Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual Medicago truncatula

M. SIOL, *I. BONNIN, †I. OLIVIERI, ‡J. M. PROSPERI* & J. RONFORT*

*UMR 1097 Diversité et Adaptations des Plantes Cultivées, INRA Montpellier, Domaine de Mełguil, Mauguio, France
†UMR (INRA-CNRS-UPS-INAPG) de Génétique Végétale, Ferme du Moulon, Gif/Yvette, France
‡Institut des Sciences de l’Evolution, Université Montpellier II, Place Eugène Bataillon, Montpellier Cedex, France

Introduction

The effective population size \((N_e)\) is a major concept in evolutionary and conservation biology, because it relates the theory of population genetics to the real world of natural populations (Wright, 1969). The magnitude of \(N_e\) determines the amount of gene sampling errors between generations that cause genetic drift and thus allows predictions about the rate of genetic differentiation between populations as well as inferences about the relative effects of genetic drift and selection at non-neutral loci. For instance, populations with small \(N_e\) values show reduced capacity to respond to changing environmental pressures and suffer from the accumulation of deleterious alleles (Higgins & Lynch, 2001). \(N_e\) equals the census size of the population only if the population behaves as an ‘ideal’ population, i.e. individuals mate at random, the number of progeny per parent follows a Poisson distribution and size is constant through time (Wright, 1931). Natural populations however seldom satisfy these criteria. Rather, a wide range of demographic and reproductive features like inbreeding, variance in reproductive success between individuals larger than for a Poisson distribution, or fluctuations in population size are expected to cause departures from the ideal theoretical population. \(N_e\) then represents the number of individuals that would have the same variance in allele frequencies or level of inbreeding or any feature reflecting the effect of drift than observed in the actual population (Kimura & Crow, 1963; Vitalis & Couvet, 2001). How violations of idealized population assumptions may affect \(N_e\) has been largely considered. This has led to the development and application of models predicting effective size from demographic and reproductive data (Crow & Kimura, 1970; Nunney, 1993; Whitlock & Barton, 1997; Nunney, 1999; Wang & Caballero, 1999). An alternative to such approaches are estimates based on genetic data (see review by Caballero, 1994). Such methods stem from the direct relationship between \(N_e\) and genetic variability, described in finite population genetics models (Crow & Kimura, 1970).

Abstract

Despite its significance in evolutionary and conservation biology, few estimates of effective population size \((N_e)\) are available in plant species. Self-fertilization is expected to affect \(N_e\), through both its effect on homozygosity and population dynamics. Here, we estimated \(N_e\) using temporal variation in allele frequencies for two contrasted populations of the selfing annual Medicago truncatula: a large and continuous population and a subdivided population. Estimated \(N_e\) values were around 5–10% of the population census size suggesting that other factors than selfing must contribute to variation in allele frequencies. Further comparisons between monolocus allelic variation and changes in the multilocus genotypic composition of the populations show that the local dynamics of inbred lines can play an important role in the fluctuations of allele frequencies. Finally, comparing \(N_e\) estimates and levels of genetic variation suggest that \(H_e\) is a poor estimator of the contemporaneous variance effective population size.

Keywords:
effective population size;
genetic drift;
microsatellite;
population structure;
selfing.

Correspondence: Mathieu Siol, UMR 1097 Diversité et Adaptations des Plantes Cultivées, INRA Centre de Montpellier, Domaine de Mełguil, F-34130 Mauguio, France. Tel.: ++33/(0)4 67 29 06 37; fax: ++33/(0)4 67 29 39 90; e-mail: mathieu.siol@supagro.inra.fr
Genetic estimators have been formulated based on measures of different parameters, like allele number, gene diversity, or linkage disequilibrium between polymorphic loci. Among these methods, the so-called ‘temporal method’, based on measurements of variation in allele frequencies over generations (Nei & Tajima, 1981; Waples, 1989), has received considerable attention (Nei & Tajima, 1981; Waples, 1989; Wang & Caballero, 1999; Anderson et al., 2000). \( N_e \) is inferred from observed changes in allele frequencies through time rather than from the existing pattern of genetic diversity. As a result, estimates are not confounded by historical processes and require no assumptions about previous gene flow and selection. The temporal method has been applied to a variety of animal species (Arnaud & Laval, 2004; Meunier et al., 2004; Aspi et al., 2006; Poulsen et al., 2006), but there are still relatively few published studies estimating \( N_e \) in plants (see however Husband & Barrett, 1992; Goldringer et al., 2001).

As a result of increased homozygosity, self-fertilization is expected to reduce \( N_e \): in a population at drift–mutation equilibrium, \( N_e = N(2-s)/2 \) where \( s \) is the proportion of progeny derived from self-fertilization (excluding the selfing events because of random matings in a finite population; \( s \) is assumed constant through time and homogeneous among individuals) and \( N \) the demographic size of the (otherwise supposedly ideal) population (Pollak, 1987; Nordborg & Donnelly, 1997; Laporte & Charlesworth, 2002). Other attributes of self-fertilizing species may however interact with selfing and affect the effective population size. First, hitchhiking effects and background selection can lead to a larger reduction in neutral gene diversity than in outcrosser because of extended linkage disequilibria between selected and neutral markers. Such effects would be analogous to drift and reduce \( N_e \) (Charlesworth et al., 1993; Caballero & Santiago, 1995). Secondly, because of limited pollen flow, predominantly self-fertilizing populations are likely to be subdivided into more or less isolated lineages, which could reduce or increase the effect of drift at the whole population level depending on population dynamics and variation in the relative contribution of subpopulations (Barton & Whitlock, 1997; Whitlock & Barton, 1997; Wang & Caballero, 1999). Theory thus predicts that self-fertilizing species should exhibit reduced effective population sizes and hence reduced levels of within population genetic diversity compared with outcrossing species (see also Ingvarsson, 2002). Although reviews of empirical data on patterns of diversity generally confirm this latter expectation (Hamrick & Godt, 1990; Schoen & Brown, 1991; Charlesworth, 2003), little is known about the local dynamic of self-fertilizing populations and in particular on the magnitude of genetic drift at the population level (see however Goldringer et al., 2001 and Meunier et al., 2004 for experimental populations of wheat and a selfing animal species respectively).

