npg

ORIGINAL ARTICLE

How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*

M Siol[1], JM Prosperi[1], I Bonnin[2] and J Ronfort[1]

[1]UMR 1097 *Diversité et Adaptations des Plantes Cultivées, INRA Montpellier, Domaine de Melgueil, Mauguio, France* and [2]UMR *(INRA-CNRS-UPS-INAPG) de Génétique Végétale, Ferme du Moulon, Gif/Yvette, France*

The occurrence of populations exhibiting high genetic diversity in predominantly selfing species remains a puzzling question, since under regular selfing genetic diversity is expected to be depleted at a faster rate than under outcrossing. Fine-scale population genetics approaches may help to answer this question. Here we study a natural population of the legume *Medicago truncatula* in which both the fine-scale spatial structure and the selfing rate are characterized using three different methods. Selfing rate estimates were very high (∼99%) irrespective of the method used. A clear pattern of isolation by distance reflecting small seed dispersal distances was detected. Combining genotypic data over loci, we could define 34 multilocus genotypes. Among those, six highly inbred genotypes (lines) represented more than 75% of the individuals studied and harboured all the allelic variation present in the population. We also detected a large set of multilocus genotypes resembling recombinant inbred lines between the most frequent lines occurring in the population. This finding illustrates the importance of rare recombination in redistributing available allelic diversity into new genotypic combinations. This study shows how multilocus and fine-scale spatial analyses may help to understand the population history of self-fertilizing species, especially to make inferences about the relative role of foundation/migration and recombination events in such populations.
Heredity (2008) **100,** 517–525; doi:10.1038/hdy.2008.5; published online 20 February 2008

Introduction

Understanding how mating system affects plant genetic diversity is a major theme in evolutionary genetics. Selfing is known to reduce within-population diversity at equilibrium by a factor $(2-S)/2$, where S is the selfing rate (Pollak, 1987; Nordborg and Donnelly, 1997). Because inbreeding also reduces the efficiency of recombination (Nordborg, 2000), a further reduction in neutral polymorphism is expected from hitchhiking associated with selective sweeps (Maynard Smith and Haigh, 1974; Barton, 2000) and from background selection against deleterious mutations (Charlesworth *et al.*, 1993). Self-fertilization may also increase between-population differentiation due to reduced pollen dispersal— one of the two forms of gene flow among plant populations. Consistently, empirical surveys of allozyme and molecular variation among species have shown that high levels of self-fertilization are associated with lower within-population diversity and higher between-populations differentiation, as compared to outcrossing species (Hamrick and Godt, 1990, 1996; Schoen and Brown, 1991;

Charlesworth and Yang, 1998; Vitalis *et al.*, 2002; Meunier *et al.*, 2004; see also Charbonnel *et al.*, 2005). However, as noted by Schoen and Brown (1991) in a seminal paper, there is a higher population-to-population variation in the level of diversity in preferentially inbreeding species than in outcrossing ones. This means that although the mean level of genetic diversity is on average lower in inbreeders, some populations exhibit high genetic diversity. Since then, this result has been confirmed, for example, in several studies on *Arabidopsis thaliana* (Bergelson *et al.*, 1998; Kuittinen *et al.*, 2002; Jorgensen and Mauricio, 2004; Bakker *et al.*, 2006) in which some populations exhibited a rather high polymorphism whereas other populations were monomorphic (composed of a single multilocus homozygous genotype). This particular feature is not fully understood and no satisfying explanations have been given to date. It has been proposed that inbreeding species should be composed of ancient populations exhibiting high levels of polymorphism and of recent marginal populations derived from small samples of individuals from these large 'source' populations (Schoen and Brown, 1991). However this explanation depends critically on the dispersive abilities of the species and other explanations can be proposed. Variation in the level of selfing experienced by different populations of the same species may also explain the different levels of genetic diversity observed between populations. Finally, population subdivision is often invoked to explain the maintenance of

genetic diversity. Under rather simplistic assumptions it can be shown that population subdivision reduces genetic drift at the global scale compared to an unsubdivided population and therefore allows maintaining higher amounts of genetic diversity at the population level (Wang and Caballero, 1999).

In the present study, the mechanisms underlying the maintenance of large levels of genetic variability in self-fertilizing populations are investigated in a natural population of the selfing annual plant species, *Medicago truncatula* (Lesins and Lesins, 1979). *M. truncatula* is mainly studied as a model organism for the legume–*Rhizobium* symbiosis (Cook, 1999). It is widespread all around the Mediterranean basin and can be considered an opportunistic species, common in open areas. It is reported as a highly selfing species (Lesins and Lesins, 1979; Bataillon and Ronfort, 2006). Previous population genetic analyses in this species have shown very high F_{IS} values and high within-population structure. Both findings are consistent with a high selfing rate (Bonnin *et al.*, 1996, 2001; Bataillon and Ronfort, 2006). Indeed, under such high selfing rates, populations are expected to be composed of a number of nearly independent fully homozygous sibships descended from founder individuals or from newly migrants. In order to disentangle the relative contribution of population substructure (Walhund effect) and selfing to the high observed F_{IS} values it is necessary to gather information on the spatial distribution of individuals together with relatedness information. To date, no direct assessment of selfing rates have been conducted in this species and all estimations have been obtained through the expected relationship between heterozygote deficiency and selfing rate at inbreeding equilibrium. However, the opportunistic status of this species and the ephemeral nature of its populations should violate the equilibrium assumption and bias selfing rates estimated from F_{IS} (Bataillon and Ronfort, 2006). It is thus important to get new estimate of selfing rates that do not assume that populations are at inbreeding equilibrium. Following this goal, we report here an analysis of the mating system and of the fine-scale structure of a population of *M. truncatula* that exhibits large levels of within-population molecular variation. The mating system is investigated using progeny arrays and selfing rate estimates are compared to indirect measures based on deviation from Hardy–Weinberg genotype frequencies (F_{IS}). The study of the multilocus composition of the population allowed us to characterize the spatial distribution of inbred sibships and to evaluate the possible impact of migration and recombination on population genetic diversity.

