Tree Genetics & Genomes (2011) 7:63–77
DOI 10.1007/s11295-010-0315-9

ORIGINAL PAPER

Genotyping systems for *Eucalyptus* based on tetra-, penta-, and hexanucleotide repeat EST microsatellites and their use for individual fingerprinting and assignment tests

Danielle Assis Faria · Eva Maria Celia Mamani · Georgios Joannis Pappas Jr. · Dario Grattapaglia

Received: 5 January 2010 / Revised: 14 May 2010 / Accepted: 8 June 2010 / Published online: 22 June 2010

Abstract Eucalypts are keystone species in their natural ranges and are extensively planted worldwide for high-quality woody biomass. A novel set of 21 polymorphic and interspecifically transferable microsatellite markers based on tetra-, penta- and hexanucleotide repeats were developed and tested for high-precision genotyping of species of *Eucalyptus*. These microsatellites were characterized in population samples of four species, *Eucalyptus grandis*, *Eucalyptus globulus*, *Eucalyptus urophylla*, and *Eucalyptus camaldulensis*, representing three phylogenetic sections of subgenus *Symphyomyrtus*. These markers provide a clear advantage for accurate allele calling due to their larger allele size difference. Two multiplexed microsatellite combinations, a 14-locus/four-dye and an 18-locus/five-dye set, analyzable in single lanes were designed, providing resolution and throughput analogous to those routinely used in human DNA profiling. This set of microsatellites was shown to have high resolution for clone fingerprinting, inter-individual genetic distance estimation, species distinction, and assignment of hybrid individuals to their most likely ancestral species. These systems will be particularly useful for comparative population genetics and molecular breeding applications that require consistent allele calling across different points in time or laboratories.

Keywords Microsatellites · Multiplex · Assignment tests · *Eucalyptus*

Introduction

Eucalypts are long-lived, evergreen trees belonging to the angiosperm family *Myrtaceae* that occurs predominantly in the southern hemisphere (Ladiges et al. 2003). Plantation forestry species of *Eucalyptus* are well known for their fast growth, straight form, valuable wood properties, wide adaptability to soils and climates, and ease of management through coppicing (Potts 2004). They are now planted in more than 90 countries where the various species are grown for industrial use in cellulose pulp production, energy supply in the form of charcoal for steel manufacture, sawn timber, essential oils, as well as for firewood, shade, and shelter (Myburg et al. 2007). Besides their role in plantation forestry, eucalypts are dominant or co-dominant trees in almost all vegetation types where they occur and are considered keystone species for ecological studies in their natural ranges (Doughty 2000).

Eucalyptus subgenus *Symphyomyrtus*, the most speciose of the genus with over 300 species, includes the majority of the 20 or so commercially planted species. In temperate regions, *Eucalyptus globulus* has been the premiere choice for plantation forestry, providing fast growth and the best

Communicated by G. G. Vendramin

Electronic supplementary material The online version of this article (doi:10.1007/s11295-010-0315-9) contains supplementary material, which is available to authorized users.

D. A. Faria · E. M. C. Mamani · G. J. Pappas Jr. · D. Grattapaglia
Plant Genetics Laboratory,
Embrapa–Recursos Genéticos e Biotecnologia,
Parque Estação Biológica,
Brasília 70770-970 DF, Brazil

E. M. C. Mamani · D. Grattapaglia
Department of Cell Biology, Universidade de Brasília–UnB,
Brasília 70910-900 DF, Brazil

D. A. Faria · G. J. Pappas Jr. · D. Grattapaglia (✉)
Graduate Program in Genomic Sciences and Biotechnology,
Universidade Católica de Brasília–SGAN 916 modulo B,
Brasília 70790-160 DF, Brazil
e-mail: dario@cenargen.embrapa.br

combination of wood properties for pulp and paper production. In the tropics, on the other hand, production forestry of *Eucalyptus* is currently based on a combination of interspecific hybrid breeding and clonal propagation, with *Eucalyptus grandis* as the pivotal species. Traits such as fast growth, wide adaptability, disease resistance, and tailored wood properties for specific end products are coalesced into elite clones which are in turn propagated for large-scale plantations (Grattapaglia and Kirst 2008).

The hypervariability and simple inheritance of microsatellites provide a powerful system for the unique identification of individuals for fingerprinting purposes, parentage testing, germplasm characterization, and population genetic studies. The individual identification of elite clones is currently a widespread application of molecular markers in tree breeding and production forestry. Quality control and quality assurance of large-scale clonal plantation operations becomes a crucial aspect in forestry, especially in vertically integrated production systems where the pulp mill plans on the availability of wood from specific clones with specific wood properties at specific times. Correct clonal identity also has important implications in several breeding procedures such as seed orchard management or controlled pollination programs affecting the expected gains of breeding cycles (Grattapaglia and Kirst 2008). Additionally, microsatellite markers coupled to Bayesian model-based clustering procedures (Pritchard et al. 2000) have been used in animals (Koskinen 2003; Kumar et al. 2003; Tadano et al. 2008) and increasingly in plants to assign individuals to species/populations or to estimate the most likely ancestral composition of admixed individuals, especially when the phenotypic differentiation between the species/populations in question is difficult and pedigrees are unavailable or ambiguous (Honjo et al. 2008; Millar et al. 2008; Muir and Schlotterer 2005; Sampson and Byrne 2008; Sarri et al. 2006).

The power of microsatellites for individual identification and population genetic studies in *Eucalyptus* has been demonstrated in a number of reports (Chaix et al. 2003; Grattapaglia et al. 2004b; Kirst et al. 2005; Ottewell et al. 2005). However, all the microsatellites currently available for clonal identification and population analysis are derived from genomic sequences containing di- or trinucleotide repeats (Brondani et al. 2006, 1998; Glaubitz et al. 2001; Ottewell et al. 2005; Steane et al. 2001). These markers, while providing powerful discrimination, do not provide high-precision genotyping needed for comparative multilocus profiling across laboratories or even at different times in the same equipment. This is due to the small base pair differences among alleles and to the well-known phenomenon of stuttering during PCR that renders allele calling challenging especially in dinucleotide repeat microsatellites (Litt et al. 1993). In human forensic DNA, a consensus was reached several years ago that for individual identification, tetranucleotide repeat markers should be used as the gold standard (Bar et al. 1995; Gill et al. 1994). Currently, only tetra- and pentanucleotide repeat microsatellites are acceptable for routine human forensic casework (Holt et al. 2002; Krenke et al. 2002).

In vertebrates in general, motifs of length equal or higher than tetranucleotides are relatively frequent and have been commonly used for marker development (Sharma et al. 2007). In plants, while tetranucleotide repeats have been observed at relatively high frequency both in mono- and dicotyledonous genomes (Morgante et al. 2002), only very few recent studies have reported the development of markers based on longer simple sequence repeats from expressed sequence tags (EST; Feng et al. 2009; Yi et al. 2006). In *Eucalyptus*, descriptive studies of existing EST databases (Ceresini et al. 2005; Rabello et al. 2005; Yasodha et al. 2008) confirmed the abundance of microsatellites seen previously in genomic library screening (Brondani et al. 2006). To date, however, no targeted marker development has been made from these EST resources. In this study, we report the development and characterization of a set of 21 polymorphic microsatellite markers based on tetra-, penta-, and hexanucleotide repeats derived from a large collection of ESTs of *Eucalyptus*. Four of the most widely planted species worldwide that also represent contrasting phylogenetic sections within *Symphyomyrtus* were used to develop this set of microsatellite markers, evaluate their interspecific transferability, and assess their genetic information content for population analyses, individual fingerprinting, and assignment tests.

Materials and methods

EST database mining and primer design Tetra-, penta-, and hexanucleotide repeat microsatellites were mined in a database that had approximately 88,000 phred-20 filtered 5′-sequenced ESTs generated during a sequencing effort in the Genolyptus project (Grattapaglia et al. 2004a). EST sequences from leaf and developing xylem RNA were mostly from *E. grandis*, although approximately 30% was derived from three other species: *Eucalyptus urophylla*, *Eucalyptus pellita*, and *Eucalyptus globulus*. With an optimized microsatellite pipeline based on the software MREPS (Kolpakov et al. 2003), simple sequence repeats were identified under the following parameters: two to six base SSR motifs, perfect structure, i.e., no microvariant interruptions, and a minimum core of three repeated units. The microsatellite markers were derived from the alignment of a variable number of ESTs from the four species using *E. grandis* as the reference sequence. Primer pairs flanking these microsatellites were designed targeting expected PCR products between 80 and 450 bp.

