Heredity (2019) 123:176–191
https://doi.org/10.1038/s41437-019-0182-6

the geneticssociety

ARTICLE

Structure of multilocus genetic diversity in predominantly selfing populations

Margaux Jullien[1] · Miguel Navascués[2,3] · Joëlle Ronfort[1] · Karine Loridon[1] · Laurène Gay[1]

Received: 11 September 2018 / Revised: 28 December 2018 / Accepted: 8 January 2019 / Published online: 22 January 2019

Abstract

Predominantly selfing populations are expected to have reduced effective population sizes due to nonrandom sampling of gametes, demographic stochasticity (bottlenecks or extinction–recolonization), and large scale hitchhiking (reduced effective recombination). Thus, they are expected to display low genetic diversity, which was confirmed by empirical studies. The structure of genetic diversity in predominantly selfing species is dramatically different from outcrossing ones, with populations often dominated by one or a few multilocus genotypes (MLGs) coexisting with several rare genotypes. Therefore, multilocus diversity indices are relevant to describe diversity in selfing populations. Here, we use simulations to provide analytical expectations for multilocus indices and examine whether selfing alone can be responsible for the high-frequency MLGs persistent through time in the absence of selection. We then examine how combining single and multilocus indices of diversity may be insightful to distinguish the effects of selfing, population size, and more complex demographic events (bottlenecks, migration, admixture, or extinction–recolonization). Finally, we examine how temporal changes in MLG frequencies can be insightful to understand the evolutionary trajectory of a given population. We show that combinations of selfing and small demographic sizes can result in high-frequency MLGs, as observed in natural populations. We also show how different demographic scenarios can be distinguished by the parallel analysis of single and multilocus indices of diversity, and we emphasize the importance of temporal data for the study of predominantly selfing populations. Finally, the comparison of our simulations with empirical data on populations of *Medicago truncatula* confirms the pertinence of our simulation framework.

Introduction

Most angiosperms are hermaphrodite (70%; Yampolsky and Yampolsky 1922). This co-occurrence of both male and female reproductive organs on the same individual allows self-pollination and, in the absence of self-incompatibility mechanisms, self-fertilization. Indeed, about 40% of flowering plants do self at various rates, and about 15% of them reproduce predominantly through selfing, with outcrossing rates lower than 10% (Igic and Kohn 2006). Such high selfing rates are expected to have strong consequences on the genetic diversity of natural populations and on its organization.

Theoretical studies have examined the consequences of selfing for the genetic diversity of populations, particularly in terms of effective population size N_e. The effective population size is defined as the size of an ideal Wright–Fisher population experiencing the same rate of genetic drift as the population under consideration (Crow and Kimura 1970). As reviewed in Charlesworth (2009), the effective population size can be affected by several factors, including the mating system. Self-fertilization reduces the number of independent gametes sampled for reproduction, which directly decreases N_e (Pollak 1987). Demographic events are also likely to affect the effective size of selfing populations where founder effects can be frequent, e.g., through the establishment of a new population by a single

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-019-0182-6) contains supplementary material, which is available to authorized users.

✉ Margaux Jullien
margauxjullien3@gmail.com

[1] AGAP, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

[2] CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Université de Montpellier, Montpellier, France

[3] Institut de Biologie Computationnelle IBC, Montpellier, France

Table 1 Ranges of temporal F_{ST} and estimates of effective population sizes in the literature of predominantly selfing populations

Reference	Species	Number of populations	τ	F_{ST}	$\widehat{N_e}$	H_E
Siol et al. (2007)	*Medicago truncatula*	4	3–6	–	9.8–153[a]	0.18–0.47
Gomaa et al. (2011)	*Arabidopsis thaliana*	9	1–4	0.033–0.264	1–12[a]	0.00–0.28
Frachon et al. (2017)	*Arabidopsis thaliana*	1	8	0.0215	91[b]	–
Bomblies et al. (2010)	*Arabidopsis thaliana*	14	1	0.03–0.129	2–8[b]	0.00–0.32
Lundemo et al. (2009)	*Arabidopsis thaliana*	11	1	0–0.626	1–250[b]	0.00–0.17
Meunier et al. (2004)	*Lymnea truncatula*	5	1–4	0.002–0.47	1–25[b]	0.05–0.49
Trouvé et al. (2005)	*Galba truncatula*	6	1–3	0.005–0.175	1–65[b]	0.00–0.54
Viard et al. (1997)	*Bulinus truncatus*	12	1–6	0.06–1	1–27[b]	0.10–0.75
Barrière and Félix (2007)	*Caenorabditis elegans*	7	2–72[c]	0.159–0.963	1–95[b]	0.00–0.88

The temporal estimate of N_e, $\widehat{N_e}$, is given as the number of diploid individuals. τ stands for the number of generations between the two samplings, H_E is the range of genetic diversity across populations. When not directly computed in the study, $\widehat{N_e}$ is computed from temporal F_{ST} values according to Frachon et al. (2017) (see Methods).

[a]Estimated in the literature using F_C

[b]Estimated from temporal F_{ST}

[c]Using generation time in laboratory conditions of 1 generation every 2 weeks

individual (Baker 1967). In addition, according to the "dead-end hypothesis" (Stebbins 1957), selfing populations are expected to accumulate deleterious mutations (Lynch et al. 1995; Abu Awad et al. 2014), and could lack the genetic diversity required to adapt to changing environmental conditions (Charlesworth and Charlesworth 1995; Lande and Porcher 2015; Abu Awad and Roze 2018). Therefore, we can expect frequent catastrophic demographic events, with strong bottlenecks or even extinctions followed by recolonization accompanied by strong founder effects (Schoen and Brown 1991). Such metapopulation dynamics (Ingvarsson 2002) are expected to further reduce N_e (Pannell and Charlesworth 2000). Finally, selective effects can also reduce the effective size in selfing populations. Indeed, the increase in homozygosity due to selfing (Caballero and Hill 1992) reduces the effective recombination (Golding and Strobeck 1980; Nordborg 2000). The reduction of N_e due to selective sweeps or background selection is thus expected to extend by hitchhiking to larger linked regions of the genome or, in some extreme cases, to the whole genome (Charlesworth 2009). These different effects of the mating system on the effective population size have been summarized by Glémin (2007) as

$$N_e = \frac{\alpha N}{1 + F}, \quad (1)$$

where N is the census size, $F = \sigma/(2 - \sigma)$ is Wright's equilibrium fixation index with a selfing rate σ, and α summarizes the reduction of the effective population size due to demographic effects and hitchhiking ($\alpha \in [0; 1]$). Overall, this formalizes the predominant role of genetic drift in shaping genetic diversity within selfing populations.

Empirical observations confirm these theoretical expectations, notably through estimations of N_e and genetic diversity in natural predominantly selfing populations. The contemporary effective size of a population can be estimated based on the temporal changes in allele frequency (F_C, Waples 1989) as was done by Siol et al. (2007) and Gomaa et al. (2011) (Table 1). More recently, Frachon et al. (2017) extended the original method to use the temporal differentiation (F_{ST} instead of F_C). When we apply this method to published temporal F_{ST} values, we find that N_e estimates in predominantly selfing populations are generally lower than 100 (Table 1). For comparison, Palstra and Ruzzante (2008) reviewed temporal estimates of N_e in 83 studies concerning different taxa (including the aforementioned selfing species) and found a median N_e of 260. Published estimates of N_e (or temporal F_{ST}) therefore support theoretical predictions of a reduced effective population size in predominantly selfing populations compared to outcrossing ones. Reviews on allozyme data (Schoen and Brown 1991; Hamrick and Godt 1997), as well as on sequence polymorphism (Glémin et al. 2006), showed convincing evidence for lower genetic diversity (as measured by H_E) in predominantly selfing populations compared to outcrossing ones. Yet, Schoen and Brown (1991) also reported larger variability in levels of genetic diversity among selfing populations than among outcrossing populations. Empirical estimates for genetic diversity reported in the temporal studies we reviewed for estimates of effective size are consistent with these findings (Table 1), with some monomorphic populations ($H_E = 0$) along with highly diverse populations (up to $H_E = 0.88$). Substantial genetic diversity can therefore persist in some predominantly selfing populations, suggesting that evolutionary forces other

than genetic drift may play a significant role in shaping the genetic diversity of natural populations reproducing predominantly through selfing.

Besides the level of genetic diversity, the within-population genetic structure is also affected by selfing. Indeed, due to reduced effective recombination (Nordborg 2000), we expect populations to be organized in homozygous lineages, where some multilocus genotypes (thereafter called MLGs) can reach a high frequency (Hartfield et al. 2017). Empirical studies confirm this expectation, for example in the nearly obligate selfing species *Lobelia inflata* where substantial genetic differentiation was found between completely homozygous lineages co-occurring within populations (Hughes and Simons 2015). Similar population genetic structures have also been observed in several other predominantly selfing plant or animal species (e.g., Barrière and Félix 2007; Montesinos et al. 2009; Siol et al. 2008). Because this multilocus genetic structure is specific to predominantly selfing populations, we believe that the comparison of single and multilocus indices of diversity can be relevant to separate the effects of selfing and genetic drift due to small population sizes or demographic processes such as population size changes or migration.

