JrnlID 13595_ArtID 87_Proof# 1 - 15/05/2011

Annals of Forest Science
DOI 10.1007/s13595-011-0087-x

ORIGINAL PAPER

A case study of *Eucalyptus globulus* fingerprinting for breeding

Maria Margarida Ribeiro · Leopoldo Sanchez · Carla Ribeiro · Fátima Cunha · José Araújo · Nuno M. G. Borralho · Cristina Marques

Received: 17 August 2010 / Accepted: 21 December 2010

Abstract

• ***Introduction*** Tree genetic improvement programs usually lack, in general, *pedigree* information. Since molecular markers can be used to estimate the level of genetic similarity between individuals, we genotyped a sample of a Portuguese *Eucalyptus globulus* breeding population—a reference population of 125 individuals—with 16 microsatellites (SSR).

• ***Materials and methods*** Using genotypes from the reference population, we developed a simulation approach to recurrently generate (10^5 replicates) virtual offspring with different relatedness: selfed, half-sib, full-sib and unrelated individuals. Four commonly used pairwise similarity coefficients were tested on these groups of simulated offspring. Significant deficits in heterozygosity were found for some markers in the reference population, likely due to the presence of null alleles. Therefore, the impact of null alleles in the relatedness estimates was also studied. We conservatively assumed that all homozygotes in the reference population were carriers of null alleles.

• ***Results*** All estimators were unbiased, but one of them was better adjusted to our data set, even when null alleles were considered. The estimator's accuracy and precision were validated with individuals of known pedigree obtained from controlled crosses made with the same reference population's parents. Additionally, a clustering algorithm based on the estimator of choice was constructed, in order to infer the relatedness among 24 *E. globulus* elite individuals. We detected four putatively related elite individuals' pairs (six pairs considering the presence of null alleles).

• ***Conclusions*** This work demonstrates that in the absence of pedigree information, our approach could be useful to identify relatives and minimize consanguinity in breeding populations.

Keywords Microsatellites · *Eucalyptus globulus* · Null alleles · Relatedness

Handling Editor: Christophe Plomion

Electronic supplementary material The online version of this article (doi:10.1007/s13595-011-0087-x) contains supplementary material, which is available to authorized users.

M. M. Ribeiro (✉)
Departamento de Recursos Naturais e Desenvolvimento Sustentável, Escola Superior Agrária,
6001-909 Castelo Branco, Portugal
e-mail: mataide@ipcb.pt

L. Sanchez
INRA Centre d'Orléans, Unité Amélioration, Génétique et Physiologie Forestière,
45166 Olivet, France
e-mail: leopoldo.sanchez@orleans.inra.fr

C. Ribeiro · F. Cunha · J. Araújo · C. Marques
RAIZ-Direcção de Investigação Florestal,
Herdade de Espirra,
2985-270 Pegões, Portugal

N. M. G. Borralho
BorralhoIDea,
Urbanização S. Francisco, 18,
2070-220 Cartaxo, Portugal

1 Introduction

Eucalyptus globulus ssp. *globulus* (hereafter *E. globulus*) is an economically important species for pulpwood production, actively bred in many countries (Eldridge et al. 1994), including Portugal, where the first formal breeding program for the species began in 1966 (Borralho et al. 2007). In general,

the foundation of breeding populations aims to capture, as close as possible, the genetic diversity of the original population. However, breeding activities will rapidly reduce genetic diversity due to selection intensity, linkage and random drift in finite populations (Lefèvre 2004). Moreover, inbreeding depression is known to be severe in this species (Hardner and Potts 1995; Costa e Silva et al. 2010). To ensure that levels of coancestry and inbreeding among selected trees are kept to a minimum, it would be advantageous to know the relatedness among parents of unknown *pedigree*, particularly in early stages of breeding programs (Ballou and Lacy 1995). In the absence of known pedigree information, estimates of relatedness between individuals can be obtained through the use of molecular markers. Codominant microsatellite markers (SSR) are particularly suitable for this purpose, as they can be used to estimate individuals' pairwise relatedness, based on probability ratios of identity in state between individuals and an unrelated reference population. These estimates are very useful to infer the level of relatedness among sub-populations of elite material, to assure the deployment of unrelated elite clones and/or for the design of controlled crosses between putatively unrelated parents.

Estimators of pairwise relatedness were first considered for DNA data by Lynch (1988). This first estimator was modified by Li et al. in order to accommodate codominant markers ~~(1993)~~. Band sharing by chance is difficult to separate from band sharing by descent, and a method-of-moments (MM) estimator for pairwise relatedness was developed by Queller and Goodnight (1989). Afterwards, more accurate and precise MM estimators were developed by Ritland (1996) and Lynch and Ritland (1999). Recently, Wang (2002) introduced a new estimator, an improved version of the one proposed by Li et al. (1993), but Csillery et al. (2006) demonstrated that its performance was poor. Other estimators, including maximum likelihood methods (ML), were proposed to estimate relatedness in the absence of known pedigree structure (Queller and Goodnight 1989; Li et al. 1993; Lynch and Ritland 1999; Wang 2002; Milligan 2003; Thomas 2005; Oliehoek et al. 2006) and were used in different areas of research (reviewed by Blouin 2003 and Thomas 2005). Their performance was compared in several studies using simulated and empirical datasets (Lynch and Ritland 1999; Van de Casteele et al. 2001; Wang 2002; Milligan 2003; Csillery et al. 2006). These studies agree in that no single estimator is universally superior to the others in terms of bias and variance and that the performance rank order of the estimators depends on the estimation of the true relatedness value, the informativeness of the markers (number of *loci* and number and frequencies of alleles per *locus*) and the sample size used to estimate allele frequencies. For the commonly available markers in most studies (~ 5 to 20 microsatellites), the MM estimators are preferred because the ideal properties of ML methods are only achieved asymptotically (Lynch and Ritland 1999; Wang 2002; Milligan 2003). Additionally, the presence of null alleles in SSR markers can introduce a bias in the estimation of relatedness (Wagner et al. 2006). However, little is known on the actual impact of null alleles on the behaviour of relatedness estimators.

In this study, we compared three commonly used MM coefficients to estimate pairwise similarity: Ritland (1996) (R), Queller and Goodnight (1989) (Q) and Lynch and Ritland (1999) (LR), and a band sharing method: Li et al. (1993) (L), in the context of a Portuguese *E. globulus* breeding population. We followed a Monte Carlo simulation strategy and, unlike previous studies in the literature, considered two different criteria to identify the best performing estimator: (1) smaller average overlapping areas between every two density distribution relatedness categories and (2) smaller impact from the presence of null alleles.

We have used 16 publicly available SSR markers to screen 125 putatively unrelated individuals from an elite breeding population of *E. globulus*. The assumption of Hardy–Weinberg equilibrium in breeding populations of artificial origin might not hold true. However, this issue was overcome by measuring relatedness on the randomly generated in silico individuals from the existing parents in the reference breeding population.

In order to define a threshold to transform the continuous range given by the pairwise methods into genealogical relatedness (e.g. Blouin et al. 1996; Kozfkay et al. 2008), the density distributions of the simulated selfed, half-sib, full-sib and unrelated offspring were obtained. The selected threshold corresponds to the interception of the probability distribution curves of the unrelated and the half-sib individuals. This critical value is only coincident with the cut-off defined by Blouin et al. (1996) when the density distributions are absolutely symmetric, which is not always the case (e.g. Kozfkay et al. 2008). An additional population of 24 elite trees from the genetic improvement program was genotyped, as a practical application of the methodology developed here.

The objectives of this study are to provide estimates of the genetic parameters of the SSR used, including its discriminant power (*D*), to select the better suited relatedness estimator across unrelated (UR), half-sib (HS), full-sib (FS) and individuals generated by selfing a single parent (SF), to validate the estimator's precision and accuracy with individuals of known pedigree (HS, FS and SF), and to study the impact of null alleles in the relatedness estimates.