Microsatellite marker selection and preliminary screening Only tetranucleotide or higher order motifs were targeted for marker development. No selection was practiced regarding the potential location of the microsatellite in the expressed sequence, base composition of the motif, or BLAST hit identity. Besides being a tetranucleotide or higher repeat motif, priority was given to longer microsatellites, i.e., with a larger number of repeated units based on the assumption that these would likely be more polymorphic at the population level. Screening of primer pairs for simple amplification, polymorphism, and interspecific transferability was carried out in a panel of six unrelated trees involving five different species of *Eucalyptus*, four of them as pure species (*Eucalyptus urophylla*, *E. grandis*, *E. globulus*, *Eucalyptus calmadulensis*) plus *Eucalyptus dunnii* in one of the two hybrid combinations (*E. dunnii* × *E. grandis* and *E. urophylla* × *E. globulus*). Regular primers at small scale were synthesized (AlphaDNA, Montreal, CA, USA) and used for PCR amplification with a common touchdown PCR thermal profile: a hot start for 5 min at 96°C; 10 cycles of 94°C for 1 min, 64°C for 1 min, and 72°C for 2 min; 20 cycles of 94°C for 1 min, 56°C for 1 min, and 72°C for 2 min; and a final elongation step at 72°C for 7 min. The same reaction composition was used as described earlier (Brondani et al. 2006). High-resolution agarose (3.5%) gel electrophoresis and ethidium bromide detection were used for PCR product visualization. Microsatellite markers were classified as transferable when amplification was observed in all five species and tentatively polymorphic when at least one difference in product size was observed among the individuals in the screening panel.

Plant material A population sample of 16 unrelated trees of each one of the four target species, *E. grandis*, *E. urophylla* (section *Latoangulatae*) *E. globulus* (section *Maidenaria*), and *E. camaldulensis* (section *Exsertaria*), were used for microsatellite characterization and to establish preliminary reference species data sets of allele frequencies for the assignment tests. A sample size $n = 16$, i.e., 32 alleles, provides a coefficient of variation of the mean squared error of the expected heterozygosity below 10% (Kirst et al. 2005) adequate for the purpose of this characterization analysis. Within each species, eight trees from each of two different provenances were sampled, Atherton (17°15′ S, 145°28′ E) and Coffs Harbor (30°18′ S, 153°07′ E) for *E. grandis*, Jeeralang (38°24′ S, 146°28′ E) and Flinders Island (40°00′ S, 148°07′ E) for *E. globulus*, Flores Island (8°39′ S, 122°15′ E) and Timor Island (9°37′ S, 124°10′ E) for *E. urophylla*, and Walsh River (17°17′ S, 144°88′ E) and Kennedy River (15°43′ S, 144°17′ E) for *E. camaldulensis*. These have been some of the most widely employed provenances in breeding programs in Brazil and thus most relevant for the evaluation of the assignment tests. A set of 24 elite public clones commercially planted in Brazil were used as a test set to evaluate the power of the microsatellite markers to assign individuals to their most likely source species. Four of the 24 clones had a documented hybrid origin as they were produced by controlled crosses of *E. camaldulensis* × (*E. urophylla* × *E. globulus*) and thus served as control cases for admixed individuals. The remaining 20 clones were of unknown origin, although anecdotal reports suggest a hybrid origin for several of them involving mainly *E. grandis* and *E. urophylla*.

Microsatellite genotyping DNA extractions from expanded leaves of the target trees and microsatellite genotyping by fluorescence detection was carried out as described earlier (Missiaggia et al. 2005), with some modifications in the PCR protocol. PCR reactions in multiplexed systems were carried out in 10 µl volumes containing 1 µl of 10× Qiagen Multiplex PCR Buffer (Qiagen Inc., Valencia, CA, USA), equal concentration (0.1 µM) of all primers for all microsatellite markers co-amplified, and 2.0 ng of genomic DNA. The recommended Qiagen Multiplex PCR Handbook cycling protocol was used with an annealing temperature of 60°C and 30 PCR cycles. PCRs were carried out in hexaplex or heptaplex systems combining markers in such a way that loci whose alleles migrate in the same size range were labeled with different fluorochromes either 6-FAM (blue), NED (yellow), or HEX or VIC (green). To assist in the design of the multiplexed genotyping systems, primer pairs for all selected microsatellites were screened for potential cross-reactivity (i.e., primer dimer and hairpin structures) using the web-based version of the software AutoDimer (Vallone and Butler 2004). Default parameters were used and primer pairs that displayed primer dimer structures with score value >7 (i.e., number of matches minus number of mismatches) were avoided when choosing loci to be co-amplified. An aliquot of 1 µl of PCR was mixed with 1 µl of freshly prepared ROX-labeled size standard (Brondani and Grattapaglia 2001) and 10 µl of Hi-Di formamide (Applied Biosystems, Foster City, CA, USA). The mixture was electroinjected in an ABI 3100 genetic analyzer and data collected under dye set D spectral calibration using Genescan and analyzed with Genotyper (Applied Biosystems).

Microsatellite characterization The following parameters of genetic information content were estimated for each microsatellite marker and species separately: (1) number of alleles (A); (2) allele size range; (3) observed (H_o) and expected (H_e) heterozygosity and a p value of an exact test for Hardy–Weinberg equilibrium; (4) polymorphism information content (PIC; Botstein et al. 1980); (5) probability of identity (PI) that corresponds to the probability of two random individuals displaying the same genotype; and (6) paternity exclusion probability (PE) that corresponds to the power with which the locus excludes an erroneously selected individual tree as being the parent of an offspring. This last parameter was estimated taking into account frequent situations when using microsatellites for paternity analysis in forest trees: (PE_1) paternity exclusion probability for one candidate parent given the genotype of a known parent, a common situation when paternity is investigated in open-pollinated progeny individuals with maternal control, and (PE_2) paternity exclusion probability for a candidate parent pair, a common situation when paternity and maternity needs to be checked in progeny individuals derived from controlled crosses, i.e., with maternal and paternal control. The software Cervus (Kalinowski et al. 2007) was used to estimate A, H_o, H_e, PIC, PI, and both versions of PE, and Powermarker (Liu and Muse 2005) was used to carry out an exact test for Hardy–Weinberg equilibrium for each microsatellite marker. Considering that *Eucalyptus* species are known for operating largely under a mixed mating model (Burczyk et al. 2002; Gaiotto et al. 1997), the frequency of null alleles at the 21 loci in the four species was estimated using an individual inbreeding model with the software INEST (Chybicki and Burczyk 2009). To account for missing data due to PCR failure, this analysis also provided a probability estimate (β) for absence of alleles due to random amplification failure as opposed to null allele homozygosity. The combined multilocus paternity exclusion probabilities and the probability of identity were also estimated for different combinations of multiplexed systems of microsatellites for genotyping applications.

Evaluation of microsatellites for genetic distance and population structure analysis Multilocus genotypes for the 21 microsatellites were used to estimate pairwise individual-level genetic distances among the 88 individuals (64 pure species individuals, 4 known hybrids, and 20 suspect hybrid elite clones) to specifically assess the effective discrimination ability for fingerprinting purposes. For the co-dominant data, a shared allele distance (D_{SA}) was calculated based on the infinite allele model. DSA is estimated by $1 - P_{SA}$, where P_{SA} is the proportion of shared alleles averaged across loci (Bowcock et al. 1994). Distance matrices (1,000 bootstrap replicates) were calculated using MICROSAT (Minch et al. 1995). The matrix of genetic distances was then used to graphically represent distance relationships between the 88 individuals with an unweighted pair group method with arithmetic mean (UPGMA) consensus tree (majority rule, strict) constructed using the NTSYS 2.0 package (Exeter Software, USA). Based on the genotype data at the 21 microsatellite loci, the 64 individual trees of the reference species were assigned probabilistically to a given number of populations inferred with a Bayesian approach without any prior population information using STRUCTURE 2.1 (Pritchard et al. 2000). The tests were done based on an admixture model where the allelic frequencies were correlated and applying burn-in period of 50,000 and 100,000 iterations for data collection. The analysis was run with K ranging from two to eight inferred clusters (four species and two provenances per species) performed with five independent runs each. The model choice criterion to detect the most probable value of K was ΔK (Evanno et al. 2005). Average results of ten runs at the most likely K were entered into DISTRUCT (Rosenberg 2004) to provide a graphic display of population structure. Pairwise estimates of population differentiation (F_{st}) between the four *Eucalyptus* species were also obtained using the software Arlequin (Excoffier et al. 2005).

Evaluation of microsatellites for assignment tests Assignment of the individuals of the test set, i.e., the 24 elite clones, to the clusters created based on the reference species sets was carried out with STRUCTURE 2.1 (Pritchard et al. 2000) using both the reference and test sets combined. Assignments were tested using prior population information for individuals from the reference data set and an admixture model to allow for more flexibility to deal with the complexities of these populations. The number of clusters (K) was set to the most likely value determined in the previous structure analysis. STRUCTURE 2.1 was thus used to estimate the posterior probability that each test individual (elite clone) belongs to a given cluster corresponding to each one of the *Eucalyptus* species under consideration. Average results of the posterior probabilities of ten independent runs at the most probable K were used to estimate the most likely hybrid composition of each individual elite clone. These values were entered into DISTRUCT (Rosenberg 2004) to provide a graphic display of the ancestral composition of each elite clone.