Materials and methods

Sampling design
Seeds were collected in summer 1999 on a young fallow (about 10 years old), located near Perpignan (Southern France; 02°55′N, 42°50′E). The choice of this population was motivated by its size (approximately 80×50 m) and the large number of pods available. Pods were sampled at corners of a 50×50 cm² space. We sampled two pods per square, for a total of 100 mapped squares. Pods were threshed and the extracted seeds were sown in a greenhouse in order to obtain two different data sets: (1) a first sample (hereafter referred to as Sample1) obtained using one seed per pod (total of 200 seeds) was sown to study the genetic structure of the population, (2) among the 200 pods, 160 were randomly chosen for progeny array analyses: for that purpose, two additional seeds from each of these pods were sown, for a total of 180 individuals (that is, 60 sibships of three half-sib or full-sib individuals, hereafter referred as Sample2).

Microsatellites analyses
Total DNA was extracted according to Tai and Tanksley (1990) from 2 g of young leaves previously frozen in liquid nitrogen. Polymorphism was assayed on the first sample (200 individuals) at seven loci previously developed by Baquerizot-Audiot *et al.* (2001): MAA660456, MTR58, MTSA6, MTSA5 MTPG85C, MAA660749 and MTR52. For Sample2, only five of these loci (MAA660456, MTSA5, MTSA6, MTR58 and MTPG85C) were used. As detailed in Table 1, this set of microsatellite loci includes dinucleotide, trinucleotide and composed repeat loci. These loci are relatively well dispersed over the eight linkage groups of *M. truncatula* (Table 1). Amplification reactions were performed in a final volume of 20 μl as described in Baquerizot-Audiot *et al.* (2001). PCR products were loaded on 6% denaturing polyacrylamide gels and revealed through classical silver staining. Allele sizes were determined with M13 as a size control sequence.

Enzyme electrophoresis
Enzymatic polymorphism was assayed on Sample1. Extracts were prepared from 80–100 mg of young leaves. Samples were ground in 1 M Tris-HCl buffer pH 7.2.

Table 1 Microsatellite loci used for the analysis and summary of microsatellite data: observed number of alleles per locus (*A*, with range of allele size), unbiased expected heterozygosity (H_E), observed heterozygosity (H_O) and F_{IS} values

Locus	Repeat motif	LG	A (size range in bp)	H_E	H_O	F_{IS}
MTSA5	$(TC)_9$	7	5 (239–247)	0.629	0.005	0.992
MTR52	$(AG)_{11}$	5	4 (58–67)	0.634	0.025	0.961
MTR58	$(TTG)_5(AG)_{12}$	1	4 (157–188)	0.561	0.010	0.982
MTSA6	$(CT)_{16}$	5	3 (141–163)	0.568	0.025	0.956
MAA660743	$(GAA)_{11}$	?	2 (266–272)	0.227	0.000	1.000
MTPG85C	$(TC)_{11}(AC)_8$	?	2 (135–139)	0.501	0.015	0.970
MAA660456	$(TTC)_8$	4	2 (102–120)	0.077	0.000	1.000
	Overall value (s.d.)		3.14	0.457 (0.217)	0.011 (0.011)	0.973 (0.003)

Abbreviation: LG, linkage group.

Filter paper wicks (Whatman) were dipped into supernatant of centrifuged samples and these were inserted into 12% horizontal starch gel prepared with the appropriate buffer. Seven enzymatic systems (Wendel and Weeden, 1989b) were studied using Tris-citrate gel buffer pH 7 and lithium-borate buffer pH 8.3. Shikimate dehydrogenase (SDH, E.C. 1.1.1.25), 6-phosphogluconate dehydrogenase (PGD, E.C. 1.1.1.44) and phosphoglucomutase (PGM, E.C. 2.7.5.1) isozymes were resolved with TC system. Glutamate oxaloacetate transaminase (GOT, E.C. 2.6.1.1), endopeptidase (ENP, E.C. 3.4.-.-), leucine amino peptidase (LAP, E.C. 3.4.11.1) and mannose phosphate isomerase (MPI, E.C. 5.3.1.8) isozymes were resolved using LB buffer. Recipes for electrophoresis and staining procedures were adapted with minor modifications from Wendel and Weeden (1989a).

Statistical analysis

Genetic diversity: For each locus, the number of alleles and gene diversity H_E were estimated (Nei, 1987). Both values were compared between allozymes and microsatellites using Mann–Whitney's U-test with loci as replicates. For each locus, departure from Hardy–Weinberg expectations was tested through permutations of alleles among individuals and Wright's F-statistics F_{IS} estimated according to Weir and Cockerham (1984) using GENETIX version 4.05 (Belkhir et al., 1996–2004). Genotypic linkage disequilibrium was measured for each pair of loci and tested through Fisher's exact test using GENEPOP version 3.2 (Raymond and Rousset, 1995) and applying sequential Bonferroni-type corrections to test for significance.

