Heredity (2010) 105, 358–369

www.nature.com/hdy

ORIGINAL ARTICLE

Using linked markers to estimate the genetic age of a volunteer population: a theoretical and empirical approach

M-F Ostrowski[1], Y Rousselle[1], A Tsitrone[1], S Santoni[1], J David[1], X Reboud[2] and M-H Muller[1]
[1]*UMR DIA-PC 1097, INRA Montpellier, Domaine de Melgueil, Mauguio, France and* [2]*UMR 1210, 'Biologie et gestion des Adventices', INRA, Dijon, France*

Volunteers deriving from unharvested seeds of a crop can lead to persistent feral populations and participate in genetic exchanges across the agro-ecosystem, both between crop varieties and between crops and their wild relatives. A first step to understand the importance of volunteers is to characterize their capacity to reproduce autonomously for several generations. For that purpose, we constructed and evaluated a maximum-likelihood method to estimate the genetic age of a population deriving from one of the most common field crop type: an F1-hybrid variety. The method estimates the number of reproduction cycles that occurred since the cultivation of that variety. It makes use of genotypic data at a number of linked microsatellite loci pairs, thus exploiting the recombination of parental haplotypes, which is expected to occur as the population is reproducing. Estimates with moderate bias and variance were found for a broad range of parameter values in simulations, and the method revealed robust to some deviations from the assumptions of the underlying model. We propose a specific procedure to test the hypothesis of persistence, that is has a given volunteer population experienced more than one cycle of reproduction since the F1-hybrid state? The method was applied to both an experimental and a natural sunflower volunteer population and revealed promising, considering these ideal case studies. Possible further developments toward more complex natural systems are discussed.
Heredity (2010) **105,** 358–369; doi:10.1038/hdy.2009.156; published online 9 December 2009

Keywords: molecular markers; linkage; *Helianthus annuus*; maximum-likelihood estimation; feral populations; gene flow

Introduction

Crops result from the domestication of wild plants species and are often claimed to be maladapted in a natural environment. However, crop-derived plants developing without having been intentionally sown are commonly observed in fallows, field margins and within the field itself. Such plants, referred to as volunteers, originate generally from seed loss at harvest or during transport, the most popular example being oilseed rape (Crawley and Brown, 2004). The population dynamics of volunteers is still poorly documented, but an increased focus was recently given to both their potential for autonomous evolution and their function in genetic exchanges across the agro-system. Volunteers may lead to management problems for farmers. They can also be involved in gene flow across the agro-ecosystem through their contribution to the pollen pool (Devaux *et al*., 2005). For instance, volunteers acting as a genetic bridge between different members of the crop–weed–wild complex (Reagon and Snow, 2006) could impede any desired genetic isolation between GM and non-GM varieties. Moreover, if volunteers naturalize and constitute self-perpetuating feral populations, they could freely evolve and develop weedy characters (Londo and Schaal, 2007; Bagavathiannan and Van Acker, 2008).

The persistence of feral populations has been assessed through multi-year demographic surveys. This way, Crawley and Brown (2004) showed that oilseed rape populations are not self-replacing and that they rely on the introduction of new seeds (for example, losses from trucks). However, using biochemical markers, Pessel *et al*. (2001) showed that some feral populations presented original genetic characteristics that readily differ from any varieties cultivated in the studied area over the last 8 years, showing that a variety can influence the composition of the agro-system many years after its cultivation.

To explore the potential for autonomous genetic evolution, one decisive question can be formulated as follows: is a given volunteer population a 1-year transitory crop descent, or is it resulting from more than one cycle of reproduction? Molecular markers and population genetics tools can be helpful to address these questions. Indeed, methods relying on multilocus genotypic data have been developed to investigate the recent history of natural populations (for example, Cornuet *et al*., 1999; Wilson and Rannala, 2003; Excoffier *et al*., 2005). Cultivated plants differ from their wild relatives in their genetic structure. Nowadays, a typical field crop consists of a population of genetically homogeneous individuals resulting from the initial cross of two inbred

Correspondence: Dr M-H Muller, UMR DIA-PC 1097, INRA Montpellier, Domaine de Melgueil, Mauguio F34130, France.
E-mail: Marie-Helene.Muller@supagro.inra.fr
Received 22 September 2008; revised 9 October 2009; accepted 15 October 2009; published online 9 December 2009

lines (F1 hybrid, for example sunflower, oilseed rape, maize). As a consequence, volunteers deriving from a given field crop will typically develop into an out-of-equilibrium population. Methods of inference have then to be adapted or specifically designed (for example Devaux *et al*., 2005; Devaux *et al*., 2007). Fortunately, one advantage of these cultivated species is that a great deal of molecular information is often available, such as a wealth of mapped molecular markers (Koopman *et al*., 2007).

This paper describes a model-based method to estimate the genetic age of a volunteer population deriving from an F1-hybrid field crop, that is the number of reproduction cycles that occurred since the cultivation of the F1-hybrid individuals. This method relies on the genotypic information from markers linked by a known genetic distance and was applied to both an experimental and natural volunteer population of sunflower. The performances and potential limits of the method depending on various parameters of the model are interpreted and discussed.

Materials and methods

F1-hybrid model: assumptions and bilocus expectations

An F1-hybrid variety results from the cross between two inbred lines. Both parental lines are theoretically genetically fixed and homozygous at all loci; the F1-hybrid variety is thus expected to include a single genotype, heterozygote for the loci that are polymorphic among the parental lines. An interesting property is that for a pair of linked polymorphic loci, perfect association, that is maximum linkage disequilibrium, is expected between alleles inherited from each of the two inbred lines. When the variety reproduces, this linkage disequilibrium decreases from generation to generation and it is possible to give the theoretical expectation of the bilocus genotypic frequencies as a function of both the recombination rate between the loci and the selfing rate of the population.

Considering two loci with two alleles, let the gametic phase (the haploid two-locus genotype in each parental line) be [AB] and [ab]. The genotype of the F1 hybrid is thus AB/ab. There are 10 possible two-locus genotypes {AB/AB, ab/ab, Ab/Ab, aB/aB, AB/Ab, AB/aB, ab/Ab, ab/aB, AB/ab, Ab/aB} in frequencies $\{x_{1,t}, x_{2,t}, ..., x_{10,t}\}$ at generation t. Let r be the recombination rate between the two loci and s the selfing rate of the population. At generation $t+1$:

$$x_{t+1} = (1-s)\varphi(M_1 \cdot x_t) + sM_2 \cdot x_t \quad (1)$$

with x_t corresponding to the column vector of genotypic frequencies at time t, M_1 and M_2 to the transition matrix of x_t because of outcrossing and selfing, respectively, and φ to an application defined from $[0,1]^4$ (that is from four gametes) to $[0,1]^{10}$ (that is to 10-bilocus genotypes). Details are given in the appendix. Using recursion Equation (1) and any given initial state, the expected bilocus genotypic frequencies can be computed for any time t, assuming no genetic drift, no mutation and no gene flow from other populations. A representation of the expected changes in frequencies is given in Figure 1 for contrasting parameter values. By convention, the initial state (F1 hybrid) corresponds to $t=1$. In our model, t will be referred to as the genetic age of the volunteer population.

Presentation of the method

The method considers a sample drawn from a volunteer population and analyzes the observed genotypic frequencies at K independent pairs of physically linked loci. The principle is to compute the probability of the observed frequencies given the genetic age t using the expected bilocus genotypic frequencies at time t (Equation (1)). We defined $\hat{t}$, an estimator of the genetic age, as the t value maximizing the probability of data. Box 1 provides the definitions of all parameters.

