Mol Breeding (2016) 36:97
DOI 10.1007/s11032-016-0522-7

SSRs, SNPs and DArTs comparison on estimation of relatedness and genetic parameters' precision from a small half-sib sample population of *Eucalyptus grandis*

Eduardo P. Cappa · Jaroslav Klápště · Martín N. Garcia · Pamela V. Villalba · Susana N. Marcucci Poltri

Received: 2 September 2015 / Accepted: 27 June 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Simple sequence repeats (SSR) are the most widely used molecular markers for relatedness inference due to their multi-allelic nature and high informativeness. However, there is a growing trend toward using high-throughput and inter-specific transferable single-nucleotide polymorphisms (SNP) and Diversity Arrays Technology (DArT) in forest genetics owing to their wide genome coverage. We compared the efficiency of 15 SSRs, 181 SNPs and 2816 DArTs to estimate the relatedness coefficients, and their effects on genetic parameters' precision, in a relatively small data set of an open-pollinated progeny trial of *Eucalyptus grandis* (Hill ex Maiden) with limited relationship from the pedigree. Both simulations and real data of *Eucalyptus grandis* were used to study the statistical performance of three relatedness estimators based on co-dominant markers. Relatedness estimates in pairs of individuals belonging to the same family (related) were higher for DArTs than for

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0522-7) contains supplementary material, which is available to authorized users.

E. P. Cappa (✉)
Bosques Cultivados, Instituto de Recursos Biológicos, Centro de Investigación en Recursos Naturales, Instituto Nacional de Tecnología Agropecuaria (INTA), De Los Reseros y Dr. Nicolás Repetto s/n, 1686 Hurlingham, Buenos Aires, Argentina
e-mail: cappa.eduardo@inta.gob.ar

E. P. Cappa
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

J. Klápště
Department of Forest and Conservation Sciences, University of British Columbia, 2424 Main Mall, Vancouver, BC V6T 1Z4, Canada

J. Klápště
Department of Genetics and Physiology of Forest Trees, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamycka 129, 165 21 Praha 6, Czech Republic

M. N. Garcia · P. V. Villalba · S. N. Marcucci Poltri
Instituto de Biotecnología, Centro de Investigación en Ciencias Veterinarias y Agronómicas, Instituto Nacional de Tecnología Agropecuaria (INTA), De Los Reseros y Dr. Nicolás Repetto s/n, 1686 Hurlingham, Buenos Aires, Argentina

Present Address:
J. Klápště
Scion (New Zealand Forest Research Institute Ltd.), Private bag 3020, 49 Sala Street, Rotorua 3046, New Zealand

Published online: 08 July 2016

SNPs and SSRs. DArTs performed better compared to SSRs and SNPs in estimated relatedness coefficients in pairs of individuals belonging to different families (unrelated) and showed higher ability to discriminate unrelated from related individuals. The likelihood-based estimator exhibited the lowest root mean squared error (RMSE); however, the differences in RMSE among the three estimators studied were small. For the growth traits, heritability estimates based on SNPs yielded, on average, smaller standard errors compared to those based on SSRs and DArTs. Estimated relatedness in the realized relationship matrix and heritabilities can be accurately inferred from co-dominant or sufficiently dense dominant markers in a relatively small *E. grandis* data set with shallow pedigree.

Keywords Molecular markers · Marker-based relationship matrix · Relatedness · Heritability · *Eucalyptus grandis*

Introduction

Knowledge of the relationships between individuals in a population is essential in many areas of genetics (Santure et al. 2010). For example, relationships between individuals are required to estimate heritabilities, a key parameter in natural and breeding populations of animals and plants. Precision of genetic parameter estimates, such as additive genetic variance and heritability, requires accurate information of the genetic relationship between individuals within the population under study (Rodríguez-Ramilo et al. 2007). In the classical individual-tree mixed model (i.e., animal model), the average numerator relationship matrix (NRM) based on information from pedigrees (Wright 1922) is used to appropriately consider additive genetic relationship between any pair of individuals (Henderson 1984). However, estimations based on models using the average NRM suffer from some potential limitations (Ødegård and Meuwissen 2012). First, they ignore relationships beyond the known pedigree (i.e., historical relatedness accumulated in populations from which the base population is derived). Additionally, the NRM contains only expected (based on the expected proportion of alleles identical-by-descent, IBD) rather than actual, or realized relationships. Assuming unrelated parents, the relationship coefficient between any pair of half-sibs will always be 0.25 when estimated from pedigree (i.e., the expected proportion of the genome that is IBD equals to 0.25). However, due to sampling during meiosis, two half-sib individuals may actually share more or less than 25 % of the genome by IBD. This means that the expected relationship matrix derived from the pedigree cannot capture the effect of Mendelian sampling produced during meiosis (Hayes and Goddard 2008), and estimates of genetic (co)variance component are estimated based only on between-family variation.

The use of molecular marker information to infer the realized relationship matrix was proved as an efficient alternative to constructing the average NRM when the pedigree is incomplete or missing (e.g., Kumar and Richardson 2005; Bessega et al. 2011; El-Kassaby et al. 2012). A marker-based relationship matrix may better estimate the exact (i.e., realized) proportion of alleles IBD shared between individuals with high degree of precision (Villanueva et al. 2005). The idea of estimating quantitative genetic parameters using relatedness estimates derived from molecular markers has been initially investigated by Ritland (1996) and further explored by a number of authors (e.g., Mousseau et al. 1998; Thomas and Hill 2000; Thomas 2005). Estimation of genetic parameters using marker-based relationship matrix is especially useful in studies of wild populations where pedigree information is not known (Frentiu et al. 2008; Sillanpää 2011) or small breeding populations with limited pedigree information and/or few available parents (Ødegård and Meuwissen 2012).

There are numerous different types of molecular markers used in plant genetic analyses to estimate the coefficients of relatedness and infer the realized relationship matrix. Genetic markers can be classified as co-dominant or dominant, depending on their ability to distinguish allelic status of a heterozygote from a dominant homozygote. Microsatellites or simple sequence repeats (SSR) are short tandem repeated DNA sequences, widely distributed throughout the eukaryotic genomes. SSRs are probably the most widely used genetic markers for relatedness inference because they typically display many alleles per locus (i.e., are highly polymorphic) and are co-dominant by their nature, resulting in highly informative markers (e.g., Hardy 2003). The single-nucleotide polymorphisms (SNP) are a more recently alternative

to SSR markers and represent single-point base mutation. For a review of many technical and statistical uses of SSRs and SNPs see Vignal et al. (2002). SNPs are bi-allelic markers which may limit their resolving power per locus (Glaubitz et al. 2003). However, lower single-locus power can be compensated for by increase the number of loci assayed using high-throughput next generation sequencing technologies (e.g., "genotyping-by-sequencing") (Elshire et al. 2011). Alternatively, precise pairwise relatedness estimates might also be obtained using large number of dominant markers (Hardy 2003). Diversity Arrays Technology (DArT) is a genotyping platform based on genome complexity reduction, followed by hybridization to microarrays that offer a rapid and efficient method for high-throughput DNA marker analysis (Kilian et al. 2012). Particularly, in *Eucalyptus*, DArTs have been developed with a wide genome coverage and inter-specific transferability (Sansaloni et al. 2010). DArT dominant bi-allelic markers are scored as either present or absent (i.e., 1 or 0); therefore, DArTs provide less genetic information for a given locus than the co-dominant SSRs and SNPs (Simko et al. 2012). However, dominant markers can be developed relatively easily even for species for which no prior genomic information is available, and the cost per sample is much lower than current SNP genotyping platforms for an equivalent number of markers (Sansaloni et al. 2010). Therefore, these dominant markers may be viable alternatives for the estimation of relatedness between individuals (Hardy 2003).

Given the trend toward the increasing use of SNPs and DArTs in forest genetics, it is of interest to compare the performance of the multi-allelic co-dominant SSR with the bi-allelic co-dominant SNP and dominant DArT markers for estimating both relatedness coefficients and genetic parameters on the same set of individuals. DArT markers have been applied in the estimation of pairwise relatedness in cereal species such as barley (e.g., Wenzl et al. 2004) and wheat (e.g., Crossa et al. 2007). However, they have not been used to estimate pairwise relatedness in forest tree species to date. We are aware that only two other studies have explored the utility of DArT markers for the study of relationships (Steane et al. 2011; Przyborowski et al. 2013). Nevertheless, these studies were focused on relationships between species of *Eucalyptus* (Steane et al. 2011) and *Salix* (Przyborowski et al. 2013). Additionally, limited studies compared the dominant DArT markers with co-dominant SSR and SNP. Simko et al. (2012) compared different numbers of three types of markers: SSR, SNP and DArT for estimating the genotype diversity, clustering varieties into populations and assigning a single variety into the expected population in a set of 54 hybrid varieties of sugar beet. Lamara et al. (2013) used SSR and DArT markers for comparing different genetic diversity estimation methods among 92 Canadian barley cultivars. Similarly, Laidó et al. (2013) compared DArT and SSR markers to evaluate genetic diversity and population genetic structure of 230 accessions in seven tetraploid *Triticum turgidum* L. subspecies.