For each individual, a multilocus genotype was defined combining the genotypic information of the polymorphic loci. Multilocus diversity was then measured using the Simpson index model corrected for finite sample size (Pielou, 1969):

$$SI = 1 - \left[\sum_i \frac{n_i(n_i - 1)}{N(N - 1)} \right]$$

where n_i denotes the number of individuals with multilocus genotype i and N the total sample size. The term in brackets reflects the probability that two randomly chosen individuals are identical (Nei, 1987).

Population structure

Spatial autocorrelation analyses were used to describe the spatial organization of the genetic diversity in this population, using Spagedi version 1.2 (Hardy and Vekemans, 1999, 2002). This software estimates conditional kinship coefficients between individuals as a function of their spatial distance. Kinship coefficients were calculated as described in Loiselle et al. (1995) using each individual multilocus genotype. To estimate the level of within-population subdivision, we divided the sampled area according to a grid of decreasing mesh size. Seven grids were defined, dividing the population in 4 (2×2) to 25 (5×5) squares. The genetic differentiation between the subpopulations thus defined was estimated by the overall F_{ST} values among subpopulations. F_{ST} estimates were computed according to Weir and Cockerham (1984) and their significance was tested using 1000 permutations of individuals within the population using GENETIX (Belkhir et al., 1996–2004).

Mating system analyses

Three different methods were used to assess the mean selfing rate (S) of this population. S was first inferred using the commonly used relationship between Wright's within-population inbreeding coefficient and the selfing rate, $F_{IS} = S/(2 - S)$. This relation assumes that the selfing rate has been constant for a sufficient number of generations, that the population is at inbreeding equilibrium, and that selfing is the major cause of departure from Hardy–Weinberg genotypic frequencies (no spatial structure, no fitness difference between selfing and outcrossing progenies). A jacknife over loci was used to obtain confidence intervals for S. As a second measure of S, we used the maximum likelihood estimator developed by Enjalbert and David (2000). This method uses multilocus individual heterozygosity and provides selfing rate estimates and confidence intervals for the two or three last generations assuming no selection, no allelic frequency changes among generations, linkage equilibrium between loci and that outcrossing gametes meet at random.

Finally and in order to disentangle selfing effects and population structure, we applied progeny arrays analysis to the sib family dataset (Sample2). For this purpose, we used the MLTR software, version 0.9 (Ritland, 2002) to obtain maximum likelihood estimates of single (t_s) and multilocus (t_m) outcrossing rates. This method relaxes the assumption of hypothesis of inbreeding equilibrium and differences between t_s and t_m allows inferring the amount of inbreeding due to mating between relatives vs selfing (Ritland and Jain, 1981). The MLTR program also estimates the variance in selfing rates between maternal individuals and correlation between mating parameters, namely the correlation of selfing within progeny arrays. A lack of correlation of selfing ($r_s = 0$) indicates that the selfing rate does not vary among families, whereas a correlation suggests that sibships are either all selfed or all outcrossed sibs. Between the two likelihood optimization algorithms available in MLTR, namely the Expectation-Maximization (EM) method and the Newton–Raphton (NR) method, we chose the EM method because it is more suitable for highly inbred species (Ritland, 1986). In all our computations the maximum number of iterations was used and sampling variance estimates were obtained for each parameter using 1000 bootstraps.

Results

Monolocus genetic diversity

The number of alleles and the level of gene diversity observed for each microsatellite locus and each allozyme marker are reported in Tables 1 and 2, respectively. Among the seven microsatellite loci, five showed relatively high levels of diversity, displaying between two and five alleles each, and gene diversities as measured by H_E ranging from 0.5 to 0.63. The two remaining loci were less polymorphic with two alleles segregating at unbalanced frequencies. Among the eight enzymatic loci assayed, seven were polymorphic. All loci, except ENP, displayed two alleles and H_E values lower than 0.5. Both the mean number of alleles and the mean gene diversity were lower for allozymes compared to microsatellites but these differences were not significant (Mann–Whitney's U-tests, $P > 0.10$ for both tests).

A significant departure from Hardy–Weinberg equilibrium was detected for all the loci studied ($P<0.0001$). The mean F_{IS} value based on the whole set of polymorphic loci was 0.978 (s.d. $= 0.006$) with a slightly (but not significantly) lower value estimated using only microsatellite markers (Table 1) as compared to the average value obtained with allozymes (Table 2).

Spatial autocorrelation analyses using Loiselle's coefficient (Loiselle et al., 1995; Hardy and Vekemans, 2002) as a measure of the genetic relatedness between individuals indicated a significant genetic relationship between individuals located up to 8 m apart (Figure 1). Similar autocorrelograms were obtained when using microsatellite loci or allozymes only (data not shown). When subdividing arbitrarily the population into squared units, F_{ST} values as large as 0.35–0.30 were obtained especially when the number of subdivisions used was large. When increasing the size of the subdivisions, the level of differentiation progressively decreased (Figure 2). For instance, when the population was subdivided in four squares of $25 \times 25\,m^2$, the variation among subpopulations accounted for approximately 15% of the overall population variation (Figure 2).