Along with the effect of selfing, Allard (1975) interpreted this distinctive genetic structure of diversity in repeated genotypes as a result of selection favoring locally adapted MLGs, which can then reach high frequencies in the population. The reduction in gene flow through pollen dispersal in selfing populations could indeed promote local adaptation, as suggested by Hereford (2010), even if no significant effect of the mating system on local adaptation was found in his meta-analysis. However, given the strong incidence of genetic drift expected in populations undergoing high and recurrent selfing, the efficacy of selection is questionable and the role of local adaptation as opposed to genetic drift in shaping the MLGs composition of these populations remains to be assessed. We propose to test whether neutral processes alone can be responsible for the peculiar genetic structure observed in highly selfing populations (high-frequency MLGs) in the absence of selection. Answering this question requires analytical predictions for multilocus diversity indices such as those available for single locus diversity. Such predictions are lacking, and little is known about the expected range of values for multilocus diversity indices under high and recurrent selfing. Overall, a formal description of the multilocus genetic diversity expected in predominantly selfing populations evolving under neutral scenarios is still lacking and limits interpretations of empirical data.

In addition, the organization of predominantly selfing populations in MLG lineages offers the possibility to follow the changes in MLG frequencies through time. Such temporal surveys can give additional insight into the processes shaping diversity. In particular, they are useful to measure the strength of genetic drift through the estimation of the effective population size (e.g., Table 1). Although data gathered from time series are frequent in experimental populations evolving under artificial selection, they are more rarely available for natural populations (Bailey and Bataillon 2016), in particular predominantly selfing populations (Table 1). Temporal studies of natural selfing populations have found that MLGs can be maintained within population over time (Siol et al. 2007; Bomblies et al. 2010; Gomaa et al. 2011). Nonetheless, the last two studies have also found populations in which all the MLGs changed over time, which led the authors to propose extinction–recolonization dynamics to explain their observations. Yet, because there are no theoretical predictions for the trajectory of MLG frequencies over time in populations evolving neutrally, it is not clear how demographic events (from changes in population size to extinction–recolonization) affect the persistence of multilocus genotypes over time without selection.

Here, we propose to use simulations to explore how predominant selfing shapes single locus and multilocus genetic diversity in neutrally evolving populations. The goals of our study are threefold. First, we use simulations to provide neutral expectations for multilocus indices of diversity and determine whether neutral scenarios can explain the peculiar population genetic structure (with high frequency and persistent MLGs) observed in predominantly selfing species without selection. Second, we examine how combining single and multilocus indices of diversity may be insightful when studying the evolutionary trajectory of predominantly selfing populations to distinguish the effects of selfing, population size, and more complex demographic events such as bottlenecks, migration, admixture, or extinction–recolonization. Third, we use changes in allele frequency through time to examine whether we can estimate effective sizes as small as those reported in the literature, and we consider the influence of complex demographic scenarios such as bottlenecks, admixture, and extinction–recolonization on the trajectory of MLG frequencies through time. We compare our simulation results with observations from temporal data on nine populations of the highly selfing plant species *Medicago truncatula*. These nine temporal datasets for *M. truncatula* natural populations can be viewed as a reality check (independent iterations of evolution in a selfing population across 20 generations). As such, they validate the pertinence of our simulation framework as we find genetic diversity patterns similar to our simulations.

Material and methods

Simulation model and scenarios explored

We performed individual-based simulations of diploid hermaphroditic populations using SLiM 2.5 (Haller and Messer 2017). In order to be able to qualitatively compare the simulation results with our empirical data, we fixed some simulation parameters such as the type of genetic markers, the number of loci, and the time span between the temporal samples. We simulated the evolution of 20 independent loci (with a recombination rate of 0.5). SLiM output was processed in R (R Core Team 2018) in order to transform the mutations that occurred on each of the 20 predefined loci into microsatellite allele sizes following the stepwise mutation model (Ohta and Kimura 1973). Briefly, we randomly attributed an effect to each mutation (±1 repeat unit) and the effects of all the mutations occurring at a given locus in a given individual were summed in order to obtain microsatellite allele size. Mutations were neutral and occurred at a rate $\mu = 10^{-3}$ per generation and per locus, which is a realistic rate for plant microsatellites (Thuillet et al. 2002; Marriage et al. 2009). To produce the next generation, new zygotes were built as a combination of two gametes sampled either from two different individuals for outcrossing, or from the same individual for selfing, according to a fixed selfing rate (σ). Each simulation comprised two periods. A first period of 25 N generations (with N the demographic population size, measured as the number of diploid individuals) allowed the populations to reach the mutation-drift equilibrium. At this stage (time $t_0 = 0$), 100 diploid individuals were randomly sampled. Twenty generations later (t_{20}), a second sample of 100 individuals was drawn to obtain temporal sampling.

Five demographic scenarios were considered. In the first one, we simulated a single isolated population with a constant demographic size N. Four demographic population sizes were considered: $N \in [50; 100; 250; 1000]$ and combined to five different values of selfing rate (σ): 0 (completely outcrossing population), 0.5 (partially selfing population), 0.95, 0.98 (predominantly selfing population), and 1 (completely selfing population). To disentangle the effects of selfing from those of genetic drift, we also simulated populations of the same effective size with different selfing rates by setting $N = 2N_e/(2 - \sigma)$ for $\sigma \in [0; 0.5; 0.95; 0.98; 1]$ and $N_e \in [100; 250]$. To examine the impact of sampling effect, we reiterated the analysis for one of the simulations with $N = 100$ and $\sigma \in [0; 0.95; 1]$ after reducing the sample size at t_0 and t_{20} to 5, 10, 20, 30, or 50 individuals. Each sampling was repeated independently 100 times. In the following scenarios, we considered only predominantly selfing populations ($\sigma = 0.95$). In a second scenario, we explored the combined effects of predominant selfing and a bottleneck. To this aim, we simulated an isolated population of size $N = 250$ and a selfing rate $\sigma = 0.95$ undergoing at time t_{10} a drastic demographic size reduction (to $N' = 1$, 5, or 25 diploid individuals) for one generation. In a third scenario, we evaluated the effects of migration by simulating an island model with ten subpopulations of constant size $N \in [50; 100; 250]$ exchanging diploid migrants at a constant rate. Three values of migration rate (m) were simulated: 2×10^{-4}, 2×10^{-3}, and 2×10^{-2} per generation. Samples were taken from a single deme (the focal population) in these structured scenarios. The effects of more drastic migration events were investigated in a fourth scenario, the admixture scenario, where a fraction of the focal subpopulation was replaced by individuals from another single population. The metapopulation was again simulated with an island model with a migration rate $m = 2 \times 10^{-3}$ per generation. At time t_{10} a single admixture event was simulated, with an admixture rate of 50%, 75%, or 100%. Note that 100% admixture is equivalent to a local extinction and recolonization scenario without change in population size. The focal population was sampled at generations t_0 and t_{20}, as in the previous scenarios. Because extinction–recolonization events may be associated with founder events, we evaluated a final set of scenarios with a bottleneck concomitant with an extinction–recolonization event. During the bottleneck, the focal population size was reduced to 1, 5, or 25 diploid individuals. After one generation, the population size was restored to $N = 250$ individuals and 100 diploid individuals were sampled at t_{20}. For each simulation scenario described above (and summarized in Table S1), 1000 independent replicates were performed. SLiM simulation scripts for each of these scenarios as well as R scripts are available on the INRA dataportal. https://doi.org/10.15454/VYPXIJ.

Diversity indices

Diversity analyses were performed using the Hierfstat package in R (Goudet 2005). The genetic diversity of each simulated population was assessed on the t_{20} sample using the average gene diversity across loci (H_E, Nei 1973), the variance in allele size (V), the average number of alleles per locus (n_A), and the number of polymorphic loci (PL). In an isolated random mating population at mutation-drift equilibrium, H_E and V measured on microsatellite markers evolving under the stepwise mutation model are expected to vary with the effective population size N_e and the mutation rate μ as

$$H_E = 1 - \sqrt{\frac{1}{2\theta + 1}} \quad (2)$$

$$V = 2N_e\mu \quad (3)$$

where $\theta = 4N_e\mu$ (Kimmel et al. 1998).

The deviation from Hardy-Weinberg proportions was measured using the inbreeding coefficient F_{IS} with the R package Hierfstat. The percentage of pairs of loci showing significant linkage disequilibrium (*LD%*) was calculated using Genepop (Rousset 2008) with a significance threshold of 0.05. The identity disequilibrium (*g2*), which is expected to depend on the selfing rate following $g_2 = \frac{1-\sigma}{\left(1-\frac{\sigma}{4}\right)\left(1-\frac{\sigma}{2-\sigma}\right)^2} - 1$, (David et al. 2007) was also computed using the R package inbreedR (Stoffel et al. 2016). The R package Poppr (Kamvar et al. 2014) was used to identify the number of private alleles (p_A), to group individuals with identical combinations of alleles (multilocus genotypes, MLG), compute the number of distinct MLGs, their frequency and their repartition over time in the two samples (t_0, t_{20}). The multilocus diversity was characterized by the Shannon's index, computed as $H = -\sum p_i ln(p_i)$, where p_{i_p} is the frequency of the *i*th MLG. The frequency of the most frequent MLG (*MFMLG*) was also computed and we analyzed the correlation between *MFMLG* and *H* through Spearman correlations using R.

