard sampling regimes may lend themselves to unconscious user bias, whereas a truly random approach, preselecting the random X–Y coordinates to be sampled is preferable.

Use of the simulated seagrass *Cymodocea nodosa* clonal landscape (see details in the Supplementary material) to explore the consequences of different sampling geometries on the assessment of clonal diversity revealed high discrepancies in the estimates obtained on the basis of distinct geometries (Table 3). Indeed, belt transects (i.e. almost linear transects) resulted in significantly higher genotypic richness ($R = 0.39 \pm 0.02$) than that estimated on the basis of square and circular ($R = 0.29 \pm 0.02$ and 0.31 ± 0.02, respectively) sampling geometries (unilateral t-test at the 5% level); rectangular plots revealed intermediate values ($R = 0.34 \pm 0.02$) but still significantly higher than in circular and square shapes, which best approached the real genotypic richness introduced in the simulation. These results indicate that, as expected, narrow transects overestimate clonal diversity, due to the larger perimeter to surface ratio, which leads to the presence of many apparently single MLLs that in fact correspond to possibly large clonal lineages extending largely outside the boundaries of the transect. The increased apparent clonal diversity with increasing perimeter to area ratios in the sampling domain emphasizes the importance of considering edge effects when examining clonal diversity in clonal plant populations. However, only about 60% of the approximately 250 studies of genetic structure of clonal plants reviewed, and 55% of studies on other organisms, mention the shape of the sampling area (Table 1), with a total of 22% using a rectangular design and 10% using linear transects. None of those studies considered the occurrence of edge effects, probably because no test to reveal the likelihood of such effects was available.

To address this gap, we propose here a permutation procedure to test whether some apparently unique or rare MLLs located near the periphery of the sampled area may derive from edge effects, rather than a small clonal lineage size. This involves an examination of the geographic distance between unique MLLs relative to the geometric centre of the sampling domain compared to that between all of the sampling units and the centre (Box 5). An edge-effect index that tests whether the apparent unique or rare MLLs tend to be distributed towards the edges of the sampling area, thereby suggesting edge effects and consequent overestimation of clonal diversity, can therefore be calculated as described in Box 5.

A significant edge effect was observed (with an alpha of 5%) in four out of 10 linear subsamples of the simulated *C. nodosa* data set, and in five out of the 10 rectangular subsampled areas, whereas no such bias was detected for circular or square samples of the same test population (Table 3). These results, as well as examination of the causes of such edge effects, indicate that sampling geometries with low perimeter-to-area ratios, such as circular or square shapes, are least prone to edge effects. Linear or nearly linear transect designs suffer, as shown above, the highest risk of incurring bias derived from edge effects, thereby overestimating genotypic richness, and should be avoided altogether in studies aimed at elucidating clonal diversity.

Box 5 Spatial components of clonality

Edge effect

In order to test whether for the sampling design used, apparent unique or rare MLLs are more distributed towards the edges of the sampling area, thereby inducing a possible overestimation of clonal diversity, the following index can be estimated:

$$E_e = \frac{(D_u - D_a)}{D_a},$$

with D_u the average geographic distance between unique MLLs and the centre of the sampling area, and D_a the average geographic distance between all sampling units and the centre of the sampling area. The significance of such index is tested against the null hypothesis of random distribution of unique and multiply represented MLLs. In practice, the likelihood of the observed difference $D_u - D_a$ being only due to chance and not to edge effect can be tested for by permuting x times the positions of the samples (i.e. randomly reassigning the sample unit to the sampling coordinates), and calculating the index for each permutation to obtain an empirical distribution of E_e. If the observed E_e value lies beyond the critical value (function of the chosen alpha) in the distribution of E_e in the permuted data, then a significant edge effect is present that may cause indices of clonal diversity to overestimate the population diversity.

Aggregation index

In order to test for the existence of spatial aggregation of clonemates, or MLGs belonging to identical MLLs, the aggregation index A_c can be estimated as follows:

$$A_c = \frac{(P_{sg} - P_{sp)}}{P_{sg}}$$

with P_{sg} being the average probability of clonal identity of all sample unit pairs and P_{sp} the average probability of clonal identity among pairwise nearest neighbours; these are estimated from the respective observed proportions in the sample. This index will typically range from 0, when the probability between nearest neighbours does not differ on average from the global one, to 1 when all nearest neighbours preferentially share the same MLL, in a situation of spatially distant distinct clonal lineages. The statistical significance of the calculated aggregation index can be tested against the null hypothesis of spatially random distribution of samples using a resampling approach, whereby the individuals sampled are randomly assigned to the existing sampling coordinates.