JrnlID 13595_ArtID 87_Proof# 1 - 15/05/2011

2 Material and methods

2.1 Plant material and DNA extraction

The *E. globulus* population of 125 putatively unrelated individuals (hereafter reference population, RP), includes 12 individuals used in controlled crosses to produce the validation population. The remaining 113 were putatively unrelated *E. globulus* individuals representative of the genetic improvement population of RAIZ (Forestry and Paper Research Institute, Portugal) (Borralho et al. 2007). This group includes 47 trees originally selected in plantations in Portugal (referred herein as "Portuguese land race") and 66 trees from 13 Australian native races (classification follows Dutkowski and Potts (1999). The validation population comprised three half-sib families, three full-sib families and four selfed families (each family with individuals generated by selfing a single parent), from controlled crosses made between 12 putatively unrelated individuals of the Portuguese land race. Each family had six offspring. An extra set of 24 elite clones was also genotyped. These 24 elite trees were used as a practical application of the proposed methodology. They were selected from RAIZ *E. globulus* breeding population and are to be used for deployment. Total genomic DNA was extracted as in Marques et al. (1998). DNA concentration was estimated by comparison of the fluorescence intensities of ethidium bromide-stained samples to those of λDNA standards, on 1% agarose gels.

2.2 SSR, PCR conditions and sizing of PCR products

Sixteen publicly available eucalypt SSR (Appendix 1[1]) were selected for its allele number and effective number of alleles (Table 1). SSR primer design was described elsewhere (EMBRA 1–20 in Brondani et al (1998), EMCRC1-12 in Steane et al. (2001) and EMBRA 21–70 in Brondani et al. (2002)). Each SSR marker was assigned to a consensus linkage group based on *E. globulus* genetic linkage maps (unpublished results) and a consensus map of a *Eucalyptus grandis*×*Eucalyptus urophylla* pedigree (Brondani et al. 2006). EMCRC5 was the only unmapped marker in this study. Three SSR (EMBRA 6, EMBRA 11 and EMBRA 12) mapped to the same linkage group (no. 1, see Appendix 1), but in different locations (unpublished results). The remaining seven SSR mapped to different linkage groups. Despite the fact that we expect high SSR synteny in the eucalypt *Symphyomyrtus* subgenus (Marques et al. 2002), we performed linkage disequilibria tests for all loci combinations with the Genepop version 4.0.7 (Rousset 2008). The *p* values were obtained by the contingency table approach (Fisher's exact test), and the number of dememorization steps was 10,000, with 1,000 batches and 100,000 iterations per batch. The significance level, with a probability of type I error of 1%, took into account the number of tests performed by using the Bonferroni correction (Sokal and Rohlf 1997). The Hardy–Weinberg test was made by estimating the exact *p* values by the Markov chain method, with the same dememorization steps, batches and iterations per batch referred in the foregoing. The null allele frequencies per loci were estimated by using a maximum likelihood EM algorithm. Both were computed with the Genepop software.

Polymerase chain reaction amplification of SSR loci was carried out in 96-well V-bottom plates. Each reaction contained 0.2, 0.15 and 0.1 μM of primer (for SSR in groups 1, 2 and 3, respectively—Appendix 1), 0.5 U of Taq DNA polymerase (Promega, Madison, WI, USA), 0.2 mM of each dNTP (otherwise as specified in Appendix 1, Promega, Madison, WI, USA), 1× reaction buffer (Promega, Madison, WI, USA), 2 mM of $MgCl_2$ (Promega, Madison, WI, USA), DMSO 5.0% (Sigma) and 20 ng of template DNA in a final 10-μl volume. Forward primers were IRD800 (5′-fluoresceine) labelled. Reactions were cycled in an MJ Research PT-100 Thermal Controller with a heated lid, 94°C for 30 s, followed by 15 cycles of variable annealing temperature ("touch down"): 94°C for 30 s, 30 s of annealing (from 56°C, with a decrease of 0.2°C every cycle), and 72°C for 45 s; then 20 cycles of 94°C for 30 s, 53°C for 30 s and 72°C for 45 s; and finally 72°C for 7 min. Amplification products were denatured by adding 10 μL of formamide buffer (98% formamide deionized, 10 mM EDTA pH 8.0, 60 mg bromophenol blue), heated 5 min at 70 C (Termomixer Confort, Eppendorf), and 0.8 μL of the samples was loaded in 6% acrylamide denaturating gel (50% Long-Ranger–, with 10.5 g Urea and 2.5 ml TBE (10×)). Fragments were separated using a LI-COR automatic DNA sequencer (model 4200 Gene Readir) at 1,500 V, 25 W constant power, 45°C of plate temperature and a 1× TBE running buffer, for approximately 2 h. RFLPscan was used to retrieve the gel image, and the presence of the bands was visually scored with the help of a LA4000-44B LI-COR ladder.

2.3 Relatedness estimators

The coancestry coefficient (θ) between individuals *x* and *y* is the probability that two randomly chosen homologous alleles are identical 'by descent' (Lynch and Walsh 1998). In a diploid mating system, the coefficient of coancestry multiplied by 2 equals the coefficient of relatedness, r_{xy}, which is the expected fraction of alleles identical by

[1] Appendix is available online only at www.asf-journal.org.

Q6t1.1 **Table 1** Diversity parameters for the 16 SSR loci in the reference population, ordered according to its discriminant power (*D*)

	N_a	N_e	H_e	H_o	F_{is}	Sig.	Null	*D*	
EMBRA23	21	12.8	0.93	0.89	0.04	NS	0.031	0.991	t1.3
EMBRA12	19	13	0.93	0.89	0.04	NS	0.025	0.991	t1.4
EMCRC8	18	12.8	0.93	0.84	0.09	S	0.049	0.987	t1.5
EMBRA18	21	11.5	0.92	0.90	0.01	NS	0.011	0.987	t1.6
EMCRC11	16	8.9	0.89	0.83	0.07	NS	0.032	0.981	t1.7
EMBRA6	15	8.8	0.89	0.78	0.12	S	0.055	0.976	t1.8
EMCRC10	18	8.6	0.89	0.65	0.26	S	0.130	0.960	t1.9
EMBRA11	21	9.4	0.90	0.87	0.02	NS	0.029	0.960	t1.10
EMBRA2	15	6.2	0.84	0.76	0.1	NS	0.044	0.959	t1.11
EMBRA8	14	6.2	0.84	0.76	0.1	NS	0.046	0.956	t1.12
EMCRC7	14	4.8	0.79	0.70	0.11	NS	0.048	0.932	t1.13
EMBRA20	13	4.7	0.79	0.62	0.21	S	0.091	0.929	t1.14
EMCRC2	15	4.5	0.78	0.62	0.2	S	0.107	0.915	t1.15
EMBRA5	21	5.2	0.82	0.50	0.34	S	0.158	0.898	t1.16
EMCRC5	21	5.5	0.81	0.53	0.37	S	0.165	0.898	t1.17
EMBRA19	6	3.4	0.71	0.54	0.24	S	0.155	0.855	t1.18
Mean	16.8	7.9	0.85	0.73	0.15		0.074	0.948	t1.19

Sig. refers to the significance resulting from the HWE test (after Bonferroni correction, where NS means not significant and S significant), and null refers to null allele frequency estimates

N_a number of alleles per locus, N_e effective number of alleles, H_e expected heterozygosity, H_o observed heterozygosity, F_{is} fixation index

descent between two (related) individuals. Alleles are identical by descent if they recently descend from a single ancestral allele. Alleles that are identical by state (IBS) might not be identical by descent if they coalesce further back than the reference pedigree or arose independently via mutation (see Blouin 2003 for details). In fact, the estimated relatedness measures how much higher (or lower) the probability of recent coalescence is for any given pair (*x*, *y*), relative to the average probability for all pairs. The expected relatedness is 0.67 for selfed, 0.5 for full-sibs, 0.25 for half-sibs and 0 for unrelated individuals. For example, on average, a pair of siblings (FS) shares one out of two alleles identical by descent (Squillace 1974; Falconer and Mackay 1996; Blouin 2003).

