Conserv Genet (2009) 10:1343–1347
DOI 10.1007/s10592-008-9648-4

SHORT COMMUNICATION

Population estimators or progeny tests: what is the best method to assess null allele frequencies at SSR loci?

Sylvie Oddou-Muratorio · Giovanni G. Vendramin · Joukje Buiteveld · Bruno Fady

Received: 8 February 2008 / Accepted: 24 July 2008 / Published online: 15 August 2008

Abstract Nuclear SSRs are notorious for having relatively high frequencies of null alleles, i.e. alleles that fail to amplify and are thus recessive and undetected in heterozygotes. In this paper, we compare two kinds of approaches for estimating null allele frequencies at seven nuclear microsatellite markers in three French *Fagus sylvatica* populations: (1) maximum likelihood methods that compare observed and expected homozygote frequencies in the population under the assumption of Hardy-Weinberg equilibrium and (2) direct null allele frequency estimates from progeny where parent genotypes are known. We show that null allele frequencies are high in *F. sylvatica* (7.0% on average with the population method, 5.1% with the progeny method), and that estimates are consistent between the two approaches, especially when the number of sampled maternal half-sib progeny arrays is large. With null allele frequencies ranging between 5% and 8% on average across loci, population genetic parameters such as genetic differentiation (F_{ST}) may be mostly unbiased. However, using markers with such average prevalence of null alleles (up to 15% for some loci) can be seriously misleading in fine scale population studies and parentage analysis.

Keywords Microsatellites · *Fagus sylvatica* · Null alleles · Progeny · Population genetics

Nuclear SSRs are notorious for having relatively high frequencies of null alleles, i.e. alleles that consistently do not amplify during PCR. Null alleles are not detected in heterozygous diploid form, as genotypes appear as homozygous for the scorable allele (Dakin and Avise 2004). Null alleles usually result from changes in flanking region sequence (e.g. mutation, insertion or deletion) thus preventing primer binding to the DNA strand, and PCR amplification (Callen et al. 1993). Null alleles thus correspond to true allelic forms and do not result from a technical laboratory failure (e.g., lack of amplification, allelic drop out), from which they are undistinguishable in diploid homozygous form. Although undetected null alleles can lead to biased estimations of population genetic parameters (Chapuis and Estoup 2007), their effect is most serious in parentage analysis where they may lead to false exclusion of a significant number of true parents (Dakin and Avise 2004), and thereby to overestimates of migration rate from outside the studied population.

Different maximum-likelihood approaches have been proposed for estimating null allele frequencies using population genotypic data (Brookfield 1996; Chakraborty et al. 1992; Kalinowski and Taper 2006; Summers and Amos 1997). Their basic assumption is that all homozygote excess in a population (relative to Hardy-Weinberg equilibrium) is due to an excess of false homozygotes caused by null alleles. Available approaches differ in the treatment of individual samples without visible bands that can be interpreted either as true homozygote null genotypes, or genotyping failures due to human error, both processes

S. Oddou-Muratorio (✉) · B. Fady
INRA URFM, Ecologie des Forêts Méditerranéennes, Domaine Saint Paul, Site Agroparc, 84914 Avignon Cedex 9, France
e-mail: oddou@avignon.inra.fr

G. G. Vendramin
Plant Genetics Institute, CNR, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze, Italy

J. Buiteveld
Alterra Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands

being accounted for by Kalinowski and Taper (2006). Most methods rely on the EM algorithm of Dempster et al. (1977) to find numerically the maximum-likelihood estimates of null allele frequencies. The main drawback of these methods is that they can not separate homozygote excess due to null alleles from that caused by factors such as population substructure (e.g. Wahlund effect) or selection.

Experimental designs using progeny data provide a powerful, yet seldom used (but see De Sousa et al. 2005) opportunity to assess the performance of population level estimators of null allele frequencies (Dakin and Avise 2004). Null allele frequencies can indeed be estimated directly from progeny data when both parent and offspring genotypes are available from a sufficient sample of progeny arrays. Here, we estimated null allele frequencies using SSR genotypic data from three adult beech (*Fagus sylvatica* L.) populations from France for which we also have open-pollinated progeny array data. We used three different maximum-likelihood approaches to estimate null allele frequencies in our populations (Kalinowski and Taper 2006; Rousset 2008; Summers and Amos 1997). We then compared these null allele frequency estimates with the "true" ones obtained by directly counting segregating alleles in progeny.

Material and methods

European beech (*F. sylvatica*) is an economically important, widespread and widely studied tree species (Magri et al. 2006) for which few microsatellite markers are available (Pastorelli et al. 2003).