Results and discussion

Microsatellite development The data mining and microsatellite pipeline used for microsatellite marker development revealed 1,261 potential markers that met the specified constraints and for which primer pairs could be designed. Details of that study will be the subject of a separate publication. Out of the set of 1,261 potentially useful markers, the number of microsatellites that displayed at least three repeated units as the core microsatellite were 83, 51, and 116, respectively, for tetra-, penta-, and hexanucleotide repeats. This distribution reflects the higher frequency of tetra- and hexanucleotide repeats seen in genic regions when compared to pentanucleotides in *Eucalyptus* (Rabello et al. 2005), in line with previous reports in both mono- and dicotyledonous plants (Morgante et al. 2002; Zhang et al. 2004). Preliminary marker screening for amplification success and polymorphism detection was carried out for 50 tetra-, 18 penta-, and 24 hexanucleotide repeat microsatellites that displayed the largest number of tandemly repeated units in silico. From the preliminary screening, 36 primer pairs (19 tetranucleotide, 5 pentanucleotide, and 12 hexanucleotide) showed robust amplification and indication of polymorphism. These were selected for high-resolution fluorescence-based screening. Except for one locus that was removed from any further screening steps, all those that were deemed polymorphic based on the low-resolution agarose gel screening did in fact display more than one allele when tested in high-resolution electrophoresis. For the purpose of this study, where the objective was to select a set of polymorphic and transferable microsatellites across the four target species, a relatively stringent threshold was set that markers had to be polymorphic in at least three of the four species, i.e., display at least two alleles in a limited sample of 16 trees per species. This constraint was met by 11 tetranucleotide-based microsatellites, three pentanucleotides, and seven hexanucleotides for which full information is presented including the motif, the expected amplicon size, forward and reverse primer pairs, Genbank accession number of the original sequence from which the microsatellite primer pairs were designed, and the database of sequence tagged sites (dbSTS) Id (Table 1). BLASTx functional annotation returned highest hits to *Ricinus communis* (nine loci), followed by *Vitis vinifera* (five loci), *Populus trichocarpa* (five loci), and one each to *Arabidopsis thaliana* and to *Carica papaya*. Most genes where these microsatellites are contained have not yet been functionally characterized (data not shown). All these microsatellites are located in the 5′-untranslated (UTR) region of the gene. The general abundance of microsatellites in the 5′-UTR of plant genes has been described (Morgante et al. 2002). Observations of gradients of microsatellite density along the direction of transcription in rice and *Arabidopsis* ESTs (Fujimori et al. 2003) and in the rice genome when introns were scrutinized (Parida et al. 2009) suggest that some genic non-coding microsatellites might take part in regulating gene expression. These observations were reported almost exclusively for di- and trinucleotide repeat microsatellites, not higher order repeats and for a relatively small proportion of genes. It is therefore unlikely that the microsatellites developed in our study would be under selective pressure to a point that would jeopardize the premise of neutrality for population genetics studies.

Microsatellite characterization The 21 microsatellites spanned a wide range of allele sizes (Table 2) which later proved very useful to design multiplexed sets of markers that allow an optimized and higher throughput genotyping. The size range of the alleles for most loci matched the expected size of the in silico predicted amplicon. However, for loci EMBRA943 and EMBRA1374, the observed size was significantly larger than the expected one, strongly suggesting amplification across intronic sequences. The allele size range did not vary much across the four species, an important aspect to design more generalized multiplex genotyping panels (Table 3). The average number of alleles varied across loci with three monomorphic microsatellites in *E. urophylla* (EMBRA1456, EMBRA1463, and EMBRA1945) and a maximum of ten alleles observed for EMBRA813 and EMBRA1364 in *E. camaldulensis*. The average number of alleles overall species and markers was 4.43, slightly higher for *E. camaldulensis* (5.10), although not significantly different among the four species (F = 1.408, p = 0.246). Expected heterozygosities (H_e) in both species were nominally larger than the observed heterozygosity (H_o) for several loci. A goodness-of-fit test for Hardy–Weinberg equilibrium (HWE) revealed only eight (three in *E. grandis*, two in *E. globulus*, two in *E. urophylla*, and one in *E. camaldulensis*) out of the 84 tests significantly deviated from expectations at $\alpha < 0.01$ (Table 2). However, more markers could show deviations as these tests have low power due to the relatively limited sample size. Marker EMBRA1945 did not fit HWE expectations for *E. grandis* and *E. globulus*, in both cases with a significant deficiency of heterozygotes, and monomorphic in *E. urophylla*, suggesting the occurrence of null alleles.

The frequencies of null alleles for all 21 loci in all four species were estimated under the individual inbreeding model (IIM; Table 2). This model provides a useful approximation for species with a mixed mating system as it allows for an accurate estimate of null allele frequency regardless of the sample size, the number of loci, or the actual inbreeding coefficient (Chybicki and Burczyk 2009). It should be noted, however, that under this model, an estimate of null allele frequency should only be considered significantly different from zero when the locus deviates from HWE expectations and not purely based on its absolute estimated value. We found that an estimated frequency of null allele between 0.1 and 0.3 was observed in those few cases (8 in 84 locus × species combinations) where a significant deviation from HWE ($p<0.01$) was observed due to an excess of homozygotes. Furthermore, the probability (β) of random amplification failure for these

Table 1 Basic properties of the 21 microsatellite markers developed and evaluated in this study including the identification number in the dbSTS

Marker locus	SSR type	Motif	Expected amplicon size (bp)	Forward primer (5′-3′)	Reverse primer (5′-3′)	GenBank accession no.	dbSTS_Id
EMBRA813	Tetra	(CTCC)15	90	ATCTCTCTCGCCGATCTCAA	CGGAGAGATCAAAGGCATGT	GF101851	1232932
EMBRA850	Hexa	(CGCCCC)23	280	GGTGGTTCCTTGAAGATGGA	TCCTCAGGGGATTGTAGACG	GF101852	1232933
EMBRA915	Tetra	(CCCT)21	213	GGAGGAGGAGGAACAGGAAC	GCACCCGGTTCTTAAATCAA	GF101866	1232947
EMBRA925	Tetra	(TCCT)14	249	ATCCATCCCACCAAGGAAAT	CGTAGAACTTGGCGAGGAAG	GF101853	1232934
EMBRA943	Hexa	(CGCAAC)29	150	GTCTTCCTCCTCGCCTTCTT	ATCTTCTTCACGTCGTCGCT	GF101867	1232948
EMBRA954	Tetra	(CTGC)17	172	TGTTCCTGCTTCTCCCATTC	AAAAATACTCCTCCGCCTCC	GF101854	1232935
EMBRA1008	Tetra	(ATCG)16	171	AAGCTCGCAGCTCAGAAAAA	GTACTTGTCCTCCGCCATGT	GF101855	1232936
EMBRA1040	Penta	(CTCCT)18	323	TCCGCACAAACACACAAAAC	GCACACACCCCCAATTTTAG	GF101856	1232937
EMBRA1307	Hexa	(GCTCCC)18	348	TTGATTCCAAATCTCGCCTC	CAACCAAACAGCTTCGAGGT	GF101857	1232938
EMBRA1364	Tetra	(CTCC)15	328	CGTTTTCGCTCCTCTCTCTC	TGTAGAGATCGGGGTCCTTG	GF101858	1232939
EMBRA1374	Hexa	(CGCCGT)26	271	GTCTGAACTCGGCTTCCTTG	TTCTTCCCGTTGTAAATCCG	GF101859	1232940
EMBRA1456	Hexa	(CCGCCT)19	263	TTCCGACGGTTATTTGAAGG	GAAACGATTTCTTGGCTTGC	GF101868	1232949
EMBRA1463	Penta	(GCTCG)21	253	GCGCAGAACAACAAGAAGAA	CAACGCAGGAAGAGAACCTC	GF101860	1232941
EMBRA1616	Hexa	(TCTCCA)24	142	GGACACTCTGCAACCCTCTC	GACGAGGTGGAACCTGTAGC	GF101861	1232942
EMBRA1757	Tetra	(TCGA)18	179	TTCTCGCTGGGAATCAATTT	CGAGAGGAGTCCATAGCTGG	GF101869	1232950
EMBRA1811	Hexa	(CTCCTG)26	298	GTCGAGTTGAGTTCGCTTCC	AGTGAATCGGGAGAGGAGGT	GF101862	1232943
EMBRA1812	Tetra	(CTGA)13	272	ATTCCGAAGCCCTAAAAGGA	TTTTGCCTTATGGGAAATGG	GF101863	1232944
EMBRA1851	Tetra	(ACGG)29	120	GTCGTCGCCATTGAAGTTCT	CGATCCTATCAGGCTCAGTG	GF101870	1232951
EMBRA1945	Penta	(GTGGT)23	278	CCGGGCTAGCTCTTTCTCTC	GAACCTCTCCATCTCCTCCC	GF101871	1232952
EMBRA1977	Tetra	(GCGA)25	114	TTCGGCGATAGGGTTTATTG	AACTTGACGAGGAGGGGATT	GF101864	1232945
EMBRA2014	Tetra	(AGGA)20	124	CACCGACTTCCTCTTCTTCG	CCCCATCCCTTCTCTCTCTC	GF101865	1232946

Table 2 Descriptive statistics of the 21 microsatellites for the four *Eucalyptus* species