Let $D_j = \{n_{1,j}, n_{2,j}, \ldots, n_{9,j}\}$ be the vector of the numbers of sampled genotypes in each genotypic class, for the jth pair of loci. We consider here only nine genotypic classes, because the *trans*-heterozygotes are not distinguishable from the *cis*-heterozygotes when using usual laboratory techniques. D_j follows a multinomial distribution with parameters $n_{\cdot j}$, the total sample size for the jth pair of loci and the vector $\{p_{1,j,t}, p_{2,j,t}, \ldots, p_{9,j,t}\}$, the expected relative frequencies of the nine observable genotypic classes, at time t. For $i=1$ to 8, $p_{i,j,t} = x_{i,t}$ and $p_{9,j,t} = x_{9,t} + x_{10,t}$, where the $x_{i,t}$ are frequencies predicted from Equation (1), given r_j the recombination rate between the two loci of pair j and s the selfing rate of the population. The likelihood of

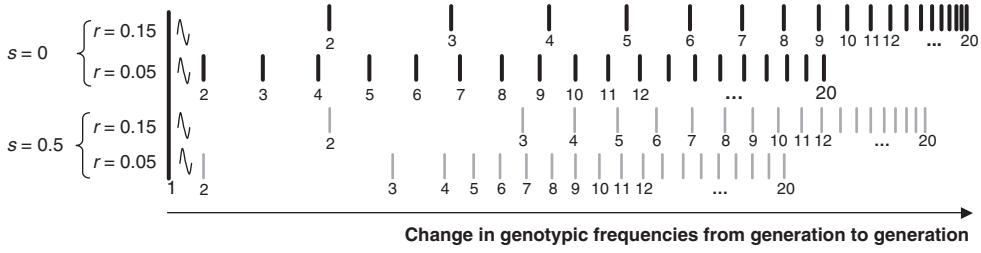

Figure 1 Representation of the change in genotypic frequencies from generation to generation for contrasting recombination and selfing rates. The numbers under vertical bars correspond to the genetic age of the population. The spacing separating vertical bar t and $t-1$ on a line was calculated as: $\sum_{i=1}^{10} |x_{i,t-1} - x_{i,t}|$, where $x_{i,t-1}$ refers to the expected frequency of the ith genotypic class at time $t-1$. Black and gray bars correspond to a population characterized with a selfing rate $s=0$ and $s=0.5$, respectively. Lines 1 and 3 and lines 2 and 4 correspond to a pair of loci characterized with a recombination rate $r=0.15$ and $r=0.05$, respectively. The sinusoidal symbol means that the spacing was truncated by the same quantity. It should be noted that this figure has only a heuristic value, as these distances are truly additive only from a given starting point in time that can be related to the Hardy–Weinberg equilibrium ($t=2$ for $s=0$) or to the approach of inbreeding equilibrium ($t=4$ for $s=0.5$). However, for $s=0.5$, the distance separating $t=2$ from $t=4$ is 96 and 91% as large the sum of the two intermediate intervals, for $r=0.15$ and 0.05, respectively.

Box 1 Definitions of parameters

K	Total number of independent pairs of loci.
n_j	Total number of genotyped individuals at the jth pair of loci ($\sum_{i=1}^{i=9} n_{ij}$).
$n_{i,j}$	Observed absolute frequency of the ith genotypic classes (with $i=1$ to 9) at the jth pair of loci (with $j=1$ to K).
D_j	Vector of the observed absolute frequencies in the nine distinguishable genotypic classes at the jth pair of loci: $\{n_{1,j}, n_{2,j}, \ldots, n_{9,j}\}$.
$p_{i,j,t}$	Expected relative frequency of the ith genotypic classes (with $i=1$ to 9) at the jth pair of loci, at time t.
$\Pr(D_j \mid t)$	Probability of the data D_j, given time t.
r_j	Recombination rate between the two loci of pair j.
r^*	Optimal recombination rate when setting $r_j = r$ for all j.
$\hat{r}_j$	Estimated recombination rate between the two loci of pair j.
s	Selfing rate of the population.
$\hat{s}$	Estimated selfing rate.
N_c	Drift parameter in simulations: number of zygotes effectively contributing to a generation.
n	Total number of genotyped individuals at any pair of loci setting $n_j = n$ for all j.
E	Sampling effort corresponding to the product n times K.
(n, K)	Sampling strategy setting $n_j = n$ for all j.
t	True age of the population.
$\hat{t}$	The estimator of t.
$\hat{t}^*$	An estimation of t.

t is defined as follows:

$$L(t|D_j) = \Pr(D_j|t) = \frac{n_{.j}!}{\prod_{i=1}^{9} n_{i,j}!} \prod_{i=1}^{9} {p_{i,j,t}}^{n_{i,j}} \quad (2)$$

For K independent pairs of loci, the likelihood of t becomes

$$L(t|D_1, D_2, \ldots D_K) = \prod_{j=1}^{K} \Pr(D_j|t) = \prod_{j=1}^{K} \frac{n_{.j}!}{\prod_{i=1}^{9} n_{i,j}!} \prod_{i=1}^{9} p_{i,j,t}^{n_{i,j}} \quad (3)$$

The maximum-likelihood estimator $\hat{t}$ is obtained by a numerical exploration over the interval [2, 40] of $t \in \mathbb{N}^*$.

Simulation study

Effects of the recombination rate, selfing rate, drift and sampling strategy on the variance and bias of the estimator were investigated using samples drawn from simulated populations. A simulated population of age t consisted of a list of K vectors of bilocus genotypic frequencies ($p_{i,j,t}$ in Equation (2)), corresponding to K independent pairs of loci. Hereafter, the term *frequencies* will be used to refer to bilocus genotypic frequencies. Simulation parameters were a selfing rate s, a drift parameter N_c, a vector of K recombination rates r_j and a list of K vectors of initial frequencies $p_{i,j,1}$ (typically, 100% of double heterozygotes). Frequencies for each pair of loci were simulated independently using an iterative procedure mimicking Equation (1) (programmed using Mathematica; Wolfram, 1996): frequencies at time $t+1$ were generated by (i) computing the frequencies in the zygotes at time $t+1$ using frequencies at time t, (ii) random sampling of N_c genotypes in the resulting multinomial frequency distribution and (iii) dividing the resulting vector of genotype numbers by N_c. The drift parameter N_c thus corresponds to the number of zygotes drawn to constitute to the next generation. In a completely outbreeding population, N_c is equivalent to the effective population size N_e, whereas N_e is lower than N_c in a selfing population ($N_c = N_e + N_e\, s/(2-s)$). When simulating a population without drift, the N_c sampling step was omitted: hereafter, this simulation modality will be referred to as the deterministic model. To produce the samples used to estimate t, n individual genotypes were drawn for each pair of loci from the simulated frequencies. We will refer to the product of n times K as the sampling effort E, and to the couple (n, K) as the sampling strategy. Variance and bias of $\hat{t}$ were estimated using 1000 simulated populations for each set of $[r_j, s, N_c, (n, K)]$ parameters.

We first studied the influence of r_j using the deterministic model and defined a value r^* minimizing the variance of $\hat{t}$ (hereafter noted $V(\hat{t})$) across both the explored interval of age t and contrasting values of selfing rates s: {0, 0.25, 0.5, 0.75}. The maximum value of s that was explored was 0.75, because F1-hybrid varieties are seldom developed for highly selfing species (but see Virmani, 1994). The effects of the other parameter values on the behavior of $\hat{t}$ were studied using r^*. For both the drift parameter N_c and the sampling effort E, three values were used: {50, 200, infinite} and {200, 400, 800}, respectively. Two sampling strategies (n, K) were contrasted for $E = 400$: {(40, 10), (80, 5)}. Finally, the explored range of variation of true genetic age t was restricted to [2, 7] mainly because our model assumes that the population is isolated from other sources of pollen, a hypothesis that does not seem reasonable in an agro-system for more than a few generations. However, we present in Supplementary Information a summary of the results obtained for older populations ($t \leqslant 14$).

The method requires estimates of s and r_j. To investigate the robustness of the estimator to errors made on these parameters, we simulated populations using some true parameter values, whereas likelihoods were computed using values deviating from the true values. We explored the effect of the uncertainty of estimates of the recombination rate r arising from mapping studies. We used (Lorieux, 1994)

$$SE_{\hat{r}} \approx \frac{1 + 2\hat{r}}{\sqrt{2}} \sqrt{\frac{\hat{r}}{L}} \quad (4)$$

as an approximation of the standard error of this parameter, with L standing for the number of recombinant inbred lines used to estimate r (Tang *et al*., 2002). The vector of recombination rates used for the simulations was obtained by drawing the K r_j values from a normal distribution with parameters (r^*, SE_{r^*}), whereas $\hat{t}$ was estimated using r^*. For the selfing rate s, we focused on the downward bias, which typically results from using F_{IS}-based estimates under a false assumption of inbreeding equilibrium (see below as well as Jarne and David, 2008).

Hypothesis testing and confidence interval

When addressing the question of whether or not a volunteer population is self-perpetuating, the relevant

point is to determine whether it has experienced more than one cycle of reproduction, the first cycle having taken place in the cultivated field. In other terms, if adult volunteer plants have been sampled, we need to test if $t>2$, and if the sample consists of seeds produced by the volunteers, we need to test if $t>3$.

