Euphytica (2010) 171:441–446
DOI 10.1007/s10681-009-0077-3

Sample size for diversity studies in tetraploid alfalfa (*Medicago sativa*) based on codominantly coded SSR markers

Doris Herrmann · Sandrine Flajoulot · Bernadette Julier

Received: 25 March 2009 / Accepted: 4 November 2009 / Published online: 15 November 2009

Abstract The number of genotypes investigated per population is important for the reliability of diversity studies. The objective of this study was to determine the sample size for the identification of differences among populations of an outcrossing autotetraploid species, alfalfa (*Medicago sativa*), using codominantly coded SSR markers. One hundred and twenty genotypes from each of two closely related populations were analysed with two markers. Twenty random subsamples for each of three sample sizes (10, 20 and 40 genotypes) were built. Compared to the populations with 120 genotypes, alleles that were no longer present in subsamples with 40 genotypes were mainly rare, whereas abundant alleles were also excluded in subsamples with 10 genotypes. F_{ST} values for pairs of subsamples between the two populations were always significantly different based on 40 genotypes, whereas for 10 genotypes more than half of the pairs were not significantly different. We concluded that 40 genotypes are a reasonable sample size for diversity studies with closely related populations of tetraploid alfalfa investigated with SSR markers. Twenty genotypes may be an economical alternative for large scale studies, but 10 genotypes were a too low number for reliable results.

Keywords Diversity · *Medicago sativa* · Sample size · SSR marker · Tetraploidy

D. Herrmann · B. Julier (✉)
INRA, UR 4, Unité de Recherche Pluridisciplinaire Prairies et Plantes Fourragères, BP 6, 86600 Lusignan, France
e-mail: bernadette.julier@lusignan.inra.fr

Present Address:
D. Herrmann
ISCB, Indo-Swiss Collaboration in Biotechnology, EPFL AI-VP CO-ISCB, 1015 Lausanne, Switzerland

S. Flajoulot
Jouffray-Drillaud, 86600 Lusignan, France

Introduction

Alfalfa (*Medicago sativa* L.), the most cultivated forage legume, has a high protein content, a suitable feeding value and a favorable environmental impact (perenniality and no nitrogen fertilizer required). This autotetraploid and allogamous species is therefore an important and valuable forage crop.

The estimation of genetic diversity within and among populations is of major importance for diversity management and conservation of genetic resources (Becker 1993). SSR markers are widely used and were proved to be an ideal marker system for this estimation due to their codominant, highly polymorphic and reproducible nature (Estoup and Angers 1998). As outcrossing species consist of many genotypes, sample size per population has to be an optimum between reasonable labour and reliability of data set. Samples should represent a maximum of rare

alleles and give a good estimation of allele frequencies so that population differentiation is accurate. Optimal sample size depends on several factors. Firstly, the marker system and its coding method may play a role (Lynch and Milligan 1994). The information is different if analyses are based on a dominant marker system such as RAPD (Kölliker et al. 1999) on a dominantly coded (to score alleles for presence and absence) codominant marker system (Labombarda et al. 2000) or on a codominantly coded (to score the dose of each allele) codominant marker system (Sardaro et al. 2008). Secondly, genetic characteristics such as the polyploidy of the investigated species have to be considered. In a single individual of an autotetraploid species, information on four alleles is available, whereas the information is reduced to two alleles in a diploid species. Consequently a smaller sample size may be sufficient to investigate diversity in tetraploid species. In contrast, within-population diversity was often reported to be even higher in tetraploid outcrossing species (Flajoulot et al. 2005) than in diploid ones (Fjellheim and Rognli 2005) and therefore a higher number of genotypes may have to be included for a reliable differentiation of tetraploid populations. Thirdly, the sample size can partly be compensated by the number of markers in the possibility to differentiate the populations (Lynch and Milligan 1994).

Forty genotypes were shown to be an adequate sample size for different marker systems to analyse diversity and differentiate populations in outcrossing diploid or tetraploid species (Herrmann et al. 2005; Labombarda et al. 2000; Gherardi et al. 1998). However, lower sample sizes were also recommended as alternatives (Bolaric et al. 2005; Labombarda et al. 2000). In diversity studies with tetraploid species, sample sizes from 10 to 40 genotypes were used (Sardaro et al. 2008; Jenczewski et al. 1999; Kölliker et al. 1999; Flajoulot et al. 2005).