Several estimators have been developed to measure pairwise relatedness coefficients between individuals for co-dominant (e.g., Queller and Goodnight 1989; Li et al. 1993; Lynch and Ritland 1999; Wang 2002, 2007) and dominant (Hardy 2003) markers, and have been used in different areas of research (see review by Blouin 2003; Thomas 2005). The statistical properties and performance of these estimators have been studied using both empirical and simulated data sets (e.g., Lynch and Ritland 1999; Van de Casteele et al. 2001; Csilléry et al. 2006; Wang 2002, 2007). These studies concluded that several aspects contribute to the performance of these estimators and resulting marker-based heritabilities. For example, the true relatedness value being estimated, the informativeness of markers utilized in an analysis (number of loci, number and frequencies of the alleles at each locus), the size of the sample in estimating allele frequencies (Wang 2007), selection at closely linked loci, genotyping errors, mutation and recent inbreeding (because of the small population size and/or the mating system) (Glaubitz et al. 2003). Recent inbreeding will result in elevated pairwise relatedness. *Eucalypts* have a mixed mating system setting inbred, originating from selfing and mating between close relatives, as well as outcrossed seed. For example, estimates of outcrossing rates of open-pollinated families of *Eucalyptus grandis* (Hill ex Maiden) from natural populations and plantations averaged 84 % (Eldridge et al. 1993; Table 19.2). Therefore, depending on the proportion of offspring generated by self-pollination, a particular open-pollinated family will have a mixture of relatedness among individuals ranging from selves to half-sibs (Cappa et al. 2010). Moreover, self-fertilization in *Eucalyptus* may result

in inflated heritability estimates (e.g., Griffin and Cotterill 1988; Hodge et al. 1996; Lopez et al. 2002) and biased additive genetic correlation estimates across ages and sites (Hodge et al. 1996), without proper consideration of the mixed mating system.

The objective of this study was to compare the efficiency of two co-dominant (SSR and SNP) and one dominant (DArT) genetic markers on the estimation of relatedness coefficients between individuals, and their effects on the genetic parameters' precision, in a relatively small data set from an open-pollinated progeny trial of *Eucalyptus grandis* (Hill ex Maiden) with limited relationship from the pedigree. This data set could emulate small samples from natural populations with little relatedness. Furthermore, because the precision of both relatedness coefficients and genetic parameters estimated on the basis of the two co-dominant markers (SSR and SNP) may partly depend on the nature of the population under study and used estimator, we also compared two widely recognized moment-based methods, Lynch and Ritland (1999) (LR) and Queller and Goodnight (1989) (QG), and one likelihood-based method, Wang (2007) (W), using both simulations and *Eucalyptus grandis* data sets.

Materials and methods

Plant material and quantitative traits

A sample of 166 trees from an open-pollinated (OP) progeny trial of *Eucalyptus grandis* (Hill ex Maiden) (hereafter *E. grandis*) established at Gobernador Virasoro (latitude 28°02′S longitude 56°03′W alt. 105 m), northern Corrientes Province, Argentina, was used in this study. The trial comprised 148 OP families from native forest: 101 families from New South Wales and 47 from southeastern Queensland, Australia; and 16 OP families from two local land race sources from Concordia, Entre Rios Province, Argentina. A detailed description of this genetic material can be found in Marcó and White (2002). This trial corresponds to one of four trials following a randomized complete block design with 17–20 replications and single tree plots. The sampled population included trees from 123 OP families, represented by one or two trees per family. There were 81 families represented by one and 42 families by two trees. Consequently, expected relationships between trees from this breeding population are very sparse; however, this could be a common situation in wild forest populations.

All surviving trees were measured at 5 years after planting for growth traits: diameter at breast height over bark at 1.3 m above the ground level (DBH) and total height, and the volume was calculated according to Marcó and White (2002). Wood chemical traits extractives in ethanol and total extractives, Klason and total lignin, syringyl-to-guaiacyl ratio and wood basic density (BD) were also measured. In total, two growth (DBH and volume) and three wood property traits (total lignin, Klason lignin and BD) were investigated in this study.

Wood chemical components were estimated using near-infrared (NIR) spectroscopy. Wood samples were collected at 1.3 m above ground level and air-dried for predicting wood chemical composition. The wood sample was ground to pass through a 1-mm screen, and NIR spectra were obtained by diffuse reflectance using a Bruker Optics Co. MPA (Madison, WI, USA). Partial least squares regression (PLSR) was used for both evaluation of the NIR spectra (NIR-PLSR models) and calculation of the prediction models. These predictions were validated using chemical assays from 15 to 22 independent samples from those used to develop the model. All models were at least good enough for screening in breeding programs with a residual prediction deviation (RPD; Williams and Sobering 1993) above 2.5 (e.g., Alves et al. 2012). The RPD values were 3.2, 6.5 and 3.3 for total lignin, Klason lignin and BD, respectively.

Normality of the five traits was evaluated using PROC Univariate in SAS (SAS Institute Inc. 2002).

Molecular markers

The genomic DNA was extracted from young leaves using the CTAB method (Hoisington et al. 1994). Genetic variability was screened with three different molecular markers: simple sequence repeat (SSR), single-nucleotide polymorphisms (SNPs) and Diversity Arrays Technology (DArT).

Fifteen SSRs belonging to the 11 linkage groups (LG) defined in the *Eucalyptus grandis* genome (Petroli et al. 2012) and showing high polymorphism information content (PIC; PIC > 0.542) were selected: EMBRA11 (LG 1), EMBRA19 (LG 1), EMBRA648 (LG 2), EG131 (LG 3), EMBRA179 (LG

4), EMBRA36 (LG 4), EMBRA168 (LG 5), EMBRA5 (LG 5), EMBRA173 (LG 6), EMBRA51 (LG 6), EMBRA46 (LG 7), EMBRA47 (LG 8), EMBRA18 (LG 9), EMBRA61 (LG 10), EG024 (LG 11). EMBRA primers are described in Brondani et al. (2006) and the EG in Thamarus et al. (2002). SSRs were amplified with a fluorescent dye-labeled forward primer and separated on an ABI3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Allele assignments were made by size comparison with the standard allelic ladder, using the GeneMapper ID software version 4.0 provided by Applied Biosystems.

A panel of 384 SNPs developed by Grattapaglia et al. (2011) using the Illumina GoldenGate Genotyping assay on an Illumina BeadXpress platform with VeraCode technology was implemented for SNP genotyping. All reagents, unless stated otherwise, were provided by Illumina. Genotyping results were analyzed by GenomeStudio Genotyping module version 1.1 (Illumina, San Diego, CA, USA). SNP markers were filtered and rejected according to the following criteria: marker was monomorphic, presented GenTrain scores $\leq$0.4 (number between 0 and 1 indicating how well the samples clustered for the loci), GenCall scores $\leq$0.4 (number between 0 and 1 indicating the reliability of each genotype call) and call rate values <0.25 (percentage of targets that could be scored as 0 or 1). Finally, 181 out of 384 SNP markers were selected for further analysis. An average of 16 SNP markers per LG were assayed with a minimum of 1 marker in linkage group 2 and a maximum of 33 in linkage group 7.

A subset of 2816 DArT markers was selected after screening the high-throughput array-based genotyping system of 7680 DArT developed (Sansaloni et al. 2010). The selected markers showed call rate values >0.8, reproducibility values >0.97 (reproducibility of scoring between replicated target assays) and polymorphism with allele frequencies ranging from 0.95 to 0.05. From the *Eucalyptus* composite map (Hudson et al. 2012), 1769 of the selected 2816 DArTs had a known map location with a reasonable genome-wide coverage was provided by these markers.

In summary, the final data set comprised 15 SSRs, 181 SNPs and 2816 polymorphic DArTs. These markers were used to genotype the 166 selected trees originating from 123 families of the OP progeny trial of *E. grandis*.

Different diversity genetic parameters were estimated for each SSR, SNP and DArT marker with GenAlEx 6.5 software (Peakall and Smouse 2012) including number of observed alleles per locus (N_a; calculated as the number of different alleles), effective number of alleles per locus (N_e; calculated as $1/\sum p_i^2$, where p_i is the frequency of the ith allele), Shannon's Information Index ($\boldsymbol{I}$, calculated as $I = -\sum p_i \ln(p_i)$), observed heterozygosity (H_o; calculated as the number of heterozygotes per locus divided by the number of individuals typed) and the expected heterozygosity (H_e; calculated as: $H_e = 1 - \sum p_i^2$).

Due to high computational cost, a subset of 800 random DArT markers and the 15 SSRs were selected to study the population structure using the Bayesian method in the STRUCTURE software (Pritchard et al. 2000). STRUCTURE analyses were performed assuming an admixture model with default settings (i.e., no informative priors were used). STRUCTURE was run from 1 to 30 genetic clusters (K) with 10 replicates for each K, each run starting with a burn-in period of 50,000 steps followed by 100,000 Markov chain Monte Carlo replicates. We selected $K = 3$ according to the ΔK method (Evanno et al. 2005). Consequently, the STRUCTURE analysis revealed three subpopulations (Online Resource 1), which coincided with the broad geographical origin in Australia. Only seven from the 166 trees had membership probabilities set below 0.6 and had to be assigned to more than one subpopulation. Subpopulation 1 included 49 trees (29.5 %), belonging to the seven natural provenances in Queensland, Australia. Subpopulation 2 included 94 trees (56.6 %) belonging to the four natural provenances in New South Wales, Australia. Subpopulation 3 included 23 trees (13.9 %) belonging to the two local land race sources from Concordia, Entre Rios, Argentina. Population structure based on 800 random DArTs and 15 SSR did not show significant differences ($R^2 = 0.86$).