For that purpose, we propose constructing the empirical distribution of $\hat{t}$ under the appropriate null hypothesis H_0, including the uncertainty on the parameters r_j and s. The H_0 distribution is obtained by estimating t in a large number of simulated populations (say $\geqslant 1000$) of age $t=2$ (or $t=3$). Each population is simulated using the same sampling strategy (n, K) than the studied population and drawing randomly s and all r_j values in a normal distribution with parameters ($\hat{s}$, $SE_{\hat{s}}$) and ($\hat{r}_j$, $SE_{\hat{r}j}$), respectively. The critical value t_{crit} of the H_0 distribution is then determined, so that the density of probability of all $\hat{t}\geqslant t_{crit}$ is $\leqslant 5\%$. If the estimated age noted $\hat{t}^*$ is $\geqslant t_{crit}$, then the test is considered significant at the 5% level.

The effect of drift can be included in the testing procedure. However, because estimates of the effective population size N_e are rarely available, we propose to compare H_0 distributions obtained simulating populations with contrasting values of the drift parameter N_c.

We compared the H_0 distributions obtained with 1000 simulated populations, setting (n, K) = (40, 10) and $r_j = r^*$ for all j, and all possible combinations of $N_c \in \{50, 200, \text{infinite}\}$, $s \in \{0, 0.5\}$, $SE_r \in \{0, 0.036\}$ and $SE_s \in \{0, 0.1\}$ (for $s=0.5$ only). The type I error (the false positive rate) and the power of the test (1 minus the false negative rate) were empirically determined for samples drawn from populations of true age $t \in [3, 7]$ simulated using all the above sets of parameter values.

The change in type I error arising from using a downwardly biased F_{IS}-based estimates of selfing rate was empirically determined for a specific case. We simulated samples from $t=2$ and 3 populations using $s=0.75$. F_{IS}-based estimates would typically yield a value of $\hat{s}=0.55$ in the $t=3$ samples. We then applied the testing procedure to these samples by simulating the appropriate H_0 distributions using $\hat{s}=0.75$ vs $\hat{s}=0.55$. Other simulation parameter values were (n, K) = (40, 10), $SE_r = 0.036$, $SE_s = 0.1$.

In addition, the 95% confidence interval (CI) of $\hat{t}^*$ may be constructed by estimating t in a large number of simulated populations of successive ages: $t=\hat{t}^*-1$, $\hat{t}^*-2$, … ($\hat{t}^*+1$, $\hat{t}^*+2$, …, respectively). The lower and upper bound of the CI is then the smallest and largest, respectively, t value for which $<97.5\%$ of the simulations yielded estimates $<\hat{t}^*$ and $>\hat{t}^*$, respectively.

Empirical study system

Almost all sunflower (*Helianthus annuus*) varieties cultivated nowadays in Western Europe are F1 hybrids. As opposed to the wild auto-incompatible *H. annuus*, these sunflower varieties can self at an unknown rate (Gandhi *et al*., 2005).

We applied the method to two volunteer populations of sunflower. We first analyzed three generations of an experimental population conducted as follows. In 2001, the F1-hybrid variety Prodisol (DEKALB) was cultivated in a 0.6 ha field of the experimental domain of INRA (Epoisses, France). The harvest of this variety was called generation G2. In 2002, soybean was sown on this field and volunteer sunflowers grew at a density varying from 0.5 to 4 plants per m². These volunteers originated from seeds lost at harvest and were then representative of generation G2. The harvest of 335 of these plants was bulked and constituted generation G3. During these 2 years, no other sunflower field occurred at a distance of <400 m. In 2004, 350 seeds of generation G3 were sown under a pollen-proof tunnel at INRA Mauguio (France). Honeybees were introduced during the flowering period to ensure free intercrossing of the plants. The harvest of these plants was bulked and constituted generation G4.

In 2004, a natural volunteer population (FR001) was sampled in a fallow close to Saint Laurent d'Aigouze (France). Hundreds of volunteers occurred at varying densities on an area of ~ 3 ha. Independent maternal families were sampled from 44 plants all over the population.

About 60 to 71 seeds per generation (G2 to G4) and one seed per maternal family for FR001 were sown for analysis. DNA was isolated from about 100 mg of plant leaves according to the Dneasy Plant Mini kit (Qiagen, GmbH, Hilden, Germany) with the following modification: 1% of polyvinylpyrrolidone (PVP 40 000) was added to buffer AP1.

Both independent microsatellite loci and pairs of loci located at different genetic distances were selected from Tang *et al*. (2002). Care was taken to choose loci with no earlier evidence of null allele (Tang and Knapp, 2003; Tang *et al*., 2003). Twenty-three loci were screened on eight G2 individuals and only polymorphic loci were retained. Seventeen loci have been used on the whole data set (Table 1).

The amplification reaction consisted of 50 ng DNA, 4 pmol of unlabeled reverse primer, 2 pmol of forward primer, fluorescently labeled with NED, HEX or FAM, 1 × reaction buffer, 2 mM $MgCl_2$, 200 μM dNTP, 0.25U Taq DNA polymerase in a total volume of 25 μl. The amplification method was 95 °C for 2 min, 36 cycles of 94 °C for 30 s, Tx for 30 s (Tx is initially 63 °C and

Table 1 Microsatellite loci analyzed in this study

Locus	*Linkage group*	*Position (cM)*	*Independent loci*	*Loci pairs for AEM in G2–G4*	*Loci pairs for AEM in FR001*
ORS610	1	3.4	x		
ORS371	1	44.2	x		
ORS423	2	1.7		A	A
ORS925	2	5.9	x	A	A
ORS432	3	42.3	x		
ORS309	4	75.5	x	B	
ORS674	4	100.8		B	
ORS887	9	38.4	x		
ORS437	10	59.9		C	B
ORS380	10	69.7	x	C	B
ORS1085	12	71.7	x		
ORS899	16	0		E	C
ORS303	16	10.2	x	D,E	
ORS656	16	26.1		D	C,D
ORS788	16	46.2	x		D
ORS297	17	29.1	x	F	
ORS677	17	57.1		F	

Abbreviation: AEM, age estimation method.
Independent loci indicates the loci used for average genetic statistics estimation. The last two columns report the loci pairs used for the AEM in each population. Each capital letter denotes a particular pair.

decreases to 1 °C per cycle for the six first cycles, until it reaches 57 °C) and 72 °C for 45s, followed by a final extension for 20 min at 72 °C. Electrophoresis was performed on an ABI 3130xl Genetic Analyser. Samples were prepared by adding 3 µl of diluted PCR products to 6.875 µl formamide and 0.125 µl of GenScan 400HD Rox size standard. The GENEMAPPER software (Applied Biosystems, Foster City, CA, USA) was used to analyze the DNA fragments and to score the genotypes.

Data analysis

Mean number of alleles per locus, multilocus heterozygosity and the fixation index F_{IS} were estimated for each generation over a subset of 11 independent loci, separated by at least 36 cM according to Tang *et al*. (2002) (Table 1), using GENETIX (Belkhir *et al*., 2001). The significance and the sampling variance of the estimated F_{IS} values were assessed using 1000 permutations of alleles among individuals and a jackknife procedure over the loci, respectively.

For both the experimental and natural volunteer population, estimates of s and approximated standard errors were computed on the assumption of inbreeding equilibrium (Jarne and David, 2008), namely

$$\hat{s} = \frac{2F_{IS}}{1 + F_{IS}} \tag{5}$$

$$SE_{\hat{s}} \approx \sqrt{\frac{4\mathrm{Var}(\hat{F}_{IS})}{(1 + \hat{F}_{IS})^4}} \tag{6}$$

To determine the magnitude of the directional error that was made when assuming inbreeding equilibrium, the selfing rate between two successive generations of the experimental population was also estimated using these F_{IS} values and the recurrence formula on the deterministic evolution of F_{IS} (Crow and Kimura, 1970):

$$F_{IS}(t+1) = \frac{s}{2}(1 + F_{IS}(t)) \tag{7}$$

For the experimental population, the variance effective population size N_e was estimated using the temporal variation of allelic frequency at the 11 independent loci. We used the estimator of Waples (1989) based on F_c, the standardized variance in allelic frequencies between sampled generations. The 95% CI on N_e was obtained from the simulation of the actual distribution of F_c based on the estimated N_e, as suggested by Goldringer and Bataillon (2004). This method was implemented in the program kindly provided by Mathieu Siol and described in Siol *et al*. (2007). N_e was estimated for two pairs of samples: G2 and G3, G3 and G4. Rare alleles were pooled. For the natural population of volunteers FR001, we used 1/10 of the demographic size as an order of magnitude of N_e (Frankham, 1995), using one individual per 5 m² as a conservative estimate of the mean density.

Application of the age estimation method

We first discarded genotypes carrying rare alleles, which were present in only one or two individuals over the whole data set; they were interpreted as contamination (gene flow from distant sunflower field) or mutation.