Genetic diversity of alfalfa populations evaluated with molecular markers is large (Sardaro et al. 2008; Jenczewski et al. 1999; Kölliker et al. 1999; Flajoulot et al. 2005; Herrmann et al. 2008), but differentiation among populations is difficult. In a study involving 10 cultivars, each represented by 40 individuals and eight codominantly coded SSRs, all pairs of cultivars were significantly different (Herrmann et al. 2008). Two of these cultivars and two SSR markers were chosen to determine the optimal sample size in alfalfa populations to differentiate closely related populations, while avoiding significant random differences among homologous populations that could purely arise from limited sample size. One hundred and twenty plants for two alfalfa varieties and two markers were investigated and random subsamples of different sample sizes were compared.

Materials and methods

The two French alfalfa varieties Mercedes (Limagrain Genetics, France) and Symphonie (Desprez Veuve et fils, France) of the Flamande type, with a dormancy adapted to Northern France and Europe, were investigated. The parental polycross of Mercedes included five half-sib families and the parental polycross of Symphonie was composed by two series of 64 plants resistant to stem nematode (*Dytilenchus dipsaci*), each series being selected within a registered cultivar. Two SSR markers were used, MAA660456 and MTIC432, mapped on chromosomes 4 and 7, respectively (Julier et al. 2003). They were chosen because of very clear gel profiles so that allelic dosage could be determined with high accuracy. In a study including 10 cultivars, the F_{ST} between the two cultivars Mercedes and Symphonie was 0.013, equal to the global F_{ST} over all 10 cultivars. Information given by these two SSRs was similar to that obtained with six other SSRs (Herrmann et al. 2008).

For each of the two populations, DNA was extracted of 120 plants from young leaves and analysed with the two markers on a LI-COR IR2 (LI-COR Inc.) DNA sequencer (Flajoulot et al. 2005). Markers were codominantly coded, i.e. not only for presence and absence but also the dose of each allele was taken into account, as described in Flajoulot et al. (2005).

Of the total of 120 genotypes, 20 random subsamples of different sizes (40, 20 or 10 genotypes) were generated without replacement. The number of alleles for the subsamples was derived from allele frequencies calculated in Gene4x software (Ronfort et al. 1998) and the average number of alleles was determined for each sample size. In addition, the coefficient of variation for the number of alleles was calculated. Very rare alleles were identified for each marker and similar calculations were performed, taking into account only alleles which occurred at a frequency of at least 0.5 or 1%. These frequencies correspond to

alleles that were present at least twice or four times over 120 genotypes (480 alleles), respectively. The methodology for estimating the sampling size was synthesized (Crossa 1989). The probability to observe at least one specific genotype in a sample of size n is given by: $P = 1 - (1 - B)^n$, with B the proportion of the specific genotype in the population. Conversely, sample size can be estimated as: $n = [\ln(1 - P)]/[\ln(1 - B)]$, and the proportion of a specific genotype is: $B = 1 - \exp\left[\frac{1}{n}\ln(1 - P)\right]$. These formulae can be applied to individuals or gametes (Crossa 1989), which in our case corresponds to four gametes/alleles per individual/genotype. The formulae were used to interpret the results.

Fixation index F_{ST} (Weir and Cockerham 1984) was determined in order to evaluate population differentiation. Gene4x software (Ronfort et al. 1998) performs F_{ST} calculation assuming no double reduction as well as a corresponding index ρ under the hypothesis of double reduction. In alfalfa, low occurrence of double reduction was proved (Ayadi et al. 2005), so F_{ST} values and their significance were computed for the 400 pairs of subsamples between Mercedes and Symphonie and for the 190 pairs of subsamples within Mercedes and Symphonie, respectively. Average and standard deviation of these F_{ST} values were calculated for each sample size separately.