Marker	*A*	Size range	H_o	H_e	*p* value	$\mathrm{null}_{\mathrm{IIM}}$	PIC	PE_1	PE_2	PI	*A*	Size range	H_o	H_e	*p* value	$\mathrm{null}_{\mathrm{IIM}}$	PIC	PE_1	PE_2	PI
	E. grandis										*E. globulus*									
EMBRA813	4	83–95	0.600	0.720	0.584	0.104	0.644	0.444	0.620	0.144	7	83–111	0.500	0.706	0.508	0.134	0.648	0.466	0.664	0.136
EMBRA850	4	265–283	0.800	0.729	0.386	0.058	0.656	0.457	0.636	0.135	4	265–283	0.563	0.700	0.398	0.118	0.621	0.418	0.591	0.161
EMBRA915	4	203–219	0.625	0.639	0.669	0.084	0.559	0.356	0.520	0.205	3	207–215	0.250	0.232	0.955	0.107	0.210	0.112	0.198	0.616
EMBRA925	5	246–266	0.538	0.711	0.095	0.156	0.634	0.439	0.622	0.149	4	230–246	0.143	0.267	0.029	0.194	0.246	0.138	0.247	0.562
EMBRA943	3	422–452	0.286	0.265	0.017	0.107	0.240	0.131	0.232	0.570	3	446–464	0.750	0.750	0.046	0.288	0.582	0.362	0.510	0.193
EMBRA954	5	172–188	0.688	0.683	0.882	0.068	0.606	0.408	0.586	0.170	3	164–176	0.400	0.536	0.567	0.136	0.451	0.263	0.403	0.299
EMBRA1008	4	159–175	0.333	0.595	0.139	0.220	0.512	0.323	0.492	0.242	4	159–171	0.231	0.348	0.039	0.176	0.317	0.185	0.318	0.461
EMBRA1040	3	299–309	0.400	0.352	0.816	0.089	0.314	0.177	0.299	0.461	5	304–329	0.700	0.763	0.678	0.108	0.681	0.489	0.671	0.120
EMBRA1307	2	344–350	0.063	0.175	0.011	0.194	0.155	0.078	0.138	0.703	3	326–344	0.333	0.307	0.948	0.220	0.269	0.148	0.255	0.525
EMBRA1364	7	311–347	0.357	0.767	0.010	0.238	0.708	0.535	0.731	0.099	7	311–343	0.625	0.819	0.039	0.130	0.766	0.605	0.791	0.070
EMBRA1374	6	333–369	0.750	0.752	0.520	0.065	0.683	0.494	0.677	0.119	7	333–375	0.625	0.631	0.859	0.065	0.585	0.410	0.616	0.177
EMBRA1456	3	250–262	0.583	0.638	0.522	0.119	0.535	0.326	0.474	0.227	3	244–256	0.385	0.557	0.500	0.159	0.428	0.233	0.351	0.324
EMBRA1463	3	249–259	0.357	0.320	0.882	0.099	0.286	0.159	0.273	0.500	2	254–259	0.273	0.368	0.458	0.171	0.290	0.145	0.230	0.483
EMBRA1616	3	148–160	0.214	0.320	0.003	0.177	0.286	0.159	0.273	0.500	4	142–160	0.333	0.395	0.957	0.128	0.347	0.198	0.328	0.417
EMBRA1757	4	145–165	0.500	0.521	0.729	0.088	0.468	0.294	0.464	0.282	3	169–181	0.313	0.280	0.908	0.098	0.248	0.132	0.228	0.554
EMBRA1811	3	287–299	0.375	0.433	0.369	0.113	0.354	0.190	0.301	0.402	4	287–305	0.273	0.610	0.008	0.256	0.533	0.343	0.516	0.224
EMBRA1812	4	255–271	0.563	0.696	0.185	0.111	0.621	0.420	0.595	0.159	2	267–271	0.188	0.272	0.248	0.155	0.229	0.114	0.191	0.577
EMBRA1851	4	102–122	0.800	0.756	0.923	0.061	0.682	0.481	0.655	0.122	8	94–126	0.688	0.758	0.077	0.074	0.703	0.530	0.729	0.102
EMBRA1945	5	251–276	0.385	0.680	0.008	0.143	0.619	0.435	0.632	0.155	5	251–281	0.400	0.701	0.000	0.201	0.638	0.451	0.644	0.143
EMBRA1977	3	96–112	0.375	0.446	0.737	0.100	0.378	0.212	0.338	0.377	5	100–116	0.500	0.520	0.951	0.091	0.447	0.269	0.420	0.303
EMBRA2014	4	113–132	1.000	0.662	0.020	0.035	0.579	0.373	0.537	0.191	5	109–125	0.938	0.665	0.073	0.037	0.588	0.388	0.561	0.183
Average	3.952	–	0.504	0.565	–		0.501	0.328	0.481	0.282	4.333	–	0.448	0.533	–		0.468	0.305	0.451	0.316
	E. urophylla										*E. camaldulensis*									
EMBRA813	5	75–91	0.933	0.782	0.426	0.039	0.715	0.528	0.710	0.100	10	59–95	0.857	0.897	0.484	0.072	0.850	0.729	0.890	0.033
EMBRA850	5	259–283	0.813	0.700	0.902	0.046	0.638	0.450	0.643	0.143	3	265–277	0.267	0.480	0.234	0.190	0.383	0.207	0.320	0.368
EMBRA915	3	199–207	0.385	0.385	0.197	0.107	0.324	0.176	0.287	0.443	8	187–215	0.929	0.886	0.089	0.050	0.837	0.705	0.871	0.039
EMBRA925	5	238–258	0.688	0.764	0.584	0.093	0.694	0.500	0.678	0.113	8	234–262	0.467	0.853	0.118	0.213	0.801	0.651	0.828	0.054
EMBRA943	5	416–464	0.375	0.516	0.032	0.158	0.474	0.306	0.489	0.276	6	422–464	0.750	0.783	0.701	0.081	0.712	0.530	0.717	0.100
EMBRA954	5	168–184	0.813	0.740	0.379	0.057	0.664	0.463	0.637	0.133	5	160–188	0.750	0.685	0.589	0.058	0.602	0.400	0.570	0.175
EMBRA1008	6	159–187	0.313	0.706	0.003	0.245	0.638	0.450	0.639	0.146	4	155–179	0.533	0.634	0.429	0.109	0.534	0.331	0.485	0.228
EMBRA1040	5	299–324	0.750	0.808	0.107	0.079	0.748	0.571	0.750	0.082	5	304–319	0.667	0.708	0.313	0.082	0.649	0.464	0.660	0.135
EMBRA1307	3	338–350	0.267	0.349	0.405	0.146	0.309	0.172	0.290	0.467	2	338–344	0.067	0.067	0.894	0.144	0.062	0.031	0.059	0.877
EMBRA1364	7	315–339	0.563	0.821	0.312	0.155	0.766	0.602	0.786	0.071	10	303–347	0.643	0.886	0.317	0.136	0.838	0.709	0.876	0.038
EMBRA1374	4	339–363	0.400	0.526	0.201	0.122	0.466	0.287	0.450	0.284	3	339–351	0.286	0.373	0.120	0.155	0.331	0.188	0.314	0.439
EMBRA1456	1	250	0.000	0.000	–	0.160	0.000	0.000	0.000	1.000	3	250–262	0.143	0.262	0.136	0.187	0.234	0.125	0.219	0.577
EMBRA1463	1	249	0.000	0.000	–	0.159	0.000	0.000	0.000	1.000	3	244–254	0.200	0.191	0.980	0.122	0.175	0.093	0.169	0.674
EMBRA1616	6	130–160	0.800	0.733	0.892	0.052	0.674	0.492	0.688	0.120	4	130–148	0.714	0.738	0.684	0.078	0.661	0.459	0.635	0.134

Table 2 (continued)

Marker	A	Size range	H_o	H_e	p value	null$_{IIM}$	PIC	PE_1	PE_2	PI	A	Size range	H_o	H_e	p value	null$_{IIM}$	PIC	PE_1	PE_2	PI
EMBRA1757	7	145–185	0.563	0.633	0.001	0.097	0.588	0.415	0.624	0.174	7	145–177	0.417	0.815	0.008	0.253	0.751	0.582	0.769	0.079
EMBRA1811	5	275–299	0.438	0.782	0.038	0.204	0.718	0.532	0.714	0.098	6	281–311	0.538	0.652	0.252	0.117	0.599	0.422	0.628	0.167
EMBRA1812	2	263–267	0.063	0.063	0.897	0.144	0.059	0.029	0.056	0.884	2	263–291	0.077	0.077	0.885	0.152	0.071	0.036	0.067	0.860
EMBRA1851	7	94–137	0.538	0.803	0.069	0.172	0.736	0.559	0.742	0.088	6	98–122	0.769	0.834	0.069	0.084	0.774	0.611	0.793	0.067
EMBRA1945	1	251	0.000	0.000	–	0.331	0.000	0.000	0.000	1.000	3	251–271	0.900	0.595	0.082	0.059	0.482	0.283	0.422	0.272
EMBRA1977	4	96–108	0.538	0.772	0.445	0.160	0.695	0.495	0.668	0.114	5	86–108	0.571	0.804	0.269	0.154	0.740	0.560	0.741	0.086
EMBRA2014	4	117–129	0.462	0.714	0.153	0.176	0.637	0.438	0.617	0.147	4	105–117	0.467	0.526	0.210	0.111	0.466	0.287	0.450	0.284
Average	4.333	–	0.462	0.552	–	–	0.502	0.498	0.641	0.328	5.095	–	0.524	0.607	–	–	0.550	0.400	0.547	0.271

Estimates are reported for each species separately including: (A) number of alleles; allele size range in base pairs; (H_o) observed heterozygosity; (H_e) expected heterozygosity; p value of an exact test for Hardy–Weinberg equilibrium; (null$_{IIM}$) estimate of null allele frequency given the individual inbreeding model-based estimator; (PIC) polymorphism information content; (PE_1) paternity exclusion probability for one candidate parent given the genotype of a known parent; (PE_2) paternity exclusion probability for a candidate parent pair; and (PI) probability of identity (see text for explanations)

few cases that deviated from HWE was always lower than 0.6%, indicating that the data are not consistent with random amplification failure but rather with the occurrence of true null alleles. No consistent pattern of deviation from HWE and frequency of null allele was seen across species. In other words, no specific locus can be pointed out as being more prone to the occurrence of null alleles in all four species. Rather it seems to follow a specific locus by species interaction. These results taken together indicate that the occurrence of null alleles is relatively rare and not a significant issue for these microsatellites selected for higher transferability. In the few particular cases where null alleles frequencies be considered significantly different from zero due to HWE deviation, the corresponding primers could be redesigned to attempt alternative flanking priming sequences. However, this might prove challenging due to the very high nucleotide diversity around 1 SNP every 30 bp for *E. globulus* and up to 1 SNP every 16 bp for *E. camaldulensis*, as recently described in a range wide re-sequencing survey of 23 genes (Kulheim et al. 2009).