We calculated the pairwise genetic distances between individuals at generation t_{20} as the number of allele differences (between 0 and 40) between each pair of individuals, regardless of the allele size. We used two indices to characterize the distributions of distances: the mean pairwise genetic distance (D_{mean}) and the maximum pairwise genetic distance (D_{max}). The correlation between some indices (D_{mean} and H_E, or D_{max} and *LD%*) was measured through Spearman correlations using R.

To summarize the trajectories of MLG frequencies through time, we considered the joint MLG frequency spectrum (*MLGFS*, by analogy with the allele frequency spectrum) as the matrix containing the proportion of MLGs found at the corresponding individual counts in each generation, averaged over simulation replicates. For *K* simulation replicates, we have $MLGFS[i,j] = \frac{\sum_{k=1}^{K} \frac{MLG(i,j)}{MLG_k}}{K}$, where MLG($i$, j) is the number of MLGs found in i individuals at t_0 and in j individuals at t_{20}, and MLG_k is the total number of different MLGs in replicate k. The *MLGFS* therefore allows to follow the evolution of the frequency of MLGs overtime.

The relative temporal differentiation between the two samples was assessed with Weir and Cockerham's F_{ST} (1984), estimated using the R package Hierfstat (Goudet 2005). The effective population size was estimated based on the temporal differentiation between samples (temporal F_{ST}) as outlined in Frachon et al. (2017):

$$\widehat{N_e} = \frac{\tau(1 - F_{ST})}{4F_{ST}}, \quad (4)$$

where $\widehat{N_e}$ is the estimate of the effective population size and τ is the number of generations separating the two sampling events. This method assumes that the population is isolated (no migration), of constant size and that no mutation occurs between samplings. We estimated the focal population effective size in our different simulation scenarios in order to examine whether the deviations from theoretical assumptions (e.g., admixture or bottlenecks) can lead to N_e estimates as small as those reported in the literature. N_e estimates in scenarios of isolated populations were compared with the theoretical expectations given by Eq. (1) (assuming $\alpha = 1$: $N_e = \frac{N}{1+\frac{\sigma}{2-\sigma}}$) for an isolated population with no change of population size, where N is the demographic size of the population and σ is the selfing rate (Pollak 1987); and $\frac{1}{N_e} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_e^t}$ for an isolated population undergoing bottlenecks, where $N_e{}^t$ is the effective population size at generation t (Crow and Kimura 1970).

Medicago truncatula natural populations

Medicago truncatula is an annual, predominantly selfing species of the legume family (Fabaceae), found around the Mediterranean Basin. Maternal progeny analyses have shown very low levels of residual outcrossing (Siol et al. 2008). Between 1986 and 2014, nine natural populations located in Spain (SP1–SP3), Corsica (CO1–CO3), and southern France (FR1–FR3) were sampled two or three times each (locations of the different populations can be found on a map in Figure S1). In order to avoid oversampling the progeny of a single individual, pods were sampled along transects running across the populations, with at least 1-m distance between each collected pod. This sampling strategy also allows to limit spatial effects due to the very fine spatial structure observed in *M. truncatula* natural populations (Bonnin et al. 2001). Sample sizes varied between 31 and 232 individuals. Hereafter, each temporal sample will be denominated by its population code followed by the sampling year.

DNA was extracted from 50 mg of fresh leaves with the Chemagic DNA Plant Kit (Perkin Elmer), according to the manufacturer's instructions. The protocol is adapted to the use of the KingFisher Flex™ (Thermo Fisher Scientific) automated DNA purification workstation. Twenty microsatellite loci were used for genotyping. Eighteen of them have been described previously (Baquerizo-Audiot et al. 2001; Arrighi et al. 2006; Ronfort et al. 2006; Siol et al. 2007). Two new loci, 319 and DMI1-6, were developed in our team after identifying long and polymorphic simple sequence repeats in resequencing studies (319-F GTGGGATTTGAATAGGATTG, 319-R CGATATGGTCCACTTTTGTC, annealing temperature: 57 °C;

Table 2 Mean values for single locus and multilocus indices of genetic diversity for isolated populations with increasing selfing rates and a constant demographic size of $N = 250$

N	σ	N_e	F_{IS}	H_E	V	$nMLG$	H	$LD\%$	g_2	$NbLoc\_g2$
250	0	250	0.0 (0.0)	0.42 (0.00)	0.50 (0.39)	99.9 (0.0)	4.61 (0.00)	0.06 (0.00)	0.001 (0)	19.9 (0.4)
250	0.5	188	0.32 (0.00)	0.37 (0.00)	0.38 (0.25)	99.2 (1.0)	4.59 (0.00)	0.13 (0.00)	0.292 (0.004)	19.5 (0.7)
250	0.95	132	0.85 (0.00)	0.30 (0.00)	0.27 (0.14)	58.2 (35.7)	3.79 (0.03)	0.38 (0.01)	5.576 (3.607)	16.7 (1.8)
250	0.98	128	0.91 (0.00)	0.29 (0.00)	0.25 (0.14)	42.7 (33.0)	3.34 (0.05)	0.41 (0.02)	11.268 (49.45)	13.9 (2.8)
250	1	125	0.94 (0.00)	0.29 (0.01)	0.24 (0.11)	29.3 (19.9)	2.86 (0.06)	0.47 (0.04)	−0.04 (7.324)	6.2 (2.1)

Values in brackets show the variance over 1000 replicates. N_e is the effective size calculated using Eq. (1) with $\alpha = 1$; F_{IS} is the inbreeding coefficient. H_E is the estimated gene diversity, V is the variance of allele size, $nMLG$ is the number of MLGs, H is the Shannon's index, $LD\%$ is the percentage of loci with significant linkage disequilibrium, g_2 is the identity disequilibrium and $NbLoc\_g2$ is the number of loci used to compute $g2$. Expected values of F_{IS}, H_E, and V are reported in Table S2

DMI1-6-F1 TAGAAGATGAAGCGCAAACG, DMI1-6-R2 TTCACCTTAACGCGTCCAAC, annealing temperature: 60 °C). We followed the protocol of amplification reactions described in Siol et al. (2007). Samples were prepared by adding 3 µl of diluted PCR products to 16.5 µl of ultrapure water and 0.5 µl of the size marker AMM524. Amplified products were analyzed on an ABI prism 3130 Genetic Analyzer and genotype reading was performed using GeneMapper Software version 5. Individuals and loci with more than 10% missing data across all samples of a population were removed from the diversity analyses, as well as completely monomorphic loci.

For each population and year, we performed the same analyzes of single and multilocus diversity as those performed on our simulated populations. To account for variation in sample sizes, mean allelic richness per locus (Rs) and private alleles (p_A) were computed using the rarefaction method with the program ADZE (Szpiech et al. 2008). Selfing rates were estimated from F_{IS} using the classical relationship $F_{IS} = \sigma/(2 - \sigma)$ (Hartl and Clark 1998) and using a maximum-likelihood approach based on the identity disequilibrium (g_2), with the software RMES (David et al. 2007). MLG frequency spectra were computed for each population. Temporal F_{ST} estimates were used to estimate the effective population size using the method described previously (Eq. (3), Frachon et al. 2017), assuming a single generation per year. Approximate bootstrap confidence intervals for the temporal estimates of effective size were computed following DiCiccio and Efron (1996).

Results

Single-locus and multilocus genetic diversity in isolated populations

In the simulations of a single isolated population for different combinations of selfing rates and demographic population sizes, estimates of single locus indices (H_E and V) are in accordance with theoretical predictions at mutation-drift equilibrium (Table S2), showing a decrease in the neutral genetic diversity with increasing selfing rates (Table 2). When the demographic size is adjusted to keep the effective size constant while the selfing rate varies, single locus diversity indices remain around the expected value too (Table 3). These results are not new as they replicate the known effects of selfing on single locus diversity but they are helpful to validate our simulation framework.

As shown in Table 2, we found that the multilocus diversity ($nMLG$ and H) also decreases with selfing, while the homozygosity (F_{IS}) and the associations between loci ($LD\%$, $g2$) increase. For completely selfing populations ($\sigma = 1$), $g2$ is biased downwards due to extremely high homozygosity limiting the number of loci available for the estimation (as $g2$ measures the correlation of heterozygosity between loci). In completely outcrossing or low selfing populations (up to 50% selfing in our simulations), there are on average as many MLGs as individuals sampled. In contrast, $nMLG$ decreases to around two thirds of the sample in our simulations with 95% selfing and to less than one-third in completely selfing populations for $N = 250$. This loss of MLGs is even more dramatic when the population size is lower (e.g., $N = 50$, Table S2). In addition, contrary to single locus indices, multilocus diversity indices keep decreasing with increasing selfing rate even for a given N_e value (Table 3), in conjunction with the increase in linkage disequilibrium. Our analysis of the effect of sample size shows that, for a given selfing rate, both H_E and the frequency of the most frequent MLG ($MFMLG$) are biased and less precise for small sample sizes. The statistics approach the expected value with a smaller sample size for single locus compared to multilocus diversity ($N_{samp} = 20$ for H_E and $N_{samp} > 30$ for $MFMLG$, Fig. S7).