Multilocus patterns of diversity

Over the 91 tests performed for linkage disequilibrium, 81 were found significant at the 5% level. When applying Bonferroni correction, 74 tests were still significant. Among those, 20 were pairs of microsatellite loci

(for 21 comparisons), 14 pairs of allozymes (21 comparisons) and 40 concerned pairs combining an allozyme marker and a microsatellite locus (for 49 tests). Combining the genotypic data of the different microsatellite markers (respectively, allozyme markers) allowed distinguishing 26 multilocus genotypes (respectively, 23). Combining both types of markers yielded 34 multilocus genotypes, among which 9 showed one or more heterozygozous loci (Figure 3.). The relative frequency of these different genotypes was highly variable (Figure 3), with 22 genotypes observed only once, whereas 4 multilocus genotypes represented 76% of the sample analysed (152 individuals in Sample1). This pattern resulted in a Simpson index (computed over the whole set of markers) of 0.805. Similar results were obtained when considering either only allozyme ($SI = 0.76$) or microsatellite markers ($SI = 0.74$) to define the multilocus genotypes. Interestingly, among the 34 multilocus genotypes detected, 6 were sufficient to account for the total allelic variation of the population (Figure 3 and see genotypes a7, b19, c14, d5, f22 and j4 in Supplementary Table S1). A closer look at the genotypic composition of the remaining multilocus genotypes

Table 2 Allele number (A), observed (H_O) and expected (H_E) heterozygosity and F_{IS} values observed with eight allozyme loci

Locus	A	H_E	H_O	F_{IS}
pgd2	2	0.500	0.015	0.970
pgm	2	0.227	0	1
sdh	2	0.212	0	1
enp	3	0.575	0	1
lap1	2	0.077	0	1
lap2	1	0	0	—
got2	2	0.049	0	1
mpi	2	0.434	0.025	0.981
Overall value (s.d.)	2	0.259 (0.219)	0.005 (0.096)	0.981 (0.006)

$F_{IS} = 0.981$ s.d. $= 5.5 \times 10^{-3}$.

Figure 2 Variation of F_{ST} values calculated over all loci for different spatial subdivisions. Black circles give estimated F_{ST} values; thin lines show the 95% limits of the distribution under the null hypothesis of no differentiation, obtained after 100 permutations of the multilocus genotypes.
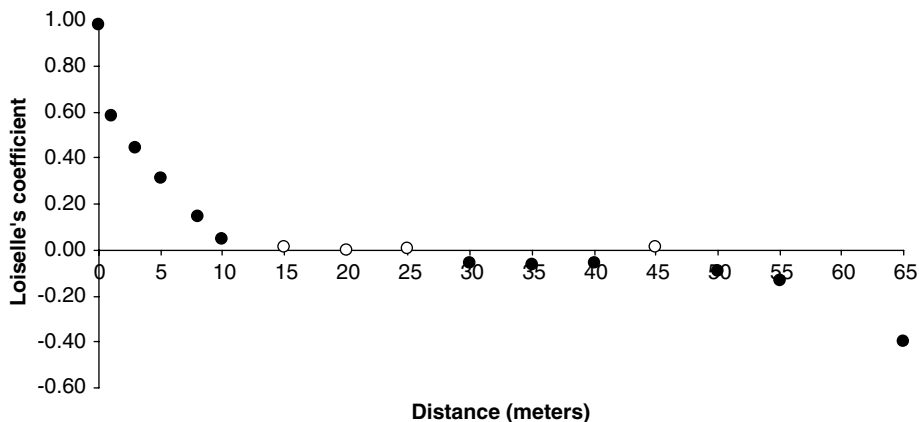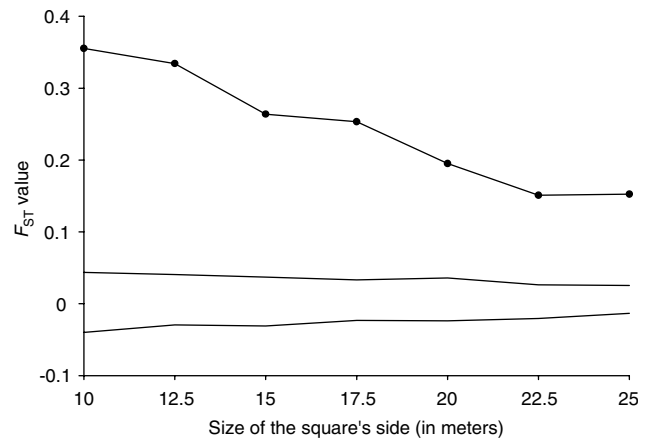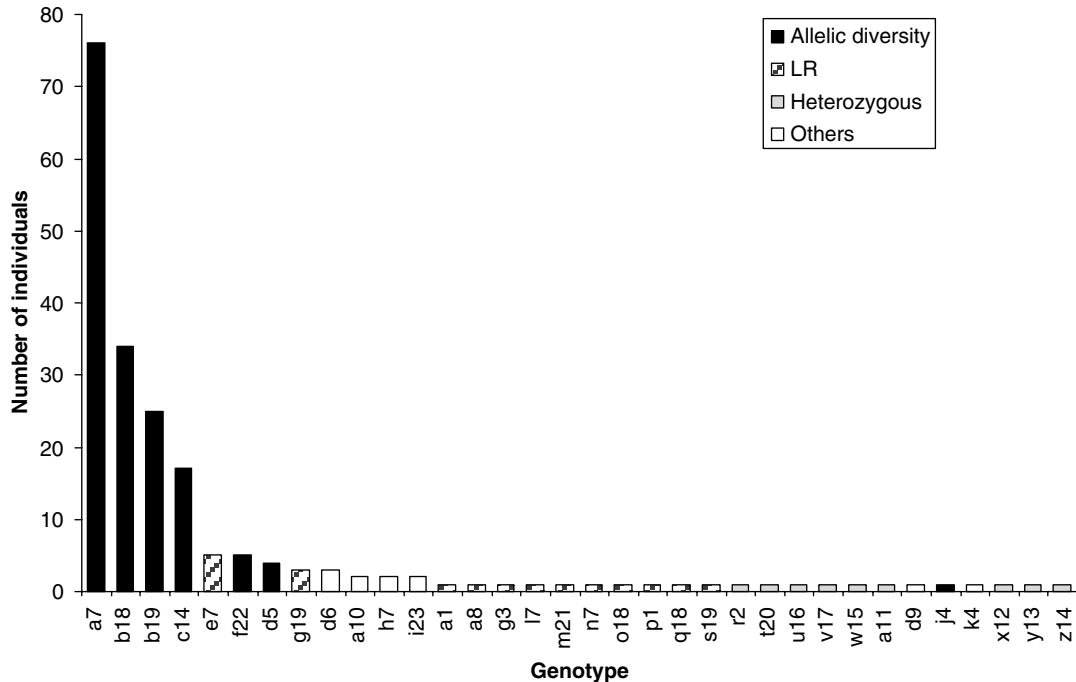