Adult trees and open-pollinated progeny arrays (seeds) were collected from 3 French *Fagus sylvatica* populations: Haye (48°40′ N, 6°04E), Sainte Baume (43°19′ N, 5°43′ E) and Ventoux (44°10′ N, 5°17′ E). In Haye (344 adult trees), a large number of progeny arrays were sampled (29), each containing a small number of seeds (13.2 seeds per array on average). In Ventoux (90 adult trees) and Sainte Baume (286 adult trees), only 5 and 4 progeny arrays were sampled respectively, but with an average of 50 seeds per array each.

DNA was isolated from buds and embryos using the Qiagen DNeasy Plant kit. DNA analysis was done using size SSR markers in Haye: FS1-46, FCM5, FS1-25, FS1-03, FS3-04, SFC-0161 and 4 SSR markers in Ventoux and Sainte Baume: FS1-46, FCM5, FS1-25, FS1-15. Details of their PCR amplification are reported in Tanaka et al. (1999); Pastorelli et al. (2003); Asuka et al. (2004). PCR products were separated on an automated 96-capillary MegaBACE^{TM} 1000 sequencer (GE Healthcare). Genotypes were sized using the internal size standards ET400 and the MegaBACE^{TM} Fragment Profiler ver. 1.2 software (GE Healthcare).

A progeny array was considered to contain null alleles when a significant number of its seeds (an arbitrary threshold of 20%) had apparently homozygous genotypes (XX, where X denotes any allele but that of the mother) incompatible with that of their apparently homozygous (AA) mother. In such progeny arrays, XX and AA individuals were assumed to be null heterozygotes (XN and AN, respectively) while missing data were considered to be null homozygotes (NN). To rule out amplification failure due to laboratory error or poor DNA quality, PCR was performed at least two times on samples without any amplification product, and DNA was re-isolated when amplification failure was systematic across loci.

To further rule out possible technical errors in genotyping that could be confused with null alleles, we tested the segregation pattern in each progeny array for goodness-of-fit to Mendelian expectations. Only seeds with non-ambiguous genotypes for the presence or absence of a null allele (that is AX, XN or NN) were used for these segregation tests. The frequency of null alleles in the overall seed population was then estimated by direct counting, after exclusion of the progeny arrays with distorted segregation pattern.

In the adult population, null allele frequencies were estimated using three different methods: (1) the maximum-likelihood (ML) estimator of Summer and Amos (1997) implemented in Cervus (Marshall et al. 1998); (2) the ML estimator based on the EM algorithm of Dempster et al. (1977) and implemented by default in GenePop 4.0 (Rousset 2007); (3) the ML estimator accounting for genotyping error implemented in ML-NullFreq (Kalinowski and Taper 2006). The first method is widely used, although the algorithm used to define and maximize likelihood has not been formally described by the authors. The third method is the only one that uses the actual genotype counts in the data instead of frequencies, and has thus been argued to be more informative (Kalinowski and Taper 2006). Observed and expected heterozygosities, fixation indices and their significance were estimated using GenePop 4.0 (Rousset 2007).

Results and discussion

At the population level, a significant heterozygote deficiency was detected in 12 population/marker combinations out of 14 (Table 1). Null allele frequency estimates ranged from 0 to 23%, and were consistent with the F_{is} estimates. The three different population estimators of null allele frequency (Marshall et al. 1998; Kalinowski and Taper 2006; Rousset 2007) provided comparable values.

Table 1 Number of sampled genotypes (N), observed (Ho) and expected (He) heterozygosities, fixation index (Fis, following Weir and Cockerham 1984), and estimated frequency of null alleles (Fnull) per locus and population using three different software : Genepop (Rousset 2007), ML-NullFreq (Kalinowski and Taper 2006) and Cervus (Marshall et al. 1998), and a direct estimate from progeny array data (Fnull progeny)

