Journal of Heredity 2010:101(4):512–520
doi:10.1093/jhered/esq024
Advance Access publication March 15, 2010

© The American Genetic Association. 2010. All rights reserved.
For permissions, please email: journals.permissions@oxfordjournals.org.

A Selected Set of EST-Derived Microsatellites, Polymorphic and Transferable across 6 Species of Eucalyptus

Danielle A. Faria, Eva Maria Celia Mamani, Marilia R. Pappas, Georgios Joannis Pappas Jr, and Dario Grattapaglia

From the Plant Genetics Laboratory, EMBRAPA—Genetic Resources and Biotechnology, PqEB, Brasília, 70770-970 DF, Brazil (Faria, Mamani, Pappas, Pappas Jr., Grattapaglia); Graduate Program in Genomic Sciences and Biotechnology, Universidade Católica de Brasília, Brasília, DF, Brazil (Faria, Pappas Jr., Grattapaglia); and Department of Cell Biology, Universidade de Brasília—UnB, Brasília, DF, Brazil (Mamani, Grattapaglia).

Address correspondence to Dario Grattapaglia at the address above, or e-mail: dario@cenargen.embrapa.br.

Species of *Eucalyptus* are keystone species for ecological studies in their natural ranges and are extensively planted in the tropical and subtropical regions of the world to supply high-quality woody biomass for various applications. We report the development of a selected set of 20 dinucleotide and trinucleotide repeat microsatellites derived from *Eucalyptus* expressed sequence tags (ESTs). These microsatellites were selected for full transferability and homogeneous rate of polymorphism across species. They were evaluated for individual fingerprinting, parentage testing, and intraspecific population structure analyses in 6 of the most extensively studied and planted species worldwide, representing key phylogenetic sections of the largest subgenus *Symphyomyrtus*. This set of markers provides exceptional resolution for population genetics and molecular breeding applications in the genus *Eucalyptus*. As they were developed from conserved transcribed regions, the transferability and polymorphism of these microsatellites will most likely extend to the other 300 or more species within the same subgenus.

Key words: *Eucalyptus, EST-SSR, tropical trees*

The genus *Eucalyptus* includes over 700 species, some of which are the most widely planted hardwoods worldwide (Potts 2004). The largest *Eucalyptus* subgenus, *Symphyomyrtus*, contains over 300 species. Three sections of this subgenus include the vast majority of the commercially planted species, such as *Eucalyptus grandis* and *Eucalyptus urophylla* (section *Latoangulatae*), *Eucalyptus globulus* (section *Maidenaria*), and *Eucalyptus camaldulensis* (section *Exsertaria*). Eucalypts are currently planted in more than 90 countries for pulp production, energy supply in various forms, sawn timber, essential oils, firewood, shade, and shelter. *Eucalyptus globulus* has been the top choice for plantations in temperate regions, providing a good combination of growth and wood properties. Tropical *Eucalyptus* forestry, on the other hand, is based on interspecific hybrid breeding and clonal propagation with *E. grandis* as the pivotal fast growing species. Traits such as growth, adaptability, disease resistance, and superior wood properties are combined into elite clones for large-scale plantation (Myburg et al. 2007).

The multiallelism and Mendelian inheritance of microsatellite markers have provided a powerful system for population genetics studies, individual identification, and parentage testing in *Eucalyptus* (Steane et al. 2001; Chaix et al. 2003; Grattapaglia, Ribeiro, and Rezende 2004; Kirst et al. 2005; Ottewell et al. 2005; Jones et al. 2008). Relatively large collections of microsatellite markers have been developed for species of *Eucalyptus* based on screening genomic libraries enriched for simple sequence repeats (SSRs) (Brondani et al. 1998, 2006; Glaubitz et al. 2001; Ottewell et al. 2005). However, all the currently available microsatellites are derived from genomic sequences that typically show unpredictable transferability across species with rates of genotyping dropout around 10–50% when transferred outside subgenera (Byrne et al. 1996; Kirst et al. 1997; Ottewell et al. 2005). This is likely due to the high nucleotide diversity found in species of *Eucalyptus*, estimated between 0.5% and 1% in transcribed regions (Poke et al. 2003; Grattapaglia and Kirst 2008; Novaes et al. 2008) and up to 3–5% when very large samples of individuals and entire genes were resequenced (Kulheim et al. 2009).