Figure 1 Results from the spatial autocorrelation analyses. The value of each point on the y axis represents the mean coefficient of relatedness (here measured following Loiselle et al., 1995) between individuals located x metres apart ($\bullet$): $P<0.05$; ($\circ$): $P>0.05$.

Figure 3 Frequency distribution of the 34 multilocus genotypes. In black are the 6 genotypes that were sufficient to explain the total allelic variation. Hatched bars represent 'recombinant lines' between the most frequent genotypes (see text). Black bars refer to genotypes with at least one heterozygous locus. Grey bars indicate genotypes that do not fit the preceding categories. The name of each genotype is indicated under each bar.

revealed that 13 of them displayed genotypes corresponding to recombinant inbred lines between two of the most frequent genotypes, that is, either a7 × b19 or a7 × b18 (Supplementary Table S1). Also 6 of the multilocus heterozygous genotypes could derive from a cross between a7 and b19 (or b18) followed by several generations of selfing (Supplementary Table S1). Mapping the different multilocus genotypes showed that (1) the dominant genotype (a7) is broadly distributed over the sampled area, (2) the other ones being more or less confined to a particular region of the population. Interestingly, small patches of identical genotypes were located at the edge of the population (see for example genotypes c14 and f22, Figure 4).

Mating system
The selfing rate inferred using Wright's inbreeding coefficient (F_{IS}) estimated over the entire population and using the whole set of markers was 0.989 (s.d. $= 10^{-3}$). Similar values were obtained when using only the five microsatellite loci used for progeny array analyses, that is, $S = 0.987$ (s.d. $= 0.0016$) or reducing the sample to the 60 plants represented in Sample2, where $S = 0.985$ (s.d. $= 5 \times 10^{-4}$). Multilocus estimates were also consistent with these values. Indeed, the algorithm developed by Enjalbert and David (2000) concluded to a selfing rate of 0.987 into the last generation and of 0.989 for the preceding couple of generations (these two values were not significantly different from one another). Progeny array analyses also yielded high selfing rates. Among the 60 families studied, 58 were composed of identical and homozygous genotypes, as expected following a selfing event on an inbred line. Overall,

outcrossing rate estimates obtained using MLT were exceedingly small ($t_m = 0.006$ (s.d. $= 0.005$) and $t_m = 0.017$ (s.d. $= 0.017$)), whatever the assumption made concerning the genotype of mother plants. Differences between t_m and t_s were always very low (for instance, $t_s = 0.014$ (s.d. $= 0.014$) when $t_m = 0.017$), suggesting that mating between relatives does not enhance the apparent selfing rate.

Discussion

Mating system
In this study, we report the first estimate of outcrossing rate using maternal progenies and likelihood methods in *M. truncatula*. Previous estimates were based on F_{IS} values, and thus assumed (1) that the population under study was at equilibrium for a fixed selfing rate and (2) that no subdivision occurred within population (Bonnin *et al.*, 2001; Bataillon and Ronfort, 2006). To estimate the bias due to deviation from these assumptions, we used two independent datasets: one devoted to the estimation of the selfing rate (that is, progeny arrays), the other documenting the fine scale spatial structure of the population and allowing two indirect measures of the selfing rate based on patterns of individual heterozygosity. Maternal progeny analyses confirmed the selfing status of *M. truncatula*, yielding a mean selfing rate of approximately 99%. Although we detected a clear pattern of within-population structure, which was expected to upwardly bias indirect estimates of the selfing rate, progeny array analyses yielded similar selfing rate estimates as indirect measures. This result is however consistent with the relationship linking F-statistics (Wright,
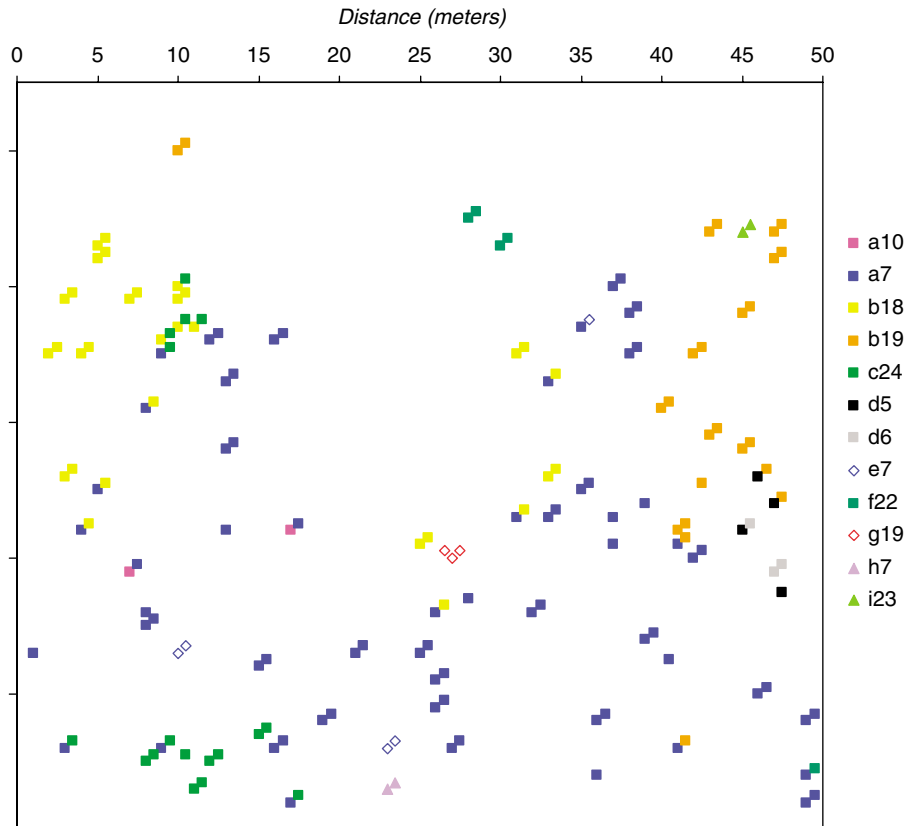
Figure 4 Map showing the location of the individuals with the more common multilocus genotypes (combining microsatellite and allozyme loci). Individuals with unique genotypes have been omitted for the sake of clarity.