The genetic model underlying the estimation method assumes that all loci are biallelic with equal initial allelic frequencies. However, many loci displayed more than two alleles and allelic frequencies were sometimes unbalanced. We interpreted these deviations as the consequence of incomplete fixation of the inbred lines used to produce the variety, as already described by Zhang *et al*. (2005). Indeed, the fixation of the parental lines is generally assessed by breeders using morphological rather than molecular markers. To uncover the haplotypic structure of the F1 hybrids, we first determined parental haplotypes and their frequencies by estimating correlation coefficients between pairs of alleles at different loci within each linkage group (LINKDIS program implemented in GENETIX, Garnier-Gere and Dillmann, 1992). We then fused some allelic classes to transform parental haplotypes into tractable initial bilocus genotypes in the F1 generation as illustrated in Figure 2a. This transformation was possible for six pairs of loci; that is after fusing the appropriate allelic classes, the loci of these pairs were biallelic with balanced allele frequencies. These six pairs involved 11 different loci (Table 1). The number of observations for each bilocus genotypic class (that is the $n_{i,j}$ in Equation 3)) was computed and used as the input for the application of the age estimation method.

The same computations were made for the natural volunteer populations. Only four loci pairs were compatible with the genetic model and were thus used for the estimation of the genetic age (Table 1).

For each sample, we determined the 95% CI of t, and tested the null hypothesis '$t=2$' for the experimental population or '$t=3$' for the natural population, as described above. The standard error of each r_j value was approximated setting $L=94$ (Tang *et al*., 2002) in Equation (4).

Interpretation of the genotypic data and robustness to deviation from the assumed genetic structure of the variety

The genotypic frequencies with several loci presenting three alleles could not fit the hypothesis of a pure F1 hybrid. To further explore the origin of the unexpected polymorphism in the experimental volunteer population, we first looked for significant associations between the less common alleles (frequency between 10 and 20%) even if the corresponding loci were not mapped on the same linkage group. Visual investigation of the data set showed that the multilocus genotypes were compatible with the following hypothesis: the F1 hybrid in the field was actually a mixture of two F1-hybrid genotypes, that is a prevailing one (Prodisol) together with a less abundant one (hereafter referred to as 'contaminant'). Their multilocus genotypes could be reconstituted (for example, Figure 2b). The multilocus genotypes in G2 could then be partitioned into three groups of progenies: Prodisol, contaminant and intercrossed between the two varieties. As this interpretation contrasted with the assumptions made on the genotype of the variety, we evaluated its consequences on the outputs of the age estimation method. We determined the actual composition of the initial field using the same recoding strategy of the data than above and simulated populations deriving from such field. Namely, we considered a

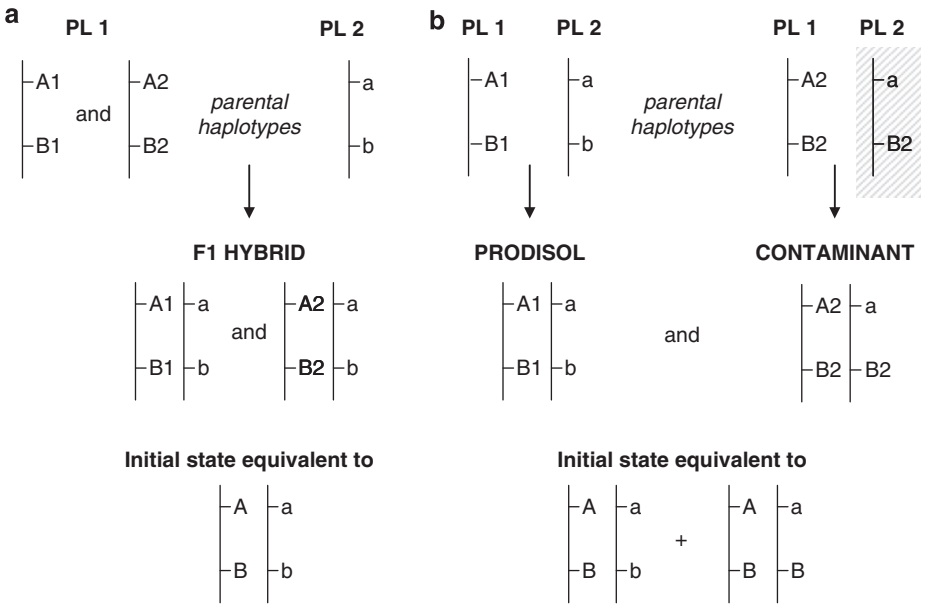

Figure 2 Two contrasting interpretations of the diversity detected in the experimental volunteer population: example with two triallelic loci. For each locus, one allele (a and b here) is approximately in frequency 50%. (**a**) F1-hybrid variety with a polymorphic parental line (PL). Three parental haplotypes are present. When fusing alleles A1 and A2, and alleles B1 and B2, the situation is equivalent to the basic model. (**b**) Mixture of two F1-hybrid varieties. One supplementary parental haplotypes (aB2) is present, although rare. It is not detected by the linkage disequilibrium analysis because it is masked by the more frequent haplotype ab. When fusing alleles A1 and A2, and alleles B1 and B2, the structure of the initial state deviates from the assumptions of the model.

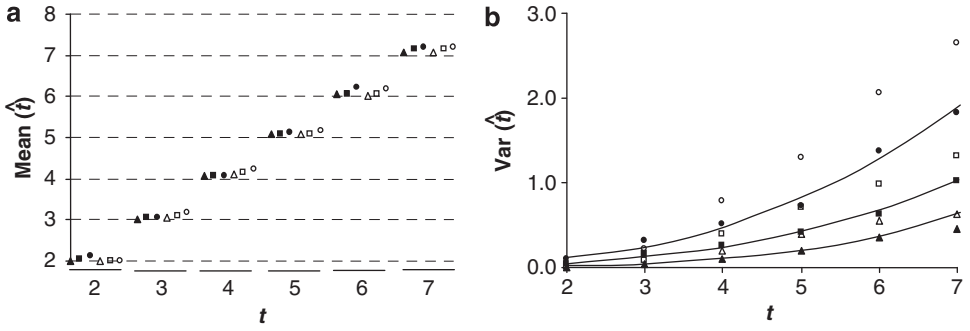

Figure 3 Mean (**a**) and variance (**b**) of $\hat{t}$ for successive values of t, for contrasted values of the sampling effort E and the selfing rate s. Symbols ▲■● and △□○ refer to $s=0$ and $s=0.5$, for $E=800$, 400 and 200, respectively. Populations were simulated using the deterministic model and a recombination rate $r_j=0.15$. Deviations from the horizontal dotted lines correspond to bias in (**a**). Polynomial curves were fitted to the $s=0$ plots in (**b**) to facilitate the reading of the figure.

variety that was an admixture of two F1 hybrids in frequencies 80/20%. The genotype of variety 1 and 2 were AB/ab and AB/aB, respectively (as in Figure 2b) for four pairs of loci, ab/ab and AB/AB for one pair, and AB/ab and AB/ab for the last one. Samples were simulated in accordance with our actual dataset ($(n, K)=(60, 6)$) and analyzed using the age estimation method (the likelihoods were computed under the assumption of a pure F1 hybrid).

Results

Simulation results

Empirically determined bias and variance: Simulations showed that $\hat{t}$ was essentially an upwardly biased estimator, although only slightly for most explored sets of parameter values (Figure 3a). Bias tended toward zero when increasing E under the deterministic model, showing that $\hat{t}$ was an asymptotically unbiased estimator (for example Figure 3a). The magnitude of the bias of $\hat{t}$ was positively correlated to its variance (not shown). A brief description of the empirical distribution of $\hat{t}$ across different sets of parameter values is given in Table 2 for $E=400$. For all explored values of $t\leqslant 7$, $s\leqslant 0.75$ and $r_j\leqslant 0.15$, the observed range of bias was $[-0.04, 0.54]$. The variance of $\hat{t}$ (also noted $V(\hat{t})$), behaved as a strictly non-linear increasing function of t (Figure 3b).