Sampling density

Definition of the appropriate area of a sampling cluster or stratum and sample size in each area requires an a priori estimation of the average sampling density (sample units per unit area) that would be high enough to encompass several repetitions of the same clonal lineages and of the average area that would be large enough to include many different clonal lineages (also see discussion of clonal subrange below). Without a priori information on clonal structure, the only guidance to design sampling strategies derives, in addition to the theoretical impact of geometry on edge effects, from knowledge on the clonal growth and demography of the species (e.g. horizontal spread rates, branching angles, lifespans), which can, alone or through the use of models (e.g. Lovett-Doust 1981; Sintes *et al*. 2005), provide expectations on the linear extent of the clonal lineages. In the absence of such information, limited pilot studies may be needed as a basis to design efficient, unbiased sampling strategies. These pilot studies should be focused on resolving clonal size structure at small spatial scales, as to ascertain the sampling density and/or 'pore' size (if relevant) of the subsequent study, since clonality is often not an issue for objectives related to the largest scales.

The consequence of various sampling densities could be assessed by using a resampling approach, whereby clonal richness indices (G and R) would be estimated for multiple random combinations of sampling units, for sample sizes ranging from 1 to N (N = total sample size in a given area; Fig. 6).

Inspection of the plot of the number of genotypes (G) vs. the number of sampling units would theoretically allow, as in the case of the selection of the number of markers required (see above), selection of the minimum sampling density yielding asymptotic R values. However, an associated feature to the power law distribution of clonal membership size observed in all studies examined (Figs 2 and 4) is that no asymptotic R value is obtained until an exhaustive sample of the population is reached. In the test case examined here, no asymptotic value could be observed in any of the samples of 40 *Posidonia oceanica* populations, except two populations with an overwhelmingly dominant single clone each (data not shown), and no asymptotic stabilization of R with increasing sampling effort was observed in the most intensively sampled population, even after

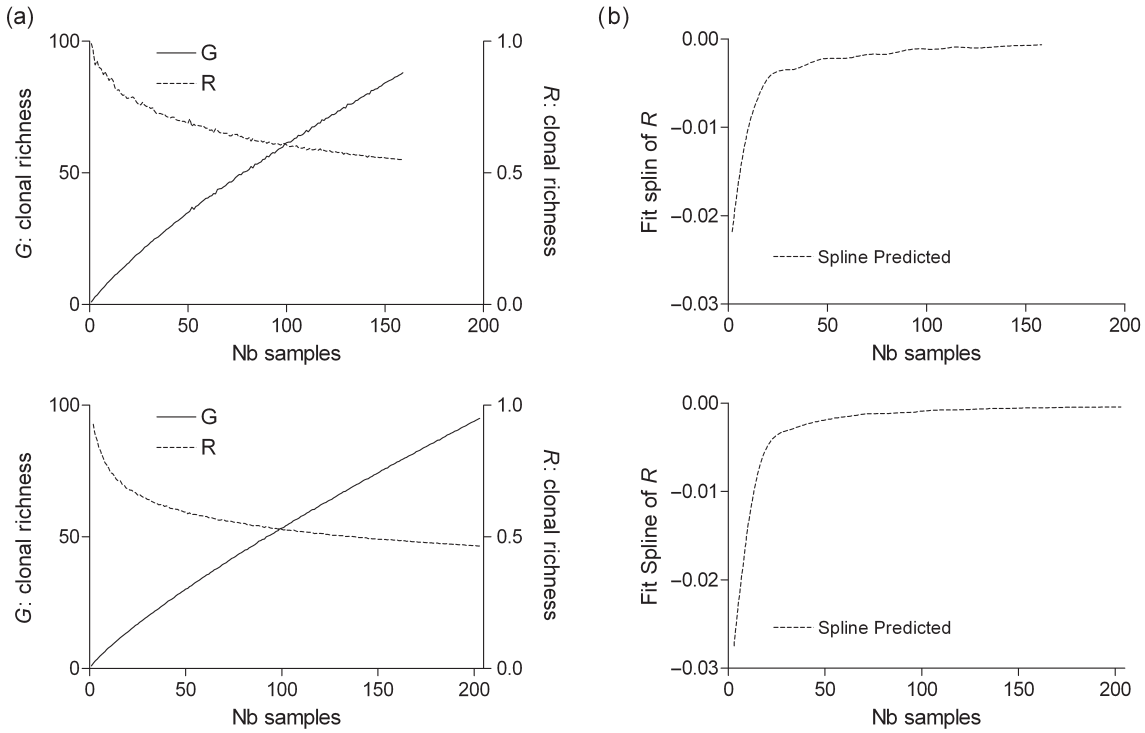

Fig. 6 (a) The relationship between the indices of clonal richness *G* and *R* and the number of sampling units *N* based on data sets of *Posidonia oceanica* ($N = 149$) and *Cymodocea nodosa* ($N = 220$), illustrating the absence of an asymptotic value of *R* and its strong dependence on sampling density. (b) Spline fit describing the rate of change in *R* with increasing *N* ($\partial R/\partial N$) for both samples (*P. oceanica*: $\lambda = 1000$; $r^2 = 0.21$, $P < 0.05$; *C. nodosa*: $\lambda = 1000$; $r^2 = 0.78$, $P < 0.001$).