Estimation of relatedness coefficients and inbreeding

The realized relationship matrices based on the co-dominant SSRs and SNPs were calculated using the moment-based estimators of Lynch and Ritland (1999) and Queller and Goodnight (1989) implemented in

SPAGeDi version 1.3a software (Hardy and Vekemans 2002), and the likelihood-based estimator of Wang (2007) was obtained with the Coancestry version 1.0.1.5 software (Wang 2011). Additionally, due to the mixed mating system of eucalypt species, the inbreeding should be accounted for in the estuation of relatedness. Therefore, the inbreeding coefficients for each individual of the *E. grandis* population were estimated from the SSR markers and the Lynch and Ritland (1999) and Wang (2007) estimators. Furthermore, we studied the impact of the inbreeding on the relatedness coefficients. In order to do this, we simulated marker and expected relationships classes mimicking the *E. grandis* population (see next section), and then, these simulations were analyzed with and without accounting for inbreeding using the Wang (2007) estimator.

Marker-based relationship matrices are often not positive definite due to several reasons: internal inconsistency resulting from lack of markers, genotyping errors or missing values. Therefore, the "nearPD" function implemented in R package "Matrix" was used to compute the nearest positive definite matrix from the original matrix. The "nearPD" function implements the algorithm of Higham (2002).

In the case of the dominant DArT markers, the realized relationship matrices were constructed with all (2816), 1500, 1000 and 500 randomly selected DArTs to examine the effect of the smaller DArTs size using the estimator defined by Hardy (2003) under the SPAGeDi version 1.3a software (Hardy and Vekemans 2002). These four groups shared between 175 and 1000 DArT markers in common.

Product-moment correlation coefficient was used to evaluate the connection between pairs of marker-based relationship matrices using each combination of marker (SSR, SNP and DArT) and estimator (LR, QG and W).

Since a high proportion of unrelated pairs over half-sib pairs of individuals were expected in our *E. grandis* population according to the information from pedigree, we investigated the discrimination of the unrelated from the related (expected) individuals. Therefore, given that the two distributions of relatedness (i.e., unrelated and related) are approximately equally symmetric, further comparisons of the different realized relationship matrices were performed using the probability that an unrelated tree would be misclassified as belonging to a related (and unrelated) tree, i.e., using the number of unrelated individuals that showed estimated relatedness coefficients above the critical value of 0.125 (midpoint between the expected means of unrelated and half-sib individuals; i.e., 0.0 and 0.25) (Blouin et al. 1996). For each combination of marker type and estimator, the amount of the overlapping areas between the density distributions (i.e., density plots) of both groups of trees was also investigated as an indicator of misclassification.

Simulations and measurement of performance

In studies of pairwise relatedness, Van de Casteele et al. (2001) and Wang (2011) recommended determining the bias and precision of different estimators using data simulated to emulate the empirical marker system and data. Therefore, to further evaluate and compare the performances of the co-dominant markers (SSR and SNP) and the different relatedness estimators, a stochastic simulation study was carried out. The effect of inbreeding was also investigated through simulation using an average inbreeding coefficient of 0.06 (i.e., the average inbreeding coefficient obtained for the *E. grandis* population for most of the SSR markers studied—11 out of 15; see below). Data were generated according to the expected genetic structure of *E. grandis* population (i.e., unrelated and half-sib), and the allelic frequencies were emulated from the empirical SSR and SNP markers. In all cases, it was assumed that the allelic frequency distribution was known without error. Simulated co-dominant multi-allelic SSR and bi-allelic SNP markers were generated and then analyzed using the LR, QG and W estimators implemented in the Coancestry version 1.0.1.5 software (Wang 2011).

The root mean squared relative error (RMSE) was used to measure the bias and precision of each marker–estimator combination, according to the following formula:

$$\left[\frac{1}{R}\sum_{l=1}^{R}(\hat{r}_l - r)^2\right]^{1/2}$$

where $\hat{r}_l$ is the relatedness estimate of the lth relatedness classes ($l = 1, 2,\ldots, R$) by a given marker–estimator combination and r is the parametric value of relatedness used to generating the R simulated related pairs of trees. For each combination of marker and relatedness estimator, a number of $R = 10{,}000$

replicates with a given relatedness classes were simulated.

Statistical analysis and estimation of heritability

The variance components and derived genetic parameters (i.e., heritabilities) were estimated by restricted maximum likelihood (REML, Patterson and Thompson 1971) implemented in the ASReml statistical program (Gilmour et al. 2006), using the following individual-tree mixed model with a total of 166 trees:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_r\boldsymbol{r} + \boldsymbol{Z}_a\boldsymbol{a} + \boldsymbol{e} \quad (1)$$

where the vector $\boldsymbol{y}$ contains the phenotypic data; the vector $\boldsymbol{\beta}$ included three subpopulations formed according to the results of the software program STRUCTURE 2.3 (see section "Molecular markers") as fixed effect; $\boldsymbol{r}$ is the vector of random replicate effects, $\boldsymbol{a}$ is the vector of random additive genetic effects of individual trees (i.e., breeding values); and $\boldsymbol{e}$ is the vector of random error; $\boldsymbol{X}$, $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_a$ are incidence matrices relating the observations ($\boldsymbol{y}$) to the model effects in $\boldsymbol{\beta}$, $\boldsymbol{r}$ and $\boldsymbol{a}$, respectively. The vector $\boldsymbol{a}$ was assumed distributed as $N \sim (\boldsymbol{0}, \boldsymbol{G}\,\sigma_a^2)$ where σ_a^2 is the additive genetic variance and $\boldsymbol{G}$ is the marker-based pairwise relationship matrix. Finally, the vector $\boldsymbol{e}$ is distributed as $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$ and σ_e^2 is the error variance. Marker-based relationship matrices from the SSR and SNP markers were corrected for the estimator bias found in the simulation study, i.e., estimates of pairwise relatedness from each marker–estimator combination were reduced (or increased) by the corresponding bias (Bush and Thumma 2013). Marker-based relationship matrices from the SSR were also calculated accounting for the inbreeding coefficient using the W estimator.

The variance component estimates relating to additive genetic effects ($\hat{\sigma}_a^2$) and error ($\hat{\sigma}_e^2$) were calculated using information from all 166 sampled trees having information on both molecular markers (i.e., SSR, SNP and DArT) and phenotypes. The narrow-sense individual heritability $(\hat{h}^2)$ was estimated as:

$$\hat{h}^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}.$$

An important limitation of the REML (co)variance estimates is that their distribution is unknown. Only an approximate measure of precision of the estimates based on asymptotic or large-sample theory can be calculated. Approximate standard errors of the heritabilities were computed with the "delta method," using an ASReml post-processing program (Gilmour et al. 2006). This asymptotic approach based on the Taylor expansion (Lynch and Walsh 1998) forces the confidence limits for (co)variances ratios to be symmetric.

Results

Marker informativeness

A total of 3012 polymorphic markers were used to genotype 166 *E. grandis* trees (Online Resource 2 and Online Resource 3). The average number of alleles per SSR locus was equal to 16.47. All SNP and DArT loci were bi-allelic. The average gene diversity or expected heterozygosity (H_e), of the SSR markers was equal to 0.82. There was high congruence between H_o and H_e at all the SSR marker loci except EMBRA61, EMBRA168, EMBRA19 and EMBRA173 (Online Resource 3). These departures from H_e could have been caused by the presence of null alleles at these four loci. Estimated null allele frequency for these markers were 0.32, 0.26, 0.24 and 0.21, respectively (calculated by CERVUS, Marshall et al. 1998). Null alleles in SSR markers (i.e., loci that fail to amplify to detectable levels via the polymerase chain reaction) are commonly encountered in population genetics studies (Chapuis and Estoup 2007). Null alleles are quite common when using SSR markers in one species that were designed for another species (Dakin and Avise 2004). The primer EMBRA19 was originally designed for *Eucalyptus urophylla*. However, the presence of null alleles for EMBRA61, EMBRA168 and EMBRA173 was unexpected given that these primer pairs were designed specifically from *E. grandis* sequences (Brondani et al. 2006). These departures from H_e could also be consequence of inbreeding. In such a case, the fraction of observed heterozygosity will be less than the fraction expected under random mating. The family-average inbreeding estimated from the 15 SSR markers and using the LR (and W) estimator in the *E. grandis* population was 0.140 (and 0.143). However, when these four SSR markers loci (i.e., EMBRA61, EMBRA168,

EMBRA19 and EMBRA173) were eliminated from the analysis, the family-average inbreeding decreased to 0.06. Similar inbreeding values were obtained for other *Eucalyptus* species based on SSR markers and trees sampled from progeny trials or native stand. For example, in *Eucalyptus cladocalyx*, a specie that is more inbred than most widely planted eucalypt species, Bush and Thumma (2013) reported inbreeding coefficient ranging from 0.06 to 0.27 and averaging 0.17. In *Eucalyptus camaldulensis*, these values ranged from 0.00 to 0.27 and averaging 0.10 (Butcher et al. 2009), and in *Eucalyptus obliqua* averaged 0.04 (Bloomfield et al. 2011). The average H_e for the SNP and DArT markers across all loci were 0.26 and 0.36, respectively. Changes in the H_e values were not observed when the number of DArTs was randomly selected (results not shown). When markers were grouped into ten classes according to their increasing level of polymorphism (Online Resource 4), the highest frequency of SSR, SNP and DArT markers were found in the classes 0.81–0.90, 0.41–0.50 and 0.41–0.50, respectively. Mean Shannon's Information Indices were 2.16 (SSR), 0.39 (SNP) and 0.53 (DArT).