The recombination rate within pairs of loci r_j had a considerable effect on $V(\hat{t})$. As illustrated in Figure 4, for a given set of other parameters values, there is an r_j value minimizing $V(\hat{t})$, a value referred to as optimal r. The optimal r behaved as a decreasing function of t, but also as a slightly increasing function of the selfing rate s. We

Table 2 Range, median, first and third quartile [25, 75%] of the empirical distribution of bias and variance of $\hat{t}$, across different set of parameters using the deterministic model and $E = 400$

	s = 0	s = 0.25	s = 0.5	s = 0.75
Range of bias	[−0.04, 0.15]	[−0.02, 0.25]	[−0.03, 0.23]	[0.00, 0.54]
Range of $V(\hat{t})$	[0.04, 1.02]	[0.00, 1.38]	[0.00, 2.25]	[0.00, 4.57]
Median of bias	0.05	0.05	0.08	0.19
Median of $V(\hat{t})$	0.39	0.49	0.69	0.80
Bias [25, 75%]	[0.04, 0.07]	[0.03, 0.09]	[0.02, 0.13]	[0.02, 0.26]
$V(\hat{t})$ [25, 75%]	[0.15, 0.58]	[0.15, 0.70]	[0.12, 1.04]	[0.02, 1.85]

For a given value of the selfing rate, the distribution pools estimates obtained from simulations performed for $t = 2$ to 7, and r_j varying from 0.025 to 0.15 by step of 0.025. Each set of parameters was replicated three times (1000 populations each time).

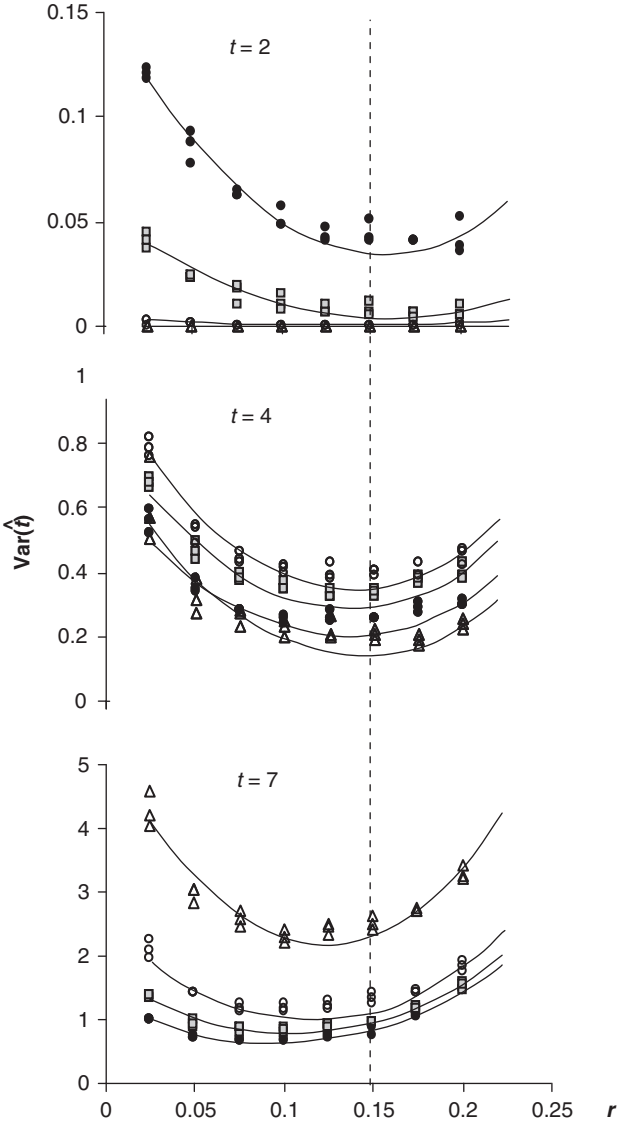

Figure 4 Variance of $\hat{t}$ as a function of recombination rate r_j for different values of t (top to down figures) and selfing rate s. Populations were simulated using the deterministic model and a sampling effort $E = 400$. Symbols ● ■ ○ △ correspond to $s = 0$, 0.25, 0.5 and 0.75, respectively. Polynomial curves were fitted to the plots to facilitate the reading of the figure. The vertical doted line corresponds to $r_j = r^* = 0.15$.

chose $r^* = 0.15$ as an approximation of the value associated to minimum $V(\hat{t})$ across the explored range of both the age $t \in [2,7]$ and selfing rate $s \in \{0, 0.25, 0.5, 0.75\}$ (Figure 4). All following results were obtained using r^*.

The selfing rate s affected the variance of $\hat{t}$ in an age-dependent manner. For small values of t, higher selfing rates resulted in lower $V(\hat{t})$. However, this relationship was reversed for larger values of t (Figure 4). The age at which the relationship switched was larger for higher selfing rates and can be related to the approach of inbreeding equilibrium. This equilibrium corresponds to the equilibrium frequency of heterozygous genotype at single loci; it depends on s and is reached earlier for lower values of s. For instance, at $t = 4$ for $s = 0.5$ and at $t = 5$ for $s = 0.75$, we noted that inbreeding equilibrium was virtually reached (that is differences between theoretical frequencies at that age and equilibrium frequencies are less than sampling noise for $E = 400$).

Increasing the sampling effort E or the drift parameter N_c resulted in a reduction of $V(\hat{t})$ and the associated 95% envelope of the distribution of the estimator (Figures 3b and 5). For a given value of E, $V(\hat{t})$ revealed sensitive to the sampling strategy (n, K), but only for populations simulated with drift ($N_c \neq$ infinite): in this case, increasing the number of pairs of loci resulted in a decrease of $V(\hat{t})$ (Figure 5).

Robustness to deviations from true parameter values: Overestimating the selfing rate s resulted in underestimating t. Conversely, underestimating s led to overestimating t, but only until inbreeding equilibrium was virtually reached; beyond this point, t was underestimated (Figure 6a). Estimating s using F_{IS} values results in an underestimation when populations are not at inbreeding equilibrium (Jarne and David, 2008). When using these downwardly biased estimates instead of the parametric s, an overestimation of t is then expected. As illustrated in Figure 6b, an excess of positive bias is indeed observed; this excess was the most pronounced for high selfing rates at early stages ($t = 3$ and 4).

Underestimating and overestimating all r_j values resulted in overestimating and underestimating, respectively, t (not shown). The magnitude of this error increased with the age of the population.

When populations were simulated drawing all r_j values from a normal distribution with parameters $\mathcal{N}$ (r_{Op}, SE ≠ 0), $V(\hat{t})$ increased relatively to simulations performed using SE = 0, and a negative bias was then observed (Figure 7).

Finally, simulating populations deviating from the hypothesized genotypic structure of the F1 hybrid as described in Materials and methods yielded a positive bias from $t = 2$ to 4, but a negative bias from $t = 5$ to 7 (Supplementary Figure S2).

Hypothesis testing and CIs: For the explored parameter values, the '$t = 2$' and '$t = 3$' H_0 distributions were only slightly affected by the value of N_c, and by the uncertainty on r_j and s (not shown, but see Figure 5). Nevertheless, the slight increase in variance of the H_0 distributions changed or come close to changing the critical value t_{crit}, thus reducing the power of the test.

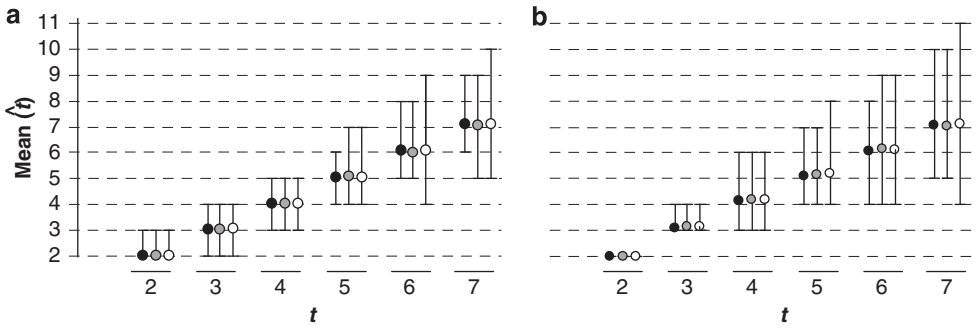

Figure 5 Mean of $\hat{t}$ for successive values of t estimated from populations simulated using different values of the drift parameter (N_c) and different sampling strategies. The sampling effort $E=(N * K)$ was fixed to 400, and all r_j to 0.15. Panels correspond to selfing rates $s=0$ (**a**) and $s=0.5$ (**b**), respectively. Symbols ● ◍ ○ are refer to $N_c=$ infinite; $N_c=200$, $(N, K)=(40, 10)$; $N_c=200$, $(N, K)=(80, 5)$, respectively. Vertical bars correspond to the 95% envelope of the empirical distribution of $\hat{t}$.