Medicago truncatula Gaertn. (Fabaceae) is an annual, diploid plant species which reproduces almost exclusively through selfing (selfing rates are typically around 96–98%. Bataillon & Ronfort, 2006). It is a widespread species native to the Mediterranean basin. Previous marker-based genetic analyses have shown that large levels of genetic diversity can be maintained at the population level in this species and that large levels of genetic differentiation can occur even at a very fine spatial scale (Bonnin et al., 1996a, 2001). In this paper we estimate the current effective population size of two natural populations of \( M. \) truncatula differing in their local structure using measurements of temporal variation in allele frequencies at neutral molecular markers. Our aims were first to quantify the contemporaneous effect of drift in these self-fertilizing populations and secondly to get insight into how population subdivision and demographic parameters could affect the effective population size and levels of diversity.

**Materials and methods**

**Populations studied and plant material**

Two populations located in the French Mediterranean region were studied: the Aude population (43°07′N, 3°04′E) and the Salses population (42°50′N, 2°55′E). The Aude population which has already been studied for genetic diversity and population structure (Bonnin et al., 1996a, 2001), is a discontinuous population, composed of three subpopulations hereafter referred to as Aude1, Aude2 and Aude3. Aude1 and Aude3 are bordering a vineyard and are approximately 30 m apart. Aude2 is bordering an old abandoned vineyard and is located 50 m from the others (Bonnin et al., 1996a). Previous molecular analyses reported a high level of genetic diversity in this population, large genetic differentiation between the three subpopulations and a fine scale population subdivision in Aude2 (Bonnin et al., 1996a, 2001). In springs of 1990, 1993 and 1996, fresh leaves were sampled for one plant every meter along a 30-m transect in each subpopulation. Sample sizes per subpopulation and year varied from 19 to 31 (Table 3). One hundred and thirty-nine, 127 and 114 flowering individuals were counted in 1996 in Aude1, Aude2 and Aude3 respectively (Bonnin et al., 2001). These values will be used as the census sizes against which we compare our \( N_e \) estimates. To evaluate between individuals variation in the reproductive effort, we carried out in situ measurements during June 1996 on all the plants sampled for this year. On each plant, we scored the number of green pods at the beginning of flowering.

The population of Salses is located near Perpignan (approximately 35 km from the Aude population) on a fallow bordering orchards and vineyard. This population covers a large surface (approximately 80 × 60 m) and is composed of a large number of individuals (the census...
population size was estimated to lie between 2000 and 5000) continuously distributed over the sampled area. Seeds were collected from pods fallen on the ground (in this species pods are released from senescing plants). Pods were sampled over the whole population in summer 1999 and 2004 and the spatial coordinates of each sampled pod were recorded in order to carry out a fine-scale spatial analysis. Pods were threshed and one seed per pod was sown in a greenhouse for molecular analyses.

Microsatellite analyses

For the Aude population, microsatellite polymorphism was assayed using the same loci and following the same protocol as in Bonnin et al. (2001) and see Table S1. For Salses population, 13 microsatellites loci were used (Baquerizot-Audiot et al., 2001 and Table S1). DNA was extracted from 200 mg of frozen leaves. Amplification reactions were performed in a final volume of 20 μL in the presence of 50 ng of template DNA, 4 pmol of the reverse primer and 1 pmol of the forward primer, 0.2 mM of each deoxynucleotide, 2 mM of MgCl₂ and 0.2 unit Taq polymerase. The forward primer was 5’ labelled with one of the three fluorophores (6FAM, NED or HEX). After 10 min at 94 °C, 35 cycles of 30 s at 94 °C were performed, 1 min at either 50 or 55 °C depending on the locus, 1 min at 72 °C followed by the final extension step of 10 min at 72 °C. Amplification products were analysed on an ABI prism 3100 Genetic Analyser. Samples were prepared by adding 3 μL of diluted PCR products to 7.85 μL formamide and 0.15 μL of size marker AMM524. Analyses were performed using the GENESCANGENESCAN 3.1 and GENOTYPER 2.5 softwares (Applied Biosystems, Foster City, USA).

Statistical analyses

Gene diversity

Allele frequencies, Nei’s gene diversity (Hₑ) and departure from Hardy–Weinberg expectation (as measured through Fₛ) were calculated for each sample (nine samples for Aude, two for Salses) using genetix 4.05 (Belkhir et al., 1996–2004). Fₛ values were estimated by the unbiased f estimator proposed by Weir & Cockerham (1984) and their significance was tested using 1000 permutations of alleles among individuals. The selling rate for each sampled population and subpopulation was then inferred using the classical formula Fₛ = s/ (2–s) which assumes that the population is at equilibrium for a fixed selling rate s and no within-population structure. For the Aude population, we tested for homogeneity in Nei’s gene diversity between subpopulations considering the three sampling dates. To do so, we used the Scheirer–Ray–Hare extension of the Kruskall–Wallis test which is analogous to a two-way ANOVA (Sokal & Rohlf, 1995). In a second step, differences between subpopulations were considered independently for each sampling date through Wilcoxon’s signed rank test using loci as replicates. Temporal variation in Nei’s genetic diversity and allelic richness were first tested at the global scale (i.e. pooling the three subpopulations) using 3000 permutations as implemented in fstat 2.9.3. At the subpopulation level and for the population of Salses, comparisons between pairs of sampling dates were examined using a Wilcoxon’s signed rank test for paired observations (Sokal & Rohlf, 1995).