Overall, this set of microsatellites has a lower information content when compared to dinucleotide repeat markers that typically display on average ten alleles per locus and heterozygosities in the range of 0.70–0.80 (Brondani et al. 2006; Ottewell et al. 2005). A lower variability of the tetra-, penta-, and hexanucleotide repeats was expected as the rate of mutation for longer simple sequence repeats has been reported to be lower in general for animals and plants (Chakraborty et al. 1997; Vigouroux et al. 2002). In spite of the lower number of alleles, the allele frequency distribution was such that good discrimination power both for parentage testing and individual identification could be reached for the majority of the microsatellites in both species (Table 3). Average PIC, paternity exclusion probabilities (PE_1 and PE_2), and PI for the set of loci were not substantially different among the four species, and these parameters were within the same range for several loci in all four species. However, for some markers, these parameters differed among species, reflecting the difference in number of alleles and/or their frequency distributions. An example was EMBRA915, highly informative in *E. camaldulensis* but not so in *E. globulus*. The impact of allele frequency distribution could be visibly recognized in locus EMBRA1851 where, in spite of the higher number of alleles in *E. globulus* (A = 8), when compared to *E. grandis* (A = 4) the allele frequency distribution was such that the genetic information content of this locus was very similar in both species. Again, as expected, these tetra-, penta-, and hexanucleotide microsatellites are evidently less powerful when it comes to parentage and individual identification when compared to dinucleotide repeat microsatellites. However, their clear advantage arises when it comes to the precision of the allele calling, a key aspect for several

Table 3 Proposed multiplex systems for high-throughput genotyping with the developed microsatellites

Marker locus		Allele size range					
		E. grandis	*E. globulus*	*E. urophylla*	*E. camaldulensis*	14-plex 4-DYE labeling	18-plex 5-DYE labeling
EMBRA943		422–452	446–464	416–464	422–464	FAM	FAM
EMBRA1374		333–369	333–375	339–363	339–351	FAM	FAM
EMBRA850		265–283	265–283	259–283	265–277	FAM	FAM
EMBRA915		203–219	207–215	199–207	187–215	FAM	FAM
EMBRA954		172–188	164–176	168–184	160–188	FAM	FAM
EMBRA2014		113–132	109–125	117–129	105–117	FAM	FAM
EMBRA1040		299–309	304–329	299–324	304–319	NED	NED
EMBRA925		246–266	230–246	238–258	234–262	NED	NED
EMBRA1008		159–175	159–171	159–187	155–179	NED	NED
EMBRA1851		102–122	94–126	94–137	98–122	NED	NED
EMBRA1811		287–299	287–305	275–299	281–311	–	PET
EMBRA1945		251–276	251–281	251	251–271	–	PET
EMBRA1757		145–165	169–181	145–185	145–177	–	PET
EMBRA1977		96–112	100–116	96–108	86–108	–	PET
EMBRA1364		311–347	311–343	315–339	303–347	VIC/HEX	VIC/HEX
EMBRA1812		255–271	267–271	263–267	263–291	VIC/HEX	VIC/HEX
EMBRA1616		148–160	142–160	130–160	130–148	VIC/HEX	VIC/HEX
EMBRA813		83–95	83–111	75–91	59–95	VIC/HEX	VIC/HEX
Combined PE_1 (%)	(14-plex)	99.8832	99.7781	99.9674	99.9924		
Combined PE_2 (%)	(14-plex)	99.9988	99.9972	99.9999	99.9999		
Combined PI	(14-plex)	1.4009E−10	2.1055E−10	8.7047E−12	5.0393E−13		
Combined PE_1 (%)	(18-plex)	99.9703	99.9492	99.9955	99.9994		
Combined PE_2 (%)	(18-plex)	99.9999	99.9998	99.9999	99.9999		
Combined PI	(18-plex)	9.2804E−13	7.2404E−12	1.6921E−14	1.5552E−16		

Allele size ranges and dye labels are included to show their compatibility and the differences in size ranges between each consecutive marker locus. Combined performances of the 14-plex and 18-plex systems for paternity exclusion (PE_1 and PE_2) and PI are listed below for each species

applications in genetic analysis. Still some markers such as EMBRA813 (tetra), EMBRA1364 (tetra), EMBRA1374 (hexa), and EMBRA1851 (tetra) displayed very comparable genetic information content to the average dinucleotides (Kirst et al. 2005), with probability of paternity exclusion around 0.6–0.7 and probability of identity around 0.1–0.2. These could be pointed out as the most informative microsatellites across all four species. This result is important as it indicates that further screening of larger sets of tetra- and hexanucleotide repeat microsatellites is warranted and could lead to the discovery of several other markers with similar behavior. With the upcoming availability of the full genome sequence for *E. grandis*, it will be possible to develop a larger set of such higher order repeat markers that consolidate high information content and high-precision genotyping quality.

Microsatellite multiplexed systems for high-throughput genotyping Besides evaluating the information content of these new markers, we were interested in providing a practical toolkit to apply them in routine genotyping in *Eucalyptus*. The relatively small number of alleles per locus turned out to be an advantage when it came to the ability of multiplexing loci in single electrophoretic runs. Narrower allele size ranges allow fitting more markers in the same fluorescence detection spectrum. Out of the 21 loci, 18 were selected with the best compromise of genetic information content across the four species. Loci EMBRA1307, EMBRA1456, and EMBRA1463 were left out as they had the lowest PIC values when all species together were considered and two of them were monomorphic in *E. urophylla*. Marker EMBRA1945, although monomorphic in *E. urophylla,* was kept for the 18-locus multiplex version as it is informative for the other three species. Given that different laboratories are used to different fluorescence dye sets, two multiplex options were designed, one based on a four-dye set and a second one on a five-dye set. The larger five-dye system with 18 loci (18-plex) is simply an extension of the first one, with 14 loci (14-plex), by the addition of four markers labeled with a fifth fluorescence. The proposed multiplex designs provide flexibility to use only some or all the markers and whatever fluorochrome combination desired. The designs respect the compatibility of the allele size range, leaving usually between 20- and 30-bp difference between loci labeled with the same dye. Although new, rarer alleles will likely be detected as the number of individuals genotyped increases, such a difference should provide sufficient buffering capacity to accommodate several new alleles without overlapping. No significant allele dropout issues were seen for this set of markers by using the high-quality PCR reagents specifically designed for multiplexed amplification. It should be noted that the maximum size range was 422–452 bp for EMBRA943 (Table 2), which is well within the amplifiable range. Furthermore, even if some larger alleles emerge as more individuals are genotyped, it is unlikely that their size will be much larger than two or three further repeat units due to the slower evolution of such longer repeats.

These two different multiplex systems have a very high combined power of paternity exclusion above 99.9% in both parentage testing situations with a higher power of the 18-plex over the 14-plex reaching >99.999%. They also provide a low combined probability of genetic identity between unrelated trees, below 10^{-10} in all species for the 14-plex and below 10^{-12} for the 18-plex (Table 3). Using the protocol described in this work, we successfully amplified all loci in two separate PCR reactions in the case of the 14-plex and three PCRs in the 18-plex. When evaluating which microsatellites could be suitably co-amplified in the same PCR reaction, the following microsatellite combinations were avoided due to primer–dimer primer interactions with score values (number of matches minus number of mismatches) greater than the threshold value of 7 recommended for designing multiplex PCR reactions (Vallone and Butler 2004): EMBRA2014 with EMBRA813 (score = 8); EMBRA1977 with EMBRA813 (score = 7); EMBRA954 with EMBRA915 (score = 11); EMBRA943 with EMBRA915 (score = 9); and EMBRA943 with EMBRA925 (score = 7). No hairpin structures were detected in the analysis. All loci in each multiplex were analyzed in a single electroinjection providing a very high data throughput analogous to the one routinely used in human DNA profiling (Krenke et al. 2002). It should be noted, however, that due to the potential incidence of indels immediately flanking these microsatellites, likely to occur in the highly diverse genome of *Eucalyptus*, difficulties in the multiplexing approach may arise as more individuals, populations, and species are genotyped.