Lynch (1988) relatedness estimator based on band sharing and modified by Li et al. (1993) (L) is:

$$r_{xy} = \frac{s_{xy} - s_0}{1 - s_0} \text{ and } s_0 = \sum_{i=1}^{n} p_i^2(2 - p_i), \quad (1)$$

where S_{xy} is the similarity index $S_{xy}=n_{xy}/2(1/n_x+1/n_y)$, n_{xy} is the number of shared alleles between individuals *x* and *y*, n_x is the number of alleles of *x*, n_y is the number of alleles of *y* and S_0 is the number of shared alleles in the reference population, based on the allele frequencies (p_i is the frequency of the *i*th allele).

Ritland (1996) (R) coancestry estimator of individuals $X=(A_1,A_2)$ and $Y=(A_3,A_4)$ can be written as:

$$\theta_{xy} = \frac{1}{4(n_i - 1)} \times \left[\left(\frac{\delta(A_1,A_3) + \delta(A_1,A_4)}{p(A_1)}\right) + \left(\frac{\delta(A_2,A_3) + \delta(A_2,A_4)}{p(A_2)}\right) - 1\right] \quad (2)$$

where δ, the Kronecker operator, is defined for alleles A_i and A_j: $\delta(A_i,A_j)=1$ if $A_i=A_j$, and $\delta(A_i,A_j)=0$ if $A_i \neq A_j$. We have six operators to compare two individuals (two within and four between individuals) in the same locus, $p(A_i)$ being the frequency of the *Ai* allele in the considered locus and reference population and n_i the total number of alleles in the considered locus and reference population (Ritland 2000).

The Queller and Goodnight (1989) (Q) relatedness estimator is based on the same Kronecker operator and is described as:

$$r_{xy} = \frac{(\delta(A_1,A_3) + \delta(A_1,A_4) + \delta(A_2,A_3) + \delta(A_2,A_4) - p(A_1) - p(A_2))}{2(1 + \delta(A_1,A_2) - p(A_1) - p(A_2))} \quad (3)$$

Still based on Kronecker operators, Lynch and Ritland (1999) developed another relatedness estimator (LR) which is defined as follows:

$$r_{xy} = \frac{(p(A_1)\delta(A_2,A_3) + \delta(A_2,A_4)) + (p(A_2)\delta(A_1,A_3) + \delta(A_1,A_4)) - 4(pA_1)p(A_2)}{(1 + \delta(A_1,A_2))(p(A_1) + p(A_2)) - 4p(A_1)p(A_2)} \quad (4)$$

2.4 Estimation of genetic parameters and simulation methods

For each SSR locus in the RP, the number of alleles (N_a), the effective number of alleles ($N_e=1/(1-H_e)$), the observed heterozygosity (H_o) and the expected heterozygocity (H_e) (Nei 1987) were computed with a FORTRAN program developed in this study, hereafter called Zeta (available upon request from LS). The fixation index (F_{is}) (Weir and Cockerham 1984) was estimated with the Genepop software version 4.0.7.

The distribution of relatedness *r*-values estimated with the L, R, Q and LR coefficients was obtained by generating 10^5 replicates of UR, HS, FS and SF individuals, from where mean and sampling variance values were calculated. Each replicate consisted of two in silico individuals. These individuals were obtained assuming free recombination and segregation out of parental SSR genotypes. Parents were sampled at random. In the UR group, four distinct parents were sampled and single-pair mated in order to obtain two unrelated offspring. For the HS group, three distinct parents were sampled, and one of them mated to the other two, in order to obtain one offspring from each mating. With the FS group, only two parents were sampled and mated, in order to obtain two full-sib individuals. Finally, in the SF's group, one parent was sampled and selfed twice, in order to get two offspring.

The relatedness between any two in silico individuals, measured in each replicate, the *r*-value (r_{xy}), was computed using a weighted multilocus average:

$$\bar{r}_{xy} = \frac{\sum_i r_{xy(i)} / \mathrm{Var}\left(r_{xy(i)}\right)}{\sum_i 1 / \mathrm{Var}\left(r_{xy(i)}\right)},$$

where $r_{xy(i)}$ is the estimator's value for the *i*th locus, according to one of the four estimators (L, R, Q or LR, in Eqs. 1, 2, 3 or 4, respectively), and $\mathrm{Var}(r_{xy(i)})$ is the Monte Carlo sampling variance for the same locus over replicates, which was used as a weighting factor for the multilocus average. Therefore, variable loci will account for less in the average, compared to less variable loci. A sampling variance was also calculated for the multilocus average ($\mathrm{Var}\, r_{xy}$), as the Monte Carlo variance over replicates.

In order to evaluate the informativeness of each SSR marker for fingerprinting, we estimated its discriminant power (*D*) by using *Zeta*. *D* was the number of replicates in which a given marker was able to discriminate between two simulated individuals with a given level of relatedness, over the total number of 10^5 replicated pairs. The discriminant power was obtained for each marker and for each relatedness group. This indicates the likelihood of discrimination of any two individuals derived from the reference population, over relatedness classes.

To study the impact of null alleles, we assumed an extreme simulation scenario where each putative homozygote in the RP was a carrier of one null allele. Pairwise relatedness estimators (r_{xyn}) were obtained with the procedure explained before and were compared to the corresponding cases without null alleles.

From each *r*-value distribution obtained from *Zeta,* based on 10^5 replicates, we randomly sampled 10,000 replicates (one tenth) and used them to draw density distributions. For each L, R, Q or LR estimator, we placed the four resulting density distributions from each relatedness group along the same axis and calculated the overlapping areas, i.e. UR–HS, UR–FS, UR–SF, HS–FS, HS–SF and FS–SF. The total overlapped area obtained per relatedness estimator was an indicator of its resolving power in distinguishing among relatedness classes. A similar procedure was carried out assuming null alleles. Based on the density distribution curves, we have also computed the exact percentiles at 2.5% and 97.5% to frame the simulated multilocus *r*-values for each relatedness group and coefficient. Density distributions and corresponding overlapping areas were computed with density functions written in the R statistical package (R Development Core Team 2008).

Most pairwise methods provide estimates within a continuous range that need to be converted into genealogical relatedness (UR, HS, FS and SF). This can be done through the use of arbitrary thresholds between relatedness classes, usually the midpoint between means of two consecutive relatedness classes (e.g. 0.125: UR–HS) (Blouin et al. 1996). We established the relatedness groups by looking at the overlapping area between density distributions and defining the relatedness value according to the interception point between any two overlapping distributions. This interception point was taken as the threshold between the two given relatedness classes (subsequently called the 'critical value'). This critical value minimizes both β and α errors (β is the overlapping area to the left of the critical value and α is the one to the right) (Kozfkay et al. 2008). Given that our interest was to know whether a given pair of individuals was unrelated or related to some extent, only one threshold between UR and the rest of the relatedness classes was obtained per estimator (L, R, Q or LR). The decision of accepting or rejecting the null (H_0: 'the pair are unrelated individuals') or the alternative hypotheses (H_1: 'the pair are half-sib individuals') was made comparing the observed *r*-value to the threshold. The threshold value was used to decide which pairs of the 24 trees from the elite population were related to some extent, at least at the half-sib level

(indicated by the comparison of the estimated pairwise *r*-value with the threshold value), using the pairwise LR values (Fig. 5).

The relatedness estimator with the smallest percentage of overlapping density probabilities and lower impact from the presence of null alleles was selected for further analysis with the 24 individuals of the elite population.