The availability of fully transferable and polymorphic microsatellites across the most extensively studied and

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

planted species of *Eucalyptus* has become an increasingly important aspect for both comparative population genetics studies and molecular breeding applications especially with the increasing interest in exploiting hybrid combinations. In contrast to microsatellite markers derived from random genomic sequences, those developed from transcribed regions typically display superior robustness, better allele resolution, and higher interspecific and intergeneric transferability, merits that significantly enhance their utility (Varshney et al. 2005). With a nucleotide diversity estimated at approximately 0.5% (1 single nucleotide polymorphism (SNP) every 192 bp) in transcribed portions of the genome (Novaes et al. 2008), it is expected that a higher interspecific conservation of sequences flanking genic microsatellites will allow the development of more robust markers. Besides facilitating the analytical routine of laboratories that work with a group of related species or genera, these markers can often be used as anchor markers for comparative mapping, population and evolutionary studies. In this study, we report the development and detailed characterization of a selected set of 20 microsatellite markers based on dinucleotide and trinucleotide repeats derived from a large collection of ESTs of *Eucalyptus*. We targeted 6 broadly planted species worldwide encompassing 3 key sections of *Symphyomyrtus* from the breeding and plantation perspective to select markers with full interspecific transferability and high genetic information content.

Materials and Methods

EST Database Mining and Primer Design

Dinucleotide and trinucleotide repeats were mined in a database that had roughly 88 000 phred-20 filtered, 5′-sequenced ESTs generated during a sequencing effort in the Genolyptus project (Grattapaglia 2004; Grattapaglia, Alfenas, et al. 2004). EST sequences from leaf and developing xylem RNA were mostly from *E. grandis* although approximately 30% were derived from 3 other species: *E. urophylla*, *Eucalyptus pellita*, and *E. globulus*. With an optimized microsatellite pipeline based on the software MREPS (Kolpakov et al. 2003), SSRs were identified under the following parameters: 2–6 bases SSR motifs, perfect structure, that is, no microvariant interruptions, and a minimum core length of 12 bp. The microsatellite markers were derived from the alignment of several ESTs from the 4 species using *E. grandis* as the reference sequence. Primer pairs flanking these microsatellites were designed targeting expected polymerase chain reaction (PCR) products between 80 and 450 bp, that is, within an adequate detection size range for analysis in automatic sequencers.

Microsatellite Marker Selection and Preliminary Screening

No selection was practiced regarding the potential location of the microsatellite in the expressed sequence, base composition of the motif, or basic alignment search tool hit identity. Priority was given to longer microsatellites, that is, with a larger number of repeated units based on the assumption that these would likely be more ancient and thus more polymorphic in the different species. Initial screening of primer pairs for amplification, polymorphism, and interspecific transferability was carried out in a panel of 12 unrelated trees involving the 6 target species of *Eucalyptus* (*E. urophylla, E. grandis, E. globulus*, *Eucalyptus calmadulensis*, *Eucalyptus dunnii*, and *Eucalyptus saligna*). Regular primers at small scale were synthesized (AlphaDNA, Montreal, CA) 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 staining were used for PCR-product visualization. Microsatellite markers were classified as transferable when amplification was observed in all 6 species and tentatively polymorphic when at least 1 difference in product size was observed among the individuals in the screening panel.