sampling about 150 shoots in a quadrat limited to 50 × 50 m (Fig. 6). A similar lack of asymptotic stabilization of *R* with increasing sampling effort was observed in both *Cymodocea nodosa* populations with about 220 sampling units in a 20 × 60 m area each (Fig. 6). Although the value of *R* does not reach an asymptote with increasing sampling effort, the rate of change in *R* with increasing sample size *N* declines as *N* increases following a law of diminishing returns described by the spline fit of *R* vs. *N* (Fig. 6), such that sample sizes of 50 units provide, in the test cases examined here, an adequate approximation of *R* (Fig. 6). The fact that the value of *R* does not reach an asymptote with increasing *N* is a consequence of the Pareto distribution of clonal membership, as the lack of a statistically defined mean value, unless the entire population is sampled, is a property of power law distributions such as the Pareto distribution. Since the review presented here shows that a power law distribution of MLL size (in terms of number of replicates) appears to be a universal feature of clonal organisms (Figs 2 and 4), sample sizes should be as large as possible to ensure that *R* values are as stable as possible (e.g. $N = 50$ for the examples in Fig. 6b). We therefore recommend collecting excess samples to test whether *R* stabilizes for a given sampling size using a subsampling approach, genotyping additional samples until a modest change in *R* with further sample size increments is achieved. In any case, the limitations imposed by this undesirable behaviour of *R* should be considered when comparing across-studies using different sample sizes or sampling density. It is also important to mention that this property being due to the power law distribution of replicates among clones, the parameters describing the Pareto distribution are mostly unaffected by sampling density, once enough genotypes have been sampled to allow the construction of a robust regression. The Pareto distribution may therefore be much more adequate to compare properties of clonal diversity among sites and studies with different sampling density, provided the sampling areas are comparable.

Clearly, the sampling strategy, density and spatial design strongly affect the estimates of clonal diversity. Remarkably however, almost half (49%) of the published reports reviewed did not include any mention of the geometry, area or procedure (random or haphazard vs. regular) used in sampling. Among those that provided sufficient detail, the geometry ranged from circles (only six studies, representing less than 4% of the studies) to rectangles or squares (40% and 18% of the studies, respectively) and linear transects (19% of the studies). Sampling was more often random or haphazard (38% of the records) than regular (27% of the records), and a minimum spacing between

sample units was set at a scale depending on a priori knowledge of species average clonal size in 31% of the studies (Table 1). Comparative analyses among studies are rendered cumbersome, if not impossible, by the diversity of methods used, and by the lack of information on the sampling strategy used, particularly on the sampled area.

Spatial autocorrelation

Spatial autocorrelation analyses have been used to ascertain the scale-dependence of clonal diversity in clonal populations, including those of clonal plants (about 22% of studies on genetic structure of clonal plant populations, and 8% in other clonal organisms). These inferences were derived from spatial autocorrelation analyses representing the average genetic distance or kinship coefficient (Loiselle *et al*. 1995; Ritland 1996; Epperson & Li 1997; Rousset 2000) between pairs of individuals within specific ranges of geographic distance, weighed against the average genetic distance or kinship coefficient between all paired samples in the population, plotted against distance (Fig. 7). Autocorrelograms are commonly used to infer properties not specific to clonal plants, such as dispersal scale and neighbourhood size. However, spatial autocorrelograms can also be used to infer properties specific to the clonal nature of the studied organism (Reusch *et al*. 1999). Spatial autocorrelograms have been applied to clonal plants in the past including or excluding replicated MLLs, depending on the specific question addressed. The comparison of spatial autocorrelograms including and excluding replicate MLLs (Fig. 7) and the estimation of the probability of clonal identity have recently been shown to allow inferences on the linear spatial domain over which clonality affects the genetic structure of the population, referred to as the 'clonal subrange' of the population (Harada *et al*. 1997; Alberto *et al*. 2005). Using these approaches, the clonal subrange can be operationally described as the spatial scale below which the spatial autocorrelograms derived either including or removing pairs among identical MLLs converge (Fig. 7, cf. Alberto *et al*. 2005), and at which the probability of clonal identity approaches zero (Harada *et al*. 1997). This clonal subrange represents the characteristic maximum size of the clonal lineages in the sample, and is the spatial scale beyond which clonality does not affect genetic structure. Application of these techniques have inferred linear clonal subranges of 20–25–30–35 m for a Mediterranean seagrass species (*C. nodosa*, Alberto *et al*. 2005), and 140–190 m for two terrestrial species, *Carpobrotus* sp. (Suehs *et al*. 2004) and *Aechmea magdalenae* (Murawski & Hamrick 1990), respectively. For clonal plants, autocorrelation analysis have reported important spatial structure in the distribution of clones in space, both including all sampling units in the analysis (about 80% of the studies report significant autocorrelation) or excluding the effect of clonality in the sample by removing replicates or distances among pairs of the same MLLs (61% of the studies report significant autocorrelation). The few studies reporting estimates of neighbourhood size for clonal plants show very limited linear dispersal scales, of the order of tens to hundreds of metres (Table 2).