The validation population (three HS, three FS and four SF families) relatedness estimators were calculated using the SPAGeDi version 1.2 software (Hardy and Vekemans 2002). The pairwise relatedness matrix of the LR coefficient estimates for the 24 elite clones was used to perform an unweighted pair group method with arithmetic mean (UPGMA) dendogram. The UPGMA tree topology was tested by comparing the elite clones LR pairwise matrix and the correspondent cophenetic matrix through a Mantel test (Sokal 1979). A normalized *Z* test was performed. The observed value after 1,000 permutations should be significantly larger than that expected by chance, in order for an association to be accepted. NTSYSpc version 2.1 (Rohlf 2000) was used to compute the UPGMA and the Mantel test.

3 Results

3.1 SSR loci

The effective number of alleles per loci (N_e) in the reference population ranged from 6 to 21, with an average of 16.8 (Table 1). The observed heterozygosity (H_o) values ranged from 0.5 to 0.9. Loci with the same number of alleles (N_a) exhibited different effective number of alleles (N_e) and also different discriminant power (*D*) (*loci* with 21 alleles show N_e ranging from 5.2 to 12.8, Table 1). As an example, the allele frequency distributions of loci EMBRA5 and EMBRA23 (same N_a different N_e) are displayed in Appendix 2. Locus EMBRA23 has a more even allele frequency distribution compared to locus EMBRA5, which results in differences in N_e, though they have the same N_a. EMBRA5 has few high frequent alleles and many alleles with very low frequencies. Loci that displayed higher values of N_a/N_e also showed higher values in H_e/H_o ratio (i.e. EMBRA5, EMCRC5, EMBRA20 and EMCRC2) and are among the loci with lowest *D*.

High F_{is} values—the loss of heterozygosity due to non-random mating of parents —reflected differences between observed and expected heterozygosity. We need to note here that the reference population included individuals selected in stands after phenotypic evaluation and without pedigree information. Loci displayed different deviations from Hardy–Weinberg expectations (HWE), and half of them were not under HWE. The presence of null alleles is one complementary hypothesis for departures from HWE. Table 1 shows null allele frequencies above 5% and significant HWE deviations for EMBRA6, EMCRC10, EMBRA20, EMCRC2 EMBRA5, EMCRC5 and EMBRA19. All loci combinations gave non-significant *linkage disequilibrium* values after the Bonferroni correction. The only locus without mapping information (EMCRC5) appeared not linked to any other marker. Therefore, we assumed that all the markers used in this study have independent segregation.

3.2 Relatedness estimators

All estimators revealed similar levels of upward bias (the distance between the expected relatedness value and the observed mean) (Fig. 1), more evident in the higher relatedness class (FS and SF). Despite these biases, expected values fell well within exact percentiles at 2.5% and 97.5% for all four estimators and relatedness classes. R showed a different behaviour, with overlapping exact percentiles at 2.5% and 97.5% for all the relatedness classes. According to this information, unrelated individuals could be distinguished from FS and SF individuals, and HS could be distinguished from SF individuals, for all estimators except R. The LR estimator produced slightly smaller exact percentiles at 2.5% and 97.5% (confidence percentiles=CP) than the Q estimator, in particular the UR class. The L estimator had slightly smaller confidence percentiles than LR, but not in the case of the unrelated individuals. Considering the percentage of overlapping areas of the density distributions of *r*-values (without taking into account the presence of null alleles), on average, the R coefficient had the highest mean overlapping distributions' area (OD) across relatedness groups (20.8%) and the LR estimator the lowest (11.6%), as shown in Table 2. The percentage of overlapping area was higher, for the comparison between FS–SF (36.5%), followed by the HS–FS and the UR–HS. The lowest OD was found in the UR–SF, with no overlapping areas for LR and Q estimators. Therefore, the overlapping area for LR was generally the lowest, with the exceptions in the comparison UR–HS where it equalled R and in HS–FS where the L coefficient had a slightly better performance.

Considering nonparametric tests the overlapping areas, the worst behaving coefficient is R. LR proved to be the best overall performing relatedness estimator displaying the smallest average percentage of overlapping areas (11.6%), when compared with the other estimators' ODs (Table 2).

In Fig. 2, the density distributions for all relatedness estimators, without null alleles, are represented. L, Q and

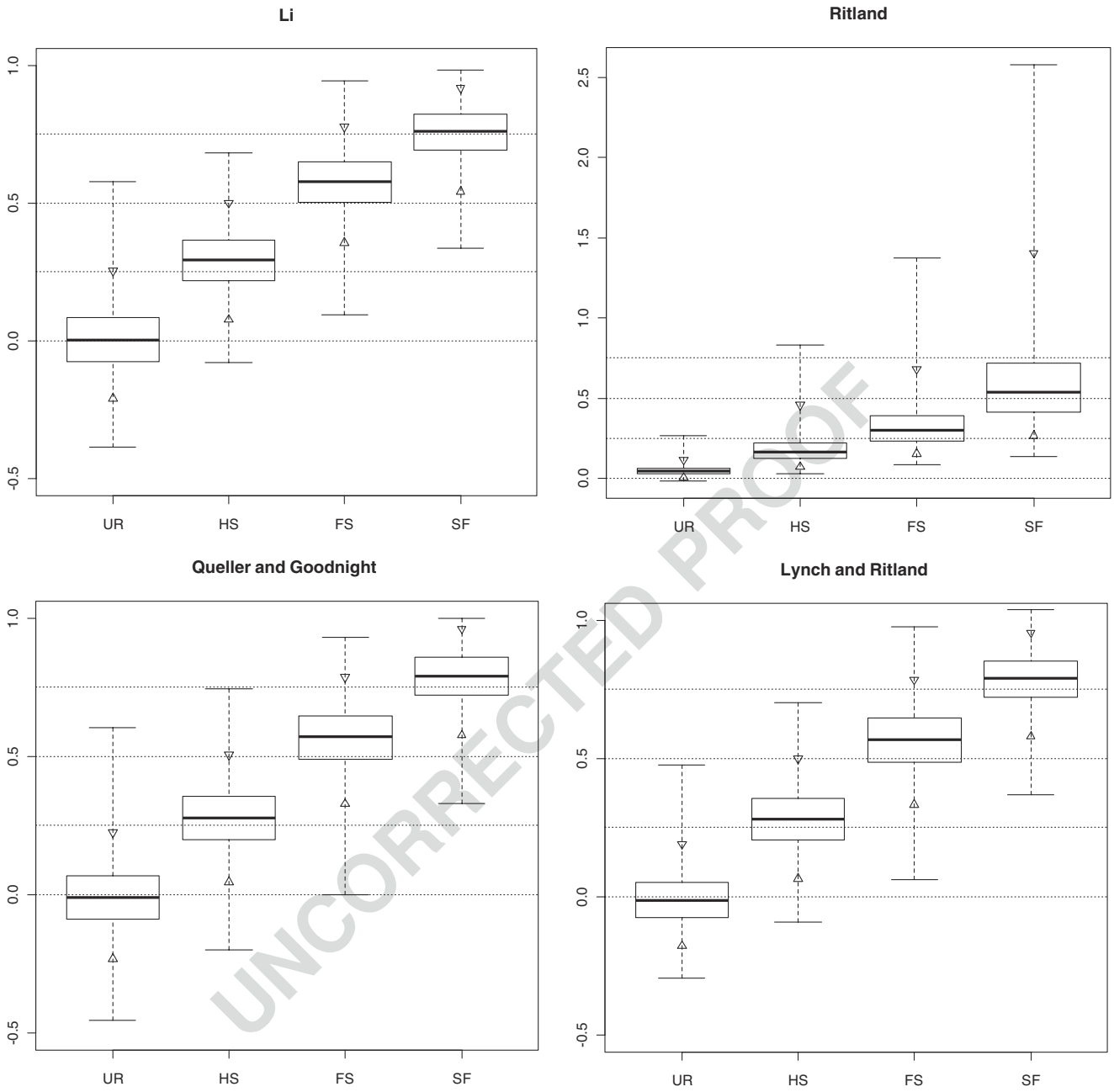


Fig. 1 Distribution of simulated multilocus *r*-values (*whiskers* for maxima and minima, *triangles* for exact percentiles at 2.5% and 97.5%, *bottom and top of the box* for the lower and upper quartiles, respectively, and *band near the middle of the box* for the median) in the different relatedness groups (unrelated, half-sibs (HS), full-sibs (FS) and individuals generated by selfing a single parent (SF) for different relatedness/coancestry estimators: Li et al. (1993) (L), Ritland (1996) (R), Queller and Goodnight (1989) (Q) and Lynch and Ritland (1999) (LR)