For each sample, we tested linkage disequilibrium between all pairs of loci using permutations of alleles among individuals and applying a Bonferroni correction as implemented in fstat 2.9.3 (Goudet, 1995; Goudet et al., 1996). Given the selfing status of M. truncatula, we were also interested in the multilocus composition of the population. Combining genotypic data among loci, we thus inferred the multilocus genotype of each sampled individual and examined the temporal evolution in the multilocus composition of each population.

Spatio-temporal variation in allele frequencies

For the population of Aude, variation in allele frequencies were first analysed using a hierarchical AMOVA accounting for both spatial and temporal variation between sampling dates and subpopulations, using the software Arlequin (Excoffier, 1998–2006). Details on the spatial structure of Aude population were then obtained computing Weir & Cockerham’s (1984) estimator of Fₛ, between each pair of subpopulations for the three sampling dates separately. For each pairwise value, we calculated a 95% confidence interval (CI) from 1000 bootstraps over loci. Each pairwise value was then tested using a G-test (as proposed by Meunier et al., 2004) permuting multilocus genotypes among populations using fstat 2.9.3 (Goudet, 1995; Goudet et al., 1996). Adjusted P-values for Fₛ were obtained applying Bonferroni correction for multiple comparisons. A similar analysis was used to compute temporal Fₛ for three types of samples: the two samples from Salses, pooling the Aude samples, and for each Aude’s subpopulation. As above we calculated a 95% CI and tested the pairwise values using a G-test (Goudet, 1999).

Effective population size estimation

We focused on the variance effective size using the variance in allele frequencies through time. We first used the classic moment estimator (Waples, 1989) defined as:

$$\hat{N}_e = \frac{t}{2 \left( \hat{F}_C - \left( \frac{1}{S_0} + \frac{1}{S_t} \right) \right)}$$

where t is the number of generations separating two samples, S₀ and Sₜ are the sample sizes at time 0 and time t respectively and $\hat{F}_C$ is the estimator of the standardized variance in allele frequencies between generation 0 and t. $\hat{F}_C$ was estimated using a weighted sum of $F_C$ values.
over loci following Waples (1989). From the above equation, it should be noted that in cases where the sampling variance is larger than the variance induced by drift (i.e. \(1/2S_0 + 1/2S_t > F_{Ct}\)), the method returns negative values of \(N_e\), implying that effective size is too large compared with the sample size. The 95% CI on \(N_e\) was derived from that of \(F_C\) using a classic \(\chi^2\) approximation (Waples, 1989). However, as noted by Goldringer & Bataillon (2004) the \(\chi^2\) approximation may be inaccurate, especially when marker loci are multiallelic and allelic frequencies are unbalanced as in our dataset. Therefore as they suggested, we simulated the actual distribution of \(F_C\) based on the estimated \(N_e\) value to find the 95% CI. We implemented their procedure in a C++ program in order to reduce computational time compared with the original procedure (details about the simulation procedure are given in Appendix and the program is available upon request to M.S.).

The above estimator of \(N_e\) is derived assuming the so-called sampling scheme II (sensu Waples, 1989) where sampled individuals do not reproduce, so that there is no correlation between estimated allele frequencies and \(N_e\). For Salses population we can use these estimators as sampling was made on pods. However, sampling in Aude population involved leaves collected on individuals that could have contributed to the next generation. According to Waples (1989), this corresponds to a sampling scheme I, where estimated allele frequencies are expected to be correlated to allele frequencies in the last generation. Following Waples (1989), \(N_e\) values for Aude population were estimated as:

\[
\hat{N}_e = \frac{t}{2 \left[ F_C - \left( \frac{1}{N_e} + \frac{1}{N_i} - \frac{1}{N} \right) \right]}
\]

where \(N\) is the demographic size of the population. As we do not have accurate estimates of \(N\) for each sampling date, we used the demographic population sizes measured by Bonnin et al. (2001) in 1996 as a proxy. As the sampling design is different between the two populations and only two loci were genotyped in both populations, direct comparisons of the results are to be considered with caution.

Effective sizes were also estimated on the same datasets using the coalescent-based two-samples method described in Anderson (2005) and implemented in the program CONE (downloadable at http://www.swfsc.noaa.gov/textblock.aspx?Division=FED&id=3436). This method is based on the assumption that the genealogy of the sampled gene copies follows a neutral coalescent process. Tavare (1984) gives the probability that \(n_0\) lineages have \(n_t\) ancestral lineages \(t\) generations in the past given \(N_e\). It is thus possible to compute the likelihood of the effective size of a population under a coalescent model between two sampling dates. One should note that this method deals only with the gene genealogy and not with the population gene frequencies (which are considered as nuisance parameters). As the genealogy is better approximated by the neutral coalescent as the population size becomes large, this method should lead to better estimates for large populations. However, even with small \(N_e\) the genealogy might be well approximated by the coalescence approach (Berthier et al., 2002).

These two estimation procedures were applied to the two sampling dates available for Salses, i.e. 1999 and 2004 and to the three available couples of years (i.e. 1990/1993, 1993/1996, 1990/1996) for Aude. In this latter case, we estimated \(N_e\) for the whole population, pooling data by year class across subpopulation samples; in a second step, each subpopulation was considered independently to obtain local \(N_e\) estimates.

Among loci variation in \(F_C\) values and effect of linkage disequilibrium

Heterogeneity in \(F_C\) values between loci can provide evidence for selection. This stems from the fact that drift acts over the whole genome whereas natural selection will affect differently each locus (Lewontin & Krakauer, 1973). For the present study, checking for heterogeneity among loci in their rate of change in allele frequency was possible only for Salses population for which data on 13 loci were available. To do this, we followed the method of Goldringer & Bataillon (2004) who proposed to compare each individual (i.e. by locus) \(F_C\)-value to the expected \(F_C\) distribution based on the effective size estimated from the remaining loci and conditional on the observed initial allele frequencies. For each locus, we used 3000 independent simulations to build the expected \(F_C\) distribution. For a given locus, the distribution obtained in this way was then used to compute the probability \(P\) for the temporal variance of allele frequencies to be greater than or equal to the observed value at this locus.