Table 3 Simulated populations with increasing selfing rates and demographic sizes adjusted to keep the effective size (N_e) constant and equal to 250

N	σ	N_e	F_{IS}	H_E	V	$nMLG$	H	$LD\%$	g_2	$NbLoc\_g2$
250	0	250	0.0 (0.00)	0.42 (0.002)	0.50 (0.39)	99.9 (0.0)	4.61 (0.00)	0.06 (0.00)	0.001 (0.000)	19.9 (0.4)
333	0.5	250	0.32 (0.00)	0.42 (0.002)	0.50 (0.44)	99.4 (0.7)	4.60 (0.00)	0.14 (0.00)	0.29 (0.004)	19.8 (0.4)
475	0.95	250	0.89 (0.00)	0.42 (0.002)	0.50 (0.40)	71.1 (26.2)	4.11 (0.01)	0.54 (0.01)	5.875 (3.138)	18.8 (1.2)
490	0.98	250	0.94 (0.00)	0.42 (0.002)	0.50 (0.42)	57.0 (27.5)	3.80 (0.02)	0.63 (0.01)	12.75 (45.025)	16.5 (2.6)
500	1	250	0.98 (0.00)	0.42 (0.002)	0.53 (0.63)	40.8 (22.5)	3.33 (0.03)	0.75 (0.03)	−0.151 (4.641)	6.4 (2.1)

The corresponding N is set according to Eq. (1) with $\alpha = 1$; the expected value for H_E is 0.42 (Eq. (2)), and 0.5 for V (Eq. (3))

Structure of multilocus genetic diversity in more complex scenarios

The frequency of the *MFMLG* summarizes the increase of repeated multilocus genotypic combinations, because it increases with the selfing rate, especially for $\sigma \geq 0.95$ (Fig. 1a, Table S2) and is highly correlated with Shannon's index H ($P < 2.2 \times 10^{-16}$, $r^2 = 0.91$, Fig. S2). In the following, we will use *MFMLG* as an indicator of multilocus diversity variations. Our simulations with lower population size, bottlenecks or metapopulation dynamics highlight that extreme patterns of MLG repetition ($MFMLG > 30\%$) are observed only for very low demographic population sizes ($N = 50$) combined with high selfing rates ($\sigma \geq 0.95$), or with strong bottlenecks (reduction to fewer than five individuals), associated with predominant selfing (Fig. 1b, Table S2). These extreme patterns of MLG repetition are nevertheless highly variable among simulation replicates. Figure 1b also illustrates that this increase of *MFMLG* with low population size or bottlenecks is associated with a decrease in single locus diversity (H_E). In addition, Fig. 1b shows that changes in single locus diversity due to migration are greater than changes in the frequency of the *MFMLG* (migration and admixture scenarios on Fig. 1b). Similar patterns can be observed when comparing Shannon's index (H) and H_E (Fig. S3).

The mean distance between two individuals within a population (D_{mean}) is highly correlated to H_E ($P < 2.2 \times 10^{-16}$, $r^2 = 0.98$, Fig. S4). We used the maximum distance found between two individuals in a population (D_{max}) to describe the genetic divergence accumulated between MLGs. D_{max} increases with the genetic diversity (with increasing N) and with the selfing rate (for populations with the same effective size, $N_e = 250$ or $N_e = 100$ in Table S2). For predominantly selfing populations ($\sigma \geq 0.95$), D_{max} is strongly correlated with $LD\%$ ($P < 2.2 \times 10^{-16}$, Fig. S5). As a consequence, D_{max} is the highest in scenarios involving migration, either at a constant rate or after drastic admixture (Fig. 2b). Another interesting result is that only very low demographic population sizes or very strong bottlenecks in isolated populations result in low D_{max} (<0.5).

Temporal changes

We measured the change in allele frequencies through time (samples separated by 20 generations) with the relative genetic differentiation between temporal samples using F_{ST}. We observed that F_{ST} estimates and their variance increase with selfing (Table S2). Whereas migration equilibrium (at rates ≤ 0.002) does not affect the temporal differentiation much, occasional large migration events (admixture) raise the differentiation through time. Genetic differentiation is particularly strong in extreme demographic scenarios such as extinction–recolonization because of population replacement.

We used the temporal F_{ST} to estimate the effective population size according to Eq. (4), ignoring the fact that some of our simulated scenarios do not meet the assumption of the underlying theoretical model (isolated populations). Figure 3a shows that, as expected in isolated populations of constant size (see Eq. (2)), $\widehat{N_e}$ estimates increase with H_E. Despite a large variance between replicates, average $\widehat{N_e}$ estimates in isolated populations are close to the theoretical expectations (see Table S2), except for large populations sizes or complete selfing ($\sigma = 1$). In these cases, the variance of $\widehat{N_e}$ is extreme, due to sampling variance and linkage disequilibrium, respectively. Under complex demographic scenarios involving strong migration events or extinction–recolonization, $\widehat{N_e}$ estimates are remarkably low compared to the simulated demographic population sizes. Those estimates disagree with the levels of genetic diversity (H_E) observed in these populations given the expectations from Eq. (2). This could be caused by departures from the model's assumptions (see discussion). Indeed, the effects of migration are visible through the increase in the number of private alleles at t_{20} (p_A, Table S2).

Figure 4 shows patterns of MLGs persistence over time (after 20 generations in our simulations) in the different scenarios we investigated. Under complete outcrossing, the conservation of a MLG after 20 generations is extremely rare, as expected due to recombination (Fig. 4a). In contrast, under predominant selfing (Fig. 4b), MLGs are frequently shared between temporal samples. Moreover, MLGs that reach a

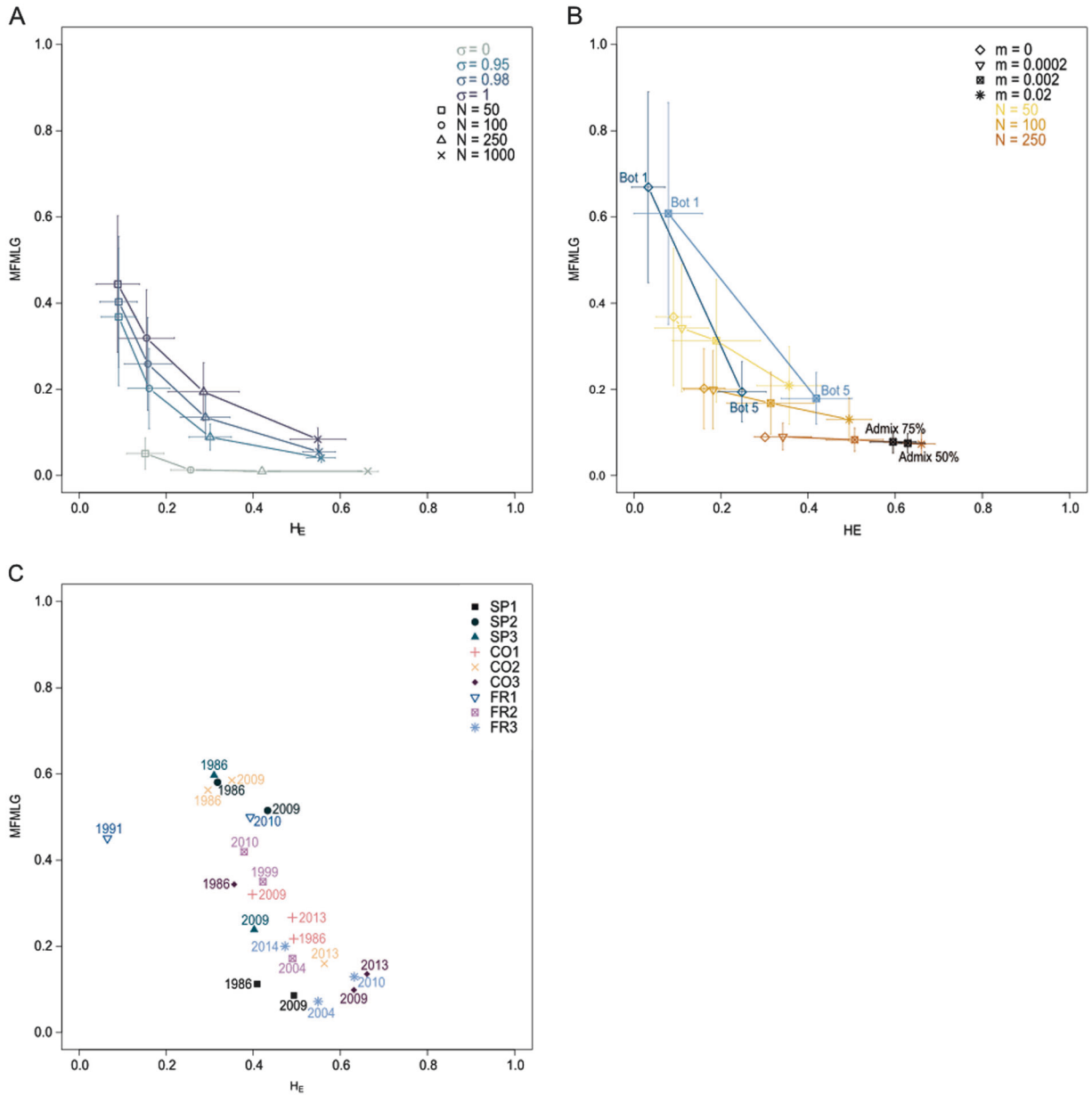

Fig. 1 Covariation between the frequency of the most frequent MLG (*MFMLG*) and the gene diversity (H_E) across simulated scenarios. **a** Scenarios of isolated populations with varying demographic sizes N and selfing rates σ; **b** scenarios of migration (orange), admixture (black), bottleneck (blue), and extinction–recolonization (light blue) with $\sigma = 0.95$; **c** natural populations of *Medicago truncatula* for each sampling date. For **a** and **b**, points indicate means and horizontal and vertical bars stand for the standard deviation across the 1000 replicates

high frequency within the first generation (measured by the abscissa for t_0), tend to remain at high frequency at t_{20}. This pattern is amplified when the population size is low (Fig. 4c). Strong bottlenecks (Fig. 4d) raise the frequency of some MLGs independently of their frequency in the first generation. Scenarios with migration and admixture slightly reduce the occurrence of conserved high-frequency MLGs (Fig. 4e, f) while scenarios including extinction–recolonization produce spectra with fewer MLGs conserved over time (Fig. 4g, h for extinction–recolonization with a bottleneck).