Locus	Population	N	*Ho*	*He*	F_{IS}[a]	*Population estimator (Fnull)*			*Progeny estimator*
						Genepop [CI][b]	ML-NullFreq	Cervus	(*Fnull progeny*)
FS1-15	Ste Baume	286	0.696	0.840	0.1720***	0.0819 [0.0560–0.1122]	0.080	0.0924	0.000
	Ventoux	89	0.562	0.658	0.1464*	0.0622 [0.0145–0.1260]	0.077	0.0697	0.000
FS1-46	Haye	128	0.742	0.838	0.1142**	0.0611 [0.0265–0.1073]	0.059	0.0587	0.042
	Ste Baume	286	0.584	0.813	0.2825***	0.1302 [0.1007–0.1626]	0.127	0.1629	0.000
	Ventoux	90	0.611	0.722	0.1547***	0.0698 [0.0267–0.1294]	0.070	0.0770	0.086
FCM5	Haye	127	0.732	0.901	0.1882***	0.0939 [0.0587–0.1390]	0.105	0.1001	0.053
	Ste Baume	286	0.808	0.868	0.0698***	0.0395 [0.0198–0.0654]	0.037	0.0337	0.000
	Ventoux	90	0.789	0.876	0.0997***	0.0464 [0.0125–0.0970]	0.046	0.0487	0.000
FS1-25	Haye	129	0.732	0.901	0.1795***	0.0945 [0.0563–0.1426]	0.107	0.0926	0.065
	Ste Baume	286	0.455	0.718	0.3674***	0.1580 [0.1262–0.1922]	0.158	0.2256	0.293
	Ventoux	90	0.633	0.757	0.1640***	0.0781 [0.0325–0.1388]	0.067	0.0865	0.096
FS1-03	Haye	130	0.600	0.677	0.1136*	0.0465 [0.0095–0.0980]	0.054	0.0609	0.083
FS3-04	Haye	127	0.551	0.485	−0.1378	0.0000 [no CI]	0.000	−0.0698	0.000
SFC-0161	Haye	128	0.844	0.806	−0.0492	0.0000 [no CI]	0.000	−0.0273	0.000
Average	*Haye*	129.2	0.700	0.768	0.0805***	0.0493	0.0542	0.0359	0.0405
	Ste Baume	286	0.636	0.810	0.216***	0.1024	0.1005	0.1287	0.0733
	Ventoux	89.8	0.649	0.753	0.139***	0.0641	0.0650	0.0705	0.0455

[a] *p*-value of the score test for heterozygote deficiency, with : * 1% < *p*-value < 5%; **0.1% < *p*-value < 1%; *** *p*-value < 0.1%

[b] 95% confidence intervals (CI) for null allele frequencies provided by Genepop

In the progeny arrays, null allele frequency estimates ranged from 0% to 29% (Table 1). Segregation patterns generally supported our hypothesis that true null alleles are present in our data (Table 2). Progeny array estimates were generally consistent with population estimates, except when null allele frequencies were estimated based upon a small number of families (such as in Ventoux and Sainte Baume). In this case, it seems that null alleles were often not detected in the progeny arrays.

Overall, five markers out of the seven investigated appeared to be affected by null alleles. This relatively high prevalence of null alleles is expected in species with large effective population size such as trees (Chapuis and Estoup 2007). Surprisingly, one of the two markers apparently not affected by null alleles (SFC-0161) was transferred from *Fagus crenata* (Asuka et al. 2004), the Asian vicariant of *F. sylvatica*. This suggests that further investigation is warranted of the frequency of null alleles in *F. sylvatica* for the 15 other markers developed for *F. crenata*. Indeed it is expected that microsatellite primers transferred closely related, congeneric species are more likely to display null alleles than those designed directly in the study species, as

Table 2 Segregation patterns of null alleles in open-pollinated maternal progeny arrays from three *Fagus sylvatica* populations

Locus	Population	Progeny	NbProg	NbProgNull	NbProgNoNull	*P*-value
FS1-46	Haye	B307	17	6	7	0.782
FS1-46	Haye	C214	11	5	4	0.739
FCM5	Haye	C132	13	7	4	0.366
FCM5	Haye	C145	14	5	4	0.739
FCM5	Haye	C196	12	5	7	0.564
FS1-25	Haye	B163	16	9	4	0.166
FS1-25	Haye	B38	14	8	5	0.405
FS1-25	Haye	C214	13	6	4	0.527
FS1-25	Ste Baume	1	51	15	8	0.144
FS1-25	Ste Baume	2	43	21	10	0.048*
FS1-25	Ventoux	1	46	14	12	0.695
FS1-03	Haye	C262	15	7	3	0.206
FS1-03	Haye	B307	15	10	1	0.007**
FS1-03	Haye	B308	15	7	3	0.206
FS1-03	Haye	C336	6	3	2	0.655

Progeny: code name of progeny array in which null alleles were detected, Nbprog: total number of offspring analyzed, NbProgNull: number of offspring carrying a null allele, NbProgNoNull: number of offspring not carrying a null allele. Offspring with ambiguous genotypes (homozygous maternal genotype) were excluded. *p*-value: *p*-value of the goodness-of-fit chi square test for deviation of Mendelian segregation for progeny

See Table 1 for levels of significance

nucleotide substitutions and insertions/deletions in microsatellite flanking regions should be more frequent between versus within species.

The average null allele frequency was between 7% and 8% for population estimators and 5% for the progeny estimator (Table 1). Our study thus demonstrates that population and progeny array based estimates of null allele frequencies can perform quite similarly. This is reassuring, considering that population estimators may be biased due to deviation from Hardy-Weinberg equilibrium caused by other processes than null alleles. Despite the limitation of gene flow by distance and the fine-scale genetic structure detected in the populations under study (Oddou-Muratorio et al, unpublished), maximum-likelihood population approaches do not seem to provide over-estimates of null allele frequencies as compared to the progeny array approach.