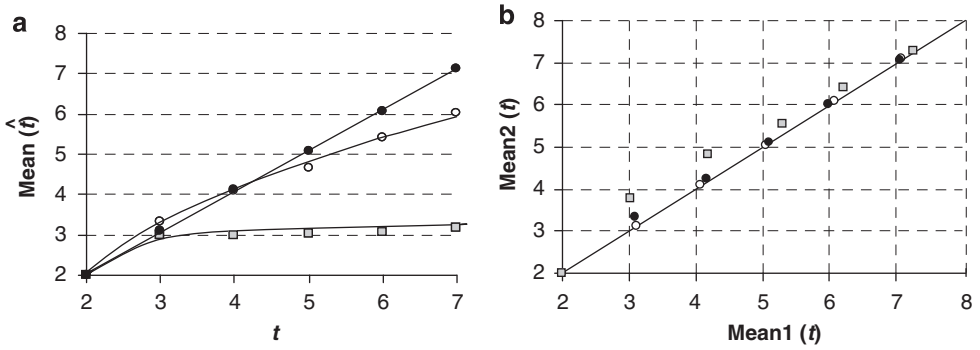

Figure 6 Robustness to directional errors on selfing rate. (**a**) Mean of $\hat{t}$ estimated for successive values of t from populations simulated with a selfing rate of $s=0.5$, but using $s=0.25$, 0.5 and 0.75 (○ ● ■, respectively) to compute the likelihoods. The $s=0.5$ plot corresponds to the case for which the age was estimated without error on the selfing rate. Polynomial curves were fitted to facilitate the reading of the figure. (**b**) Mean of $\hat{t}$ estimated from populations simulated using $s=0.25$, 0.5 and 0.75 (○ ● ■, respectively), but using either the parametric selfing rate (Mean 1 ($\hat{t}$)) or the F_{IS}-based estimate (Mean 2 ($\hat{t}$)) to compute the likelihoods. The F_{IS}-based estimates were calculated using Equation (5) and the expected F_{IS} values at age t. Dots falling on the $y=x$ line correspond to cases yielding equivalently biased estimates.

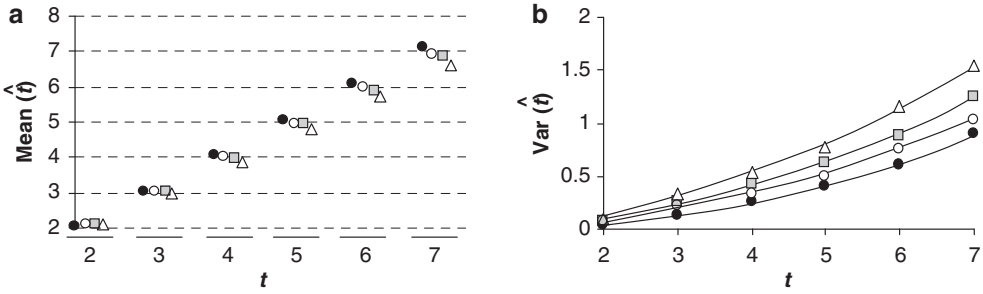

Figure 7 Robustness to randomly distributed errors on recombination rates. Mean (**a**) and variance (**b**) of $\hat{t}$ estimated from populations simulated for successive values of t. All sets of simulations were performed sampling r_j in a normal distribution $\mathcal{N}$ (0.15, s.e.). Other parameters were fixed to $s=0$, $N=40$ and $K=10$. Symbols ● ○ ■ △ correspond to s.e. = 0, 0.027, 0.036 and 0.05, respectively, corresponding to $L=$ infinite, 200, 100 and 50 (see Materials and methods for details). In (**a**), deviations from the horizontal dotted lines correspond to bias. In (**b**), polynomial curves were fitted to the plots to facilitate the reading of the figure.

All together, simulations showed that the tests were reasonably powerful even in the presence of considerable drift and that the power was greater under partial selfing (Table 3). The H_0 '$t=2$' and '$t=3$' could always be rejected whenever $\hat{t}^* \geqslant 4$ and $\hat{t}^* \geqslant 5$, respectively, with an empirically determined type I error of $P<0.015$ and $P<0.04$, respectively, Table 3. The H_0 '$t=2$' distribution exhibited a lower variance than the H_0 '$t=3$' one. Accordingly, the power of the test under the former H_0 was shown considerably higher than under the latter (Table 3).

The lower bound of the CI was frequently equal or 1 year less than that of the 95% envelope of $\hat{t}$ for simulated populations of $t=\hat{t}^*$. In contrast, the CI's upper bound was frequently considerably larger than the corresponding bound of the 95% envelope (not shown).

When using the downwardly biased estimates $\hat{s}=0.55$ instead of the parametric value $s=0.75$, the change in

Table 3 Empirically determined power of test and type I error associated to the null hypothesis H_0 '$t=2$' and '$t=3$' obtained for successive values of t, different values of the drift parameter N_c and selfing rate s

t	N_c	H_0 't=2'				H_0 '$t=3$'			
		s=0		s=0.5		s=0		s=0.5	
			Type I		Type I		Type I		Type I
3	∞	+ 0.126		0.995		—	—	—	—
4	∞	+ 0.788		1.000		0.164		0.282	
5	∞	+ 0.986	0.001	1.000	0.001	0.719	0.002	0.669	0.006
6	∞	+ 1.000		1.000		0.964		0.898	
7	∞	+ 1.000		1.000		0.997		0.971	
3	200	0.148		1.000		—	—	—	—
4	200	0.761		1.000		0.185		0.296	
5	200	0.979	0.000	1.000	0.000	0.705	0.004	0.644	0.012
6	200	0.998		1.000		0.932		0.857	
7	200	1.000		1.000		0.991		0.966	
3	50	0.182		0.982		—	—	—	—
4	50	0.723		1.000		0.232		0.343	
5	50	0.951	0.000	1.000	0.006	0.616	0.011	0.618	0.039
6	50	0.995		1.000		0.866		0.817	
7	50	0.999		1.000		0.964		0.919	

The sampling strategy was set to $(N, K)=(40, 10)$, all $r_j=r^*$, $SE_r=0.036$ and $SE_s=0.1$ (for $s\neq 0$). Each distribution was obtained independently simulating 1000 populations.

type I error associated to the H_0 '$t=2$' was negligible. Conversely, the type I error associated to H_0 '$t=3$' increased from 0.002 to 0.107.

Empirical results

The number of alleles per locus varied from 2 to 5 in the experimental population and from 2 to 6 in the natural population. Some of these alleles were observed only once or twice. Diversity statistics estimated over 11 independent loci are presented in Table 4.

Estimates of selfing rates are presented in Table 4. Interestingly, the selfing rates estimated in the experimental and the natural populations were both non-null and of the same order of magnitude (both about 0.4). It was not possible to estimate s between F1 and G2: indeed, as an F1-hybrid variety is theoretically composed of a unique, heterozygote genotype, the expectancy of genotypic frequencies resulting from pure selfing or pure outcrossing of the F1 individuals is the same. Accordingly, F_{IS} was not significantly different from zero in G2 (Table 4).

Estimates of effective population size are 54.0 [10–infinite] between G2 and G3 and 55.4 [12–infinite] between G3 and G4. No upper limit on the CI was obtained, indicating that the sampling variance was too large relative to the genetic drift. We, therefore, estimated CIs and tested the '$t=2$' null hypothesis considering $N_c=50$ and 200, respectively.

The genetic age and the corresponding CIs estimated using two contrasting values of the drift parameter ($N_c=50$, 200) in the three successive generations of the experimental population were $\hat{t}=$ 2 {[2, 2], [2, 2]}, 3 {[2, 7], [3, 5]} and 5 {[3, 19], [3, 11]}, for G2, G3 and G4, respectively. The '$t=2$' null hypothesis was rejected for both the G3 and G4 samples, but only for $N_c=200$ for the G3 sample.

Table 4 Genetic diversity statistics and F_{IS}-based estimates of selfing rate

Sample	H_e	A	F_{IS}	$\hat{s}_{eq}$	$\hat{s}_{rec}$	N
G2	0.465	2.64	−0.023 NS	—	—	70
G3	0.440	2.55	0.195***	$0.326_{(0.065)}$	0.40	65
G4	0.460	2.55	0.268***	$0.423_{(0.081)}$	0.45	60
FR001	0.449	3.00	0.242***	$0.389_{(0.07)}$	—	44

Symbols H_e, A and N stand for multilocus heterozygosity, mean number of alleles per locus and sample size, respectively. Symbols $\hat{s}_{eq}$ and $\hat{s}_{rec}$ refer to the selfing rate estimated assuming inbreeding equilibrium and to the % of selfed individuals estimated using recursion Equation (7), respectively. Numbers in parenthesis correspond to standard errors. Multilocus F_{IS} values were obtained using 11 independent loci (NS, non-significant; ***$P<0.001$).