Empirical data

In the nine natural populations of *M. truncatula* studied, F_{IS} values are high, ranging between 0.88 and 1. This translates into very high-selfing rate estimates for all populations ($\sigma_{FIS} > 0.9$, Table 4). Selfing rate estimates with RMES are sometimes lower but remain well above 0.8 (Table 4). The number of MLGs is generally low and compatible with high selfing rates.

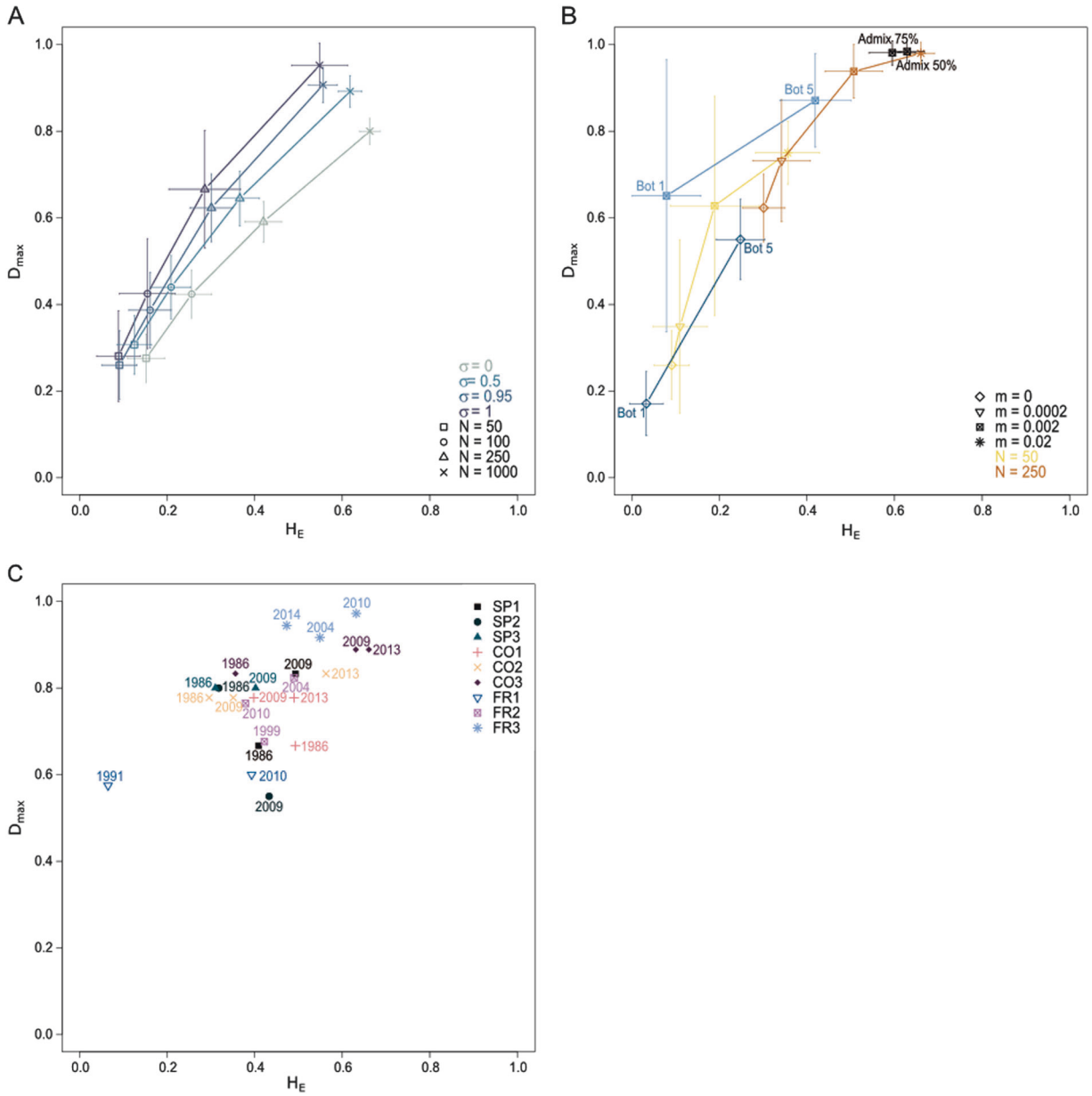

Fig. 2 Covariation between the maximum distance between pairs of individuals in a population (D_{max}) and H_E across simulated scenarios. **a** Scenarios of isolated populations with increasing demographic sizes N and selfing rates σ; **b** scenarios of migration (orange), admixture (black), bottleneck (blue) and extinction–recolonization (light blue) with $\sigma = 0.95$; **c** natural populations of *Medicago truncatula* for each sampling date. For **a** and **b**, points indicate means and horizontal and vertical bars stand for the standard deviation across the 1000 replicates

Single and multilocus diversity

The mean gene diversity within population and year (H_E) is remarkably high (higher than 0.3, Table 4). The maximum genetic distance between two individuals, D_{max}, is also always high (higher than 0.5, Fig. 2c). Most populations therefore seem to be distributed within a parameter space more limited than the one explored by our simulations (high H_E associated with high D_{max}). Only sample FR1_1991 presents both low single and multilocus diversity (Fig. 1c, Fig. 2c). *MFMLG* values are highly variable, with extreme patterns of MLG repetition (*MFMLG* higher than 30%) in nearly half of the populations studied. Accordingly, these populations also display the lowest values of Shannon's H (Table 4). However, such a combination of high single-locus diversity and extremely low multilocus diversity was not observed in any of our simulated scenarios (Fig. 1c).

Temporal dynamics of diversity

The MLG frequency spectra highlight two different types of dynamics of MLGs through time. In populations SP1, SP2,