Results and discussion

Sample size is an important factor to consider for a reliable investigation of diversity and differentiation of alfalfa populations. The 120 genotypes of two populations analysed with two markers provided a solid basis to determine this size.

As expected, we observed an increase of number of alleles with larger sizes of subsamples (Table 1): with a higher number of individuals, there is a high probability to include rare alleles (Crossa 1989). However, even with 40 genotypes, an important decrease of number of alleles was observed compared to all 120 genotypes. For diversity studies, rare alleles may bias estimation of some parameters and were recommended to be excluded from calculations (Lynch and Milligan 1994). Rare alleles were identified for each population. In Mercedes, eight alleles were present at a frequency lower than 0.5% in the

Table 1 For two SSR markers (MAA660456 and MTIC432) and two populations (Mercedes and Symphonie), average number of alleles observed among 20 random subsamples for different sample sizes (40, 20, 10 genotypes) for all alleles (Average all) and for alleles occurring at a frequency higher than 0.5% (Average >0.5%) or 1% (Average >1%) in the 120 genotypes, respectively

	Mercedes			Symphonie		
	No of alleles			No of alleles		
	MAA660456	MTIC432	Both	MAA660456	MTIC432	Both
Subsamples with 40 genotypes						
Average >1%	6.95	8.95	15.90	7.90	8.95	16.85
Average >0.5%	8.65	10.50	19.15	7.90	11.85	19.75
Average all	9.85	12.45	22.30	8.50	13.80	22.30
CV of all (%)	12.9	9.9	8.4	6.0	10.1	7.8
Subsamples with 20 genotypes						
Average >1%	6.85	8.55	15.40	7.55	8.25	15.80
Average >0.5%	8.00	9.45	17.45	7.55	10.25	17.80
Average all	8.55	10.45	19.00	7.85	10.90	18.75
CV of all (%)	12.9	12.6	8.4	8.6	17.6	11.1
Subsamples with 10 genotypes						
Average >1%	6.15	7.55	13.70	6.85	7.00	13.85
Average >0.5%	6.75	8.20	14.95	6.85	8.10	14.95
Average all	7.00	8.65	15.65	7.00	8.45	15.45
CV of all (%)	13.9	20.6	13.6	11.3	17.8	12.5
All 120 genotypes						
>1%	7	9	16	8	9	17
>0.5%	9	11	20	8	13	21
All	12	16	28	9	17	26

Coefficient of variation (CV of all) was calculated for number of all alleles

120 genotypes, and four other alleles were present at a frequency between 0.5 and 1%. In Symphonie, five alleles were present at a frequency lower than 0.5%, and four other alleles were present at a frequency between 0.5 and 1% (Table 1). When taking into account only alleles whose frequency was higher than 0.5%, the deficit for 40 genotypes compared to 120 genotypes was reduced and the deficit became nil when including only alleles of a frequency higher than 1%. In contrast, even if considering only alleles which occurred with a frequency higher than 1%, a decisive decrease was still observed for samples with 10 genotypes (Table 1).

Using the formula of Crossa (1989), we calculated that to sample an allele present at a 1% frequency in a population with a probability of 95%, 89 genotypes must be sampled. Therefore, in subsamples of 40 genotypes, most excluded alleles were rare, whereas for 10 genotypes abundant alleles were also lost. Considering a sample size of 160 alleles (40 genotypes), the alleles present at a frequency of 1.9% were sampled with a probability of 95%. With a sample size of 80 alleles (20 genotypes) and 40 alleles (10 genotypes), the alleles were sampled if they were present at a frequency of 3.7 or 7.2%, respectively. Theoretical results were thus in accordance with observed data.

The number of alleles is often compared within and among studies to investigate which populations or groups of populations shows a higher diversity (Flajoulot et al. 2005; Sardaro et al. 2008; Patto et al. 2008). The coefficient of variation of number of alleles was decisively reduced when comparing subsamples with 40 genotypes to 10 genotypes but a moderate decrease was still observed between 40 and 20 genotypes (Table 1). Consequently the possibility of detecting a false positive difference for number of alleles between populations is crucially reduced for samples with 40 genotypes compared to 10 genotypes.