For the FR001 population, the drift parameter was set to $N_c=600$ and infinite, respectively. The genetic age and the corresponding CIs were estimated to $\hat{t}=$ 4 {[2, 11], [3, 9]}. The $t=3$ null hypothesis could not be rejected in either case.

Discussion

Parameters influence on the efficiency of the method

Simulations showed that $\hat{t}$ was slightly positively biased and that its variance was increasing from generation to generation with a pattern depending on both the recombination and selfing rates. Interestingly, these results may be explained by the expected dynamics of genotypic frequencies as depicted in Figure 1. One important feature of this figure has to be pinpointed: the size of the steps (that is the magnitude of expected frequencies differences between successive generations) is decreasing when the population is aging. As a consequence, for any given age t, the step is larger between t and $t-1$ than between t and $t+1$.

These features can be expressed in terms of bias and variance of the estimator. Namely, for a population of age t, random samples will more frequently be assigned to generations deviating from the true age when the steps separating the generations are smaller. Then the decreasing size of the steps necessarily results in an increase of the variance of $\hat{t}$ with time. This also explains the increase in type I error between the tests of the '$t=2$' and '$t=3$' null hypothesis. Moreover, if we assume that the sampling process generates symmetrically distributed deviations, more samples are expected to be assigned to $\hat{t}>t$ than to $\hat{t}<t$, making $\hat{t}$ an upwardly biased estimator, as was mainly observed in the simulations.

A similar reasoning can explain how recombination rate and selfing rate affect the estimator. Considering recombination rate, Figure 4 shows that a careful choice of the genetic distance between markers of a pair is crucial. As the true age of a sampled population is unknown, it is impossible to choose a universal optimal recombination rate. However, as long as the goal is to address the self-perpetuation of these populations by testing if $t>2$ or 3, the method performed well using optimal recombination rates found over the $t\in[2, 7]$, which were approximately between $0.1\leqslant r_j\leqslant 0.15$. They could be chosen between $0.05\leqslant r_j\leqslant 0.10$ if relatively isolated older populations were to be expected ($t\in$ [8,14], Supplementary Figure S1).

Sampling effects

The results showed that for a fixed sampling effort E, the sampling strategy (n, K) can affect $V(\hat{t})$: for finite values of N_c, doubling the number of pairs of loci K was associated to lower variance than doubling the number of genotyped individuals n. This effect may be explained by the action of drift, which generates random deviations of the genotypic frequencies independently at each pair of loci. Increasing K reduces the probability of estimating frequencies that were drift deviated in the same direction, which in turn reduces the discrepancy of estimated ages.

In practice, natural populations of volunteers are often characterized with rather small effective population sizes and thus are likely affected by a consequent amount of drift. More accurate estimations and powerful tests can then be obtained by favoring a large number of loci pairs for a given sampling effort.

Robustness to deviation from assumed parameter values

Most of the considered deviations affecting s led to underestimating the genetic age, making our testing procedure generally conservative. A large underestimation of the selfing rate will, however, occur when using F_{IS}, as long as the population is far from inbreeding equilibrium (Figure 6), which may lead to a spurious rejection of the null hypothesis ('$t=3$'). One solution to solve this problem is to analyze the distribution of the individual level of heterozygosity, which allows estimating the selfing rate in the earlier generation while relaxing the assumption of inbreeding equilibrium (Enjalbert and David, 2000). We implemented such an approach on the G3 and G4 samples of our experimental population and the resulting estimated selfing rates were found very similar to those estimated using the F_{IS} recursion Equation (7) (not shown).

Directional deviations from true recombination rates may induce large errors on the estimation (not shown). However, as long as several pairs of loci are used, this risk may be substantially reduced. Simulating normally distributed deviations of recombination rate yielded an essentially negative bias and an increased variance. This underlines the need to use information from accurate genetic maps.

Moreover, we showed that the uncertainty on s and r can be taken into account when constructing the null hypothesis and/or the 95% CI of the estimated age, which reduces the risk of rejecting spuriously H_0. Simulating H_0 using a conservative value of N_c is an appropriate way to take the uncertainty on N_e into account, at the detriment of the power of test. A careful observation of the population on the sampling site may provide valuable insights about this parameter (Frankham, 1995).

Application to empirical data

The method yielded a correct estimation of the genetic age of the two first generations (G2 and G3) of our experimental population of volunteers. For the third generation (G4), the true age was included in the 95% CI. For G3 and G4, the '$t=2$' null hypothesis was appropriately rejected in both cases. The genetic age of the FR001 natural population was estimated to $\hat{t}^*=4$. As we used genotypic data obtained from the offspring rather than from the volunteer plants, the appropriate null hypothesis was '$t=3$', which could not be rejected even when using a moderately large value of N_c (600) and possibly underestimating s. We, therefore, cannot exclude that the plants actually observed in this fallow were just first-generation offspring of the plants earlier cultivated in this field. As the variance of the H_0 '$t=2$' is smaller than the H_0 '$t=3$', genotyping the plants sampled in the field rather than their offspring (seedlings) might have been a more powerful procedure. However, this result is compatible with the emerging picture of the distribution of volunteer populations of sunflower in studied French areas. Indeed, recent expedition trips have shown that such populations are rather both small and rare (Muller *et al*., 2006), contrasting with other countries such as Argentina (Cantamutto *et al*., 2008). The restricted place left in French agro-systems and also the limited spontaneous seed shattering of cultivated sunflower compared with, for example, oilseed rape probably lowers the potentiality of self-perpetuation of a volunteer population.

Imperfectness of empirical data always raises the question of the applicability of estimation methods, and perhaps more acutely when these tools are model based. To this regard, the age estimation method seemed surprisingly robust in several aspects. Indeed, despite the low effective population size, and the moderate number of markers that were used when compared with what is available in crop species such as sunflower (for example Kane and Rieseberg, 2008), the method performed reasonably well on real populations of known age (G2, G3 and G4). In addition, deviations from the theoretical genetic structure revealed tractable, provided that the data are appropriately transformed. To this respect, the consistent results obtained using samples from the experimental population are informative in two ways: (i) pooling haplotypes from distinct hybrids of the same genetic age does not distort the information as long as the parental haplotypes can be recognized. This suggests that reliable estimations could be obtained using data corresponding to more complex varietal structure (for example three ways hybrid in maize or more complex mixture of F1 hybrids sown the same year) and (ii) the elimination of rare alleles does not hinder the estimation procedure or the result.

Nevertheless, migration rates may be sufficiently large (especially through pollen import) to rapidly dilute the information about the original founders of a volunteer population. To explore the robustness of our method, we modeled a situation in which a small number of migrating gametes sharing alleles with the resident volunteer population were arriving each generation (not shown). Our simulations showed that for small migration rate (<0.15), removing all genotypes carrying the less frequent alleles (that is keeping only genotypes carrying the two most frequent alleles at each locus of any given pair) before estimating genetic age yielded consistent estimates of the genetic age of the resident population. The phase could be recognized with fairly good power by choosing the most frequent double homozygote genotype after discarding genotypes carrying the less frequent alleles.

Anyhow, the method we proposed here may be considered as a first step toward more sophisticated

ones, taking explicitly migration into account. Indeed, as genetic exchanges may constitute key events in weedy evolution, it would be of primary interest to incorporate gene flow as a parameter to be estimated, using a more comprehensive model.

We developed and evaluated the properties of a maximum-likelihood method to estimate the genetic age of a volunteer population derived from an F1-hybrid variety, and proposed a test to answer this simple question: is the sampled volunteer population a 1-year transitory crop descent or the result of more than one autonomous cycle of reproduction? This method provides an example of how the available information from linked markers can be successfully exploited to analyze the very recent history of crop-derived populations. Although generic methods are continuously designed to analyze the huge amount of genotypic information now potentially available (for example Falush *et al*., 2003), we feel that the investigation of crop-relative evolution in the agro-system requires an adaptation of these methods to crop genetic structure, to low intervarietal differentiation and to specific questions (for example Devaux *et al*., 2007), in addition to the contribution of other approaches such as demographic modeling (Pivard *et al*., 2008).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank the Domaine experimental INRA d'Epoisses for production of the two first generations and Muriel Latreille for technical assistance. This work was funded by the Bureau des Ressources Génétiques.

References

Bagavathiannan MV, Van Acker RC (2008). Crop ferality: implications for novel trait confinement. *Agric Ecosyst Environ* **127**: 1–6.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2001). *GENETIX 4.02, Logiciel Sous WindowsTM Pour la Génétique des Populations*. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II: Montpellier, France.

Cantamutto M, Poverene M, Peinemann N (2008). Multi-scale analysis of two annual *Helianthus* species naturalization in Argentina. *Agric Ecosyst Environ* **123**: 69–74.