Microsatellite performance for individual discrimination A preliminary evaluation of the power of these microsatellites for genetic distance analyses showed that this set of microsatellites provides high resolution for individual discrimination (Fig. 1). Estimates of average shared allele distances (D_{SA}) among individuals within *E. grandis* (D_{SA} = 0.478), *E. globulus* (D_{SA} = 0.474), *E. urophylla* (D_{SA} = 0.502), and *E. camaldulensis* (D_{SA} = 0.536) were in the same range, although significantly different (F = 12.99; p = 3.59 × 10^{-8}), with significantly larger values for *E. camaldulensis* when compared to the next highest estimate of D_{SA} (*E. urophylla*; t = 3.24, p = 0.00068), reflecting the higher information content of this set of loci in this species. When all 64 individuals of the species samples were analyzed jointly, the average distance increased to D_{SA} = 0.694 and ranged from 0.214 to 0.972. The average distance among the 20 elite clones of unknown origin was D_{SA} = 0.513 and ranged from 0.214 to 0.792. When plotted in a phenogram, all individuals could be clearly discriminated. Individuals belonging to each species clustered together in clearly separate branches as expected, while the elite clones formed a separate cluster positioned closer to *E. urophylla*. This distance-based analysis yielded essentially the same clusters as the population-based STRUCTURE analysis (see below). The basal nodes separating the major species clusters were, however, not supported by significant bootstrap values. Only a few bootstrap values >50 were obtained and only for nodes between subgroups inside the major species clusters or between individuals at the end of the branches (data not shown). This is in agreement with the observation that microsatellite data become less informative for phylogenetic analyses among distantly related taxa (Bowcock et al. 1994).

Provenances did not form discrete clusters, with the exception of *E. grandis* where individuals from Atherton and Coffs Harbor formed two separate clusters. All the elite clones but clone Sem3 formed a separate cluster closer to the *E. urophylla* branch. Clone Sem3 clustered directly with the *E. urophylla* individuals, suggesting a possible pure species origin, while the other 19 clones more likely have a hybrid composition with a predominance of *E. urophylla*. The 20 elite clones displayed unique multilocus genotypes with a minimum of nine allelic differences and an average of 21.5 differences out of the 42 alleles compared (2×21 loci) in all pairwise comparisons. These microsatellites provide high-resolution and high-quality fingerprints useful for clonal protection or quality control procedures in breeding and deployment operations.

Microsatellite performance for population structure analysis The multilocus data for the species reference sets were subject to a population structure analysis to test for the

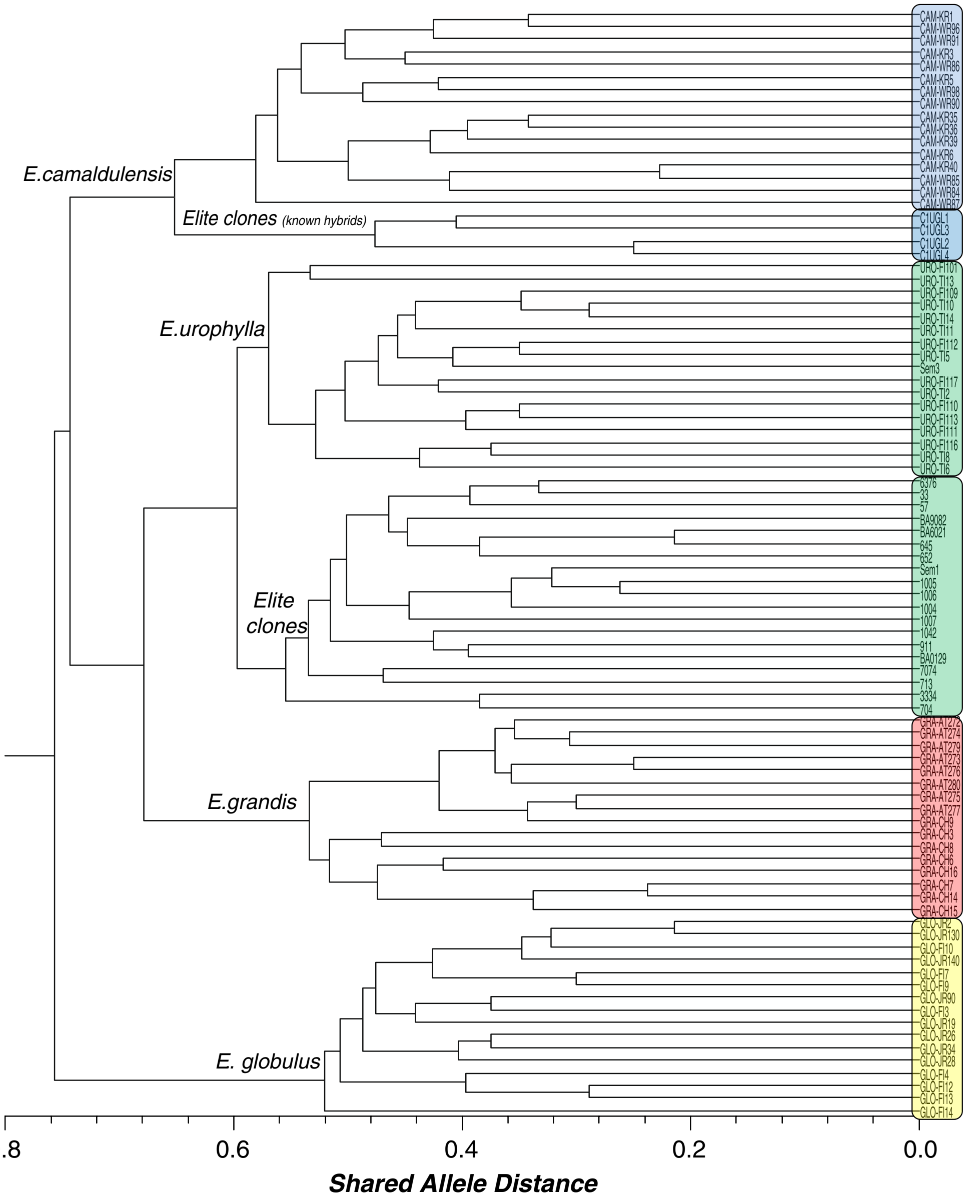

Fig. 1 UPGMA phenogram based on pairwise estimates of shared allele distance among all 88 individual *Eucalyptus* trees, 64 of them from eight provenances of four species (*CAM-KR E. camaldulensis* Kennedy River, *CAM-WR E. camaldulensis* Walsh River, *URO-FI E. urophylla* Flores Island, *URO-TI E. urophylla* Timor Island, *GLO-JR E. globulus* Jeeralang, *GLO-FI E. globulus* Flinders Island, *GRA-AT E. grandis* Atherton, *GRA-CH E. grandis* Coffs Harbor) and 24 elite clones identified by their usual codification. Species and elite clones clusters are indicated corresponding to the population structure analysis (see Fig. 2)

optimal number of clusters under an admixture model. Consistent with the peak of Evanno's ΔK at K=4 (Electronic supplementary materials (ESM) Fig. S1), the genetic structure of the species reference data set was partitioned into four groups (Fig. 2a). Each cluster correctly included all the individuals of each species. The estimates of the average proportion of membership (Q) of the predefined reference populations to clusters corresponding to the species were above 0.975 for all species, showing the very robust resolution possible using Bayesian analysis with data from these microsatellites (ESM Table S1). Estimates of F_{st} between species based on this set of microsatellites were also high and significant for all pairwise comparisons (ESM Table S2). By using prior population information regarding provenance origin and assuming K = 8 (four species versus two provenances per species), only the two provenances of *E. grandis* (Atherton and Coffs Harbor) could be separated (Fig. 2b). This result is consistent with the way that individuals clustered in the phenogram (Fig. 1) and indicates that very little detectable genetic variation with this set of microsatellites exists between the two sampled provenances of each species, possibly with the exception of *E. grandis*, although it is important to note that there is no statistical support for a structure analysis at K = 8 (ESM Fig. S1). Atherton and Coffs Harbor are the only two provenances separated at a large geographical scale, latitude-wise, by over 2,000 km. *E. globulus* provenances Jeeralang and Flinders Island are nearby populations at ~220 km, albeit the first at the southern tip of the continental coast of Australia and the second an island population to the south. Our results for *E. globulus* are thus consistent with the results of a range-wide