The strong level of linkage disequilibria between loci expected under selfing is likely to influence the power of the estimation of \(N_e\) as each locus may no longer be independent from the others. To check for such an effect in Salses, we generated the distribution of the variance in \(F_C\) values expected among 13 loci using 3000 simulations assuming independence between loci and conditional on the \(N_e\) value estimated and the observed initial allele frequencies. This distribution was used to verify if the observed variance in \(F_C\) values was included in the 95% of this distribution. In addition, we estimated the effective size from the variation in multilocus genotype frequencies over time since under complete linkage disequilibrium, each multilocus genotype then being considered as a single allele. The comparison between the effective size computed in this way and the effective size computed as a weighted sum of monolocus allelic frequency variation allows disentangling the contribution of the multilocus genotype frequency variation into allelic frequencies variation.
Results

Gene diversity

The number of alleles per locus, Nei’s gene diversities ($H_e$) and $F_{IS}$ values are reported in Tables 1 and 2 (Salses and Aude). In Salses, all microsatellite loci used were polymorphic, with two to six alleles detected per locus in 1999 and a significant increase in the mean number of alleles per locus in 2004 compared with 1999 (Wilcoxon’s signed rank test, $P < 0.01$). In 1999 and 2004, the level of gene diversity ($H_e$) in this population was large for all microsatellite loci except ENPB1 and MAA660456. Over all loci, the variation of allele frequencies between sampling dates was not significant ($F_{ST} = 0.01$ ($P > 0.05$) and the difference in the level of gene diversity between 1999 and 2004 not significant ($P > 0.05$ Wilcoxon’s signed rank test). As already observed in 1996 by Bonnin et al. (2001), population Aude as a whole also exhibited a large molecular diversity both in terms of gene diversity and of number of alleles per locus. Also consistently with previous studies (Bonnin et al., 1996a, 2001), Aude1 appeared slightly less variable than Aude2 and Aude3, whatever the sampling year (Table 2). This difference was however only significant between Aude1 and Aude3 and for the 1990 sampling (Wilcoxon’s test, $P = 0.03$).

Comparing the sampling years, we observed a general although not significant increase in the level of gene diversity and in the allelic richness between 1990 and 1996 for all subpopulations ($P = 0.65$ and $P = 0.11$ respectively). All the samples analysed displayed a significant departure from Hardy–Weinberg equilibrium (global test, $P < 0.001$, see Tables 1 and 2). For Aude1, the $F_{IS}$ estimated in 1990 and 1993 reflected lower selfing rates than in 1996 and compared with values found in the two other subpopulations (Table 2). For Salses, the estimated $F_{IS}$ value in 1999 was 0.98, which agrees with the selfing rate estimated using progeny array analyses in this population ($s = 0.98$, data not shown).

Multilocus composition

Given the selfing rate of M. truncatula, potentially high levels of linkage disequilibria were expected even between loosely linked loci. Indeed, for Salses population,
61 pairs of loci out of 78 were found to be in linkage disequilibrium (for a type I error fixed at 5% when rejecting the equilibrium hypothesis) in 1999 (52 such pairs were observed in 2004). In this population, we found 16 multilocus genotypes in 1999, four of them representing 80% of the whole sample (Fig. 1). Basically, four or five inbred lines with very different genotypes can be observed along with some recombinant lines between most common lines. These recombinant lines are generally represented by one or two individuals in the population. A last category was represented by recombinant genotypes with some loci remaining heterozygotes. In 2004, we found twice more multilocus genotypes (32). New multilocus genotypes were either recombinant lines between the most frequent lines or lines differing only at one locus from previously observed genotypes. Only one genotype (labelled mig in Fig. 1) carrying new alleles at more than one locus was found. The three most frequent genotypes in 2004 were already the most frequent in 1999, despite a substantial decrease in frequencies for two of them (b18, b19, Fig. 1). Finally, the spatial distribution of the multilocus genotypes was relatively stable over time (Fig. 1).

Consistently with previous studies (Bonnin et al., 2001), we found a rather low linkage disequilibrium between marker loci in the subpopulations of Aude, with less than seven pairs of loci out of 10 displaying significant linkage disequilibrium whatever the sampling year. Aude2 displayed a similar multilocus composition as Salses with a set of four genotypes occurring at intermediate frequencies (Table 3). In contrast, Aude1 and Aude3 displayed a single dominant multilocus genotype and several multilocus genotypes observed only once (Table 3). The low number of multilocus genotypes detected could be due to the low number of markers (more genotypes may have been detected with more markers). The multilocus composition of Aude1 was quite stable between 1990 and 1996. Conversely, Aude2 showed the largest variation among generations, with new multilocus genotypes replacing multilocus genotypes that were frequent in 1990. Particularly interesting is the case of genotype 19 which was absent in 1990, and progressively increased in frequency.

Fig. 1 Location (spatial coordinates) of the multilocus genotypes revealed in Salses for the two sampling dates. Multilocus genotype labels are depicted in each panel. All labels beginning with LR indicates that the corresponding genotype is a recombinant line between either a7 and b18 or a7 and b19. The label LR-1 gather different recombinant lines that were represented only once in our samples. Note that none of those observed in 1999 was still present in 2004. Hyb represent recombinant genotypes between a7 and either b18 or b19 displaying heterozygosity at some loci. In (b), triangles of a given colour refer to the multilocus genotype represented by the same colour except for one locus. Axes units are metres.

Table 3 Multilocus composition of the three Aude’s subpopulations. For simplicity, only the multilocus genotypes represented more than once are represented.