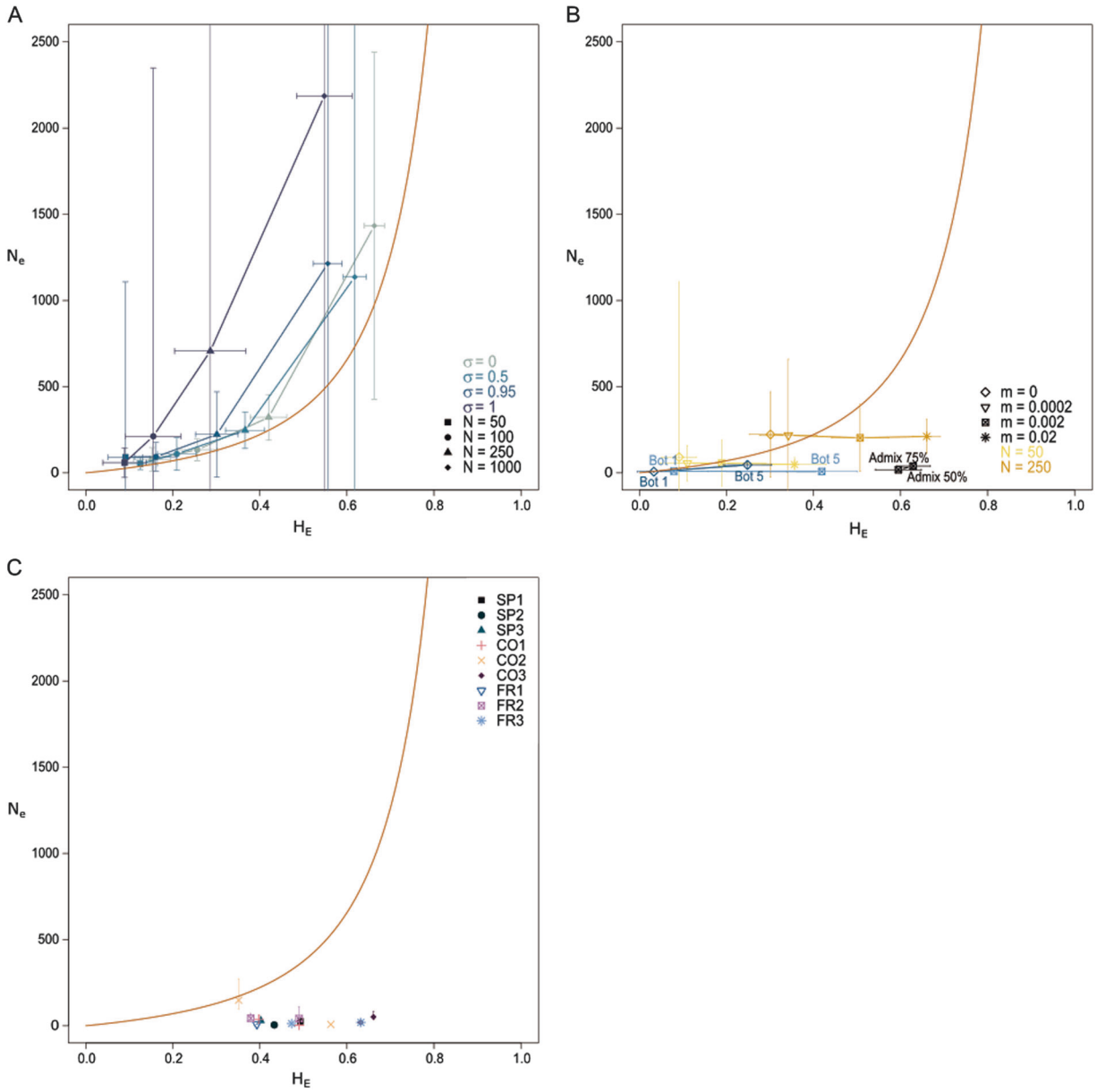

Fig. 3 Temporal estimation of effective population size ($\widehat{N_e}$) compared to gene diversity (H_E). The red curve corresponds to the expected relationship between H_E and N_e assuming a constant isolated population (Eq. (2)) and a mutation rate of 0.001. **a** Scenarios of isolated populations with varying selfing rates σ and demographic sizes N; **b** scenarios of migration, admixture, bottleneck and extinction–recolonization with $\sigma = 0.95$; **c** temporal $\widehat{N_e}$ estimates in natural populations of *Medicago truncatula*, with vertical bars representing the 95% confidence interval. For **a** and **b**, horizontal and vertical bars stand for the standard deviation across the 1000 replicates

and CO1, a single low-frequency MLG or none at all remain over time (Fig. S6). In our simulations, such dynamics of multilocus diversity are obtained with extinction–recolonization events only (Fig. 4g, h). In the other populations, several MLGs are conserved through time (SP3, CO3, FR3, CO2, FR2, and FR1, Fig. S6). Among these populations, FR3 and CO3 are the most diverse ($H_E > 0.5$) and present extreme patterns of multilocus diversity with *MFMLG* lower than 0.2 and D_{max} higher than 0.8 (Figs. 1c and 2c). Such patterns were also observed in our simulations with strong migration and admixture (Figs. 1b and 2b).

Temporal differentiation between sampling years, as estimated by temporal F_{ST}, is high for most of the populations studied (Table 4). The effective population sizes estimated using the temporal F_{ST} method are variable but consistently low: all estimates except one are lower than 100 (with a maximum of 150 for population CO_2). These estimates are too low compared with the observed single locus diversity given the expectations from eq. 2 (Fig. 3c).

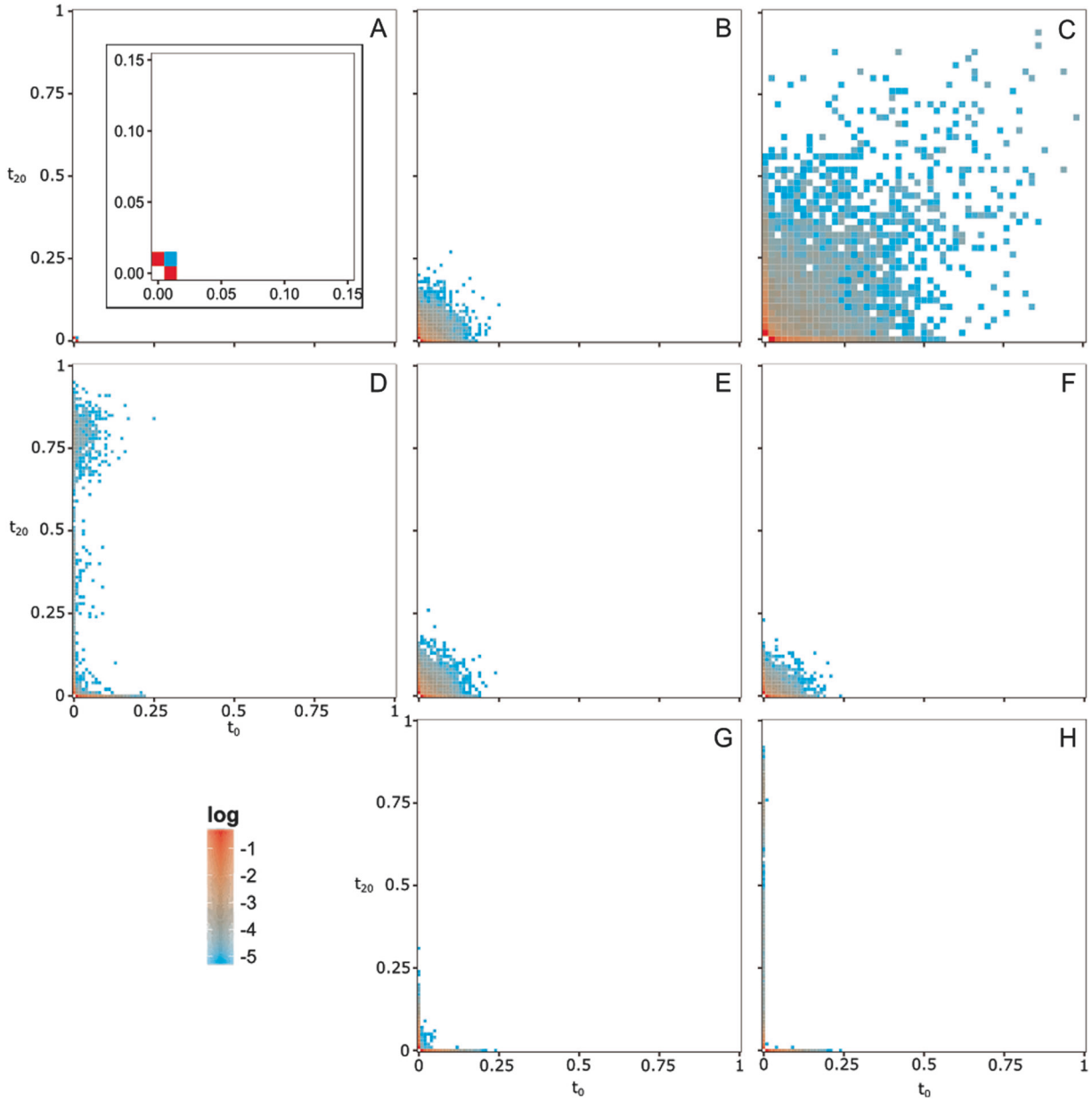

Fig. 4 Joint MLG frequency spectra for each demographic scenario. The horizontal axis represents the frequency at which a MLG is found in the first sample (t_0), the vertical axis represents the frequency of this same MLG in the second sample (t_{20}). The color gradient represents the $\log_{10}$ of the frequency at which each case is observed in 1000 simulation replicates. **a** Isolated outcrossing population ($\sigma = 0$) of 250 individuals. The inset is a zoom of frequencies between 0 and 0.15; **b** isolated predominantly selfing population ($\sigma = 0.95$; $N = 250$); **c** isolated predominantly selfing population ($\sigma = 0.95$; $N = 50$); **d** isolated predominantly selfing population ($\sigma = 0.95$; $N = 250$) undergoing a bottleneck of one individual at t_{10}; **e** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$); **f** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing 50% admixture with constant population size at t_{10}; **g** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing extinction–recolonization with constant population size at t_{10}; **h** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing extinction–recolonization by one individual at t_{10}

In our simulations, we observed a similar mismatch between H_E and N_e values for migration and admixture scenarios. This resemblance with migration and admixture scenarios is also visible in the high number of private alleles we observe in recent compared to older temporal samples in our populations (Table 4, Table S2).

Discussion

Our work aimed at describing the consequences of high-selfing rates and metapopulation dynamics on the structure of genetic diversity in populations, and how it can change over time. We argue that the classical single locus diversity

Table 4 Genetic diversity in *M. truncatula* populations

Population	Year	n	F_{IS}	σ_{FIS}	σ_{RMES}	H_E	$nMLG$	p_A	H	$MFMLG$	F_{ST}
SP1	1986	71	0.96	0.98	0.93 (0.03)	0.41	31	–	3.10	0.11	0.186
	2009	93	0.99	0.99	0.92 (0.06)	0.49	39.2	1.87 (0.40)	3.55	0.09	
SP2	1986	31	1.00	1.00	–	0.32	4	–	0.90	0.58	0.536
	2009	66	0.97	0.98	0.98 (0.01)	0.43	9.3	2.31 (0.32)	1.76	0.52	
SP3	1986	67	0.97	0.98	0.99 (0.01)	0.31	12	–	1.51	0.60	0.261
	2009	88	0.98	0.99	0.99 (0.01)	0.40	24.7	1.67 (0.38)	2.71	0.24	
CO1	1986	46	1.00	1	–	0.49	18	–	2.43	0.22	
	2009	78	0.98	0.99	0.99 (0.01)	0.40	12.6	0.22 (0.10)	2.08	0.32	0.13
	2013	60	0.96	0.977	0.96 (0.02)	0.49	23.7	0.76 (0.12)	2.87	0.27	0.243
CO2	1986	64	0.95	0.97	0.96 (0.02)	0.30	20	–	1.89	0.56	
	2009	94	0.96	0.98	0.99 (0.01)	0.35	16.8	1.19 (0.27)	1.83	0.59	0.03
	2013	100	0.96	0.98	0.93 (0.02)	0.56	32.6	1.42 (0.45)	3.27	0.16	0.115
CO3	1986	64	0.94	0.97	0.97 (0.02)	0.36	20	–	2.16	0.34	
	2009	81	0.96	0.98	0.99 (0.01)	0.63	41.0	0.71 (0.24)	3.65	0.10	0.226
	2013	162	0.94	0.97	0.95 (0.01)	0.66	44.6	1.20 (0.45)	4.08	0.14	0.019
FR1	1991	91	0.88	0.94	0.97 (0.02)	0.07	8.7	–	1.29	0.45	0.337
	2010	82	0.99	0.99	–	0.39	11	1.75 (0.27)	1.34	0.50	
FR2	1999	60	0.98	0.99	0.99 (0.01)	0.42	11	–	1.77	0.35	
	2004	64	0.95	0.97	0.90 (0.04)	0.49	38.9	1.32 (0.33)	3.41	0.17	0.029
	2010	93	0.96	0.98	0.88 (0.06)	0.38	17.4	0.82 (0.26)	2.21	0.42	0.033
FR3	2004	97	0.98	0.99	0.97 (0.01)	0.55	48	–	3.57	0.07	
	2010	201	0.90	0.95	0.93 (0.01)	0.63	63.0	1.23 (0.26)	4.22	0.13	0.071
	2014	135	0.97	0.98	0.97 (0.01)	0.47	46.4	0.80 (0.18)	3.46	0.20	0.071