Microsatellite Genotyping by Fluorescence Detection

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), equal concentration (0.1 μM) of all primers for all microsatellite markers coamplified, 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 pentaplex or hexaplex 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), VIC or HEX (green). An aliquot of 1 μl of PCR mixture was mixed with 1 μl of ROX-labeled size standard (Brondani and Grattapaglia 2001) and 10 μl of Hi-Di formamide (Applied Biosystems, Foster City, CA). 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

A population sample of 16 unrelated trees of each 1 of the 6 target species, *E. grandis*, *E. saligna*, and *E. urophylla* (section *Latoangulatae*); *E. globulus* and *E. dunnii* (section *Maidenaria*); and *E. camaldulensis* (section *Exsertaria*) were used to select for interspecific transferability and to carry out microsatellite characterization. For each species, individuals from a single provenance were sampled, Coffs Harbor (30°18′S, 153°07′E) for *E. grandis*, Clouds Creek (30°05′S, 152°37′E) for *E. saligna*, Jeeralang (38°24′S, 146°28′E) for *E. globulus*,

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

Timor Island (9°37′S, 124°10′E) for *E. urophylla*, and Walsh River (17°17′S, 144°88′E) for *E. camaldulensis*. These provenances have been among the main ones used in breeding programs in Brazil. The following parameters of genetic information content were estimated for the microsatellite markers for each species separately: 1) number of alleles (A); 2) allele size range; 3) observed (H_o) and expected (H_e) Heterozygosity and *P* value of an exact test for Hardy–Weinberg Equilibrium (HWE); 4) polymorphism information content (PIC) (Botstein et al. 1980); 5) probability of identity (PI) that corresponds to the probability of 2 unrelated 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) PE 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) PE for a candidate parent pair, a common situation when paternity and maternity needs to be checked in progeny individuals derived from controlled crosses, that is, 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) to carry out an exact test for HWE for each microsatellite marker. Considering that *Eucalyptus* species are known for operating largely under a mixed mating model (Gaiotto et al. 1997; Burczyk et al. 2002), we estimated the frequency of null alleles at the 20 loci in the 6 species using an individual inbreeding model (IIM) 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. Finally, the combined multilocus PEs and PI were also estimated for the combined sets of dinucleotide and trinucleotide microsatellites to provide an assessment of the use of each marker set for genotyping applications.

Evaluation of Microsatellites for Genetic Analyses

Multilocus genotypes for the 20 microsatellites were used to estimate genetic distances among the 96 individual trees. Considering the heterozygous nature of the genotype data the "Shared Allele Distance" estimator between individuals was used (Chakraborty and Jin 1993) implemented by the online calculator (http://www.biology.ualberta.ca/jbrzusto/sharedst.php). The shared allele distance between any 2 individuals is defined as one minus half the average number of shared alleles per locus. The matrix of genetic distances was then used to graphically represent the distance relationships between the 96 individuals with a UPGMA (unweighted pair group method with arithmetic mean) phenogram constructed using the NTSYS 2.0 package (Exeter Software, United States). To evaluate the selected microsatellites for species differentiation and intraspecies population structure analyses, individual trees were assigned probabilistically to a given number of clusters inferred with a Bayesian approach implemented by the STRUCTURE software (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 3 to 12 inferred clusters performed with 5 independent runs each. The model choice criterion to detect the most probable value of *K* was ΔK (Evanno et al. 2005). Results were entered into DISTRUCT (Rosenberg 2004) to provide a graphic display of population structure.

Results and Discussion

Microsatellite Development

The analytical pipeline used for microsatellite marker development resulted in 1261 potential microsatellite markers that met the specified constraints and for which primer pairs were designed. Of the set of 1261 designed primer pairs, 305 were targeting dinucleotide and 690 trinucleotide repeat microsatellites. Preliminary marker screening for amplification success and polymorphism detection was carried out for 287 dinucleotide and 375 trinucleotide repeat microsatellites that were selected for having the largest number of tandemly repeated units in silico. From the preliminary screening, 91 dinucleotide and 77 trinucleotide (total 168) microsatellite showed robust amplification across the 6 species and clear indication of polymorphism. These were selected for high-resolution fluorescence-based screening. As the objective of this study was to select a set of polymorphic and transferable microsatellites across the 6 target species, a relatively stringent threshold was set to maximize the number of alleles and PIC in all 6 species simultaneously. To streamline this microsatellite selection process, data from a parallel mapping study were used where all the 168 microsatellites were evaluated in *E. grandis* and *E. globulus,* the 2 phylogenetically most contrasting species out of the 6 under study, so that polymorphism found in these 2 would most likely reveal polymorphism in the additional species (Faria DA, unpublished data). Markers that displayed at least 4 alleles and $H_o > 0.5$ in both species were selected. Full information is presented for these 20 selected microsatellites 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 dbSTS id (Table 1). All these microsatellites were located in the 5′ untranslated region.