Finally, it may be relevant to screen for the occurrence of clonally mediated dispersal (e.g. by means of fragmentation, dispersal and re-establishment), which can be an additional efficient source of dispersal susceptible to significantly affect the genetic neighbourhood (Charpentier 2001; Hammerli & Reusch 2003). The autocorrelogram may show a non-null probability of clonal identity at large geographic distance scales preceded by null for several distance classes. This may signal the occurrence of dispersal by clonal fragmentation, although the absence of such profiles cannot be used to infer the absence of this process, which may be a rare event, requiring therefore large sampling efforts for its detection.

Aggregation

Knowledge of the spatial position of the individuals sampled also allows the examination of the extent to which clonal lineages occur segregated or intermingled in the population. The extent of intermingling of clones in a population therefore provides insight into the history of clonal growth and space occupation, and the competitive interactions among clones. A segregated distribution of the clonal lineages in space (high aggregation) may for example arise from either recent colonization, where clonal lineages are still expanding in relatively empty space or due to competitive exclusion, as observed by Cheplick (1997). Conversely, an intermingled pattern suggests either a full occupation of space by a large number of clonal lineages due to a long history following colonization and/or high density, and relatively weak competitive interactions among clones. We propose that the extent of intermingling or aggregation of the clonal lineages can be assessed by comparing the probability of clonal identity (set as 0 among replicates of the same MLL and 1 among sampling units belonging to distinct MLLs) between nearest neighbours relative to that between pairs of sampling units drawn at random from the population. A spatial clonal aggregation index, A_c, can therefore be estimated as described in Box 5.

The application of this estimator to the 34 populations of *P. oceanica* sampled across the Mediterranean showed very contrasting results spanning from 0.00^{NS} (implying high level of intermingling) to 0.68** (implying high and significant level of spatial aggregation of ramets belonging to the same MLLs). These two extremes were observed in two populations of the Balearic Islands, showing high variability in the extent of aggregation of clones in populations located relatively close to one another.