Cornuet J-M, Piry S, Luikart G, Estoup A, Solignac M (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**: 1989–2000.

Crawley MJ, Brown SL (2004). Spatially structured populations dynamics in feral oilseed rape. *Proc Roy Soc Lond* **271**: 1909–1916.

Crow JF, Kimura M (1970). *An Introduction to Population Genetics Theory*. Burgess publishing company: Minneapolis, USA.

Devaux C, Lavigne C, Austerlitz F, Klein EK (2007). Modelling and estimating pollen movement in oilseed rape (*Brassica napus*) at the landscape scale using genetic markers. *Mol Ecol* **16**: 487–499.

Devaux C, Lavigne C, Falentin-Guyomar'ch H, Vautrin S, Lecomte J, Klein EL (2005). High diversity of oilseed rape pollen clouds over an agro-ecosystem indicates long-distance dispersal. *Mol Ecol* **14**: 2269–2280.

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

Excoffier L, Estoup A, Cornuet J-M (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Frankham R (1995). Effective population size/adult population size ratios in wildlife: a review. *Genet Res* **66**: 95–107.

Gandhi SD, Heesacker AF, Freeman CA, Argyris J, Bradford K, Knapp SJ (2005). The self-incompatibility locus (*S*) and quantitative trait loci for self-pollination and seed dormancy in sunflower. *Theor Appl Genet* **111**: 619–629.

Garnier-Gere P, Dillmann C (1992). A computer program for testing pairwise linkage disequilibria in subdivided populations. *J Heredity* **83**: 239.

Goldringer I, Bataillon T (2004). On the distribution of temporal variation in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics* **168**: 563–568.

Jarne P, David P (2008). Quantifying inbreeding in natural populations of hermaphroditic organisms. *Heredity* **100**: 431–439.

Kane NC, Rieseberg LH (2008). Genetics and evolution of weedy *Helianthus annuus* populations: adaptation of an agricultural weed. *Mol Ecol* **17**: 384–394.

Koopman WJM, Li Y, Coart E, Van de Weg E, Vosman B, Roldán-Ruiz I *et al*. (2007). Linked vs unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. *Mol Ecol* **16**: 243–256.

Londo JP, Schaal BA (2007). Origins and population genetics of weedy rice in the USA. *Mol Ecol* **16**: 4523–4535.

Lorieux M (1994). 'Aspects statistiques de la cartographie des marqueurs moléculaires' in Document de travail de la mission biométrie du CIRAD n°1-94, pp 31–35.

Muller M-H, Arlie G, Bervillé A, David J, Delieux F, Fernandez-Martinez JM *et al*. (2006). Le compartiment spontané du tournesol *Helianthus annuus* en Europe: prospections et premières caractérisations génétiques. *Actes du Colloque BRG* **6**: 335–353.

Pessel FD, Lecomte J, Emeriau V, Krouti M, Messean A, Gouyon PH (2001). Persistence of oilseed rape (*Brassica napus* L.) outside of cultivated fields. *Theor Appl Genet* **102**: 841–846.

Pivard S, Adamczyk K, Lecomte J, Lavigne C, Bouvier A, Deville A *et al*. (2008). Where do the feral oilseed rape populations come from? A large-scale study of their possible origin in a farmland area. *J Appl Ecol* **45**: 476–485.

Reagon M, Snow AA (2006). Cultivated *Helianthus annuus* (Asteraceae) volunteers as a genetic 'bridge' to weedy sunflower populations in North America. *Am J Bot* **93**: 127–133.

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.

Tang S, Kishore VK, Knapp SJ (2003). PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci in cultivated sunflower. *Theor Appl Genet* **107**: 6–19.

Tang S, Knapp SJ (2003). Microsatellites uncover extraordinary diversity in native American landraces and wild populations of cultivated sunflower. *Theor Appl Genet* **106**: 990–1003.

Tang S, Yu J-K, Slabaugh MB, Shintani DK, Knapp SJ (2002). Simple sequence repeat map of the sunflower genome. *Theor Appl Genet* **105**: 1124–1136.

Virmani SS (1994). *Monographs on Theoretical and Applied Genetics 22: Heterosis and Hybrid Rice Breeding*. Springer: Verlag.

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

Wilson GA, Rannala B (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**: 1177–1191.

Wolfram S (1996). *The Mathematica Book*, 3rd edn. Wolfram media Cambridge University Press: Cambridge, UK.
Zhang LS, Le Clerc V, Li S, Zhang D (2005). Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. *Can J Bot* **83**: 66–72.

Appendix

Let $x_t = (x_{1,t},\ldots,x_{10,t})^{\mathrm{T}}$ be the column vector containing the 10 genotypic frequencies (order: AB/AB, ab/ab, Ab/Ab, aB/aB, AB/Ab, aB/ab, AB/aB, Ab/ab, Ab/aB, AB/ab) at time t. T stands for, 'Transpoae'.

Let $\mathbf{M}_1$ be the 4×10 transition matrix relating vector x_t to the vector $(y_1,y_2, y_3, y_4)^{\mathrm{T}}$, containing the four gametic frequencies (order: AB, ab, Ab, aB):

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & r/2 & (1-r)/2 \\ 0 & 1 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & r/2 & (1-r)/2 \\ 0 & 0 & 1 & 0 & 1/2 & 0 & 0 & 1/2 & (1-r)/2 & r/2 \\ 0 & 0 & 0 & 1 & 0 & 1/2 & 1/2 & 0 & (1-r)/2 & r/2 \end{pmatrix}$$

Let φ be the application providing the expected genotypic frequencies under panmixia from a vector of gametic frequencies. φ is defined from $[0,1]^4$ to $[0,1]^{10}$ such that

$$\begin{aligned} \varphi &= (y_1, y_2, y_3, y_4) \\ &= (y_1^2, y_2^2, y_3^2, y_4^2, 2y_1y_3, 2y_2y_4, 2y_1y_4, 2y_2y_3, 2y_3y_4, 2y_1y_2)^{\mathrm{T}} \end{aligned}$$

The expected genotypic frequencies under panmixia at $t+1$ are then given by $\varphi\,(\mathbf{M}_1 \cdot x_t)$.

Let $\mathbf{M}_2$ be the 10×10 transition matrix relating genotypic frequencies at time t to the genotypic frequencies at the next generation, under selfing.

$$\mathbf{M}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 0 & \frac{r^2}{4} & \frac{(1-r)^2}{4} \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/4 & \frac{r^2}{4} & \frac{(1-r)^2}{4} \\ 0 & 0 & 1 & 0 & 1/4 & 0 & 0 & 1/4 & \frac{(1-r)^2}{4} & \frac{r^2}{4} \\ 0 & 0 & 0 & 1 & 0 & 1/4 & 1/4 & 0 & \frac{(1-r)^2}{4} & \frac{r^2}{4} \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & \frac{r(1-r)}{2} & \frac{r(1-r)}{2} \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & \frac{r(1-r)}{2} & \frac{r(1-r)}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & \frac{r(1-r)}{2} & \frac{r(1-r)}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & \frac{r(1-r)}{2} & \frac{r(1-r)}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{(1-r)^2}{2} & \frac{r^2}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r^2}{2} & \frac{(1-r)^2}{2} \end{pmatrix}$$

Let $\mathbf{s}$ be the selfing rate (assumed equal for all genotypes). The recursion equation for genotypic frequencies under mixed mating is given by:

$$x_{t+1} = (1-s)\varphi(\mathbf{M}_1.x_t) + s\mathbf{M}_2 \cdot x_t \qquad \text{(A1)}$$

It is difficult to get a general expression relating x_t to x_1 (genotypic frequencies at first generation) from Equation (A1). However, under complete selfing, Equation (A1) leads to

$$x_t = \mathbf{M}_2^{t-1} x_1$$

And under complete outcrossing, Equation (A1) leads to

$$x_{1,t} = x_{2,t} = \frac{1}{16}\left((1-r)^{t-2}(1-2r)+1\right)^2$$

$$x_{3,t} = x_{4,t} = \frac{1}{16}\left(1-(1-r)^{t-2}(1-2r)\right)^2$$

$$x_{5,t} = x_{6,t} = x_{7,t} = x_{8,t} = \frac{1}{8}\left(1-\left((1-r)^{t-2}(1-2r)\right)^2\right)$$

$$x_{9,t} = \frac{1}{8}\left(1-(1-r)^{t-2}(1-2r)\right)^2$$

$$x_{10,t} = \frac{1}{8}\left(1+(1-r)^{t-2}(1-2r)\right)^2$$

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