LR show approximately similar densities, with LR having a slightly narrower curve for UR. In general, these three estimators show symmetrical curves for UR, with asymmetry increasing progressively towards classes with higher relatedness. In SF class of *r*-values, the right tail is slightly shorter than the left tail, i.e. exhibiting negative skewness. Considering the R estimator, the density curve was extremely leptokurtic for UR *r*-values, and with increasing platykurtic properties and positive skewness towards classes with higher relatedness.

The LR pairwise relatedness values computed for the groups of individuals with known pedigree (SF, full-sibs, half-sibs and unrelated) are shown in Fig. 3, together with the corresponding exact percentiles at 2.5% and 97.5%. Observed LR relatedness appears slightly downward biased for half-sib and full-sib groups, while SF shows upward

t2.1 **Table 2** Relatedness group overlapping distribution areas excluding and accounting for null alleles (percent)

		L		R		Q		LR		Mean	
t2.3		No nulls	Nulls	No nulls	Nulls	No nulls	Nulls	No nulls	Nulls	No nulls	Nulls
t2.4	UR–HS	21.49	38.70	15.45	26.50	21.87	37.35	15.53	29.05	18.58	32.90
t2.5	UR–FS	1.32	8.95	2.23	7.10	1.40	8.08	0.67	4.27	1.40	7.10
t2.6	UR–SF	0.07	0.11	0.31	1.64	0.03	0.31	0.00	0.13	0.10	0.55
t2.7	HS–FS	19.38	40.38	44.35	53.25	21.78	40.11	21.00	36.16	26.63	42.47
t2.8	HS–SF	2.48	1.90	16.80	26.50	1.83	5.12	1.56	5.04	5.67	9.64
t2.9	FS–SF	38.35	16.17	45.50	60.40	31.13	30.23	30.87	34.02	36.46	35.20
t2.10	Mean	13.85	17.70	20.77	29.23	13.01	20.20	11.60	18.11		

See Figs. 2 and 4 for details

estimates, when compared to theoretical expectations. All observed estimates fell within the exact percentiles at 2.5% and 97.5%. Additionally, LR was also calculated for the reference population of 125 putatively unrelated trees. Results not shown graphically here indicate that 4.4% of relatedness fell beyond what would be expected to be the upper bound for unrelated pairs, based on the 97.5% exact percentile for UR, with a β error of 8%.

3.3 Impact of null alleles on relatedness estimators

ODs per relatedness coefficient and across relatedness classes when null alleles were assumed are shown in Table 2. In general, the inclusion of null alleles led to increases in OD, making it more difficult to differentiate the four relatedness classes through the use of the estimators. Only a few cases involving SF with L exhibited lower OD with null alleles than without them. Considering the resulting ODs per estimator, R remained the one with the highest overlapping areas amongst density distributions of *r*-values. The other three estimators had similar ODs, with L showing the smallest, closely followed by LR, and Q being the second largest.

Density distributions are represented for all relatedness estimators with null alleles in Fig. 4. In general, the inclusion of null alleles led to distributions of larger variances and correspondingly broader bell shapes. As a consequence of that, the overlapping areas were larger under the hypothesis of null alleles and also the mode decreased, at least for L, Q and LR, in particular for the higher relatedness classes. The only exception was the L estimator and the SF class, for which the overlapping area with other neighbouring distributions was smaller.

Therefore, in general, the presence of null alleles resulted in increased difficulties to discriminate among relatedness classes. All estimators showed this effect, though in different extents, with L being the estimator with the lowest impact in the case of the SF.

3.4 Pairwise relatedness of elite clones

After 1,000 permutations, the Mantel test showed that the simulated LR values between pairs of elite clones were larger than the observed values (r=0.65; P<0.001), a moderate correlation yet significant. The average (±SD) pairwise elite clone relatedness values computed with the LR coefficient was −0.045±0.067. Out of the 276 pairwise LR values, only four (1.4%) pairwise comparisons between elite clones had an LR estimator greater than the critical value of 0.126. Most of the other values were close to zero (Fig. 5), suggesting that levels of relatedness among selected clones are generally low. The critical value of 0.126 comes from the interception between UR and HS density distributions (Fig. 2). Therefore, pairs of individuals with relatedness above this critical value may be considered related to some degree, at least at a level close to HS. The risk here is type II error, where a pair of individuals is considered unrelated when in fact they are related to some extent. In this latter case, the type II error was 8%, i.e. the overlapping area to the left of the critical value for the UR vs. HS test. The pairs with LR greater than the critical point were CE7–CE22 (0.1316), CE5–CE13 (0.1543), CE8–CE23 (0.1701) and CE21–CE24 (0.3727). The last pair's LR value is a logic result, since it was discovered that CE21 is the mother of CE24, with an expected relatedness coefficient of 0.5.

When we account for the presence of null alleles, the critical values decreased from 0.126 to 0.088 for the UR–HS, and from 0.216 to 0.189 in the UR–FS case. Considering the new critical value (0.088), the probability of type II error increased (14.4%), as well as the number of putatively related pairs in the elite population. Two additional pairs were detected: CE17–CE20 (0.0902) and CE3–CE14 (0.0965).

All other relatedness coefficients had critical values above that for LR and therefore were less stringent in detecting related pairs of individuals.