In addition to diversity, the reliable determination of differentiation among populations and at the same time prevention of a random differentiation of populations is a main objective (Fjellheim and Rognli 2005; Sardaro et al. 2008; Herrmann et al. 2005). In general, no significant (random) difference for F_{ST}

Table 2 Test of population differentiation by F_{ST} estimates of pairs of subsamples of different sizes (40, 20, 10 genotypes) within and between two populations (Mercedes and Symphonie), based on two SSR markers (MAA660456 and MTIC432)

	Pairs within Mercedes ($N = 190$)			Pairs within Symphonie ($N = 190$)			Pairs among Symphonie and Mercedes ($N = 400$)		
	MAA660456	MTIC432	Both	MAA660456	MTIC432	Both	MAA660456	MTIC432	Both
Subsamples with 40 genotypes									
Significant F_{ST} (%)	0.0	0.0	0.0	0.0	0.0	0.0	61.5	100.0	100.0
Average F_{ST}	−0.002	−0.001	−0.001	−0.003	−0.002	−0.002	0.006	0.009	0.008
SD F_{ST}	0.003	0.004	0.002	0.003	0.004	0.002	0.006	0.008	0.004
Subsamples with 20 genotypes									
Significant F_{ST} (%)	0.5	0.0	1.1	0.5	2.6	0.5	18.0	92.0	90.3
Average F_{ST}	−0.001	−0.002	−0.001	−0.002	−0.002	−0.002	0.007	0.010	0.009
SD F_{ST}	0.008	0.006	0.006	0.009	0.007	0.005	0.011	0.012	0.007
Subsamples with 10 genotypes									
Significant F_{ST} (%)	3.7	3.2	3.2	0.0	0.5	0.0	6.3	41.1	30.9
Average F_{ST}	−0.001	−0.004	−0.002	−0.005	−0.006	−0.005	0.006	0.009	0.007
SD F_{ST}	0.022	0.012	0.013	0.018	0.011	0.010	0.021	0.016	0.013
All 120 genotypes									
F_{ST} value	–	–	–	–	–	–	0.006***	0.010***	0.008***

SD standard deviation

*** F_{ST} values significant with $P < 0.001$

values was observed for both markers and pairs of subsamples within Mercedes or Symphonie independently of sample size (Table 2). F_{ST} values were close to zero but negative. This may be explained by a bias even though a biological cause in which alleles are more related between than within populations cannot be excluded (Cockerham 1973). However, standard deviations of F_{ST} values for pairs of subsamples were at least threefold reduced when comparing subsamples with 40 and 10 genotypes, thus indicating a higher reliability for 40 genotypes. In contrast, for the differentiation between the two populations, differences among the two markers and sample sizes were observed. For all sample sizes, MTIC432 was able to better differentiate the two populations (Table 2), probably because it contains more alleles than MAA660456 (Table 1). Therefore, MAA660456 was apparently not an efficient marker to differentiate these two populations. However, for both markers, large differences in the percentage of subsamples differentiating significantly the two populations were observed for the different sample sizes. As only 30% of the pairs between Mercedes and Symphonie were significantly different (Table 2), a sample size of 10 genotypes was too small to differentiate the two populations. Twenty genotypes could be an adequate number to differentiate the two populations, although a further advantage with 40 genotypes per subsample was observed.

This study was conducted with only two markers but a large number of markers might compensate for the uncertainty of small sample size (Lynch and Milligan 1994). However, for samples with only 10 genotypes, loss of information is probably too important to be compensated by additional markers. Twenty genotypes may be an economical choice, first of all if a large number of populations are genotyped in a study. But if the differentiation between populations is expected to be low, this sample size could limit the possibility to distinguish the populations. Indeed, in a study investigating seven French alfalfa varieties, each represented by 20 genotypes, with seven codominantly coded SSR markers, around 25% of pairs of varieties were not significantly differentiable (Flajoulot et al. 2005). In an other study involving 10 cultivars, each represented by 40 genotypes, with eight SSRs, all pairs of cultivars were significantly different ($P < 0.01$) (Herrmann et al. 2008).