<table>
<thead>
<tr>
<th>Genotype identifier</th>
<th>Multilocus genotype</th>
<th>1990</th>
<th>1993</th>
<th>1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aude 1</td>
<td>7</td>
<td>215-140-112-102-214</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>241-140-112-102-214</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>235-140-112-102-214</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>235-164-112-099-216</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>n = 20</td>
<td>n = 25</td>
<td>n = 31</td>
<td></td>
</tr>
<tr>
<td>Aude 2</td>
<td>16</td>
<td>215-148-112-102-224</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>89</td>
<td>245-148-112-123-224</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>88</td>
<td>245-148-112-102-224</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>245-148-112-093-224</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>215-164-112-090-224</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>n = 19</td>
<td>n = 24</td>
<td>n = 30</td>
<td></td>
</tr>
<tr>
<td>Aude 3</td>
<td>72</td>
<td>241-160-109-099-224</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>235-164-112-102-224</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>241-160-112-099-224</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>235-164-112-099-216</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>n = 20</td>
<td>n = 23</td>
<td>n = 29</td>
<td></td>
</tr>
</tbody>
</table>

Each multilocus genotype is described as the allelic state (allele size) observed respectively at locus MTS5, MTS6, MAA660252, MAA660456 and MAA660538 (all these multilocus genotypes were homozygous for all the studied loci).

n. Sample size.
replacing other genotypes. An intermediate situation is observed in Aude3 where the dominant genotype varies over time. In agreement with the high pairwise $F_{ST}$ values between subpopulations (see below), there are very few multilocus genotypes shared by two subpopulations: one is shared by Aude1 and Aude3 in 1990 and is represented by two individuals in total and one is shared by Aude1 and Aude3 in 1996 and is represented by four individuals.

**Spatio-temporal variation in allele frequencies**

As shown in Table 4, allele frequencies varied significantly through both space and over the three sampling dates). In contrast, overall significant differentiation between years in Aude1 ($F_{ST} = 0.491^{***}$). Although we found a better agreement between moment-based and likelihood methods when rare alleles were pooled, and the likelihood curves also appeared sharper except for Aude1 between 1990 and 1996. We consequently choose to report only values obtained when pooling rare alleles. Confidence intervals based on simulations were very close to those based on the $\chi^2$ approximation and are not shown. For the population of Salses, average $N_e$ values estimated using each microsatellite locus as an independent data point were around 150–170. Although the observed variance in $F_c$ values among loci was somewhat low ($V_{obs} = 4.05 \times 10^{-4}$) compared with the average variance expected assuming independence among loci, this difference was not significant ($P = 0.37$). Furthermore, among the 13 loci assayed, we found no deviation of $F_c$ values from the expected distribution (as obtained through simulations following Goldringer & Bataillon, 2004) except for locus MAA660749 which exhibited a fairly small $P$-value ($P = 0.0183$).

Considering the Aude population as a single one, and using the largest time interval available (i.e. 1990/1996),

### Table 4 AMOVA table for the population of Aude.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>Variance components (%)</th>
<th>$F$-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subpopulations</td>
<td>234.6</td>
<td>0.792 (44.8)</td>
<td>$F_{CT} = 0.448^{***}$</td>
</tr>
<tr>
<td>Between years within subpopulations</td>
<td>27.11</td>
<td>0.076 (4.3)</td>
<td>$F_{SC} = 0.078^{***}$</td>
</tr>
<tr>
<td>Between individuals</td>
<td>379.6</td>
<td>0.898 (50.8)</td>
<td>$F_{ST} = 0.491^{***}$</td>
</tr>
<tr>
<td>Total</td>
<td>641.3</td>
<td>1.766</td>
<td></td>
</tr>
</tbody>
</table>

$***P < 0.001$

### Table 5 Effective population size estimates and their confidence intervals.

<table>
<thead>
<tr>
<th>Population</th>
<th>$n_s$</th>
<th>$t$</th>
<th>$F_c$</th>
<th>$N_e_{Waples}$</th>
<th>$N_e_{Anderson}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aude-90/96</td>
<td>59–90</td>
<td>6</td>
<td>0.033</td>
<td>141.2 [46.4–458.6]</td>
<td>15.5 [59.7–646.9]</td>
</tr>
<tr>
<td>Aude-93/96</td>
<td>72–90</td>
<td>3</td>
<td>0.060</td>
<td>30 [10–68]</td>
<td>32 [14.8–75.4]</td>
</tr>
<tr>
<td>A1–90/96</td>
<td>20–31</td>
<td>6</td>
<td>0.045</td>
<td>297.2 [29–∞]</td>
<td>45.95 [14.8–∞]</td>
</tr>
<tr>
<td>A1–90/93</td>
<td>20–25</td>
<td>3</td>
<td>0.081</td>
<td>38.6 [8.7–∞]</td>
<td>54.4 [15.2–∞]</td>
</tr>
<tr>
<td>A2–90/96</td>
<td>19–30</td>
<td>6</td>
<td>0.257</td>
<td>13.6 [1.98–47.5]</td>
<td>16.8 [6.4–58.9]</td>
</tr>
<tr>
<td>A2–90/93</td>
<td>19–24</td>
<td>3</td>
<td>0.213</td>
<td>8.9 [1.5–35.2]</td>
<td>15.1 [5.8–74.1]</td>
</tr>
<tr>
<td>A3–90/96</td>
<td>20–29</td>
<td>6</td>
<td>0.045</td>
<td>265 [31.3–∞]</td>
<td>455. 8 [44.8–∞]</td>
</tr>
<tr>
<td>A3–90/93</td>
<td>20–23</td>
<td>3</td>
<td>0.265</td>
<td>6.7 [2–16.6]</td>
<td>8.5 [4.8–14.3]</td>
</tr>
<tr>
<td>A3–93/96</td>
<td>23–29</td>
<td>3</td>
<td>0.266</td>
<td>6.5 [1.4–18.5]</td>
<td>9.5 [5.2–24.9]</td>
</tr>
<tr>
<td>Salses-99/04</td>
<td>85–97</td>
<td>5</td>
<td>0.028</td>
<td>153.3 [82.6–422]</td>
<td>173.2 [75.32–548]</td>
</tr>
</tbody>
</table>

$n_s$ gives the sample size of the two generations considered, and $t$ the time between these generations. $F_c$ is the averaged variance of allele frequencies over all loci. $N_e$ is given according to the methods described in Waples (1989) and Anderson (2005). The symbol $\approx$ indicates either that the moment-based method returned a negative $N_e$ estimate or that the likelihood method returned an infinite $N_e$ estimate. Census sizes considered for each population are 2000–5000 for Salses, 139, 127 and 114 for Aude1, Aude2 and Aude3 respectively (see text). Bold values indicate estimates associated with finite confidence intervals.