1969), that is, $(1-F_{IS}) = (1-F_{IS}^{selfing}) \times (1-F_{IS}^{subdivision})$ where $F_{IS}^{selfing}$ and $F_{IS}^{subdivision}$ are departure from Hardy–Weinberg frequencies due to selfing and population subdivision (that is, Wahlund effect), respectively. From this formula, it appears that under high selfing rates (such as the one detected in this population), population subdivision will only have a reduced effect on the global F_{IS} value.

Progeny array analyses also concluded to very low levels of mating between relatives. This result should however be considered with caution. Indeed, given the particular multilocus composition of the population (with a small set of dominant inbred lines), outcrossing events involving sister lines cannot be distinguished from selfing events. This means that the proportion of outcrossing events in the studied population is probably slightly higher than estimated. A closer look at progeny arrays showed that outcrossing events were not randomly distributed among families but rather restricted to two progenies, suggesting that outcrossing results from the fertilization of a limited number of flowers. Nevertheless, in our sample, each family originated from a single pod collected on the ground. It is thus not possible to know whether outcrossing events are concentrated on a small set of plants or they are randomly distributed over the population. Further studies involving several pods per plants are needed in order to clarify this issue.

Gene diversity and population structure
Despite the large self-fertilization rate estimated in this population, the polymorphism revealed was relatively

high, especially when measured through Nei's index of gene diversity (H_E). Indeed, both microsatellite markers and allozyme loci displayed gene diversity approximately two times larger compared to mean H_E values reported for allozymes markers in self-fertilizing species (Hamrick and Godt, 1990; Schoen and Brown, 1991). Due to higher mutation rates (Jarne and Lagoda, 1996; Goldstein and Schlötterer, 1998), we expected larger polymorphism with microsatellites than with allozymes. But, although both the number of alleles per loci (A) and gene diversity (H_E) were on average larger for microsatellites than for allozymes, these differences were not significant, in contrast to other studies (Estoup *et al.*, 1995; Streiff *et al.*, 1998; Freville *et al.*, 2001). For the number of alleles, this result may be due to the large variance observed for microsatellite loci ($\sigma^2_{(A)} = 1.48$) compared to allozymes ($\sigma^2_{(A)} = 0.29$), which could reflect large differences in mutation rates among loci. For gene diversity, however, the variance among loci was large for both allozymes and microsatellites. This reflects the reduced number of multilocus genotypes occurring in the Salses population, and the fact that the most common genotypes are highly genetically differentiated. This also suggests that H_E values measured in this population may reflect a short time period (that begins at population foundation), probably too short to see the effect of different mutation rates.

Previous population studies of *M. truncatula* have shown that in this species, there is an important population-to-population variation in gene diversity (Bonnin *et al.*, 2001; Bataillon and Ronfort, 2006). This

observation is consistent with results found in other selfing species, (Schoen and Brown, 1991; Green *et al.*, 2001; Bakker *et al.*, 2006; see also Ramakrishnan *et al.*, 2006). The factors and mechanisms responsible for the high level of genetic diversity maintained in such populations remain unclear. Following different foundation events, a self-fertilizing population is likely to be subdivided into small neighbourhoods that consist of single differentiated lineages. Such subdivision is expected to reduce the effect of drift and could thus play a major role in the maintenance of genetic diversity at the whole population level (Barton and Whitlock, 1997). Another explanation could be that more variable populations display higher outcrossing rates than classically thought in this species (Jorgensen and Mauricio, 2004). For the population of Salses, our study clearly showed that large levels of genetic variation can be observed at the population level despite a very high selfing rate. Combining the different loci studied, we could show that most of the allelic variation observed in this population resulted from the co-occurrence of a limited number of highly differentiated inbred lines. Three of these lines were relatively dispersed over the populations and are thus probably the initial founders of this population. The remaining inbred lines were less common and represented as small patches generally confined around the edge of the population. These lines are thus likely to result from recent migration events followed by a few generations of reproduction through self-fertilization. All together, our results suggest that the allelic variation revealed in this population mostly results from different founding and migration events, and that the high levels of gene diversity observed seems to be due to the maintenance in relatively high frequency of the different founders (or recent immigrants). The patchy spatial organization detected in this population could then explain the persistence of these different inbred lines (and the corresponding allelic variation). Selection could act at the microhabitat level, favouring different genotypes in different parts of the site. Even without selection, reduced pollen dispersal among subpopulations should lead to the maintenance of large genetic variation at the whole population level because genetic drift occurs at the subpopulation level, allowing different alleles to be maintained in each deme (Barton and Whitlock, 1997). As a conclusion, we thus suggest that the history of foundation of a population and the spatial structure plays an important role in the maintenance of allelic variation and gene diversity at the population level. However, other factors like higher outcrossing rates or dispersal in time *via* the seed bank are possible additional sources of variation in other populations of *M. truncatula* or in other self-fertilizing plant and animal species (Bonnin *et al.*, 2001; Chauvet *et al.*, 2004; Charbonnel *et al.*, 2005).