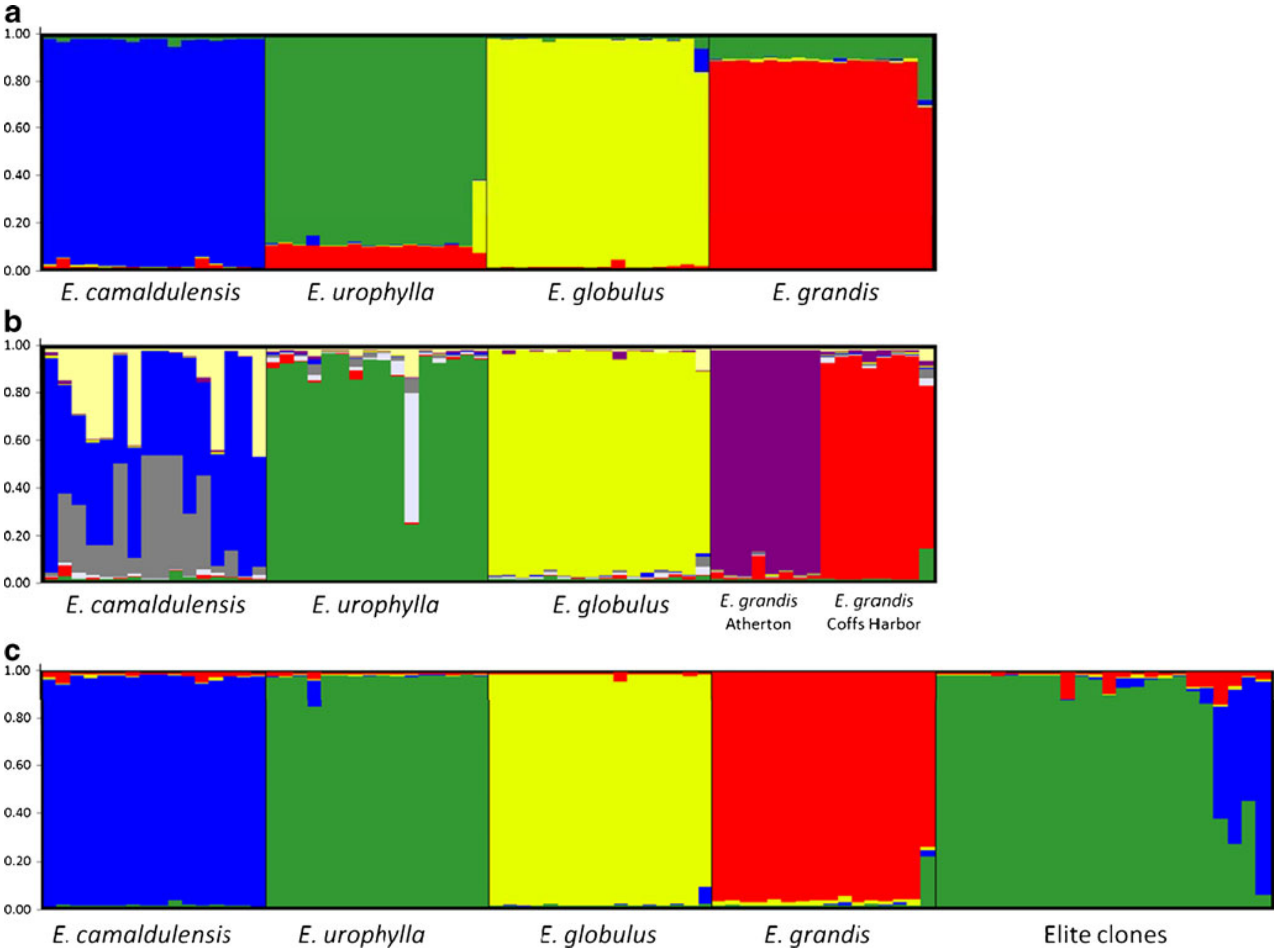

Fig. 2 Population analysis and assignment tests using STRUCTURE 2.1. Each individual is represented as a *vertical line* partitioned into *K* colored segments whose length is proportional to the individual coefficients of membership in each of the *K* inferred clusters that represent the four species of *Eucalyptus*. **a** Reference species sets using an admixture model with $\alpha = 0.0288$ for $K = 4$. **b** Reference species sets using prior population information and an admixture model with $K = 8$, i.e., four species versus two provenances per species, showing the separation of the two provenances for *E. grandis* but not of the other species. **c** Reference species sets and test individuals (elite clones) for $K = 4$ using prior population information for the species and an admixture model showing the assignment of the elite clones to the four species clusters. The last four clones to the *right* are known hybrids of *E. camaldulensis* × (*E. urophylla* × *E. globulus*) derived from controlled crosses

survey of population structure of *E. globulus* when a strong association was found between genetic similarity and geographic proximity (Steane et al. 2006). Similarly, for *E. urophylla* and *E. camaldulensis*, the two provenances within each species could not be separated, a result which is consistent with previous range-wide surveys that showed very little variation among provenances in *E. urophylla* (Payn et al. 2008) and *E. camaldulensis* (Butcher et al. 2002).

Microsatellite performance for individual assignment tests The reference data set was used to assign the test individuals (elite clones) to one of the four genetic clusters and estimate their most likely ancestral species composition (Fig. 2c and ESM Table S3). Assignments were performed using prior species information for the reference set under an admixture model, i.e., assuming that these elite clones have mixed ancestry. So, for example, there is a 99.04% posterior probability that clone Sem1 has recent ancestry in *E. urophylla* and only 0.22% in E. *globulus* and 88.56% that clone BA6021 has ancestry in *E. urophylla* and 10.87% in *E. grandis*. All 20 elite clones showed a very strong predominance of *E. urophylla* ancestry that varied between 87.18% and 99.31% (ESM Table S3). The four controlled hybrids of *E. camaldulensis* × (*E. urophylla* × *E. globulus*) showed the anticipated hybrid composition with a predominance of *E. camaldulensis* genome, although the relative proportions did not match the expectations, especially considering the small proportion of *E. globulus* contribution, theoretically expected at 25% but observed only at

1.2% for clone C1UGL3 up to 13.19% for clone C1UGL1 (ESM Table S3). This less than expected contribution of *E. globulus* in these controlled hybrids might be the result of a strong selection for adaptability to tropical environments that took place during the development of these elite clones which resulted in the preclusion of *E. globulus* genome. The predominance of *E. urophylla* in the group of 20 elite clones of unknown origin is somehow expected, although not at such high levels. *E. urophylla* was introduced in Brazil in the early 1970s and extensively used in hybrid combinations with *E. grandis* to provide higher levels of resistance to *Eucalyptus* canker caused by *Cryphonectria cubensis* (Alfenas et al. 1983; Heerden and Wingfield 2002). These hybrid clones have been anecdotally considered to be F1 hybrids of *E. grandis* × *E. urophylla*. This assignment analysis, however, performed with a set of microsatellites that provide a robust separation between these two species (ESM Fig. S1), is not consistent with this hypothesis. Potential explanations for this result include repeated backcrosses, both spontaneous and controlled to *E. urophylla,* coupled to a strong preferential selection for *E. urophylla* genome and possibly some level of misclassification of seed sources or individual trees during the breeding procedures. These results, taken together, indicate that the estimated proportions of ancestral genomic composition should be viewed as a pointer and not be unambiguously taken at face value. Furthermore, although these higher order repeat microsatellites display alleles with wide frequency differentials among the species and thus provided a clear distinction among them, the development of larger numbers of ancestry informative markers is warranted. Following intensive screening efforts, selected SNPs with contrasting allele frequencies among the populations under study have been found and used for assignment tests in admixed humans (Lins et al. 2010) and animal (Stephens et al. 2009) populations. The discovery of such SNPs in *Eucalyptus* will soon be possible by employing next-generation re-sequencing efforts of a few tens of individuals of each target species at a reasonable coverage and mapping these sequences on the forthcoming *Eucalyptus* reference genome (Grattapaglia and Kirst 2008). However, the much higher nucleotide diversity in *Eucalyptus* (Kulheim et al. 2009) when compared to humans and domestic animals may complicate the robustness of SNP genotyping assays. So, although in principle SNPs will be more powerful and automatable for assignment tests, the development of robustly assayable SNPs across *Eucalyptus* species will depend on screening several hundred SNPs whose flanking sequences are conserved enough across species to make the assay work consistently. This same approach could also be very valuable to develop microsatellites less prone to the occurrence of null alleles as they are used across species.

Conclusions

In summary, we have exploited existing resources of *Eucalyptus* ESTs to develop a fully operational set of microsatellite markers based on higher order repeats, still rare for plants in general. Although they are less variable than the existing *Eucalyptus* dinucleotide- and trinucleotide-based microsatellites, they provide a significant advantage from the practical standpoint for easier allele calling due to their larger allele size difference. Multiplexed systems with up to 18 microsatellites in two or three PCR reactions were proposed that supply very high resolution power in all four studied species. These systems will be particularly useful for clone fingerprinting and parentage testing purposes, applications that require consistent allele calling for comparative analysis across different points in time or laboratories. These markers were also shown to provide good resolution for individual identification, species distinction, and individual assignment tests for some of the main planted species worldwide. A comparison of the observed and expected results of the assignment tests indicate that the estimated proportions of ancestral composition of individuals should be viewed as reliable leads but not be taken as definitive genomic proportions. Due to their genic origin, the interspecific transferability and genetic information content of these microsatellites will likely extend to other phylogenetically close species within the same subgenus, further emphasizing their practical value for *Eucalyptus* genetics and breeding.

Acknowledgments This work was supported by the Brazilian Ministry of Science and Technology through FINEP grant 1755-01 and CNPq grant 520489/02-0 both within the Genolyptus project and EMBRAPA Macroprogram 3 project grant 03.07.01.001. DAF and EMCM were supported respectively by a CNPq post-doctoral and CAPES doctoral fellowships; GJPJr and DG have been awarded individual research fellowships from CNPq.