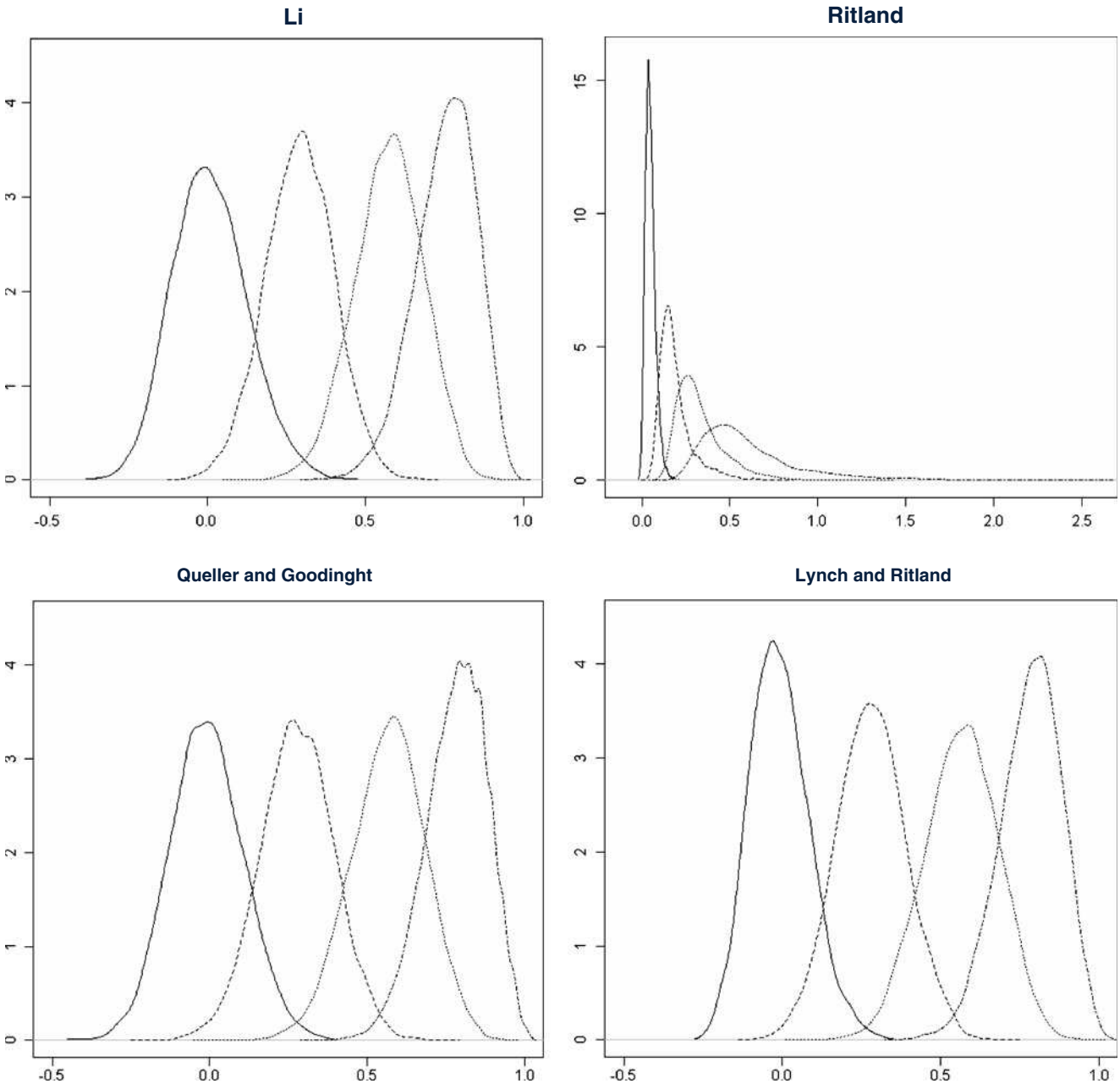
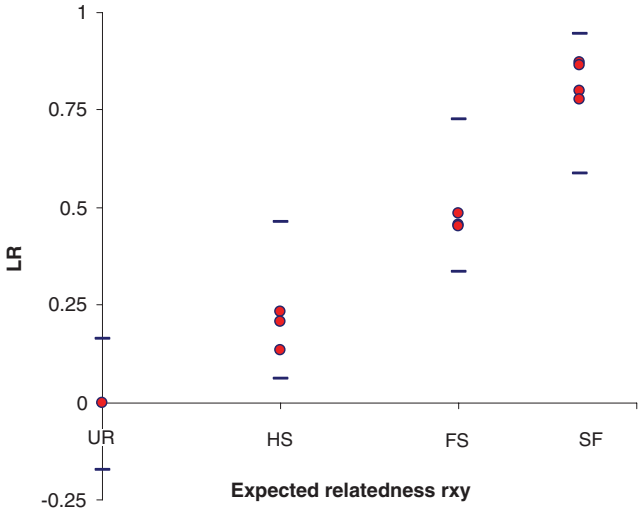


Fig. 2 The plotted values are the density distributions obtained from Monte Carlo simulations based on 10,000 replicas, excluding null alleles. In the *x-axis* the relatedness range and in the *y-axis* the density values. The overlapping distributions from *left to right* represent UR, HS, FS and SF for the different relatedness/coancestry estimators: Li et al. (1993) (L), Ritland (1996) (R), Queller and Goodnight (1989) (Q), and Lynch and Ritland (1999) (LR)

4 Discussion

4.1 SSR markers' informativeness

The average expected heterozygosity reported in the literature for *E. globulus*, using SSR markers, is similar to the value we obtained in the current study (~0.85). However, reported H_o is generally lower than our observed value (0.73): 0.66 (Steane et al. 2001) and 0.62 (Jones et al. 2002). The fact that we used an artificial population could explain, at least partly, the higher levels for H_o found in our study. In an Australian breeding population (140 individuals), Jones et al. (2006) obtained H_e=0.82 and H_o=0.71, with H_o being lower in the corresponding native populations that they studied (0.66). Astorga et al. (2004) detected similar values in *E. globulus* using 26 SSR markers with trees selected in progeny trials: H_e=0.80 and H_o=0.70. Finally, in other



Q8

Fig. 3 LR relatedness coefficient pairwise values based on real data (*filled circles*) framed by the exact percentiles at 2.5% and 97.5% (*between dashes*) from the simulated data, as in the Lynch and Ritland plot from Fig. 1, in the different relatedness groups. The reference population was used to estimate the unrelated pairs pairwise LR

studies using microsatellites in *E. grandis* and *E. urophylla*, the average observed heterozygosity was much smaller than the expected one ($H_o \approx 0.56–0.62$ and $H_e \approx 0.86–0.82$) (Brondani et al. 1998, 2002).

In terms of the amount of expected heterozygosity, Blouin et al. (1996) concluded that 10 loci with $H_e=0.75$ would accurately discriminate more than 90% of the FS from UR individuals, but 14 loci would be required to achieve the same level of discrimination between FS and HS. In this context, the circumstances of the present study are seemingly far more promising, as only one marker out of 16 had $H_e < 0.75$.

However, besides expected heterozygosity, other factors play a role in the quality of relatedness discrimination, like the number of available SSR loci, the number of segregating alleles and their spectra of frequencies. Different relatedness estimators respond differently to the available sample (Milligan 2003), making prospective studies invaluable. Ideally, marker locus should have a large number of alleles with even allelic frequencies. For instance, EMBRA23 showed the highest *D*, or discrimination power, as well as one of the flattest allele frequency distributions. Other less informative loci brought, however, additional precision to the multilocus estimates of relatedness. Dropping the less polymorphic loci, for example, if suspected of hosting null alleles, as advised by Dakin and Avise (2004), could increase the estimator's sampling variance. It is expected (Milligan 2003) that the standard error of the estimator declines with the number of loci. Furthermore, some of the less polymorphic markers with uneven allele distributions have rare alleles, which are important to discriminate some genotypes.

4.2 Relatedness coefficient selection

Marker-based relatedness estimates typically show a large error of inference (Ritland 1996; Lynch and Ritland 1999). One of the sources of variation comes from the recombination and segregation of polymorphic markers (Blouin 2003). However, there are differences between relatedness estimators, and these are usually dependent on the characteristics of the sample, such as allele frequency spectra, number of alleles per locus and the actual range of pedigree relatedness to be estimated. Van de Casteele et al. (2001) suggested the use of prospective studies to evaluate different estimators in the context of the target population, for instance, by the use of Monte Carlo simulations with actual data. Other studies of this kind used the allele frequencies obtained from real data to simulate gene pools from which to draw pairs of related individuals (e.g. Blouin et al. 1996; Lynch and Ritland 1999; Van de Casteele et al. 2001; Milligan 2003). In our study, we used the real genotypes of the reference population as a source of virtual gametes from which to obtain pairs of related and unrelated individuals in silico. The advantage of our approach is to be closer to the actual genotypic arrangements, when selecting the best fitted estimator for a particular population, and to take into account any deviation due to linkage disequilibrium between markers. Such deviations from equilibrium are common in breeding populations, which are usually artificial composites of genotypes coming from different origins.