Rare but observable recombination events
If a reduced set of well-represented and highly differentiated inbred lines accounted for approximately 80% of the population, we also revealed a large set of rare multilocus genotypes, most of them being detected only once. Interestingly, most of these unique genotypes can be seen as naturally occurring recombinant inbred lines, deriving from the segregation under self-fertilization of outcrossing events between the most frequent lines. Recombinant genotypes have already been observed in other predominantly self-fertilizing species (see for example Ramakrishnan *et al.*, 2006). Such observations suggest that although rare, pollen-mediated gene flow in this self-fertilizing species might play a major role in the organization of the genetic variation both within and among populations. This also raises the question of the possible role of outcrossing/recombination with regard to natural selection and adaptation in self-fertilizing species. Natural selection could favour outcrossed offsprings following two different mechanisms. First, hybrid progenies from crosses between differentiated inbred lines could display a better fitness compared to their parental lines because heterozygosity at individual locus may hide recessive and partially recessive deleterious mutations (Falconer and Mackay, 1996). Second, outcrossing events followed by repeated self-fertilizations should result in a set of recombinant inbred lines displaying a large panel of new allelic combinations among loci. Some of these new combinations could result in higher fitness value or in new adaptations to environmental conditions. In order to determine if the frequency of recombinant inbred lines observed in the Salses population was consistent with a neutral model, we ran deterministic simulations to obtain the expected distribution of segregating genotypes derived from the cross between two lines differing for 8 loci (as observed between a7 and b19, see details in the supplementary information S2 online). It shows that, assuming an outcrossing rate of 1% as estimated in Salses and a large effective population size, it is not necessary to invoke selection to explain the observed frequency of recombinant lines. However this model is overly simplifying since it does not consider the effects of drift. A survey of the temporal variation in allele frequencies in this population has shown that genetic drift is not too strong ($N_e \sim 150$, Siol *et al.*, 2007). Further empirical studies are however needed in order to assess the role of recombination in the evolutionary dynamic of self-fertilizing populations.

Concluding remarks and implications for sampling designs and conservation
In summary, our study show that even under very large selfing rates, genetic and genotypic diversity can be high. As could be expected, we observed different more or less related inbred lines located throughout the population. Our results suggest that the maintenance of these lines may result from the peculiar structure and from the colonization history of the population. Finally, outcrossing events, although rare, generate new genotypic combinations that could have an important role in the dynamic of genetic variation in self-fertilizing species. From a more methodological perspective, our study emphasizes that taking into account multilocus information and fine-scale structure may help greatly to understand how evolutionary factors shape genetic diversity in selfing populations. *M. truncatula* is now recognized as the model plant for the genetic and genomic of Legumes (Cook, 1999). As for *A. thaliana*, there is thus a strong interest for its reproductive biology as well as for the natural genetic variation occurring in this species (Ronfort *et al.*, 2006). Due to the reduced level of

diversity expected within population under high selfing rates, collections of naturally occurring variation in selfers are generally composed of few inbred lines per population (1 line per population in many cases). In agreement with recent results in *A. thaliana* (Bakker *et al.*, 2006) and previous population genetic analyses in *M. truncatula* (Bataillon and Ronfort, 2006), the present study indicates that sampling strategies based on a single individual per population or choosing populations at random (for germplasm conservation purposes for example) are likely to miss a large amount of diversity.

Acknowledgements

References

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J (2006). Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Mol Ecol* **15**: 1405–1418.

Baquerizot-Audiot E, Desplanques B, Prosperi JM, Santoni S (2001). Characterization of microsatellite loci in the diploid legume *Medicago truncatula* (barrel medic). *Mol Ecol Notes* **1**: 1–3.

Barton NH (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B* **355**: 1553–1562.

Barton NH, Whitlock MC (1997). The evolution of metapopulations. In: Hanski I and Gilpin ME (eds). *Metapopulation Biology: Ecology, Genetics and Evolution*. Academic Press: San Diego, CA, pp 183–210.

Bataillon T, Ronfort J (2006). Evolutionary and ecological genetics of *Medicago truncatula*. *The Medicago Handbook*. http://www.noble.org/MedicagoHandbook/pdf/EvolutionaryEcologicalGenetics.pdf

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (1996–2004). GENETIX 4.05, logiciel sous Windows TM pour la Génétique des Populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II. Montpellier (France).

Bergelson J, Stahl E, Dudek S, Kreitman M (1998). Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* **148**: 1311–1323.

Bonnin I, Huguet T, Gherardi M, Prosperi JM, Olivieri I (1996). High level of polymorphism and spatial structure in a selfing plant species, *Medicago truncatula* (Leguminosae), shown using RAPD markers. *Am J Bot* **83**: 843–855.

Bonnin I, Ronfort J, Wozniak F, Olivieri I (2001). Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). *Mol Ecol* **10**: 1371–1383.