Results of relatedness from the simulations

In terms of RMSE, which takes both bias and sampling variance into account, the likelihood-based W method outperforms the two moment-based methods (i.e., LR and QG), giving the lowest RMSE values across the two relationship classes (i.e., unrelated and half-sib pairs) and markers types (i.e., SSR and SNP) (results not shown). However, the differences in RMSE among the three estimators and two marker types were small (from 0.002 to 0.014). The higher differences were found in unrelated (0.014 and 0.010 for SSRs and SNPs, respectively) compared to half-sib (0.002 and 0.006 for SSRs and SNPs, respectively) pairwise relatedness.

The same simulated data were evaluated by the W method with and without accounting for inbreeding. In comparison, the RMSEs of the W estimator allowing for inbreeding were higher, with differences between the two approaches (i.e., W with and without allowing inbreeding) varying from 0.002 to 0.018 (again higher differences in half-sib than for unrelated pairwise relatedness). Wang (2007) also showed that allowing for inbreeding decreases the precision of the likelihood estimator mainly due to the increase in the number of parameters to be estimated from the same simulated data. Based on the results obtained in the simulation study and on the small family-average inbreeding (0.06) calculated with most of the SSR markers studied (11 out of 15), we carried out all further analyses of relatedness and estimations of heritabilities in the real data of *E. grandis* without considering the inbreeding.

Results of relatedness from the empirical data of *E. grandis*

Pairwise relatedness estimates for all pairs of individuals were split into two groups (Bessega et al. 2011): (1) both individuals within each pair belong to the same family (related) and (2) individuals within each pair belong to different families (unrelated). Overall, the three estimators (LR, QG and W) and the three types of markers (SSR, SNP and DArT) performed well in differentiating between the expected relatives (i.e., half-sibs) and unrelated individuals (Table 1; Fig. 1). As we expected, the estimated average of the pairwise coefficients were consistently higher (0.233) for related pairs of trees than for the unrelated pairs of trees (0.015). However, these averages exhibited differences across the two estimators and the three types of markers.

Table 1 Mean and standard error (SE) values of pairwise estimated relatedness for individuals from the same family (related) and to different families (unrelated) obtained from the three different markers (SSR, SNP and DArT) in *E. grandis*. For co-dominant markers (SSR and SNP) relatedness was estimated using: Lynch and Ritland (1999, LR), Queller and Goodnight (1989, QG), and Wang (2007, W). The DArT markers (DArTs) are followed by a number denoting the number of markers used to calculate the realized relationship matrix

Markers	Related		Unrelated	
	Mean	SE	Mean	SE
SSR_LR	0.182	0.019	−0.007	0.001
SSR_QG	0.225	0.018	−0.007	0.001
SSR_W	0.166	0.015	0.031	0.001
SNP_LR	0.144	0.025	−0.007	0.001
SNP_QG	0.190	0.037	−0.007	0.002
SNP_W	0.294	0.026	0.153	0.001
DArTs_2816	0.280	0.015	−0.002	0.001
DArTs_1500	0.284	0.016	−0.002	0.001
DArTs_1000	0.291	0.015	−0.002	0.001
DArTs_500	0.275	0.016	−0.002	0.001
AVERAGE	0.233	0.020	0.015	0.001

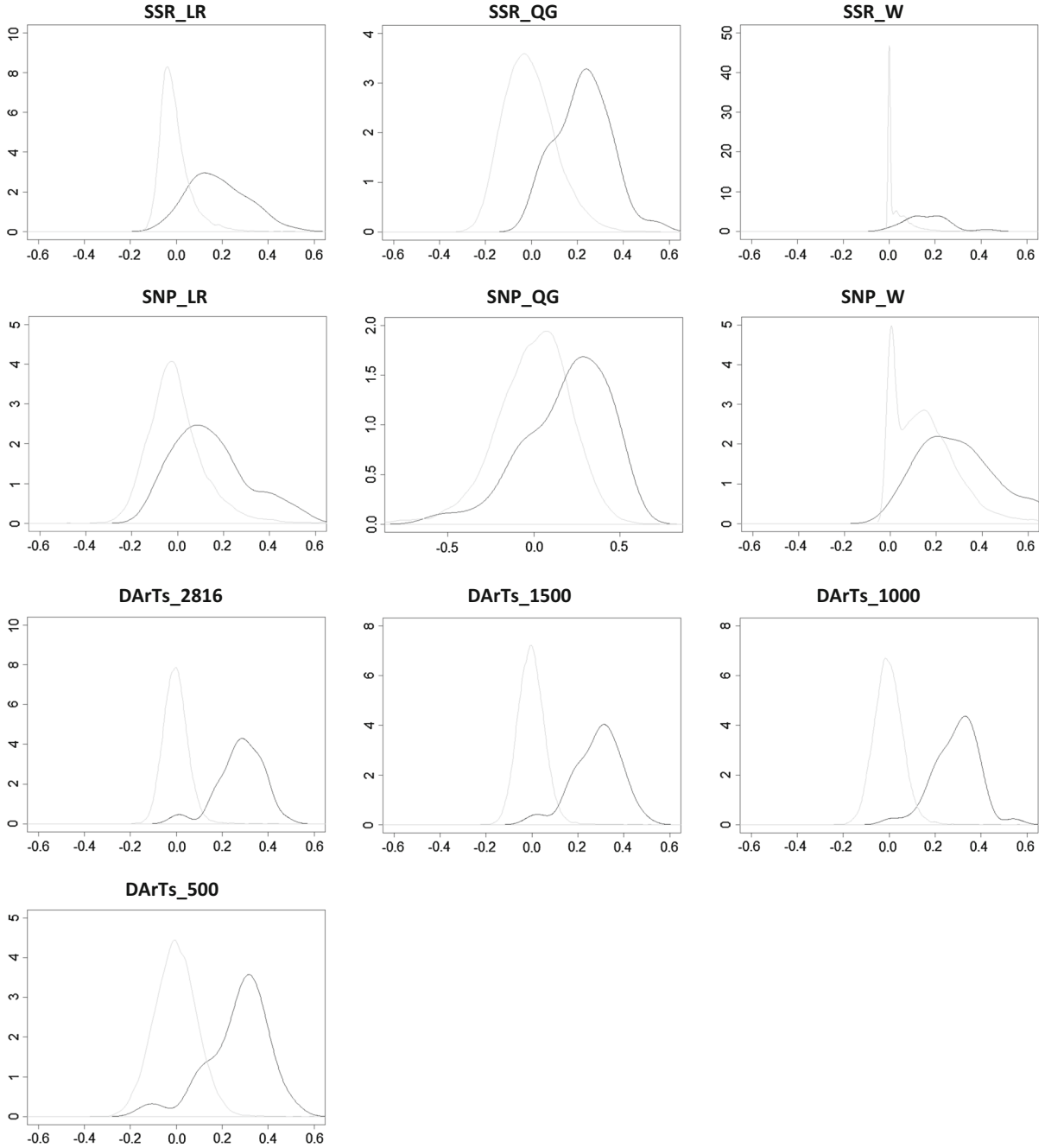

Fig. 1 Density plot of the pairwise estimated relatedness for individuals from the same family (*related*, *black line*) and to different families (*unrelated*, *gray line*) obtained from the three different markers (SSR, SNP and DArT) in *E. grandis*. For co-dominant markers (SSR and SNP) relatedness was estimated using: Lynch and Ritland (1999, LR), Queller and Goodnight (1989, QG) and Wang (2007, W). The DArT markers (DArTs) are followed by a number denoting the number of markers used to calculate the realized relationship matrix

Table 1 shows that for pairs of individuals from the same open-pollinated family (i.e., related individuals), the QG and the W estimators yielded highest estimated average of the pairwise coefficients for the SSR and SNP markers, respectively. However, except for the W estimator in combination with the SNP markers, the

pairwise coefficient estimates were lower than those expected for related half-sibs individuals (equal or higher than 0.25), indicating either that the estimators are biased and/or pedigree errors, i.e., some of the supposedly open-pollinated families are in fact unrelated and driving the estimates downward. When the estimator bias was corrected, the average of the pairwise relatedness estimated for the related individuals were slight (only differences in the third decimal place) higher (SSR markers) or lower (SNP markers). For the SSR markers, the estimated standard errors of half-sib (i.e., related) were similar across all estimators. Nevertheless, higher standard errors and, thus, broadest distribution were obtained in combination with the SNP markers and the QG estimator compared to the LR and W estimators (Fig. 1). Similarly, the averages of pairwise relatedness were computed for unrelated individuals. Meanwhile, the average of pairwise relatedness estimates obtained by LR and QG estimators for SSR and SNP markers resulted in similar values, the W estimator showed the highest estimates. However, the LR and the W estimators presented the smallest standard errors and narrower distributions than the QG estimator.