The simulation approach allowed us to select LR as the relatedness estimator best fitted for fingerprinting the population under study. LR was unbiased, more accurate, with lower percentage of overlapping values between relatedness groups and smaller exact confidence percentiles. Moreover, it demonstrated smaller impact when null alleles were present, except in the case of higher relatedness values. These features are important because they improve the ability to identify, with statistical confidence, unrelated from related individuals. Thomas (2005) refers that the regression-based relatedness estimator of Lynch and Ritland (1999) (our LR) shows the most desirable properties over the widest range of marker data. In agreement with Van de Casteele et al. (2001), the author adds that, ideally, simulations should be used to check whether this holds true for the particular population under study. Csillery et al. (2006) studied natural outbred populations that were less related than half-sibs and, in agreement with our findings, concluded that the Q estimator had smaller sampling variances in high relationship categories while LR was better in the low relationship categories. Furthermore, Blouin et al. (1996), in their study on misclassification in sheep, found that for all populations

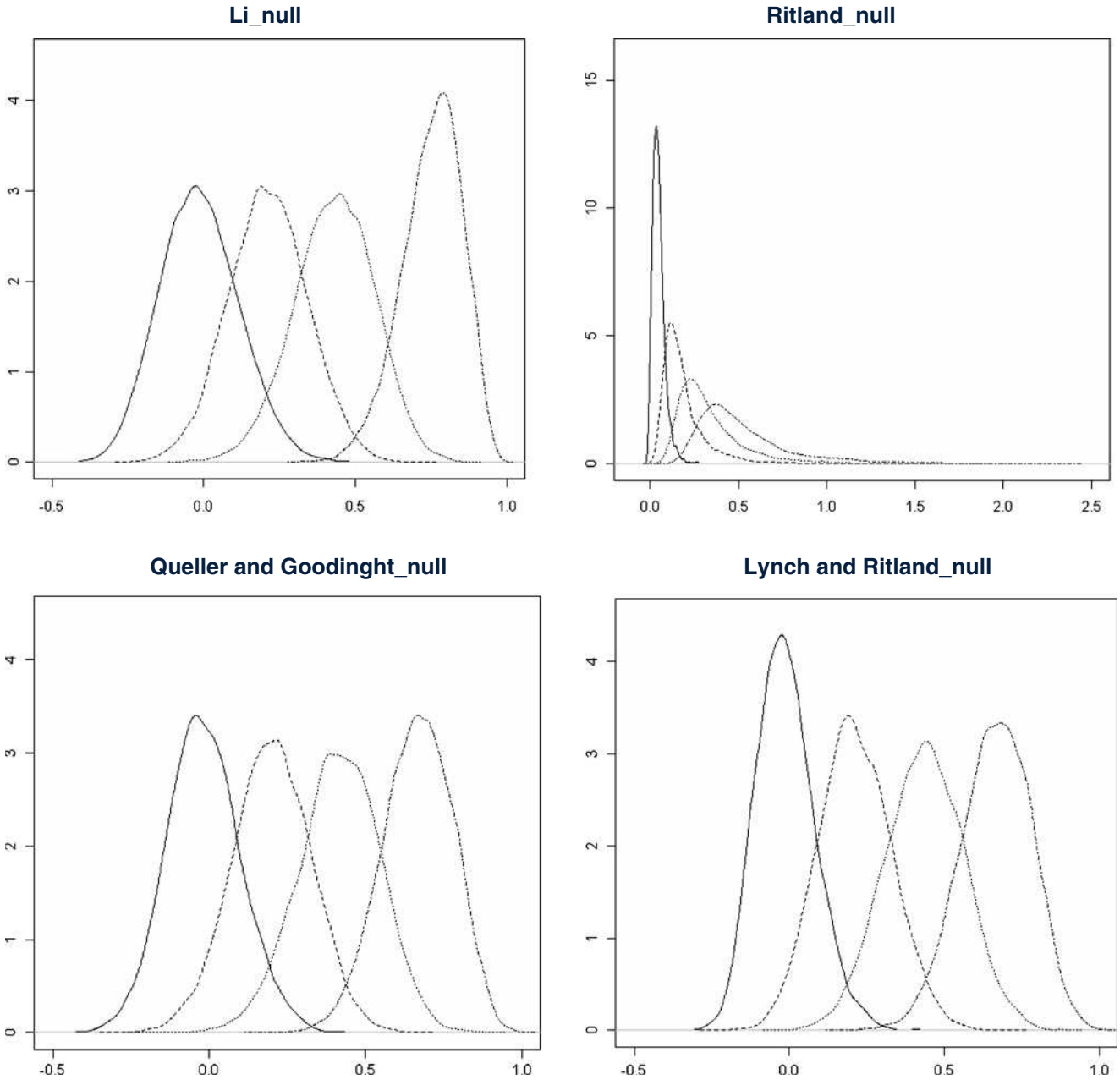


Fig. 4 The plotted values are the density distributions obtained from Monte Carlo simulations based on 10,000 replicas, accounting for the presence of null alleles. In the *x-axis* the relatedness range and in the *y-axis* the density values. The overlapping distributions from *left to right* represent UR, HS, FS and SF for the different relatedness/coancestry estimators: Li et al. (1993) (L), Ritland (1996) (R), Queller and Goodnight (1989) (Q) and Lynch and Ritland (1999) (LR)

studied, the misclassification rate was lowest with the LR estimator. They demonstrated that the highest proportion of the relatedness variance was explained with LR, reflecting the fact that this estimator had the smallest sampling variance for the UR or low-related pairs, which are more common in outbred populations (Csillery et al. 2006). In our study, we wanted to discriminate the unrelated from the related individuals and therefore needed a coefficient with higher precision for the low-related pairs of individuals.

The results from our study also confirmed those presented by Ritland and Travis (2004), where the LR estimator showed lower error variances compared with R, except for the class of unrelated individuals. Indeed, we found that the exact confidence percentiles of R increased rapidly with the expected values of coancestry, making it unsuitable for assigning a relatedness group for most of the observed *r*-values (Fig. 1). Milligan (2003) points out that the R estimator performs less well than other estimators,

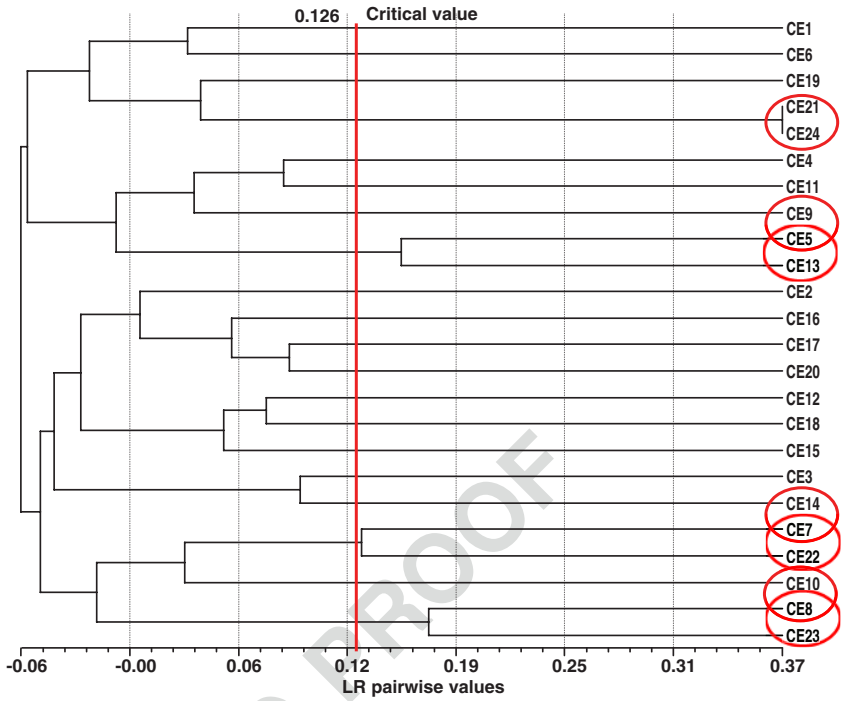


Fig. 5 Elite clones' relatedness dendrogram (UPGMA) built with the Lynch and Ritland (1999) pairwise relatedness estimator matrix. The *x-axis* represents the LR coefficient similarity distances, the *labels in the right part* of the figure, from *CE1* to *CE24*, are the elite clones' codes. The *vertical dotted lines* represent relatedness values intervals. The *vertical straight line* corresponds to the threshold (critical value=0.126) to distinguish UR from HS. The four pairs of individuals that were found to be related to a certain extend were included inside *circles*

especially under conditions of high relatedness and less polymorphic markers. In the same paper, Milligan shows that estimators of relatedness are often skewed, Q and R in particular, but in opposite directions. This was confirmed in our study, though Q was only slightly skewed to the right for high relatedness distributions (Fig. 4). This skewness may have significant impacts on the use of these estimators, as suggested by Milligan (2003), because means and modes do not match.