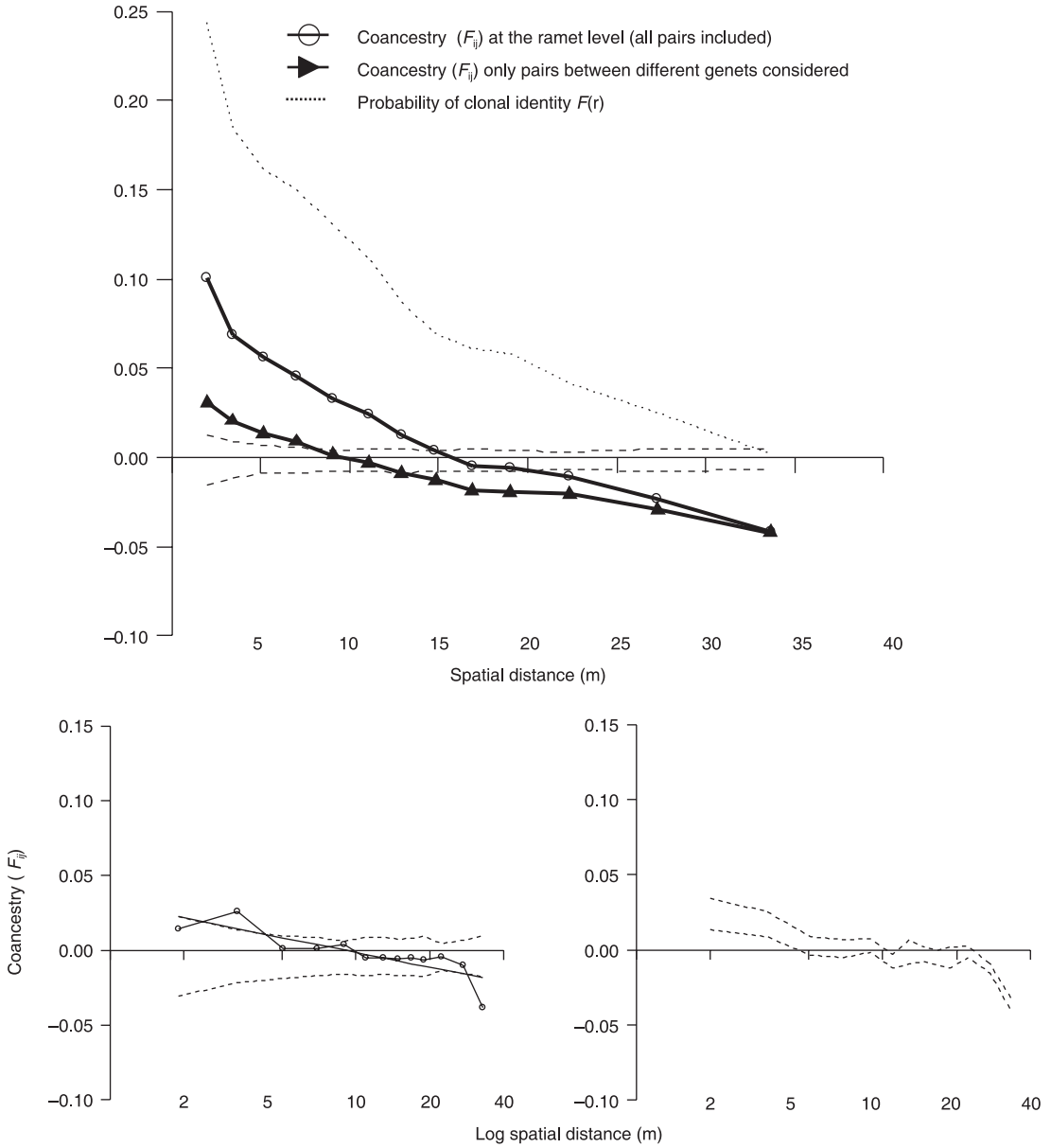

Fig. 7 Spatial autocorrelation analysis of *Cymodocea nodosa* in Alfacs Bay (from Alberto *et al*. 2005). (a) Clonal structure and subrange (on top). Kinship estimates from all ramet pairs or only for pairs between ramets showing a different multilocus genotype, and probability of clonal identity (proportion of pairs between ramets with identical multilocus genotypes), with confidence limits (for $P = 0.975$ and $P = 0.025$) based on 1000 permutations of spatial coordinates. (b) Genet level analysis (below), using a single copy for each multilocus genotype. The slope of the regression of mean kinship estimates as a function of the logarithm of spatial distance is plotted on the left, using as spatial coordinates the central zone occupied by multiramet genets, with broken lines delimiting 95% confidence limits around the null hypothesis of random distribution of genets in space. On the right side a single ramet per multiramet genet was randomly selected to create a 100-genet data file to generate the confidence limits for the correlogram.

Software

Some recently published software compute most of the indices and metrics detailed in this work (Box 6), MLGSIM (Stenberg *et al*. 2003), GENALEX (Peakall & Smouse 2006), GENOTYPE and GENODIVE (Meirmans & Van Tienderen 2004), and GENCLONE 1.0 (Arnaud-Haond & Belkhir 2007). Methods to assess clonality and clonal membership are available in all of them with slight differences, but only the last two, GENOTYPE and GENODIVE (Meirmans & Van Tienderen 2004) and GENCLONE 1.0 (Arnaud-Haond & Belkhir 2007) allow resampling procedures to test for the

Box 6 Software available to analyse molecular data on populations of clonal organisms

Data sets analysed, methods to assess clonality and clonal membership, clonal diversity estimates, and methods proposed to describe spatial components of clonal growth: GENALEX (Peakall & Smouse 2006), GENCLONE (Arnaud-Haond & Belkhir 2007), GENOTYPE and GENODIVE (Meirmans & Van Tienderen 2004) and MLGSIM (Stenberg *et al*. 2003).

		MLGSIM	GENOTYPE and GENODIVE	GENALEX	GENCLONE
Data sets					
Levels of ploidy	Haploid		✓		✓
	Diploid	✓	✓	✓	✓
	Polyploid		✓		
Markers	Dominant		✓		✓
	Codominant	✓	✓	✓	✓
Clonality and clonal membership					
MLGs discrimination		✓	✓	✓	✓
MLLs: use of frequency distribution of pairwise differences to group MLGs likely to be distinct due to somatic mutation or scoring errors	Frequency distribution		✓		✓
	Allelic distance		✓		✓
	Length distance (microsatellites)		✓		
	Custom distance		✓		
Probability of clonality, or of clonal identity	P_{gen}	✓	✓	✓	✓
	P_{sex}	✓	✓	✓	✓
	and confidence interval	✓			
	P_I (exclusion over the sample)			✓	
Subsampling frequency to test for the efficiency of the set of marker used					✓
Subsampling procedure to correct richness indices for different sample size			✓		✓
Clonal richness and diversity					
Clonal richness	G		✓		✓
	P_d		✓		✓
	R				✓
Clonal diversity and evenness	Shannon diversity and evenness		✓		✓
	Simpson diversity and evenness				✓
	Hill diversity				✓
	Pareto distribution				✓*
Spatial components of clonality					
(When coordinates are available)	Map clone			✓	✓
	Clone size			✓	✓
	Clonal subrange				✓
	Spatial autocorrelation				✓
	Edge effect				✓¥
	Aggregation index				✓¥

*available in the newly released version of GENCLONE, GENCLONE 2.0.

accuracy of the set of samples and loci used. The same two software, GENODIVE and GENCLONE 1.0, allow estimating richness and diversity indices, but only the latter allows estimating Simpson and derived indices (Hill's and evenness).