In addition, pairwise relatedness coefficients were estimated with dominant bi-allelic DArT markers using the Hardy (2003) estimator. While SSR and SNP based relationship coefficients within open-pollinated families showed comparable average values across estimators, DArT marker-based relationship coefficients reached higher average coefficients (0.280), indicating that some particular open-pollinated families will have a mixture of relatedness among individuals ranging from half-sibs to selfs. Additionally, DArT markers showed smaller average standard error (Table 1) and, in general, narrow distributions (Fig. 1) for related individuals. Decrease in the number of used DArT markers resulted in only a slight effect on the average of relatedness with very similar standard error. Our analysis showed that the average pairwise relatedness estimated by 2816 DArT markers was the closest to zero, and had a smaller standard error for pairs of expected unrelated individuals than those estimated by SSRs with the QG estimators and SNPs. Decreasing the number of DArTs produced only slight differences in estimated pairwise relatedness within the unrelated individuals (in the fifth decimal place) and no differences in its standard error.

To test whether more relationship classes were present within the unrelated individuals from the same subpopulation, the unrelated group was split into two subgroups (Bessega et al. 2011): pairs of individuals from the same subpopulation (S_S) and pairs of individuals from different subpopulation (D_S). Overall, the average marker-based pairwise relatedness estimates were comparable and close to zero for the S_S and D_S subgroups (Table 2), and showed some non-overlapping areas between both subgroups (Fig. 2). However, these averages and the areas of the density plots exhibited difference across both estimators and marker types. Specifically, in pairs of individuals from S_S subpopulation, the LR (and W) estimator produced the smallest (and the highest) average of pairwise relatedness (i.e., closest to the expected value of zero) for the SSR and SNP markers, whereas DArT markers yielded higher values (average across data set sizes 0.020) in comparison with SSR and SNP markers for the LR and QG estimators. However, DArT markers showed the highest mean differences together with the highest non-overlapping areas between the two subgroups of unrelated trees.

The number of unrelated pairs of individuals that showed estimated pairwise relatedness greater than 0.125 varied from 1.58 to 52.57 % across the estimators and markers. For unrelated pairs, the LR estimator showed a smaller number for the SSR (4.86 %) and

Table 2 Comparison of average pairwise estimated relatedness for individuals from different families (unrelated) split into two subgroups: both members of the individual pair belong to the same subpopulation (S_S) and both members of the individual pair belong to different subpopulation (D_S). Abbreviations used for the estimator and marker types were described in the text and in the caption of Table 1

	S_S	D_S	Mean difference
SSR_LR	0.006	−0.017	0.023
SSR_QG	0.015	−0.024	0.039
SSR_W	0.040	0.023	0.018
SNP_LR	0.002	−0.014	0.016
SNP_QG	0.006	−0.017	0.023
SNP_W	0.162	0.145	0.017
DArTs_2816	0.020	−0.021	0.041
DArTs_1500	0.019	−0.020	0.039
DArTs_1000	0.021	−0.021	0.041
DArTs_500	0.020	−0.021	0.041
AVERAGE	0.031	0.001	0.030

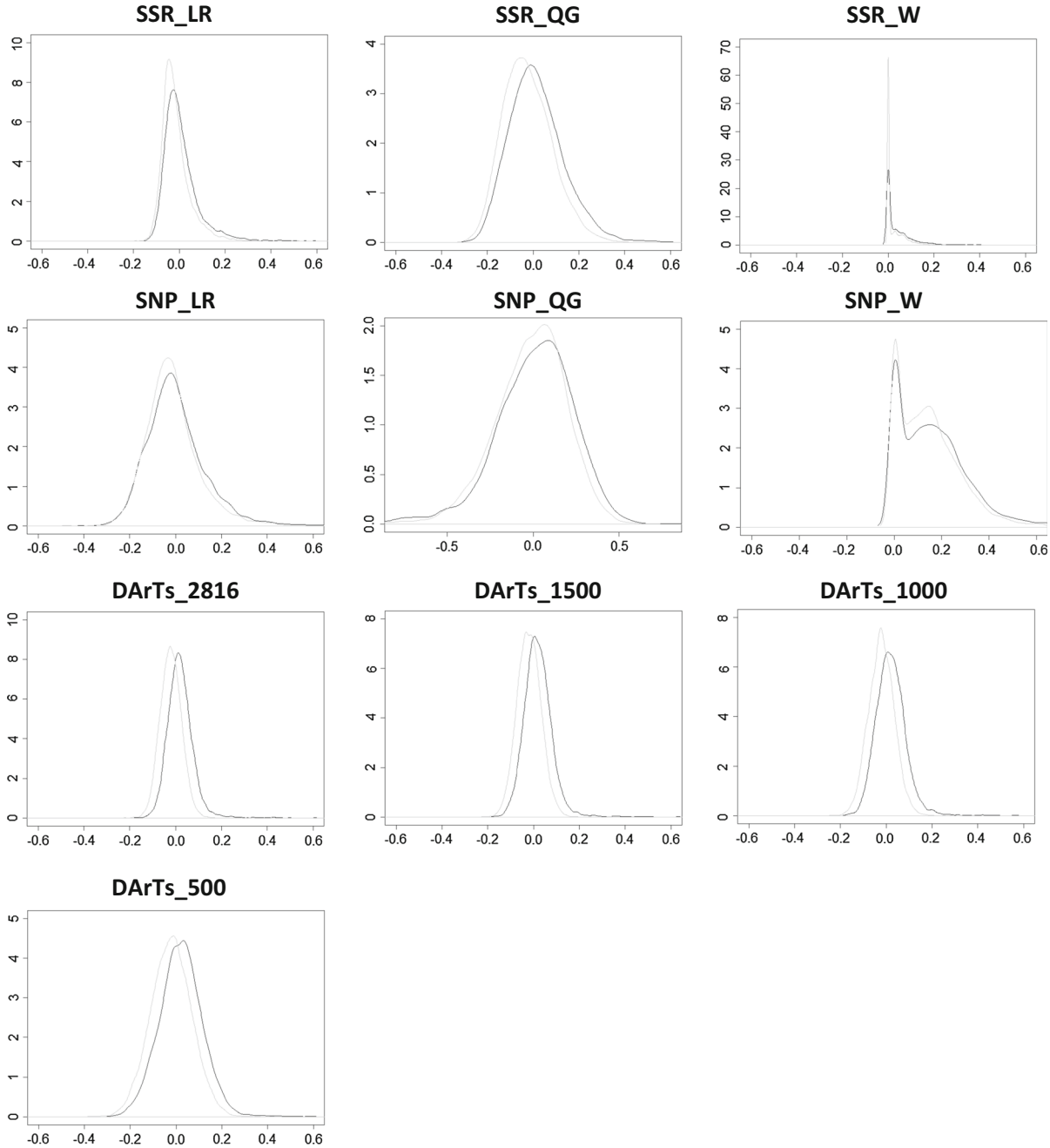

Fig. 2 Density plot of the pairwise estimated relatedness for individuals from different families (*unrelated*) split into two subgroups: both members of the individual pair are from same subpopulation (*black line*) and both members of the individual pair are from different subpopulation (*gray line*). Abbreviations used for the estimator and marker types were described in the text and in the caption of Table 1

SNP (12.63 %) markers compared to the QG (12.22 and 28.18 %, respectively) and the W (5.74 and 52.57 %, respectively) estimators. However, the smallest numbers were found for the DArT markers (average = 2.89 %), followed by SSR and SNP markers, calculated using the pairwise LR values.

Moreover, DArTs showed the smaller overlapping areas between the distributions of related and unrelated individuals (Fig. 1). The number of unrelated pairs of individuals that showed estimated pairwise relatedness greater than 0.125 increased from 1.58 to 5.48 % with the decrease in the numbers of random selected DArTs from 2816 to 500.

To identify the best combination of estimator and marker used, we also calculated the product-moment correlations between pairs of marker-based relationship matrices generated by using each estimator and marker type combination. We found high correlations that were significantly different from zero ($p < 0.0001$) between the LR, QG and W estimators for the SSR (from 0.64 to 0.73) and SNP (from 0.71 to 0.84) markers, and within DArTs comparisons when the number of markers was reduced from 2816 to 500; for example: DArTs_2816 vs. DArTs_1500 = 0.92, DArTs_2816 vs. DArTs_1000 = 0.86, DArTs_2816 vs. DArTs_500 = 0.72. However, the marker-based relationship matrices were significant but not well correlated across marker types (SSRs vs. SNPs averaged 0.04; SSR vs. DArTs_2816 averaged 0.21; SNPs vs. DArTs_2816 averaged 0.08).

Heritability

When we compared the three types of markers, on average, for the growth traits (DBH and Volume), the heritabilities estimated based on co-dominant bi-allelic SNPs were higher and with lower standard errors compared to the estimates obtained from SSRs and DArTs (Table 3). However, the total lignin and BD traits heritability estimates based on DArTs were higher than those based on SSRs and SNPs. For Klason lignin, similar heritability estimates were obtained with the DArTs and SSRs, but clearly smaller with the SNPs.

Discussion

Over the last years, relatively easy access to highly informative DNA markers has increased interest in estimates of relationship matrix in wild populations where pedigree information is not known, or in small breeding populations with limited pedigree information and/or few available parents. In this study, we compared the effects of three types of markers (co-dominant multi-allelic SSR, co-dominant bi-allelic SNP and dominant bi-allelic DArT) on the estimation of relatedness coefficients and their effects on the precision of genetic parameter estimates, using three estimators for co-dominant markers (i.e., Lynch and Ritland 1999; Queller and Goodnight 1989 and Wang 2007) and one for dominant markers (Hardy 2003), in both a simulated data set and a relatively small *E. grandis* population data with limited relationships from the pedigree.