JOURNAL COMPILATION © 2007 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY
we obtained $N_e$ values of the same order of magnitude as those estimated for the population of Salses. However, when estimated using the intermediate time intervals (1990–1993 and 1993–1996), we found $N_e$ values approximately five times smaller. Despite the small sample sizes when considering each individual subpopulation, relatively accurate estimates of $N_e$ were obtained at the subpopulation level. Out of 9 estimations, respectively 6 and 5 showed finite confidence intervals for the moment and likelihood methods (Table 5). Whatever the time interval considered, the effective population size estimated for Aude2 was around 15 individuals. $N_e$ values lower than 10 were obtained for Aude3 for the 3-year intervals but $F_c$ values estimated between 1990 and 1996 were very low which translated in a large $N_e$ value for this period. For Aude1, $N_e$ estimates appeared slightly greater than in the two other subpopulations, with only one estimate (over six values) associated with a finite confidence interval (i.e. Waples's estimate for 1993–1996). Indeed infinite confidence intervals stem from the fact that variation in allele frequencies is close to sampling variance. As sample sizes were quite similar between all three subpopulations, this tends to indicate a larger $N_e$ for Aude1.

Finally, when estimated from variation in multilocus genotype frequencies, $N_e$ in Salses was 134 ($\chi^2$ CI = [34–998]). This estimation is quite close from those obtained considering the weighted sum of monolocus $F_c$ values, suggesting that multilocus genotype (hereafter denoted lines) frequency variation accounts for an important part of the variation in allele frequencies. When applied to Aude subpopulations, this method indicated that lines dynamic was a major determinant in allele frequency variation in Aude2 as well but not in the two others subpopulations. For these, most of the detected lines differ very little from one another (because of either recombination or mutation), so that allele frequency variation is much more important than line frequency variation (see Table 5). For example between 1990 and 1996, $N_e$ estimated in this way yielded 35, 7 and 3 for Aude1, Aude2 and Aude3 respectively showing a larger discrepancy with previously shown estimates for Aude1 and Aude3 than for Aude2.

Pod production in the Aude population

Pod production in the Aude population during 1996 varied widely between individuals. Indeed, the mean number of pods scored per plant was 6.80 and the between individuals variance was 128.6. These values can be used to estimate the standardized variance in the contribution of each individual to reproduction in terms of female gametes, i.e. $V_I = \text{variance}/\text{mean}^2 = 2.78$ (Nunney, 1999). This value is approximately three times larger than the value expected under a Poisson distribution of individual contributions ($V_I = 1$). The multilocus genotype of the scored plants being available, we compared the relative success of each multilocus genotype (roughly estimated as the ratio between the line frequency in 1996 vs. 1993) and its averaged in situ pod production (data not shown). These two measures were clearly not positively correlated ($r = 0.08$, $P > 0.20$).

Discussion

In this study we have reported effective population size estimates in two populations of the highly selfing annual *M. truncatula*, together with various parameters linked to genetic diversity and spatial structure. Consistent with the selfing status of *M. truncatula*, we observed high $F_{IS}$ values for all the samples studied and we detected linkage disequilibria between most microsatellite loci even at a fine spatial scale (i.e. within Aude subpopulations). As selfing is expected to reduce $N_e$ by a factor $1+F_{IS}$ (Pollak, 1987; Nordborg & Donnelly, 1997), we expected the mating system to contribute to a loss in gene diversity at the population level and $N_e/N$ ratios around one half. Most of the estimated $N_e/N$ ratios however, were lower. As discussed below, this discrepancy is probably related to variation in demographic size but also to the particular dynamic and population structure of self-fertilizing populations.

$N_e/N$ ratios in a continuous population

For the large and continuously distributed population of Salses, average $N_e$ values were around 150. Indicating a $N_e/N$ ratio between 5% and 10%. Interestingly, this value is close to the average ratio found in Frankham’s (1995) meta-analysis on effective population size ($N_e/N \sim 10\%$). This result also shows that other factors than selfing per se must be invoked to explain this discrepancy between $N_e$ and $N$ values.