Spatial components can be analysed using either GENALEX or GENCLONE 1.0 by mapping clones or estimating maximum clone size and the latter also allows performing clonal subrange analysis and implements specific spatial autocorrelation methods adapted to the occurrence of replicates of the same genotype in the data set.

Finally, the Pareto distribution and parameter, aggregation index and edge effects are proposed in the new version of the software GENCLONE, GENCLONE 2.0 (Arnaud-Haond, Belkhir, available for download on GENCLONE website).

Prospect

The rapid growth in the research effort examining the clonal structure of populations is providing an important empirical basis to probe the implications of clonality. As this empirical basis grows ever larger, there is a need to standardize procedures to allow comparative analyses to be formulated and common patterns in the clonal diversity and structure of clonal populations to emerge. As discussed above, comparisons across studies are not straightforward as most of the descriptors of clonal structure are strongly sensitive to sampling choices; hence the need to move towards standardized procedures.

We recommend that studies of clonal diversity and structure be based on samples collected at random coordinates within sampling areas that minimize the perimeter-to-area ratio (e.g. circles or squares). We try to dissipate present ambiguity in the use of the terms by introducing the concept of clonal lineage and MLL instead of clone and MLG, in order to include in a clonal lineage (MLL) not only an MLG but also any group of MLGs characterized by very few genetic differences that appear more likely to be derived from somatic mutations or scoring errors rather than from distinct zygotes. We describe how Monte-Carlo procedures can help ascertain the number of loci required to deliver accurate assignments of clonal lineages as well as to elucidate potential sampling biases derived from edge effects, thereby delivering the most robust estimates of clonal richness possible. We recommend the use of genetic richness (R), the Simpson evenness index (V), and the complement of the slope of the Pareto distribution of clonal membership as the most parsimonious set of nonredundant indices of clonal diversity. The issue of sampling design and density has also been shown to be far from trivial, and the sensitivity of most indices to these parameters, rendering risky any comparison among studies, that may be interpreted with high caution. A spline fit describing the rate of change in R with increasing N may be used to get the most accurate possible estimates of R and the Pareto distribution may be chosen for comparative proposes, as it is the less sensitive indices to sampling density. Lastly, the preceding discussion emphasizes the critical importance of explicitly considering the distribution of clonal lineages in space, allowing the analysis of spatial clonal traits such as estimates of the clonal subrange and the extent of clonal aggregation. Most of features are now available through four principal software packages released recently (Box 6), which should facilitate the use and standardization of these methods.

The elements provided here represent a first step towards an increasing realization of the consequences of clonality in the design and analyses of studies, helping to develop a coherent framework for the study of genetic structure of clonal plant populations. We believe that consideration of the recommendations herein proposed should help move this emerging research program further, and we hope they will provide new impetus towards further exploration of the consequences of clonality for the population dynamics and evolution of species.

Acknowledgements

This work is a contribution of the EU Network of Excellence MARBEF-Marine Biodiversity and Ecosystem Function, the EU Project M&MS (EVK3-CT-2000-00044) a project funded by the Fundación BBVA, project PNAT/1999/BIA/15003/C of the Portuguese Science Foundation — FCT, and fellowships from FCT and ESF. We are grateful to Cécile Perrin, Ashwin Engelen, Onno Diekmann, Elena Varela and Elena Diaz-Almela for fruitful discussions, which helped improve this article, and to Gareth Pearson for revising the manuscript. We wish to thank the Editor and two anonymous referees for the improvements they suggested on previous versions of the manuscript.

References

Alberto F, Correia L, Arnaud-Haond S, Billot C, Duarte CM, Serrão EA (2003a) New microsatellites markers for the endemic Mediterranean seagrass, *Posidonia oceanica*. *Molecular Ecology Notes*, **3**, 253–255.

Alberto F, Correia L, Billot C, Duarte CM, Serrao EA (2003b) Isolation and characterization of microsatellite markers for the seagrass, *Cymodocea nodosa*. *Molecular Ecology Notes*, **3**, 397–399.

Alberto F, Gouveia L, Arnaud-Haond S, Pérens-Lloréns JL, Duarte CM, Serrão EA (2005) Spatial genetic structure, neighbourhood size and clonal subrange in seagrass (*Cymodocea nodosa*) populations. *Molecular Ecology*, **14**, 2669–2681.

Anderson JB, Kohn LM (1995) Clonality in soilborne, plant-pathogenic fungi. *Annual Review of Phytopathology*, **33**, 369–391.

Arnaud-Haond S, Alberto F, Teixeira S, 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**, 434–440.