n is the sample size, F_{IS} is the inbreeding coefficient, σ_{Fis} is the selfing rate estimated from the F_{IS}, σ_{RMES} is the selfing rate estimated using RMES, H_E is the mean gene diversity, $nMLG$ is the MLG number (calculated using a rarefaction method), and p_A is the mean number of private alleles per locus found in the second or third temporal samples (calculated using a rarefaction method, with standard deviation in brackets), H is the Shannon's index, $MFMLG$ is the frequency of the most frequent MLG and F_{ST} is the temporal F_{ST} between successive temporal samples. In samples SP2_1986, CO1_1986 and FR1_1991, the lack of heterozygosity in the population prevented the estimation of σ_{RMES}

indices are not sufficient to fully understand the demographic history of predominantly selfing populations, which should benefit from multilocus indices. Because of the lack of analytical expectations for such indices, we proposed a simulation approach to address the question in a theoretical framework.

Neutral scenarios can explain the multilocus population genetic structure of predominantly selfing species

Our simulations of isolated and predominantly selfing populations (with selfing rates above 0.95), show a population genetic structure organized in repeated multilocus genotypes. As for single locus diversity, multilocus diversity (measured as the number of MLGs or the haplotypic diversity H), decreases with increasing selfing rates. In addition, the nonindependence between loci increases with selfing (Nordborg 2000), as seen through the elevated linkage and identity disequilibrium, and the multilocus diversity is further reduced. The maximum distance between two individuals, D_{max}, increases with the selfing rate, highlighting the fact that self-fertilizing populations are composed of differentiated lineages. The increase in D_{max} is caused by the reduced effective recombination in selfing populations, which constrains new mutations within only one genetic background. For predominantly selfing populations ($\sigma \geq 0.95$), D_{max} is also highly correlated to $LD\%$, because they both increase with within-population structure. The empirical data obtained on natural populations of *M. truncatula* strongly support our simulation results: our estimates of selfing rates confirm *M. truncatula* as a predominantly selfing species, and we find repeated MLGs and large D_{max} in every population. In *Arabidopsis thaliana*, studies in natural populations also showed that they are composed either of identical or highly differentiated individuals (Bakker et al. 2006; Montesinos et al. 2009). Overall, these results highlight the importance of multilocus analyses for the study of natural selfing populations. Yet, such analyses are often overlooked in the literature (e.g.,

Trouvé et al. 2005; Gow et al. 2007) or are limited to reporting the number of distinct MLGs (e.g., Bomblies et al. 2010; Gomaa et al. 2011). Furthermore, our results stress out that accurate estimates of MLG frequencies are essential, especially in the presence of rare MLGs, and require larger samples than for estimates of single locus diversity (above 30 individuals, Fig. S7).

The first aim of the present study was to verify if selectively neutral scenarios with high-selfing rates could lead to repeated MLGs, sometimes at high frequency and maintained through time. Indeed, Avise and Tatarenkov (2012) argued that this peculiar genetic structure in selfing populations provided evidence for the occurrence of selective processes promoting locally adapted MLGs. Our simulation results show that strong genetic drift induced by small population size or bottlenecks may be sufficient to explain the multilocus genetic structure observed in selfing populations, without any selection. It is, however, important to stress out that our results are not sufficient to rule out selective processes in a population but only present alternative hypotheses to explain the observed structure of genetic diversity. Testing Avise and Tatarenkov (2012)'s hypothesis would require reciprocal transplants, or at least measuring the fitness of the MLGs in order to see if the locally most frequent MLG has indeed the highest fitness in the local environment.

Combining single and multilocus indices of diversity is insightful when studying the demographic history of predominantly selfing populations

We analyzed the multilocus structure of the simulated populations and focused on indices describing MLG frequency (*MFMLG*) or MLG genetic similarity (D_{max}). Those indices are especially informative when analyzed conjointly with single locus diversity (H_E) and can help disentangle the effect of selfing and demographic events (such as bottlenecks or migration) on genetic diversity in selfing populations. Indeed, high *MFMLG* combined with low H_E is characteristic of small population (constantly small or due to a bottleneck) with a high selfing rate. On the other hand, low levels of multilocus diversity while single locus diversity is high were observed with strong migration or admixture scenarios. This highlights the fact that migration restores single locus diversity faster than multilocus diversity in predominantly selfing populations (Fig. 1b). This was also visible when analyzing conjointly H_E and D_{max}: in our migration scenarios, new alleles combined within migrant MLGs were introduced in the population, resulting in both high *LD%* and D_{max} values.

Even though our simulations explored only a restricted number of scenarios (in terms of population size, sample size, selfing rate, time span between sampling, etc.), they were able to replicate patterns observed in empirical data. Except for one population (FR1), the levels of genetic diversity (H_E) in most populations of *M. truncatula* were surprisingly high compared with theory and other studies of predominantly selfing populations (e.g., Gomaa et al. 2011; Lundemo et al. 2009; Stenøien et al. 2005). In addition, high single locus diversity (H_E) was combined with repeated MLGs (high *MFMLG*), which may be consistent with populations belonging to a metapopulation with strong migration or even admixture events. Genetic diversity measured by both *MFMLG* and H_E (or D_{max} and H_E) as well as the MLG frequency spectrum suggest likely extinction–recolonization events in three populations of *M. truncatula* (SP1, SP2, and CO1). Nevertheless, a fine analysis of the spatial genetic structure should be performed to ensure that the drastic changes in genetic structure observed in these populations are not due to a shift in the location of the sampling transect. However, in a set of populations, MLG repetition (*MFMLG*) and single-locus genetic diversity (H_E) were both very high (higher than 0.4). This combination of a low number of MLGs and high single-locus diversity was never observed in our simulations. Other studies also reported a similar pattern (e.g., Barrière and Félix 2007), which is probably associated with scenarios or combinations of parameter values that were not considered in our set of simulations. This highlights the need for a more systematic exploration of the parameter space if one intends to perform statistical inferences on empirical data.

Insights from the temporal analysis of predominantly selfing populations

The temporal dimension of our analyses is one of the main particularities of this study, and is rarely examined in natural populations with high selfing rates (we found less than a dozen studies, some being reported in Table 1). Yet temporal data are useful because they allow estimating the effective population size and thus give insight into the strength of genetic drift (Waples 1989). However, the decreased effective recombination in selfing populations reduces the number of independent loci, and after several generations of predominant selfing the whole genome tends to behave as a single "superlocus". F_{ST} estimates based on few or a single locus suffer from a large sampling variance (Weir and Hill 2002) and this is visible in our simulations in which the variability of F_{ST} estimates increased with the selfing rate. Indeed, measuring F_{ST} from linked loci is equivalent to measuring it from a lower number of loci. Moreover, if metapopulation dynamics are frequent in selfing populations, it may cause departures from the assumptions of the theoretical model underlying the estimation of N_e from temporal F_{ST} (i.e., isolated

population of constant size, Waples 1989). Thus, temporal estimates of N_e in highly selfing populations should be treated with caution.

Interestingly, our simulations showed that examining the trajectory of MLG frequencies through time gives insights into the demographic history of a predominantly selfing population. In particular, the MLG frequency spectrum (*MLGFS*) describes the upper and lower bounds for the trajectory of MLG frequencies between two generations for a given demographic scenario. For example, a strong increase in *MFMLG* between two generations suggests a bottleneck. Although *MLGFS* are more difficult to interpret on empirical data because of the absence of replicates (e.g., Fig. S6), we show that they can provide support for a hypothesis of extinction–recolonization, along with large values of temporal differentiation.

Finally, our study also highlights high temporal stochasticity in *M. truncatula* natural populations. Indeed, diversity often changed over time, and the temporal samples of a given population were not clustered together in joint analyses of diversity indices (Figs. 1c and 2c). This temporal variability was also described in *Bulinus forskalii* by Gow et al. (2007), who attributed it to "highly dynamic demographic systems, including bottleneck and extinction–recolonization events". Larger variance in diversity levels among selfing compared to outcrossing populations has also been reported before (Schoen and Brown 1991).