In a nonsubdivide and monoeocious population, two other factors are expected to have major effects on $N_e$: (i) fluctuations in population size (Frankham, 1995) and (ii) variation in the contribution of individuals to the next generation, which is expected to have a much more pronounced effect under regular self-fertilization. Under partial selfing, the effective size in a population of constant size is given by $N_e = 4N/[2(1-F_{IS}) + (1+F_{IS})V_k]$ (where $V_k$ is the variance in reproductive success, Caballero & Santiago, 1995). As noted by Caballero & Santiago (1995) the variance in reproductive success may arise from what they term ‘random selection’ which encompasses environmental processes and from heritable differences in individual fitness. Data about the relative importance of environmental vs. genetic differences in fitness have not been gathered in this natural population. However seed production of genotypes collected in 1999 in Salses was measured in a greenhouse experiment (J. Ronfort, unpublished data). This study showed that the seed production of individual line is positively
correlated \((r = 0.766; P < 0.01)\) with the number of individuals displaying these genotypes in 2004 (rougly considered as the realized success of each of the considered lines). This relation seems to indicate that fitness variation exists between lines but the importance of this component to the total variance of reproductive success cannot be inferred precisely. This result is also consistent with the fact that the effective population size estimated using each multilocus genotype as an allele was close to the values found when considering the weighted sum of monolocus \(F_c\). Therefore in this population, the variation of allele frequencies is primarily determined by the frequency dynamics of the different inbred lines. Similar results have been reported in partially asexual organisms (for a review see De Meester et al., 2006). In such species and under rare sexual reproduction, it can be shown that the effective population size can be expressed as a function of the frequency of these genotypes in the population \(i.e. 1/N_e = \sum p_i^2\), where \(p_i\) is the relative frequency of the \(i\)th genotype). In our study however, because of rare outcrossing events between lines, loci are not completely linked and each locus provides complementary information in the \(N_e\) estimation. Indeed, the variance of \(F_c\) values between loci fell within the distribution simulated under the assumption of independence between loci. This means that even though linkage disequilibrium is quite strong, each locus brings some information in the \(N_e\) estimation. In conclusion, although both fluctuations in population size and variation in reproductive success could explain the reduced \(N_e/N\) ratio observed in Salses, our multilocus analysis pointed toward variation in the contribution of the different lines to the following generations. This also means that the \(N_e\) values estimated in this population may be influenced by selective differences between individuals as well as environmental variation (which could affect the recruitment of the different genotypes) and cannot be considered reliably as a neutral \(N_e\).

Our multilocus analysis showed that new lines appeared in Salses population during the course of the study. These new lines could result from either immigration or mutation events. As the mutation rate of microsatellites is not expected to exceed \(10^{-7}\) (Jarne & Lagoda, 1996; Thuillot et al., 2002), we expected mutation to be negligible over 5 years. However, comparing the multilocus genotypic composition of Salses population between 1999 and 2004 showed that the new genotypes detected in 2004 generally differ from the resident ones at no more than one or two loci. This suggests that mutation rather than migration could be responsible for the appearance of new microsatellite variation in this population (although rare multilocus genotypes could have gone unnoticed in 1999 if they occurred in very low frequency). Furthermore, as indicated on Fig. 1, only one instance of immigration was identified with an individual carrying several new alleles.

## Spatio-temporal effects in the subdivided Aude population

Given the sampling design followed in the population of Aude (three sampling dates and independent sampling of the three subpopulations), several estimates of \(N_e\) could be computed for this population. Although we detected a clear and large differentiation between the three subpopulations for all the sampling years, we first computed \(N_e\) for the whole population, pooling data across subpopulations. Pooling data in this way provided more precise temporal-method estimates of \(N_e\) by increasing the total number of individuals used in the estimation and thereby decreasing the confidence intervals around \(N_e\) (Waples, 1989). Comparisons between \(N_e\) values obtained in this way for the different time intervals available (1990–1993, 1993–1996 and 1990–1996) yielded an unexpected result: variation in allele frequencies over a 3-year period (1990–1993 and 1993–1996) indicated \(N_e/N\) values around 10% whereas the 6-year interval (1990–1996) resulted in larger effective population size estimates, with \(N_e/N\) ratios of approximately one half. Similar results were obtained for subpopulations 1 and 3 (Table 5); we thus hypothesize that the behaviour of \(N_e\) estimates at the whole population level probably reflect a particular dynamic of these two subpopulations. Such variation over time in the effective population size could be due to fluctuations in census size of the subpopulations. For Aude3 however, we showed that the multilocus composition observed in 1996 was closer to that observed in 1990 than in 1993, due mostly to a new line occurring in high frequency in 1993 and nearly absent in 1996. Such variation over years in the multilocus composition of the population could be due to variation in climatic conditions or germination abilities of the different lines. Although the significance of the seed bank in this species needs investigation, it is likely to have an effect on the population dynamic. Indeed seeds are tough in *M. truncatula* and able to germinate after several years of dormancy (see also Bonnin et al., 2001).

When pooling data across subpopulations, our interest was also to compare \(N_e\) estimates with or without explicitly incorporating population structure (Whitlock & Barton, 1997; Nunney, 1999). Assuming an idealized population and equal productivity among demes, \(N_e\) in a structured population is a function of within-deme inbreeding \((F_{IS})\) and nonrandom mating among demes \((F_{ST})\), namely, \(N_e = 4N/(1-F_{ST})(2(1-F_{IS}) + \{1-F_{IS}\}V_k)\) where \(N\) is the total number of adults summed over demes and \(V_k\) is again the variance in individuals contributions (Wang & Caballero, 1999, Equation 16). For the Aude population, following this model and assuming \(F_{IS} \approx 1\) and \(F_{ST} \approx 0.5\), we get \(N_e = 4N/V_k\). Then, assuming a Poisson distribution of individual contributions with complete selfing \((s = 1)\), \(V_k = 2 + 2s = 4\) and \(N_e\) is expected to be approximately equals to the
demographic population size \((N_e = N)\) (Caballero & Hill, 1992). Therefore, a \(N_e/N\) ratio of 0.1 means that \(V_e\) must be ten times larger than the Poisson expectation. Consistently, in situ between individuals variances in pod production during 1996 were shown to be very large in the Aude population although not sufficient in itself to explain the low \(N_e/N\) value we found for this population. However, unlike in the Salses population, the mean pod production by inbred line appeared uncorrelated with their reproductive success. Interestingly, a similar result was obtained when comparing the reproductive success of individual lines and the pod production of these lines under greenhouse conditions (data extracted from Bonnin et al., 1996b). All these results together suggest that the low effective population size observed in the subpopulations of Aude are mainly due to the occurrence of a large variance in the relative contributions of individuals to the next generations. However, unlike the results available in the Salses population, the data obtained in the population of Aude do not show a clear signal of selective differences between multilocus genotypes.