Arnaud-Haond S, Belkhir K (2007) GENCLONE 1.0: a new program to analyse genetics data on clonal organisms. *Molecular Ecology Notes*, **7**, 15–17.

Arnaud-Haond S, Diaz Almela E, Teixeira S, Alberto F, Duarte CM, Serrao EA (2007) Vicariance patterns in the Mediterranean sea: East-West cleavage and low dispersal in the endemic seagrass *Posidonia oceanica*. *Journal of Biogeography*, **34**, 963–976.

Ayre DJ, Hughes TP (2000) Genotypic diversity and gene flow in brooding and spawning corals along the Great Barrier Reef, Australia. *Evolution*, **54**, 1590–1605.

Bastidas C, Benzie JAH, Uthicke S, Fabricius KE (2001) Genetic differentiation among populations of a broadcast spawning soft coral, *Sinularia flexibilis*, on the Great Barrier Reef. *Marine Biology*, **138**, 517–525.

Charpentier A (2001) Consequences of clonal growth for plant mating. *Evolutionary Ecology*, **15**, 521–530.

Cheplick GP (1997) Responses to severe competitive stress in a clonal plant: differences between genotypes. *Oikos*, **79**, 581–591.

Chung MY, Nason JD, Chung MG (2004) Implications of clonal structure for effective population size and genetic drift in a rare terrestrial orchid, *Cremastra appendiculata*. *Conservation Biology*, **18**, 1515–1524.

Cywinska A, Hebert PDN (2002) Origins of clonal diversity in the hypervariable asexual ostracode *Cypridopsis vidua*. *Journal of Evolutionary Biology*, **15**, 134–145.

Diaz-Almela E, Arnaud-Haond S, van de Vliet MS *et al*. Feed-backs between genetic structure and perturbation-driven decline in seagrass (Posidonia oceanica) *Meadows Conservation Genetics*, doi: 10.1007/s10592-007-9288-0.

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, Dodd RS (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.

Ellstrand NC, Roose ML (1987) Patterns of genotypic diversity in clonal plant-species. *American Journal of Botany*, **74**, 123–131.

Epperson BK, Li TQ (1997) Gene dispersal and spatial genetic structure. *Evolution*, **51**, 672–681.

Fager EW (1972) Diversity: a sampling study. *American Naturalist* **106**, 293–310.

Freeland JR, Noble LR, Okamura B (2000) Genetic diversity of North American populations of *Cristatella mucedo*, inferred from microsatellite and mitochondrial DNA. *Molecular Ecology*, **9**, 1375–1389.

Gini C (1912) Variabilità e mutabilità. In: *Studi Economico-Giuridici Facolta de Giurisprundenza dell' Universita di Cagliari, A*., Vol III, parte II.

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

van Groenendael J, de Kroon H (1990) *Clonal Growth in Plants: Regulation and Function*. SPB Academic Publishers, The Hague, The Netherlands.

Halkett F, Plantegenest M, Prunier-Leterme N, Mieuzet L, Delmotte F, Simon JC (2005a) Admixed sexual and facultatively asexual aphid lineages at mating sites. *Molecular Ecology*, **14**, 325–336.

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

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

Hangelbroek HH, Ouborg NJ, Santamaria L, Schwenk K (2002) Clonal diversity and structure within a population of the pondweed Potamogeton pectinatus foraged by Bewick's swans. *Molecular Ecology*, **11**, 2137–2150.

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.

Harper JL (1977) *Population Biology of Plants*. Academic Press, London.

Hill MO (1973) Diversity and evenness: 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.

Kapralov MV (2004) Genotypic variation in populations of the clonal plant Saxifraga cernua in the central and peripheral regions of the species range. *Russian Journal of Ecology*, **35**, 413–416.

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

Le Goff-Vitry MC, Pybus OG, Rogers AD (2004) Genetic structure of the deep-sea coral *Lophelia pertusa* in the northeast Atlantic revealed by microsatellites and internal transcribed spacer sequences. *Molecular Ecology*, **13**, 537–549.

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 understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Lovett-Doust L (1981) Population dynamics and local specialization in a clonal perrenial (*Ranunculus repens*). I. The dynamics of ramets in contrasting habitats. *Journal of Ecology*, **69**, 743–755.

Loxdale HD, Lushai G (2003) Rapid changes in clonal lines: the death of a 'sacred cow'. *Biological Journal of the Linnean Society*, **79**, 3–16.

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.

Murawski DA, Hamrick JL (1990) Local genetic and clonal structure in the tropical terrestrial bromeliad, *Aechmea magdalenae*. *American Journal of Botany*, **77**, 1201–1208.

Nagamitsu T, Ogawa M, Ishida K, Tanouchi H (2004) Clonal diversity, genetic structure, and mode of recruitment in a *Prunus ssiori* population established after volcanic eruptions. *Plant Ecology*, **174**, 1–10.