Marker informativeness

One of the major drawbacks in relatedness and/or genetic parameters estimation is the statistical bias caused by small sample sizes, i.e., small number of marker loci and/or small number of individuals (Ritland 1996). Therefore, the precision with which marker-based pairwise relatedness and heritability can be estimated depends largely on the number of loci and the number of alleles at each locus (e.g., Ritland 1996; Hardy 2003; Rodríguez-Ramilo et al. 2007). The number of loci and the average number of alleles per locus in the three types of markers used in this study would be considered as adequate for the reliable estimation of relatedness, in accordance with the prerequisites indicated by Ritland (1996). The author suggested that n loci and m alleles per locus (i.e., the quantity $n \times (m - 1)$) should lie between 25 and 100 to give reliable results (i.e., SSR = 232, SNP = 181 and DArT = from 2816 to 500). However, due to the different nature of the molecular markers used in this study (co-dominant and dominant), the number of loci and the different number of alleles per locus (multi- and bi-allelic), the total number of effective alleles per locus may be a better predictor of the marker informativeness on precision of the relatedness coefficient estimation. There were a total of $7.48 \times 15 = 112.2$ effective alleles for SSRs, $1.43 \times 181 = 259.6$ for SNPs and $1.62 \times 2826 = 4557.5$ for DArTs. Therefore, the high number of the DArT markers compensates for the low number of recognizable alleles per locus (2) and demonstrates the highest marker informativeness of these markers in our data set, followed by the co-dominant bi-allelic SNPs and co-dominant multi-allelic SSRs. Similar findings were reported in sugar beet (Simko et al. 2012), wheat (Laidó et al. 2013) and barley (Lamara et al. 2013), where more bi-allelic or dominant

Table 3 Heritabilities ($\hat{h}^2$) and their approximate standard error (SE) for diameter at breast height (DBH), volume, total and Klason lignin and basic density estimated from marker-based relationship matrices from the three estimators (LR, QG and W) and three different types of markers (SSR, SNP and DArT) in *E. grandis*. Marker-based relationship matrices from the SSR and SNP markers were corrected for the estimator bias. Abbreviations used for the estimator and marker types were described in the text and in the caption of Table 1

	DBH		Volume		Total Lignin		Klason Lignin		Basic Density	
	$\hat{h}^2$	SE	$\hat{h}^2$	SE	$\hat{h}^2$	SE	$\hat{h}^2$	SE	$\hat{h}^2$	SE
SSR_LR	0.092	0.044	0.082	0.046	0.092	0.051	0.117	0.048	0.081	0.05
SSR_QG	0.087	0.038	0.074	0.041	0.092	0.045	0.121	0.040	0.047	0.049
SSR_W	0.082	0.049	0.067	0.050	0.100	0.052	0.110	0.052	0.123	0.048
SNP_LR	0.103	0.037	0.123	0.024	0.033	0.055	0.006	0.057	0.096	0.047
SNP_QG	0.065	0.021	0.074	0.014	0.016	0.049	0.009	0.050	0.029	0.034
SNP_W	0.132	0.011	0.112	0.017	0.029	0.059	0.000	–	0.087	0.045
DArTs_2816	0.063	0.064	0.043	0.066	0.132	0.059	0.130	0.060	0.132	0.056
DArTs_1500	0.081	0.061	0.064	0.063	0.123	0.059	0.126	0.059	0.139	0.053
DArTs_1000	0.049	0.065	0.019	0.068	0.117	0.057	0.111	0.058	0.101	0.057
DArTs_500	0.045	0.060	0.026	0.064	0.117	0.050	0.108	0.053	0.133	0.044

markers were necessary to compensate for the relatively large amount of information per SSR locus. Additionally, Ritland (1996) stressed that the number of pairwise comparisons to obtain reasonable estimates of relatedness coefficients and heritabilities should be greater than 10^4 (and preferably 10^5), i.e., the sample size should range between a 150 and 450 individuals, which was the case with our data set: 166 trees (13,695 pairwise comparisons).

We found that the SSRs showed heterozygosity more than three times higher than SNPs and more than two times higher than DArTs. The average H_e for the SSRs was similar to the result reported for *E. grandis* by Brondani et al. (2006) and other *Eucalyptus* species, such as in *E. globulus* (i.e., from 0.71 to 0.93; average 0.85, Ribeiro et al. 2011), *E. urophylla* (i.e., from 0.74 to 0.93; average 0.86) and *E. dunnii* (i.e., from 0.68 to 0.93; average 0.82, Zelener et al. 2005). In terms of expected heterozygosity, Blouin et al. (1996) concluded that a set of 15 SSR loci with a H_e of 0.75 would accurately discriminate more than 80 % of the half-sibs from unrelated individuals (see Figure 3 in Blouin et al. 1996). It is commonly found that higher H_e values for SSR markers than those for SNP or DArT markers, because multi-allelic markers can reach higher H_e values. However, the H_e values for SNP markers were relatively low in our study. Anderson and Garza (2006) showed in a simulation study that SNPs with a minor allele frequency (MAF) less than 0.2 (corresponding to a $H_e < 0.32$), rapidly lose power to estimate relationship using parentage inference. Also, in a simulated scenario of 100 independent SNPs, each with a minor allele frequency of 0.2 (corresponding to a H_e of 0.32), Glaubitz et al. (2003) showed that about 18 % of unrelated individuals would be misclassified as half-sibs, and this percentage increased rapidly when the MAF was smaller than 0.2 (see Figure 4B in Glaubitz et al. 2003). In particular, our study showed that 105 of the 181 loci had expected heterozygosity smaller than 0.32 (corresponding to a MAF < 0.2).

Relatedness

In forest genetic studies, the pairwise relatedness elements of the marker-based relationship matrix have been estimated using observed relationship implied by pedigree reconstruction (e.g., El-Kassaby et al. 2011; Telfer et al. 2015) or relatedness estimates from pairwise relatedness estimators (e.g., Kumar and Richardson 2005; Bessega et al. 2011; El-Kassaby et al. 2012). When we compared the two commonly used pairwise relatedness estimators (LR and QG) and the W estimator for the two co-dominant markers (SSR and SNP), our results from the simulations showed that the W estimator was the best. However,

when we compared the W estimator with the LR and QG estimators in term of differences in RMSE for the unrelated and half-sib simulated relatedness, the lowest differences were founded for the combination of LR and unrelated relationships (0.012 vs. 0.033 for unrelated and related relationships, respectively), and QG and half-sib relationships (0.004 vs. 0.046 for related and unrelated relationships, respectively). These results reflect our empirical results, where the W and QG estimators performed better for the half-sibs category (i.e., estimated relatedness were near or higher than 0.25) and the LR was a better estimator for the unrelated category (i.e., estimated relatedness were closer to the expected value of zero, smaller standard error and better discrimination of unrelated and half-sib individuals). However, when we compared estimated relatedness obtained using the method of QG, LR and W, we found high correlations within SSR and SNP markers. Additionally, in comparison with the two moment-based estimators, our empirical results showed that the W estimator estimated the highest values of relatedness between the individuals from different families (i.e., unrelated) when co-dominant markers were used. In agreement with our findings, Csilléry et al. (2006) concluded that the QG estimator had smaller sampling variances for the high relationship categories, while LR was better for the low relationship categories in five natural outbreeding populations that were less related than half-sibs. This is also in agreement with the results obtained by Ribeiro et al. (2011) in a simulation study. Van de Casteele et al. (2001), using microsatellite markers, showed that the LR estimator performs better in populations with more than 60 or 70 % unrelated pairs, while estimators with locus specific weights (i.e., Li et al. (1993) and QG), perform better in populations with more than 50 % related pairs. In summary, given the shallow pedigree in our *E. grandis* population (i.e., high proportion of expected unrelated individuals over half-sib pairs of individuals), and the need for a coefficient with higher precision for the pairs of distantly related individuals, the LR estimator is preferred.

The differences found between the three estimators studied may also be a function of the variation in the number of loci or marker informativeness. We also compared three different types of markers with different patterns of inheritance (dominant vs. co-dominant) and different numbers of alleles per locus (i.e., multi-vs. bi-allelic). The results showed that co-dominant multi-allelic SSRs are much more informative than the co-dominant bi-allelic SNP markers, an outcome previously discussed by Wang (2006), given that SSRs can reach a higher number of alleles per locus, and are often recognized as the most efficient marker for relatedness estimation (Hardy 2003). Likewise, SNPs are more informative than the dominant bi-allelic DArT markers, which scored as either present or absent, thus providing less genetic information for a given locus. However, although the dominant markers are individually less informative than co-dominant ones, our analysis of 166 trees of *E. grandis* revealed that a high number of dominant bi-allelic DArT markers (in our case 2816) obtained with a relatively simple technology and low cost (Sansaloni et al. 2010), may yield comparable relatedness coefficients and accuracies to co-dominant multi-allelic (SSR) and bi-allelic markers (SNP). The effect of the dominant DArT markers on the estimated relatedness coefficients has not been extensively studied. In our study, when we compared the estimated average of the pairwise coefficients from the three studied markers, DArT markers appeared to perform better than SSRs and SNPs to estimate pairwise relatedness coefficients from related (i.e., estimates higher than 0.25, smaller standard error) and unrelated (i.e., estimates closer to expectation value of zero, smaller standard error and, in general, narrower distributions) individuals. Moreover, the dominant bi-allelic DArT marker panel showed a higher ability than SSRs and SNPs to discriminate unrelated from related half-sib individuals (Table 2; Fig. 2). Higher rates of unrelated individuals misclassified as half-sibs (18 %) were obtained by Blouin et al. (1996) using a set of 20 SSR loci with an expected heterozygosity of 0.75 and by Glaubitz et al. (2003), using 100 independent SNPs, each with a MAF of 0.2 (corresponding to a $H_e < 0.32$). Glaubitz et al. (2003) concluded that 16–20 microsatellites with a H_e of 0.75 would be expected to provide information equivalent to that given by 100 independent SNPs, each with a minor allele frequency of 0.2. Similarly, Yu et al. (2009) suggested that an SSR-to-SNP ratio of 1:10 was required to provide robust estimates of relatedness for association mapping. Simulation performed by Wang (2006) showed that about 89 bi-allelic SNPs (i.e., total number of effective alleles equals to 178) are required to achieve the same power as 13 10-allelic SSRs (i.e.,

total number of effective alleles equals to 130) to distinguish half-sib from unrelated pairs of individuals.