**Fine-scale effects**

Although we expected reduced estimation power, samples obtained from each subpopulation were used to compute \(N_e\) estimates for each subpopulation. We found that over nine computations, six returned finite confidence intervals. This means that in most cases, the variance in allele frequencies clearly exceeded the variance because of sampling error. Furthermore in that case even if the estimated allele frequencies are rather inaccurate (because of small sample sizes) the error in the \(N_e\) estimation is low as the relation between \(F_c\) and \(N_e\) is hyperbolic. Finally, the likelihood method performs poorly in some instance, for example in Aude1 between 1993 and 1996. The flatness of the likelihood curve in this particular case indicates a poor convergence of the method (or that there is little information in the data for estimating the parameter of interest). The confidence intervals obtained with this method are broader than with the moment-based method which is an unusual result (Anderson et al., 2000; Berthier et al., 2002; Anderson, 2005). The reasons for this are unclear but may be related to the violation of some assumptions.

Considering only estimates associated with finite confidence intervals we found that estimates range from 6 to 30 among the three subpopulations. Although confidence intervals around \(N_e\) estimates were overlapping, and despite similar demographic population size, our results point to lower effective size in Aude2 compared with Aude1, indeed \(F_c\) values observed for each locus in Aude2 were clearly larger than in Aude1 (data not shown). This difference can be related to differences in subpopulation structure and demography. Actually and as already shown by Bonnin et al. (2001), subpopulation 2 is itself subdivided, exhibiting a patchy spatial distribution of different inbred lines. In contrast, Aude1 displayed a very different spatial structure and the dominant inbred line observed in this population in 1990 was maintained at high frequency through generations (congruent with a low variation in allele frequencies). These two contrasting dynamics (Aude1 vs. Aude2) can potentially be explained by environmental differences between these two subpopulations. Bonnin et al. (2001) already noted that Aude1 is bordering a vineyard and is ploughed every year which probably facilitates dispersal of seeds and maintains an open environment. In contrast, Aude2 is located in a less disturbed environment which is progressively invaded by vegetation. Therefore, competition with other plant species is probably stronger and there are consequently fewer sites available for colonization.

**Comparison with other species and concluding remarks**

Despite the importance of the concept of effective population size for both evolutionary and conservation issues, very few estimates based on genetic data have been obtained on plant species. This is even truer in selfing species; to our knowledge our study is the first reporting \(N_e\) estimates in self-fertilizing plant natural populations. Published studies in plant species however suggest \(N_e/N\) ratios of about 5–10\%. For instance, in *Eichhornia paniculata*, among 10 outcrossing populations examined, effective sizes varied from 3.4 to 70.6 and on average were about 11\% of the census number of reproductive individuals (Husband & Barrett, 1992). For wheat, which reproduces almost exclusively through selfing, \(N_e\) values in an experimental composite population were around 150–200 for a demographic population size artificially maintained at 5000 individuals \((N_e/N \approx 3–4\%)\,\text{Goldringer et al., 2001} .\) Results obtained in *M. truncatula* are close to these values. Schoen & Brown (1991) reported estimates of \(N_e\) for 17 plant species with contrasting mating systems using data on allelic richness from the literature. The values they obtained were considerably higher than in our study with mean \(N_e\) for inbreeding and outbreeding species being 3557 and 6990 respectively. It is difficult to compare these values with those obtained in *M. truncatula* because estimates of \(N\) were not available for the populations surveyed and the method used integrates historical processes. Interestingly, in our study, there was a discrepancy between \(N_e\) estimates and the level of diversity (as measured through either \(H_e\) or the allelic number). Indeed the positive association one should intuitively expect between these two parameters was not observed in the Aude dataset. This suggests that \(H_e\) is a poor estimator of the genetic drift currently at work within population. This may stem from the fact that \(H_e\) is determined by the long-term \(N_e\) through the population mutation rate \((\theta=4N_e\mu)\). We advocate here that the use of the temporal method is
suitable in conservation studies as it gives insight into the level of genetic drift actually experienced by the population over the chosen time frame. From a methodological perspective, our study shows that for highly selfing species, considering both monolocus and multilocus information yields deeper information about allele frequencies variation and provides a better understanding of the underlying processes at work. It would be of great interest to understand what causes fluctuations in inbred line frequencies in natural populations. It is likely that a combination of neutral, selective and environmental processes is involved. More work is needed to clarify these points, in particular the significance of seed bank and the existence of fitness variation between inbred lines.

Acknowledgments

The authors wish to thank T. Bataillon and A. Tsitrone for comments on the manuscript and helpful discussions, and the two anonymous reviewers for their useful comments on the first version of the manuscript. Many thanks to D. Tauzin, A. Weber, C. Poncet and S. Roussel for technical support, and to S. De Mita for programming help. This is publication ISEM 2007–52 of the Institut des Sciences de l’Evolution.

References


**Appendix**

We use the same simulation procedure as in Goldringer & Bataillon (2004). An exact multinomial sampling scheme allows to simulate genetic drift provided a vector of initial allele frequencies, an effective population size value and two sample size values and a number of generation between the two samples. A total of 3000 independent replicates allow obtaining distributions of $F_c$ values. A first step is to check if the simulated value of $N_e$ based on the simulations is accurate. Then the second step is to find the bounds of the confidence interval. The bounds are not taken from the preceding distribution of simulated $F_c$ values. Instead we look for the smallest (respectively largest) value of $N_e$ simulated for which the true (‘parametric’) value of $N_e$ is contained in the upper (respectively lower) tail of the distribution at the 2.5% level for example (in the case of a 95% confidence interval).

**Supplementary material**

The following supplementary material is available for this article:

**Table S1** Primers and characteristics of the microsatellite loci.

This material is available as part of the online article from http://www.blackwell-synergy.com/doi/abs/10.1111/j.1420–9101.2007.01409.x

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.

Received 1 May 2007; revised 12 June 2007; accepted 14 June 2007