Olsen JL, Stam WT, Coyer JA *et al*. (2004) North Atlantic phylogeography and large-scale population differentiation of the seagrass *Zostera marina* L. *Molecular Ecology*, **13**, 1923–1941.

Parker ED Jr (1979) Ecological implications of clonal diversity in parthenogenetic morphospecies. *American Zoologist*, **19**, 753–762.

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.

Peakall R, Smouse P (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Peet R (1974) The measurement of species diversity. *Annual Review of Ecology and Systematics*, **5**, 285–307.

Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.

Pielou EC (1966) Shannon's formulae as a measure of species diversity: its use and misuse. *American Naturalist*, **100**, 463–465.

Pielou EC (1969) *An Introduction to Mathematical Ecology*. Wiley-Interscience, New-York.

Pielou EC (1975) *Ecological diversity*. New York, USA, 165pp.

Piquot Y, SaumitouLaprade P, Petit D, Vernet P, Epplen JT (1996) Genotypic diversity revealed by allozymes and oligonucleotide DNA fingerprinting in French populations of the aquatic macrophyte, *Sparganium erectum*. *Molecular Ecology*, **5**, 251–258.

Reusch TBH (2001) New markers — old questions: population genetics of seagrasses. *Marine Ecology — Progress Series*, **211**, 261–274.

Reusch TBH, Hukriede W, Stam WT, Olsen JL (1999) Differentiating between clonal growth and limited gene flow using spatial autocorrelation of microsatellites. *Heredity*, **83**, 120–126.

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

Rousset F (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58–62.

Rozenfeld AF, Arnaud-Haond S, Hernández-García E *et al*. (2007) Spectrum of genetic diversity and networks of clonal populations *Journal of the Royal Society Interface*.

Schroeder M (1991) *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman, New York.

Simpson EH (1949) Measurements of diversity. *Nature*, **163**, 688.

Sintes T, Marba N, Duarte CM, Kendrick GA (2005) Nonlinear processes in seagrass colonisation explained by simple clonal growth rules. *Oikos*, **108**, 165–175.

Smith B, Wilson JB (1996) A consumer's guide to evenness measures *Oikos*, **76**, 70–82.

Sokal RR, Rohlf FJ (1995) *Biometry*. W.H. Freeman, New York.

StatSoft I (2001) *statistica (Data Analysis Software System), Version 6*. www.statsoft.com.

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

Stoddart JA (1983) A genotypic diversity measure. *Journal of Heredity*, **74**, 489–490.

Stoddart JA, Taylor JF (1988) Genotypic diversity — estimation and prediction in samples. *Genetics*, **118**, 705–711.

Suehs CM, Affre L, Medail F (2004) Invasion dynamics of two alien *Carpobrotus* (Aizoaceae) taxa on a Mediterranean island: I. Genetic diversity and introgression. *Heredity*, **92**, 31–40.

Taylor DJ, Foighil DO (2000) Transglobal comparisons of nuclear and mitochondrial genetic structure in a marine polyploid clam (Lasaea, Lasaeidae). *Heredity*, **84**, 321–330.

Tibayrenc M, Ayala FJ (2002) The clonal theory of parasitic protozoa: 12 years on. *Trends in Parasitology*, **18**, 405–410.

Tibayrenc M, Kjellberg F, Ayala F (1990) A clonal theory of oarasitic 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.

Ueda HR, Hayashi S, Matsuyama S *et al*. (2004) Universality and flexibility in gene expression from bacteria to human. *Proceedings of the National Academy of Sciences, USA*, **101**, 3765–3769.

Uthike S, Benzie JAH, Ballment E (1998) Genetic structure of fissiparous populations of *Holothuria* (*Halodeima*) *atra* on the Great Barrier Reef. *Marine Biology*, **132**, 141–151.

Van der Hulst RGM, Mes THM, Falque M, Stam P, Den Nijs JCM, Bachmann K (2003) Genetic structure of a population sample of apomictic dandelions. *Heredity*, **90**, 326–335.

Vidondo B, Prairie YT, Blanco JM, Duarte CM (1997) Some aspects of the analysis of size spectra in aquatic ecology. *Limnology and Oceanography*, **42**, 184–192.

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.

Sophie Arnaud-Haond is a researcher in IFREMER (France) and an associate researcher in CCMar (Portugal). Her research interests focus on the influence of mating system and clonality, dispersal and selection on the ecology and evolution of marine populations. Carlos M. Duarte leads a team in IMEDEA (Spain) studying marine biodiversity from the genetic, species and habitat level to global biogeochemical cycles. Fillipe Alberto is a post-doctoral researcher in CCMar, interested in marine population genetics and ecology, marine phylogeography and clonality. Ester Serrão leads a research group (CCMar) that is primarily interested in marine ecology, adaptation and population genetics.

Supplementary material

The following supplementary material is available for this article:

Table S1 Survey list of published studies using molecular markers, and information extracted from the literature

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-294X.2007.03535.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.