We tested the effect of the number of DArT markers on the precision of relatedness coefficients. The correlations between pedigree- and marker-based relatedness coefficients decreased with a reduction in the number of DArTs from 2816 to 500 (DArTs_2816 = 0.27; DArTs_1500 = 0.25; DArTs_1000 = 0.25; DArTs_500 = 0.19). Decrease in the number of dominant bi-allelic DArT markers had only a slight effect on estimated average relatedness. Though, in general, the standard error of the estimator declines with the number of loci (Milligan 2003), a random reduction in the number of DArT markers yielded the same standard error of the relatedness coefficients (at the second decimal place) for related and unrelated pairs of individuals. However, the correlation between the average pairwise relatedness based on different numbers of DArTs showed a reduction in the estimated correlations when the number of markers was reduced from 2816 to 500, particularly when the number of markers decreased to 500 DArTs. Moreover, the density plot shows a broad distribution (i.e., high variance) when only 500 DArT markers were used. Therefore, it appears that 1000 randomly selected bi-allelic dominant DArT markers are enough to achieve accurate relatedness coefficients in our *E. grandis* population.

All moment-based estimators studied assume the absence of inbreeding and genotyping errors, while likelihood-based methods as that proposed by Wang (2007) can account for both phenomena. Our simulation analyses, which mimicked the expected relationships classes (i.e., unrelated and half-sibs) in the *E. grandis* population and emulated the allele frequency characteristics of the SSR markers, showed that the RMSEs were higher than those that did not accounting for inbreeding. Moreover, our empirical data showed that the average inbreeding coefficients estimated from the 15 SSR markers were relatively low (0.140 and 0.143 for the LR and W estimators, respectively); however, this coefficient declined significantly to 0.06 when the four SSR markers with highest deficit of the heterozygosity were removed from the analysis. Wang (2007), using simulation, concluded that it seems unjustified to take inbreeding into account in the likelihood-based methods for estimating relatedness, when the inbreeding coefficients are lower than 0.15 or in samples with high proportion of unrelated pairwise relationships, except when there is ample marker information (hundreds of SSR markers). Finally, we did not consider the genotyping error in both simulated and empirical data sets. However, Wang (2007) showed that accounting for typing errors in the estimator impairs the estimates when the pairwise of relatedness in the sample are loosely related or unrelated.

Heritability

A major motivation in estimation of relatedness among individuals is to construct the additive relationship matrix, allowing the estimation of genetic parameters such as the additive genetic variance or heritability within the individual-tree mixed model framework. Estimation of genetic parameters using marker-based relationship matrices is especially useful in studies of wild populations where pedigree information is not known (Frentiu et al. 2008; Sillanpää 2011), or in small breeding populations with limited pedigree information and/or few available parents (i.e., reduced variation in the pedigree relationship; Ødegård and Meuwissen 2012).

Very few published reports compare the estimated heritabilities from different types of markers. Our results show that, in general, for growth traits (DBH and volume) co-dominant bi-allelic SNP markers yielded higher heritabilities and more precise results (i.e., smaller standard error) than co-dominant multi-allelic SSR and dominant bi-allelic DArT markers. Meanwhile, for wood properties traits: total lignin and basic density, the dominant bi-allelic DArT markers yielded the highest heritabilities and most precise results. In contrast to our findings, Bessega et al. (2011) showed that heritability estimates based on 128 dominant markers (57 AFLPs and 71 ISSR) were clearly lower compared to those based on 6 SSR. However, they concluded that more dominant markers are required to compensate for the low number (2) of detectable alleles in comparison with SSRs.

Generally, the decrease in the number of DArT markers used to derive the relationship matrix resulted in lower heritability estimates, except for the growth and the BD traits from 2816 DArTs to 1500 DArTs. Similar results were obtained by Hayes and Goddard (2008), where the estimated heritability from simulated data was 0.32, 0.30 and 0.21 when the marker-

based relationship matrices were estimated with a decreasing number of SNP markers: 9000, 5000 and 1000, respectively.

The low number of trees sampled and the low number of trees per family used in this study could limit the use of these estimates for practical applications in a breeding population of *E. grandis* or for the comparisons with other studies. Moreover, our estimates of heritability from the SSR markers for the tree estimators studied and the SNP markers using the LR and QG estimators could be inflated since the average of the pairwise coefficients for related individuals were lower than half-sib (i.e., 0.25; Table 1). However, our results show that SNP marker-based heritabilities (the highest across markers) for growth traits were 0.100 and 0.103 for DBH and Volume, respectively (average from LR, QR and W estimators). In a review of seven traits of *Eucalyptus* spp., conifers and other broadleaf forest tree species from 67 published papers, Cornelius (1994) reported higher values of heritabilities for DBH (0.19) and volume (0.18). However, the values reported by Cornelius (1994) could be overestimated, since open-pollinated families were assumed to have a coefficient of relationship equal to 0.25. Estimated average DArT marker-based heritabilities (the highest across markers) for wood properties are smaller (total lignin = 0.132, Klason lignin = 0.130 and BD = 0.139) than those previously published for other *Eucalyptus* species. Gion et al. 2011 reviewed estimates of genetic parameters of wood properties from *Eucalyptus* species and concluded that wood property traits had higher heritabilities compared to growth traits. For example, in *Eucalyptus globulus*, Stackpole et al. (2011) reported heritabilities for wood density of 0.51 and for Klason lignin of 0.27. Therefore, our results show that 2816 DArT marker-based heritabilities could be underestimated based on average estimated heritabilities reported for other *Eucalyptus* studies for wood properties traits. Ødegård and Meuwissen (2012) indicated that when a fraction of the genome is not covered by the markers, the total genetic variance (and thus the heritability estimates) will be underestimated. They suggested that this underestimation may be completely covered by including a polygenic effect in the mixed model, which has a covariance structure equal to the NRM. Including a polygenic effect in the mixed linear model (1) such an effect led to higher estimated heritabilities, except for the trait with the lowest pedigree-based heritability such as Klason lignin. However, estimated standard errors of heritabilities were higher and closer to the pedigree-based approach (results not shown).

Conclusion

Though our empirical data is far from an ideal scenario for marker-based in situ approaches because of the low number of accessions and very sparse relationships between individuals, the results demonstrated that the marker-based relatedness estimates has clear advantages over the expected categorical measure of relationships. The marker-based methods allowed the estimation of the actual relatedness between two individuals, i.e., the realized relationships. The lower standard error of the pairwise coefficients from DArT markers would indicate that these markers are more conservative to recover Mendelian sampling (likely due to their nature) than the more polymorphic (informative) co-dominant markers (SSR and SNP); however, all of them could capture hidden relationships between individuals. Our work suggests that the relatedness coefficients (i.e., realized relationship matrix) and heritability estimates can be accurately inferred from co-dominant or sufficiently dense dominant molecular markers in a relatively small *E. grandis* data set with a shallow pedigree. Results from both simulated and empirical data provided a premise to select between three estimators for co-dominant markers, the type of molecular markers and number of DArT markers that are needed for estimated relatedness coefficients and heritabilities in this *E. grandis* data set.

Acknowledgments The authors are grateful to the company Forestal Las Marias for providing the land for the test and logistical support. The authors also thank Leonel Harrand, Javier Oberschelp, Mauro Surenciski and Juan López who assisted with field work and data collection. The contribution of José Rodrigues and colleagues at the IICT lab in Portugal, in the development of specific NIR models and subsequent processing and spectral evaluation of wood samples for the wood quality traits used in this study is also greatly acknowledged. We also thank to Cintia Acuña, Andrea Puebla and María Carolina Martínez for their help with the DNA extraction and SNPs analysis, and Dario Grattapaglia, Carolina Sansaloni, Cesar Petroli and Danielle Paiva for their work for generation of DArTs and SNPs data. Finally, we would like to thank the helpful suggestions and support given by Yousry El-Kassaby and Esteban Hopp during this work, and Leopoldo Sanchez and two anonymous referees for their insightful comments on early version of this manuscript.