We also show that the congruence between estimates increases with the number of maternal open pollinated progeny arrays sampled. For a given sampling effort, it is thus more efficient to increase the number of progeny arrays rather than the sample size within each progeny array. However, many factors can confound the detection of null alleles, such as spatial genetic structure of adults and their mating system. The effort put into detecting and measuring null allele frequencies should be matched with the specific objectives of the study. In fine scale genetic studies, for example, where individual values of the parameters of interest are required, a progeny array design will have the added value of providing the precise mapping of the individuals carrying the null alleles, rather than just overall null allele frequencies.

Once null allele frequencies are estimated for a given marker set, it can be decided whether they should be or not accounted for in statistical analyses. According to the simulation study of Chapuis and Estoup (2007), ignoring the presence of null alleles within the observed range of frequency (5–8% on average across loci) will only slightly bias classical estimates of population differentiation (such as F_{ST}). Similarly, Dakin and Avise (2004) showed using simulations that this range of null allele frequency (5%–8%) equates to a less than 5% risk of falsely excluding an actual parent of a heterozygous offspring in parentage/paternity analyses. However, their results were based on expected false exclusion probabilities at a single locus and assuming panmixy (Jamieson and Taylor 1997). The risk of false exclusion of true parents, and conversely of false assignation to unrelated individuals, may thus dramatically increase with the number of markers affected by null alleles used to assign parentage. Moreover, in parentage trials within natural populations, the effective exclusion power will be reduced as compared to what expected under panmixy, in particular because of deviation from random mating and fine-scale population structure.

The solutions available to handle with null alleles fall within four categories: (1) keeping markers affected by null alleles without accounting for them, which usually results in estimation bias (Chapuis and Estoup 2007; Dakin and Avise 2004; Wagner et al. 2006); (2) removing them from the analysis, which can decrease the accuracy of estimates (Wagner et al. 2006), unless having an important number of markers at beginning (3) redesigning the primers, which is costly and may not be completely successful. The fourth solution is to correct genotypes for null alleles, for instance by changing systematically homozygous genotypes XX into null heterozygous genotypes XN, and to modify accordingly the estimator to account for null allele (see Chapuis and Estoup 2007 for the case of F_{ST} estimates; see Wagner et al. 2006 for relatedness coefficients). This last solution may actually be the most advisable when dealing with species like *Fagus sylvatica*, with a limited number of available microsatellite markers, and/or, more importantly, a high frequency of null alleles, possibly because of large population size (Chapuis and Estoup 2007).

Acknowledgements Data used for this paper were generated in the course of the European project DYNABEECH (Contract QLRT-1999-01210) and of a national project supported by the French Bureau des Ressources Génétiques (project BRG No. 88-2003-2004, AIP 223). A meeting for co-writing of this paper was supported by the European Commission through the Network of Excellence EVOLTREE (contract number: 016322) under the Sixth Framework Programme. We thank B. Jouaud, A. Roig (INRA, Avignon, France) and R. Pastorelli (CNR, Firenze, Italy) for laboratory work, as well as J. Glaubitz, and one anonymous reviewer for detailed comments on a previous version of the manuscript.

References

Asuka Y, Tani N, Tsumura Y, Tomaru N (2004) Development and characterization of microsatellite markers for Fagus crenata Blume. Mol Ecol Notes 4:101–103

Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. Mol Ecol 5:453–455

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR (1993) Incidence and origin of null alleles in the (AC)n microsatellite markers. Am J Hum Genet 52:922–927

Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56:45–57

Chapuis M-P, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol 24: 621–631

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. Heredity 93:504–509

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

De Sousa SN, Finkeldey R, Gailing O (2005) Experimental verification of microsatellite null alleles in Norway spruce (*Picea abies* [L.] Karst.): Implications for population genetic studies. Plant Mol Biol Rep 23:113–119

Kalinowski ST, Taper ML (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. Conserv Genet 7:991–995

Magri D, Vendramin GG et al (2006) A new scenario for the quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. New Phytol 171:199–221

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7:639–655

Pastorelli R, Smulders MJM, Van't Westende WPC, Vosman B, Giannini R, Vettori C, Vendramin GG (2003) Characterization of microsatellite markers in *Fagus sylvatica* L. and *Fagus orientalis* Lipsky. Mol Ecol Notes 3:76–78

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

Summers K, Amos W (1997) Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. Behav Ecol 8:260–267

Tanaka K, Tsumura Y, Nakamura T (1999) Development and polymorphism of microsatellite markers for *Fagus crenata* and the closely related species, *F. japonica*. Theor Appl Genet 99:11–15

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

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. Heredity 97:336–345