4.3 Validation with individuals of known pedigree

After selecting the most suitable estimator, LR pairwise relatedness values were computed in groups of individuals with known pedigree (UR, HS, FS and SF), for validation. All families' *r*-values were within the simulated exact percentiles at 2.5% and 97.5% for each relatedness group (Fig. 3). The slight departures of observed *r*-values from expected values are not easily explained. These departures correspond to upward biases for SF and downward biases for HS and FS. Asymmetries in the distribution of expected values do not appear to be a possible cause, as distributions for HS and FS were nearly symmetrical, while that of SF presented less values being greater than the mode. The relatively small number of families and their small size could increase the sampling effects.

4.4 Null allele impact

Our analyses revealed an important deficit of observed heterozygosity for some markers, from what would be expected from allelic frequencies in the reference population. Other studies with *E. globulus* also found deficits in observed heterozygosity (e.g. Astorga et al. 2004; Jones et al. 2006). The presence of a relatively high percentage of null alleles could be one of the main reasons for this. Based in our estimations, seven out of the 16 SSR loci had null allele's frequencies above 5%. This high number of affected loci could be partially explained by the fact that EMBRA SSR loci were originally developed for *E. grandis* (Brondani et al. 1998). The frequency of null alleles is expected to increase when transferring markers between more distantly related species. Indeed, in their study, Brondani et al. (2006) observed that the overall occurrence of null alleles was much higher in *E. urophylla* than in *E. grandis*, when using SSR originally developed from *E. grandis* libraries.

The presence of null alleles had a negative effect in all relatedness estimators, as expected from the literature (Wagner et al. 2006). Our assumption was extreme in the sense that all homozygotes were considered to be carriers of null alleles, hence being an upper bound for the expected effects of null alleles. Null alleles increased the variation associated to each estimator and consequently the overlapping areas between neighbouring density distributions of simulated *r*-values. This had the effect of increasing the associated α and β errors. Accordingly, critical values between relatedness classes decreased with null alleles. As a consequence, the probability of type II error and the number of putatively related pairs detected in the elite population increased. As a principle of precaution, and given the likelihood of null alleles when working with

transferred markers from distant species, pairs detected as related, close to the critical value, should be considered related, at the risk of falling into type II errors. Nevertheless, our results show that the Lynch and Ritland (1999) relatedness estimator proved adequate for our data set, even when all the homozygotes were considered carriers of null alleles.

4.5 Putatively related elite clones

Excluding the presence of null alleles, four pairs of putatively unrelated elite individuals were considered related to the level of half-sibs, based on the LR estimator. This represents a small portion (1.4%) of all the possible pairwise values (276) in the relatedness matrix. In the worst-case scenario, when all homozygotes were considered carriers of a null allele, we detected two additional pairs of putatively related individuals (2.2% of the total). Despite the fact that the group of elite clones had, to our knowledge, no recent common ancestors, there might have been an influx of relatedness into the Portuguese land race (Borralho et al. 2007) or mislabelling in the breeding population management. Recently established plantations may have been originated from the same seed collected on a few trees, with pollination dominated by a restricted number of males. Moreover, eucalypts have a mixed mating system, and the collected open-pollinated seeds from one mother-plant may contain a mixture of selfs (and possibly other forms of inbreeding) and unrelated crosses (Eldridge et al. 1994; Jones et al. 2006; Costa e Silva et al. 2010). This would explain why some of the elite clones, selected in different plantations, could show some level of relatedness. Decisions about the elite clones to be used in future crossings schemes should take into account not only their breeding values but also their level of coancestry. The long-term effect of inbreeding depression on traits related to fitness, such as survival and growth, is severe in *E. globulus* (Costa e Silva et al. 2010), neutralizing improvement efforts.

We present a simulation approach that allows different estimators to be evaluated in a particular context, even when a population is the result of an artificial mixture of different origins. In the absence of reliable pedigree information, LR values could be useful to avoid or limit consanguinity and to identify relatives in breeding populations. However, our goal was not simply to confirm the suitable properties of LR compared to other relatedness estimators. Indeed, some of the results could be expected given the characteristics of the population under study, notably the absence of high relatedness that could pinpoint the use of LR. Our objective was to propose a method that could be easily applied to other populations and species, confronted with the dilemma of selecting from a series of relatedness estimators.

References

Astorga R, Soria F, Basurco F, Toval G (2004) Diversity analysis and genetic structure of *Eucalyptus globulus* Labill. In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M, Tomé M (eds) Eucalyptus in a changing world. IUFRO, RAIZ, Instituto Investigação da Floresta e Papel, Aveiro, pp 351–358

Ballou JD, Lacy RC (1995) Identifying genetically important individuals for management of genetic variation in pedigreed populations. In: Ballou JD (ed) Population management for survival and recovery. Columbia University Press, New York, pp 76–111

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. TREE 18:503–511

Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 5:393–401

Borralho NMG, Almeida MH, Potts BM (2007) O melhoramento do eucalipto em Portugal. In: Alves AM, Pereira JS, Silva JMN (eds) O eucaliptal em Portugal. ISA Press, Lisbon, pp 62–110

Brondani RPV, Brondani C, Tarchini R, Grattapaglia D (1998) Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. Theor Appl Genet 97:816–827

Brondani R, Brondani C, Grattapaglia D (2002) Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. Mol Genet Genomics 267:338–347

Brondani RPV, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. BMC Plant Biol 6:20

Costa e Silva J, Hardner C, Tilyard P, Pires AM, Potts BM (2010) Effects of inbreeding on population mean performance and observational variances in *Eucalyptus globulus*. Ann For Sci 67:605

Csillery K, Johnson T, Beraldi D, Clutton-Brock T, Coltman D, Hansson B, Spong G, Pemberton JM (2006) Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics 173:2091–2101

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

Dutkowski GW, Potts BM (1999) Geographic patterns of genetic variation in *Eucalyptus globulus* ssp. *globulus* and a revised racial classification. Aust J Bot 47:237–263

Eldridge K, Davidson J, Harwood C, van Wyk G (1994) Eucalypt domestication and breeding. Clarendon Press, Oxford, 312 p

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman Group, Ltd., Essex, 480 p

Hardner CM, Potts BM (1995) Inbreeding depression and changes in variation after selfing *Eucalyptus globulus* ssp. *globulus*. Silvae Genet 44:46–54

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

Jones RC, Steane DA, Potts BM, Vaillancourt RE (2002) Microsatellite and morphological analysis of *Eucalyptus globulus* populations. Can J For Res 32:59–66

Jones TH, Steane DA, Jones RC, Pilbeam D, Vaillancourt RE, Potts BM (2006) Effects of domestication on genetic diversity in *Eucalyptus globulus*. For Ecol Manag 234:78–84

Kozfkay C, Campbell M, Heindel J, Baker D, Kline P, Powell M, Flagg T (2008) A genetic evaluation of relatedness for brood-stock management of captive, endangered Snake River sockeye salmon, *Oncorhynchus nerka*. Conserv Genet 9:1421–1430

Lefèvre F (2004) Human impacts on forest genetic resources in the temperate zone: an updated review. For Ecol Manag 197:257–271