Compliance with ethical standards

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

References

Alves A, Santos A, Rozenberg P, Paques LE, Charpentier JP, Schwanninger M, Rodrigues J (2012) A common near infrared-based partial least squares regression model for the prediction of wood density of *Pinus pinaster* and *Larix* × *eurolepis*. Wood Sci Technol 46:157–175

Anderson EC, Garza JC (2006) The power of single nucleotide polymorphisms for large-scale parentage inference. Genetics 172:2567–2582

Bessega C, Saidman BO, Darquier MR, Ewens M, Felker P, Vilardi JC (2011) Accuracy of dominant markers for estimation of relatedness and heritability in an experimental stand of *Prosopis alba* (leguminosae). Tree Genet Genomes 7:103–115

Bloomfield JA, Nevill P, Potts BM, Vaillancourt RE, Steane DA (2011) Molecular genetic variation in a widespread forest tree species *Eucalyptus obliqua* (Myrtaceae) on the island of Tasmania. Aust J Bot 59:226–237

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 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

Brondani RP, 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 22:6–20

Bush D, Thumma B (2013) Characterising a *Eucalyptus cladocalyx* breeding population using SNP markers. Tree Genet Genomes 9:741–752

Butcher PA, McDonald MW, Bell JC (2009) Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*. Tree Genet Genomes 5:189–210

Cappa EP, Pathauer PS, Lopez GA (2010) Provenance variation and genetic parameters of *Eucalyptus viminalis* in Argentina. Tree Genet Genomes 6:981–994

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

Cornelius JP (1994) Heritabilities and additive genetic coefficients of variation in forest trees. Can J For Res 24:372–379

Crossa J, Burgueno J, Dreisigacker S, Vargas M, Herrera-Foessel S, Lillemo M, Singh R, Trethowan R, Warburton M, Franco J, Reynolds M, Crouch J, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177:1889–1913

Csilléry 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

Eldridge K, Davidson J, Hardwood C, van Wyk G (1993) *Eucalyptus* domestication and breeding. Oxford University Press, New York

El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, Klápšte J, Lstiburek M (2011) Breeding without breeding: is a complete pedigree necessarily for efficient breeding? PLoS One 6(10):e25737

El-Kassaby YA, Klápšte J, Guy RD (2012) Breeding without Breeding: selection using the genomic best linear unbiased predictor method (GBLUP). New For 43:631–637

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379. doi:10.1371/journal.pone.0019379

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14:2611–2620

Frentiu FD, Clegg SM, Chittock J, Burke T, Blows MW, Owens IPF (2008) Pedigree-free animal models: the relatedness matrix reloaded. Proc R Soc B 275:639–647

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2006) ASReml user guide release 2.0. VSN International Ltd, Hemel Hempstead

Gion JM, Carouché A, Deweer S, Bedon F, Pichavant F, Charpentier JP, Baillères H, Rozenberg P, Carocha V, Ognouabi N, Verhaegen D, Grima-Pettenati J, Vigneron P, Plomion C (2011) Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *Eucalyptus*. BMC Genom 8:12–301

Glaubitz JC, Rhodes OE Jr, Dewoody JD (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol Ecol 12:1039–1047

Grattapaglia D, Silva-Junior OB, Kirst M, Marco de Lima B, Faria DA, Pappas GJ Jr (2011) High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across. BMC Plant Biol 11:65

Griffin AR, Cotterill PP (1988) Genetic variation in growth of outcrossed, selfed and open-pollinated progenies of *Eucalyptus regnans* and some implications for breeding strategy. Silvae Genet 37:124–131

Hardy OJ (2003) Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. Mol Ecol 12:1577–1588

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

Hayes BJ, Goddard ME (2008) Technical note: prediction of breeding values using marker derived relationship matrices. J Anim Sci 86:2089–2092

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph

Higham NJ (2002) Computing the nearest correlation matrix—a problem from finance. IMA J Numer Anal 22:329–343

Hodge GR, Volker PW, Potts BM, Owen JV (1996) A comparison of genetic information from open-pollinated and control-pollinated progeny tests in two eucalypt species. Theor Appl Genet 92:53–63

Hoisington D, Khairallah M, Gonzalez-De-Leon D (1994) Laboratory protocols: CIMMYT. Applied molecular genetics laboratory, 3rd edn. CIMMYT, D.F., Mexico

Hudson CJ, Freeman JS, Kullan ARK, Petroli CD, Sansaloni CP et al (2012) A reference linkage map for *Eucalyptus*. BMC Genom 13:240. doi:10.1186/1471-2164-13-240

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, Aschenbrenner-Kilian M, Evers M, Peng K, Cayla C, Hok P, Uszynski G (2012) Diversity arrays technology: a generic genome profiling technology on open platforms. Methods Mol Biol 888:67–89. doi:10.1007/978-1-61779-870-2_5

Kumar S, Richardson TE (2005) Inferring relatedness and heritability using molecular markers in Radiata pine. Mol Breed 15(1):55–64

Laidó G, Mangini G, Taranto F, Gadaleta A, Blanco A et al (2013) Genetic diversity and population structure of tetraploid wheats (*Triticum turgidum* L.) estimated by SSR, DArT and pedigree data. PLoS One 8(6):e67280. doi:10.1371/journal.pone.0067280

Lamara M, Zhang LY, Marchand S, Tinker NA, Belzilea F (2013) Comparative analysis of genetic diversity in Canadian barley assessed by SSR, DArT, and pedigree data. Genome 56:351–358

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum Hered 43:45–52

Lopez GA, Potts BM, Dutkowski GW, Apiolaza LA, Gelid P (2002) Genetic variation and inter-trait correlations in *Eucalyptus globulus* base population trials in Argentina. For Gen 9:223–237

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Marcó M, White TL (2002) Genetic parameter estimates and genetic gains for *Eucalyptus grandis* and *E. dunnii* in Argentina. For Genet 9:205–215

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

Milligan BG (2003) Maximum-likelihood estimation of relatedness. Genetics 163:1153–1167

Mousseau TA, Ritland K, Heath DD (1998) A novel method for estimating heritability using molecular markers. Heredity 80:218–224

Ødegård J, Meuwissen THE (2012) Estimation of heritability from limited family data using genome-wide identity-by-descent sharing. Genet Sel Evol 44:16–25

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537–2539

Petroli CD, Sansaloni CP, Carling J, Steane DA, Vaillancourt RE et al (2012) Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. PLoS One 7:e44684. doi:10.1371/journal.pone.0044684

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:845–959

Przyborowski JA, Sulima P, Kuszewska A, Załuski D, Kilian A (2013) Phylogenetic relationships between four *Salix* L. species based on DArT markers. Int J Mol Sci 14:24113–24125

Queller DC, Goodnight KF (1989) Estimating relatedness using molecular markers. Evolution 43:258–275

Ribeiro MM, Sanchez L, Ribeiro C, Cunha F, Araújo J, Borralho NMG, Marques C (2011) A case study of *Eucalyptus globulus* fingerprinting for breeding. Ann For Sci 68:701–714

Ritland K (1996) A marker-based method for inferences about quantitative inheritance in natural populations. Evolution 50:1062–1073

Rodríguez-Ramilo ST, Toro MA, Caballero A, Fernández J (2007) The accuracy of a heritability estimator using molecular information. Conserv Genet 8:1189–1198

Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A (2010) A high high-density Diversity Arrays Technology (DArT) microarray for genome genome-wide genotyping in *Eucalyptus*. Plant Methods 6:16–26

Santure A, Stapley J, Ball AD, Birkhead TR, Burke T, Slate J (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Mol Ecol 19:1439–1451

SAS Institute (2002) SAS user's guide: statistics Version 9.1. SAS Institute, Cary

Sillanpää MJ (2011) On statistical methods for estimating heritability in wild populations. Mol Ecol 20:1324–1332

Simko I, Eujayl I, van Hintum TJL (2012) Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. Plant Sci 184:54–62

Stackpole DJ, Vaillancourt RE, Alves A, Rodrigues J, Potts BM (2011) Genetic variation in the chemical components of *Eucalyptus globulus* wood. G3 Genes Genomes Genet 1:151–159

Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J et al (2011) Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. Mol Phylogenet Evol 59:206–224

Telfer EJ, Stovold GT, Li Y, Silva-Junior OB, Grattapaglia DG, Dungey HS (2015) Parentage reconstruction in *Eucalyptus nitens* using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. PLoS One 10(7):e0130601. doi:10.1371/journal.pone.0130601

Thamarus KA, Groom K, Murrell J, Byrne M, Moran GF (2002) A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre, and floral traits. Theor Appl Genet 104:379–387

Thomas SC (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. Philos Trans R Soc B 360:1457–1467

Thomas SC, Hill WG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics 155:1961–1972

van de Casteele T, Galbusera P, Matthysen E (2001) A comparison of microsatellite-based pairwise relatedness estimators. Mol Ecol 10:1539–1549

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305

Villanueva B, Pong-Wong R, Fernández J, Toro MA (2005) Benefits from marker-assisted selection under an additive polygenic genetic model. J Anim Sci 83:1747–1752

Wang J (2002) An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215

Wang J (2006) Informativeness of genetic markers for pairwise relationship and relatedness inference. Theor Popul Biol 70:300–332

Wang J (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. Genet Res 89:135–153

Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Mol Ecol Resour 11:141–145

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci 101:9915–9920

Williams PC, Sobering DC (1993) Comparison of commercial near-infrared transmittance and reflectance instruments for analysis of whole grains and seeds. J Near Infrared Spectrosc 1:25–32

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338

Yu J, Zhang Z, Zhu C, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome 2:63–77

Zelener N, Marcucci Poltri SN, Bartoloni N, López C, Hopp HE (2005) Selection strategy for a seedling seed orchard design based on trait selection index and genomic analysis by molecular markers: a case for *Eucalyptus dunnii* Maiden. Tree Physiol 25:1457–1467