Conclusion

The comparison of our simulation results with data obtained in a highly selfing species (*Medicago truncatula*) highlighted the pertinence of our simulation approach. Yet, the number of scenarios and parameter values we explored were limited and this could limit the generalization of our results to other datasets (e.g., other molecular markers, other sampling tempo). We expect, however, that the general patterns highlighted here using microsatellites will remain unchanged with other molecular markers such as a large number of genome-wide SNPs. In addition, the scripts are available and easily amendable to fine tune the comparison with other empirical datasets (in terms of population size, sample size, selfing rate, time span between sampling, etc.).

The demographic scenarios examined here were sufficient to show that selection is not required to explain the prevalence of repeated MLGs in predominantly selfing populations and their persistence through time. If background selection or selective sweeps are expected to reduce the effective size and would reduce single and multilocus diversity concomitantly (Glémin 2007; Kamran-Disfani and Agrawal 2014), the effects of complex selection scenarios such as local adaptation are more difficult to predict. Simulating scenarios involving selection was beyond the scope of this study but would be useful to address the question of the threshold selfing rate beyond which selection will act at the MLG (or haplotype) level rather than at the locus level, due to the severely reduced effective recombination (Neher and Shraiman 2009). Another perspective to this work will be to develop an integrated method to infer parameters such as the effective size and the selfing rate, based on the summary statistics described here (using a likelihood-free inference method, e.g., Beaumont et al. 2002; Rousset et al. 2016).

Data archiving

Genotype data of *Medicago truncatula* natural populations are available on the INRA dataportal. https://doi.org/10.15454/VCZIMR

The scripts used for the simulations and the computation of diversity indicators are available on the INRA dataportal. https://doi.org/10.15454/VYPXIJ

Acknowledgments M.J.'s PhD fellowship is funded by the INRA (French National Institute of Agronomical Research) Department of Genetics and Plant Breeding and the INRA Metaprogram ACCAF. The authors thank J.M. Prosperi for the collection of seeds as well as C. Tollon, F. Mora, E. Figuet, and J. Terraillon who contributed to the production of the microsatellite dataset. We thank Mathieu Siol, John Pannell and three anonymous reviewers for comments on a previous version of the manuscript. Analyses were performed on the CIRAD—UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.southgreen.fr/). Funding was provided by the Agence Nationale de la Recherche (ANR SEAD-ANR-13-ADAP-0011).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abu Awad D, Roze D (2018) Effects of partial selfing on the equilibrium genetic variance, mutation load, and inbreeding depression under stabilizing selection. Evol Int J Org Evol 72:751–769

Abu Awad D, Gallina S, Bonamy C, Billiard S (2014) The interaction between selection, demography and selfing and how it affects population viability PLoS ONE 9:e86125

Allard RW (1975) The mating system and microevolution. Genetics 79:Suppl, 115–126

Arrighi J-F, Barre A, Amor BB, Bersoult A, Soriano LC, Mirabella R, Carvalho-Niebel F, de, Journet E-P, Ghérardi M, Huguet T et al. (2006) The medicago truncatula lysine motif-receptor-like kinase gene family includes NFP and new nodule-expressed genes. Plant Physiol 142:265–279

Avise JC, Tatarenkov A (2012) Allard's argument versus Baker's contention for the adaptive significance of selfing in a hermaphroditic fish. Proc Natl Acad Sci 109:18862–18867

Bailey SF, Bataillon T (2016) Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol Ecol 25:203–218

Baker HG (1967) Support for Baker's law-as a rule. Evolution 21:853–856

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

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

Barrière A, Félix M-A (2007) Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. Genetics 176:999–1011

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035

Bomblies K, Yant L, Laitinen RA, Kim S-T, Hollister JD, Warthmann N, Fitz J, Weigel D (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet 6:e1000890

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

Caballero A, Hill WG (1992) Effects of partial inbreeding on fixation rates and variation of mutant genes. Genetics 131:493–507

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205

Charlesworth D, Charlesworth B (1995) Quantitative genetics in plants: the effect of breeding system on genetic variability. Evol Int J Org Evol 49:911–920

Crow JF, and Kimura M (1970). An Introduction to Population Genetics Theory (Harper & Row, New York)

David P, Pujol B, Viard F, Castella V, Goudet J (2007) Reliable selfing rate estimates from imperfect population genetic data. Mol Ecol 16:2474–2487

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11:189–212

Frachon L, Libourel C, Villoutreix R, Carrère S, Glorieux C, Huard-Chauveau C, Navascués M, Gay L, Vitalis R, Baron E et al. (2017) Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. Nat Ecol Evol 1:1551

Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177:905–916

Glémin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc R Soc B Biol Sci 273:3011–3019

Golding GB, Strobeck C (1980) Linkage disequilibrium in a finite population that is partially selfing. Genetics 94:777–789

Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, Picó FX (2011) Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. Mol Ecol 20:3540–3554

Goudet J (2005) hierfstat, a package for R to compute and test hierarchical *F*-statistics. Mol Ecol Notes 5:184–186

Gow JL, Noble LR, Rollinson D, Tchuem Tchuenté L-A, Jones CS (2007) Contrasting temporal dynamics and spatial patterns of population genetic structure correlate with differences in demography and habitat between two closely-related African freshwater snails. Biol J Linn Soc 90:747–760

Haller BC, Messer PW (2017) SLiM 2: flexible, interactive forward genetic simulations. Mol Biol Evol 34:230–240

Hamrick JL, Godt MJW (1997) Allozyme diversity in cultivated crops. Crop Sci 37:26–30

Hartfield M, Bataillon T, Glémin S (2017) The evolutionary interplay between adaptation and self-fertilization. Trends Genet 33:420–431

Hartl D, and Clark AG (1998). Principles of Population Genetics (Sinauer Associates)

Hereford J (2010) Does selfing or outcrossing promote local adaptation? Am J Bot 97:298–302

Hughes PW, Simons AM (2015) Microsatellite evidence for obligate autogamy, but abundant genetic variation in the herbaceous monocarp *Lobelia inflata* (Campanulaceae). J Evol Biol 28:2068–2077

Igic B, Kohn JR (2006) The distribution of plant mating systems: study bias against obligately outcrossing species. Evol Int J Org Evol 60:1098–1103

Ingvarsson PK (2002) A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. Evol Int J Org Evol 56:2368–2373

Kamran-Disfani A, Agrawal AF (2014) Selfing, adaptation and background selection in finite populations. J Evol Biol 27:1360–1371

Kamvar, ZN, Tabima, JF, and Grünwald, NJ (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281

Lande R, Porcher E (2015) Maintenance of quantitative genetic variance under partial self-fertilization, with implications for evolution of selfing. Genetics 200:891–906

Lundemo S, Falahati-Anbaran M, Stenøien HK (2009) Seed banks cause elevated generation times and effective population sizes of *Arabidopsis thaliana* in northern Europe. Mol Ecol 18:2798–2811

Lynch M, Conery J, Bürger R (1995) Mutational meltdowns in selfing populations. Evol Int J Org Evol 49:1067–1080

Marriage TN, Hudman S, Mort ME, Orive ME, Shaw RG, Kelly JK (2009) Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). Heredity 103:310–317

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

Montesinos A, Tonsor SJ, Alonso-Blanco C, Picó FX (2009) Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments PLoS ONE 4:e7213

Neher RA, Shraiman BI (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc Natl Acad Sci 106:6866–6871

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

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

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 89:367–370

Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Mol Ecol 17:3428–3447

Pannell JR, Charlesworth B (2000) Effects of metapopulation processes on measures of genetic diversity. Philos Trans R Soc B Biol Sci 355:1851–1864

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

R Core Team (2018) R: The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi J-M (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

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8:103–106

Rousset F, Gouy A, Martinez-Almoyna C, Courtiol A (2016) The summary-likelihood method and its implementation in the Infusion package. Mol Ecol Resour 17:110–119

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

Siol M, Prosperi JM, Bonnin I, Ronfort J (2008) How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. Heredity 100:517–525

Stebbins GL (1957) Self fertilization and population variability in the higher plants. Am Nat 91:337–354

Stenøien HK, Fenster CB, Tonteri A, Savolainen O (2005) Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. Mol Ecol 14:137–148

Stoffel MA, Esser M, Kardos M, Humble E, Nichols H, David P, Hoffman JI, Poisot T (2016) inbreedR: an R package for the analysis of inbreeding based on genetic markers. Methods Ecol Evol 7:1331–1339

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics 24:2498–2504

Thuillet A-C, Bru D, David J, Roumet P, Santoni S, Sourdille P, Bataillon T (2002) Direct estimation of mutation rate for 10 microsatellite loci in Durum wheat, *Triticum turgidum* (L.) Thell. ssp durum desf. Mol Biol Evol 19:122–125

Trouvé S, Degen, Goudet J (2005) Ecological components and evolution of selfing in the freshwater snail *Galba truncatula*. J Evol Biol 18:358–370

Viard F, Justy F, and Jarne P (1997) Population dynamics inferred from temporal variation at microsatellite loci in the selfing snail *Bulinus truncatus*. Genetics 146:973–982

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121:379–391

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

Weir BS, Hill WG (2002) Estimating *F*-statistics. Annu Rev Genet 36:721–750

Yampolsky C, and Yampolsky H (1922). Distribution of sex forms in the phanerogamic flora (Gebrüder Borntraeger)