Forty genotypes is a reasonable sample size to determine diversity and to differentiate even closely related outcrossing tetraploid alfalfa populations investigated with codominantly coded SSR markers. This optimal sample size was comparable to other optimisation studies (Labombarda et al. 2000; Gherardi et al. 1998), although these studies were based on dominant (RAPD) or dominantly coded codominant (RFLP) marker systems. A sample size of 40 genotypes can therefore be recommended for diversity studies of alfalfa for all marker systems.

Acknowledgements The authors thank F. Durand for her technical assistance in the lab. D Herrmann received a grant from the Plant Breeding department of INRA. The study was supported by the French Ministry of Agriculture, in the "Contrat de branches CB47" with ACVF (Association des Créateurs de Variétés Français).

References

Ayadi R, Barre P, Huyghe C, Julier B (2005) Estimation of the coefficient of double-reduction in autotetraploid lucerne. Wageningen. July 2005

Becker H (1993) Plant Breeding. Stuttgart

Bolaric S, Barth S, Melchinger AE, Posselt UK (2005) Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. Plant Breed 124:161–166

Cockerham CC (1973) Analyses of gene frequencies. Genetics 74:679–700

Crossa J (1989) Methodologies for estimating the sample size required for genetic conservation of outbreeding crops. Theor Appl Genet 77:153–161

Estoup A, Angers B (1998) Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. Adv Mol Ecol 306:55–86

Fjellheim S, Rognli OA (2005) Molecular diversity of local Norwegian meadow fescue (*Festuca pratensis* Huds.) populations and Nordic cultivars—consequences for management and utilisation. Theor Appl Genet 111:640–650

Flajoulot S, Ronfort J, Baudouin P, Barre P, Huguet T, Huyghe C, Julier B (2005) Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. Theor Appl Genet 111: 1420–1429

Gherardi M, Mangin B, Goffinet B, Bonnet D, Huguet T (1998) A method to measure genetic distance between allogamous populations of alfalfa (*Medicago sativa*) using RAPD molecular markers. Theor Appl Genet 96:406–412

Herrmann D, Boller B, Widmer F, Kölliker R (2005) Optimization of bulked AFLP analysis and its application for exploring diversity of natural and cultivated populations of red clover. Genome 48:474–486

Herrmann D, Flajoulot S, Barre B, Huyghe C, Ronfort J, Julier B (2008) Comparison of morphological traits and SSR markers to analyze genetic diversity of alfalfa cultivars.

North American Alfalfa Improvement Conference. North American Alfalfa Improvement Conference, http://www.naaic.org/Meetings/National/2008meeting/proceedings/Herrman.pdf

Jenczewski E, Prosperi JM, Ronfort J (1999) Differentiation between natural and cultivated populations of *Medicago sativa* (Leguminosae) from Spain: analysis with random amplified polymorphic DNA (RAPD) markers and comparison to allozymes. Mol Ecol 8:1317–1330

Julier B, Flajoulot S, Barre P, Cardinet G, Santoni S, Huguet T, Huyghe C (2003) Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. BMC Plant Biol 3:9

Kölliker R, Stadelmann FJ, Reidy B, Nösberger J (1999) Genetic variability of forage grass cultivars: a comparison of *Festuca pratensis* Huds., *Lolium perenne* L., and *Dactylis glomerata* L. Euphytica 106:261–270

Labombarda P, Pupilli F, Arcioni S (2000) Optimal population size for RFLP-assisted cultivar identification in alfalfa (*Medicago sativa* L.). Agronomie 20:233–240

Lynch M, Milligan BG (1994) Analysis of population genetic-structure with RAPD markers. Mol Ecol 3:91–99

Patto MCV, Moreira PM, Almeida N, Satovic Z, Pego S (2008) Genetic diversity evolution through participatory maize breeding in Portugal. Euphytica 161:283–291

Ronfort JL, Jenczewski E, Bataillon T, Rousset F (1998) Analysis of population structure in autotetraploid species. Genetics 150:921–930

Sardaro MLS, Atallah M, Tavakol E, Russi L, Porceddu E (2008) Diversity for AFLP and SSR in natural populations of *Lotus corniculatus* L. from Italy. Crop Sci 48:1080–1089

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 68:1358–1370