References

Alfenas AC, Jeng R, Hubbes M (1983) Virulence of *Cryphonectria cubensis* on *Eucalyptus* species differing in resistance. Eur J For Pathol 13:197–205

Bar W, Brinkmann B, Lincoln P, Mayr WR, Rossi U, Budowle B, Bell C, Carracedo A, Eisenberg A, Fourney R, Gill P, Kloosterman A, Monson K, Pascal O, Rand S, Robertson J, Vandaal A (1995) DNA recommendations—1994 Report Concerning Further Recommendations of the DNA Commission of the ISFH Regarding PCR-Based Polymorphisms in STR (Short Tandem Repeat) Systems. Vox Sang 69:70–71

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Bowcock AM, Ruizlinares A, Tomfohrde J, Minch E, Kidd JR, Cavallisforza LL (1994) High-resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Brondani RP, Grattapaglia D (2001) Cost-effective method to synthesize a fluorescent internal DNA standard for automated fragment sizing. Biotechniques 31:793–795, 798, 800

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 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 6:20

Burczyk J, Adams WT, Moran GF, Griffin AR (2002) Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. Mol Ecol 11:2379–2391

Butcher PA, Otero A, McDonald MW, Moran GF (2002) Nuclear RFLP variation in *Eucalyptus camaldulensis* Dehnh. from northern Australia. Heredity 88:402–412

Ceresini PC, Silva CLSP, Missio RF, Souza EC, Fischer CN, Guillherme IR, Gregorio I, da Silva EHT, Cicarelli RMB, da Silva MTA, Garcia JF, Avelar GA, Neto LRP, Marcon AR, Bacci M, Marini DC (2005) Satellyptus: analysis and database of microsatellites from ESTs of *Eucalyptus*. Genet Mol Biol 28:589–600

Chaix G, Gerber S, Razafimaharo V, Vigneron P, Verhaegen D, Hamon S (2003) Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. Theor Appl Genet 107:705–712

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046

Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. J Hered 100:106–113

Doughty RW (2000) The *Eucalyptus.* A natural and commercial history of the gum tree. The Johns Hopkins University Press, Baltimore and London

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

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform 1:47–50

Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2009) Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol Breed 23:85–97

Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S, Tomita M (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. FEBS Lett 554:17–22

Gaiotto FA, Bramucci M, Grattapaglia D (1997) Estimation of outcrossing rate in a breeding population of *Eucalyptus urophylla* with dominant RAPD and AFLP markers. Theor Appl Genet 95:842–849

Gill P, Kimpton C, Daloja E, Andersen JF, Bar W, Brinkmann B, Holgersson S, Johnsson V, Kloosterman AD, Lareu MV, Nellemann L, Pfitzinger H, Phillips CP, Schmitter H, Schneider PM, Stenersen M (1994) Report of the European DNA Profiling Group (Ednap)—towards standardization of short tandem repeat (Str) loci. Forensic Sci Int 65:51–59

Glaubitz JC, Emebiri LC, Moran GF (2001) Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. Genome 44:1041–1045

Grattapaglia D, Kirst M (2008) *Eucalyptus* applied genomics: from gene sequences to breeding tools. New Phytol 179:911–929

Grattapaglia D, Alfenas AC, Coelho ASG, Bearzoti E, Pappas GJ, Pasquali G, Pereira G, Colodette J, Gomide JL, Bueno J, Cascardo JC, Brondani RPV, Brommonschenkel SH (2004a) Building resources for molecular breeding of *Eucalyptus*. In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M, Tomé M (eds) International IUFRO Conference: *Eucalyptus* in a changing world. RAIZ, Instituto Investigação da Floresta e Papel, Portugal, Aveiro Portugal, pp 20–32

Grattapaglia D, Ribeiro VJ, Rezende GD (2004b) Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for *Eucalyptus*. Theor Appl Genet 109:192–199

Heerden SW, Wingfield MJ (2002) Effect of environment on the response of *Eucalyptus* clones to inoculation with *Cryphonectria cubensis*. For Pathol 32:395–402

Holt CL, Buoncristiani M, Wallin JM, Nguyen T, Lazaruk KD, Walsh PS (2002) TWGDAM validation of AmpFlSTR (TM) PCR amplification kits for forensic DNA casework. J Forensic Sci 47:66–96

Honjo M, Ueno S, Tsumura Y, Handa T, Washitani I, Ohsawa R (2008) Tracing the origins of stocks of the endangered species *Primula sieboldii* using nuclear microsatellites and chloroplast DNA. Conserv Genet 9:1139–1147

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106

Kirst M, Cordeiro CM, Rezende GD, Grattapaglia D (2005) Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations. J Hered 96:161–166

Kolpakov R, Bana G, Kucherov G (2003) MREPS: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31:3672–3678

Koskinen MT (2003) Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. Anim Genet 34:297–301

Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ, Tomsey CS, Zachetti JM, Masibay A, Rabbach DR, Amiott EA, Sprecher CJ (2002) Validation of a 16-locus fluorescent multiplex system. J Forensic Sci 47:773–785

Kulheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. BMC Genomics 10:452

Kumar P, Freeman AR, Loftus RT, Gaillard C, Fuller DQ, Bradley DG (2003) Admixture analysis of South Asian cattle. Heredity 91:43–50

Ladiges PY, Udovicic F, Nelson G (2003) Australian biogeographical connections and the phylogeny of large genera in the plant family *Myrtaceae*. J Biogeogr 30:989–998

Lins TC, Vieira RG, Abreu BS, Grattapaglia D, Pereira RW (2010) Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. Am J Hum Biol 22:187–192

Litt M, Hauge X, Sharma V (1993) Shadow bands seen when typing polymorphic dinucleotide repeats—some causes and cures. Biotechniques 15:280–284

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Millar MA, Byrne M, Nuberg I, Sedgley M (2008) A rapid PCR-based diagnostic test for the identification of subspecies of *Acacia saligna*. Tree Genet & Genomes 4:625–635

Minch E, Ruíz-Linares A, Goldstein DB, Feldman M, Cavalli-Sforza LL (1995) MICROSAT—the Microsatellite Distance Program. Stanford University Press, Stanford

Missiaggia AA, Piacezzi AL, Grattapaglia D (2005) Genetic mapping of *Eef1*, a major effect QTL for early flowering in *Eucalyptus grandis*. Tree Genet & Genomes 1:79–84

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Muir G, Schlotterer C (2005) Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). Mol Ecol 14:549–561

Myburg AA, Potts BM, Marques CM, Kirst M, Gion JM, Grattapaglia D, Grima-Pettenati J (2007) *Eucalyptus*. In: Kole C (ed) Genome mapping and molecular breeding in plants. Springer, New York, pp 115–160

Ottewell KM, Donnellan SC, Moran GF, Paton DC (2005) Multiplexed microsatellite markers for the genetic analysis of *Eucalyptus leucoxylon* (*Myrtaceae*) and their utility for ecological and breeding studies in other *Eucalyptus* species. J Hered 96:445–451

Parida SK, Dalal V, Singh AK, Singh NK, Mohapatra T (2009) Genic non-coding microsatellites in the rice genome: characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. BMC Genomics 10:140

Payn KG, Dvorak WS, Janse BJH, Myburg AA (2008) Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia. Tree Genet & Genomes 4:519–530

Potts BM (2004) Genetic improvement of eucalypts. In: Burley J, Evans J, Youngquist JA (eds) Encyclopedia of forest science. Elsevier Science, Oxford, pp 1480–1490

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rabello E, de Souza AN, Saito D, Tsai SM (2005) In silico characterization of microsatellites in *Eucalyptus* spp.: abundance, length variation and transposon associations. Genet Mol Biol 28:582–588

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. Trends Biotechnol 25:490–498

Sampson JF, Byrne M (2008) Outcrossing between an agroforestry plantation and remnant native populations of *Eucalyptus loxophleba*. Mol Ecol 17:2769–2781

Sarri V, Baldoni L, Porceddu A, Cultrera NGM, Contento A, Frediani M, Belaj A, Trujillo I, Cionini PG (2006) Microsatellite markers are powerful tools for discriminating among olive cultivars and assigning them to geographically defined populations. Genome 49:1606–1615

Steane DA, Vaillancourt RE, Russell J, Powell W, Marshall D, Potts BM (2001) Development and characterisation of microsatellite loci in *Eucalyptus globulus* (*Myrtaceae*). Silvae Genet 50:89–91

Steane DA, Conod N, Jones RC, Vaillancourt RE, Potts BM (2006) A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (*Myrtaceae*), using microsatellite markers and quantitative traits. Tree Genet & Genomes 2:30–38

Stephens MR, Clipperton NW, May B (2009) Subspecies-informative SNP assays for evaluating introgression between native golden trout and introduced rainbow trout. Molecular Ecology Resources 9(1):339–343

Tadano R, Nishibori M, Tsudzuki M (2008) High accuracy of genetic discrimination among chicken lines obtained through an individual assignment test. Anim Genet 39:567–571

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques 37:226–231

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WF, Smith JSC, Doebley J (2002) Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol 19:1251–1260

Yasodha R, Sumathi R, Chezhian P, Kavitha S, Ghosh M (2008) *Eucalyptus* microsatellites mined in silico: survey and evaluation. J Genet 87:21–25

Yi GB, Lee JM, Lee S, Choi D, Kim BD (2006) Exploitation of pepper EST-SSRs and an SSR-based linkage map. Theor Appl Genet 114:113–130

Zhang LD, Yuan DJ, Yu SW, Li ZG, Cao YF, Miao ZQ, Qian HM, Tang KX (2004) Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. Bioinformatics 20:1081–1086