Charbonnel N, Rasatavonjizay R, Sellin E, Brémond P, Jarne P (2005). The influence of genetic factors and population dynamics on the mating system of the hermaphroditic freshwater snail *Biomphalaria pfeifferi*. *Oikos* **108**: 283–296.

Charlesworth B, Morgan MT, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Charlesworth D, Yang Z (1998). Allozyme diversity in Leavenworthia populations with different inbreeding levels. *Heredity* **81**: 453–461.

Chauvet S, van der Velde M, Imbert E, Guillemin ML, Mayol M, Riba M *et al.* (2004). Past and current gene flow in the selfing, wind-dispersed species *Mycelis muralis* in western Europe. *Mol Ecol* **13**: 1391–1407.

Cook DR (1999). *Medicago truncatula*: a model in the making!. *Curr Opin Plant Biol* **2**: 301–304.

Enjalbert J, David JL (2000). Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics* **156**: 1973–1982.

Estoup A, Garnery L, Solignac M, Cornuet JM (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**: 679–695.

Falconer DS, Mackay TFC (1996). *Introduction to Quantitative Genetics*, 4th edn. Longman: Harlow.

Freville H, Justy F, Olivieri I (2001). Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). *Mol Ecol* **10**: 879–889.

Goldstein DB, Schlötterer C (1998). *Microsatellites. Evolution and Applications*. Oxford University Press: Oxford, GB.

Green JM, Barker JHA, Marshall EJP, Froud-Williams RJ, Peters NCB, Arnold GM *et al.* (2001). Microsatellite analysis of the inbreeding weed Barren Brome (*Anisantha sterilis*) reveals genetic diversity at the within and between-farm scale. *Mol Ecol* **10**: 1035–1045.

Hamrick JL, Godt MJW (1990). Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahter AL and Weir BS (eds). *Plant Population Genetics, Breeding and Genetic Resources*. Sinauer Associates Inc.: Sunderland, MA.

Hamrick JL, Godt MJW (1996). Effect of life history traits on genetic diversity in plant species. *Philos Trans R Soc Lond B* **351**: 1291–1298.

Hardy O, Vekemans X (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**: 618–620.

Hardy OJ, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.

Jarne P, Lagoda P (1996). Microsatellites, from molecules to populations and back. *Trends Ecol Evol* **11**: 424–429.

Jorgensen S, Mauricio R (2004). Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. *Mol Ecol* **13**: 3403–3413.

Kuittinen H, Salguero D, Aguade M (2002). Parallel patterns of sequence variation within and between populations at three loci of *Arabidopsis thaliana*. *Mol Biol Evol* **19**: 2030–2034.

Lesins KA, Lesins I (1979). *Genus Medicago (Leguminosae). A Taxogenetic study*. Dr W Junk b.v. Publishers: The Hague (Netherlands).

Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* **82**: 1420–1425.

Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23–35.

Meunier C, Hurtrez-Bousses S, Durand P, Rondelaud D, Renaud F (2004). Small effective population sizes in a widespread selfing species, *Lymnaea truncatula* (Gastropoda: Pulmonata). *Mol Ecol* **13**: 2535–2543.

Nei M (1987). *Molecular Evolutionary Genetics*. Columbia Univeristy Press: New York.

Nordborg M (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–929.

Nordborg M, Donnelly P (1997). The coalescent process with selfing. *Genetics* **146**: 1185–1195.

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

Pollak E (1987). On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**: 353–360.

Ramakrishnan AP, Meyer SE, Fairbanks DJ, Coleman CE (2006). Ecological significance of microsatellite variation in western

North American populations of *Bromus tectorum. Plant Species Biol* **21**: 61–73.

Raymond M, Rousset F (1995). Genepop (Version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Ritland K (1986). Joint maximum likelihood estimation of genetic and mating structure using open-pollinated progenies. *Biometrics* **42**: 25–43.

Ritland K (2002). Extensions of models for the estimation of mating systems using *n* independent loci. *Heredity* **88**: 221–228.

Ritland K, Jain S (1981). A model for the estimation of outcrossing rate and gene frequencies using *n* independent loci. *Heredity* **47**: 35–52.

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi JM (2006). Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula. BMC Plant Biol* **6**: 28.

Schoen DJ, Brown AH (1991). Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc Natl Acad Sci USA* **88**: 4494–4497.

Siol M, Bonnin I, Olivieri I, Prosperi JM, Ronfort J (2007). Effective population size associated with self-fertilization:

lessons from temporal changes in allele frequencies in the selfing annual Medicago truncatula. *J Evol Biol* **20**: 2349–2360.

Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glössl J, Kremer A (1998). Within population genetic structure in *Quercus robur* L. & *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Mol Ecol* **7**: 317–328.

Tai TH, Tanksley SD (1990). A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. *Plant Mol Biol Rep* **8**: 297–303.

Vitalis R, Riba M, Colas B, Grillas P, Olivieri I (2002). Multilocus genetic structure at contrasted spatial scales of the endangered water fern *Marsilea strigosa* (Marsileaceae, Pteridophyta). *Am J Bot* **89**: 1142–1155.

Wang J, Caballero A (1999). Developments in predicting the effective size of subdivided populations. *Heredity* **82**: 212–226.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Wendel JF, Weeden NF (1989a). *Isozymes in Plant Biology.* Chapman and Hall: London.

Wendel JF, Weeden NF (1989b). Visualization and interpretation of plant isozymes. In: Soltis DE and Soltis PS (eds). *Isozymes in Plant Biology.* Chapman and Hall: London.

Wright S (1969). *Evolution and the Genetics of Populations.* The University of Chicago Press: Chicago and London.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)