Microsatellite Characterization

The 20 microsatellites spanned a wide range of allele sizes (Supplementary Material Tables S1 and S2). The size range of the alleles for all loci but one matched the expected size of the in silico predicted amplicon. For marker EMBRA979, the observed allele range for 5 of the 6 species was

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

Table 1 Basic properties of the 20 microsatellite markers developed in this study

Marker Locus	SSR type	Motif	Expected Amplicon Size (bp)	Forward primer (5′–3′)	Reverse primer (5′–3′)	GenBank Accession #	dbSTS_Id
EMBRA928	di	$(AG)_{18}$	143	ACGATGAAGATGGGTTCTGC	CACCACCACTCCCATCTCTT	GF101874	1232955
EMBRA949	di	$(CT)_{32}$	284	CGTCCGCTCCAGTTCAAAAT	ACTTGCCGTACCAGAGGATG	GF101875	1232956
EMBRA979	di	$(CT)_{34}$	267	GGCATTAAGAAGCCCATTGA	TGGGCTTCATCTTCATCACA	GF101876	1232957
EMBRA1284	di	$(TC)_{22}$	145	GATTCAGCAAGAAAGCTGGC	GGGGAAAGAATATTATTGCACTTG	GF101878	1232959
EMBRA1445	di	$(AG)_{13}$	126	ATTGAGGGAAAAACACAGCG	CGCCTCTCTCTGCTTTTGAT	GF101901	1232982
EMBRA1468	di	$(TC)_{24}$	181	CCTCTTCTTTGTTTCTGCGG	TGACCACACAGGACACTCGT	GF101903	1232984
EMBRA1868	di	$(TC)_{36}$	297	TGTGGAGCATGGAGTAGCAG	CAAATCTCAGAGACGCCACA	GF101905	1232986
EMBRA1924	di	$(TC)_{19}$	335	TCATAAAATAAAGAAAAATATGAACCG	GAGGGGGTTGGGAATTGTAT	GF101890	1232971
EMBRA1928	di	$(AG)_{42}$	333	GGACGAAGCTGGAGAAGTTG	TCCATCTGATCACCACCAAA	GF101891	1232972
EMBRA1990	di	$(AG)_{18}$	103	CCGCTCACTTCAGACAAGC	CGATTATCACCATCCCCATC	GF101908	1232989
EMBRA878	tri	$(CGG)_{24}$	219	GAGAGCTTCCGAGGAGGAAT	TATAATCCTCGGACTTGGCG	GF101952	1233033
EMBRA904	tri	$(GAA)_{27}$	175	ACAGAGCGAGCGAGAAAGAG	GTGCAAACAACGGACTCAAA	GF101957	1233038
EMBRA914	tri	$(GCT)_{31}$	139	GCGCTTCTGAAGATTGAACC	CTTCCCTCAGAGTCACCTGC	GF101900	1232981
EMBRA975	tri	$(GGA)_{14}$	324	TCGATCCTTTGCTGCTCTCT	AGAAGGCGAAGAGGTTAGGG	GF101953	1233034
EMBRA1135	tri	$(GGA)_{13}$	177	GGAATGAGCAGACTGGCAA	GGAATGAGCAGACTGGCAA	GF101954	1233035
EMBRA1332	tri	$(GAA)_{23}$	104	TGAGGTGCTGGTTGATCTTG	GCGATCTGTCGTCTTCATCA	GF101880	1232961
EMBRA1363	tri	$(GCC)_{15}$	317	CCATAGCCCTCTGCTGATTC	AATGGAAAATGGGTTCCTCC	GF101882	1232963
EMBRA1382	tri	$(TGC)_{12}$	263	GCAGTCGCAGATGTTGAAGA	ATCCGAAAGAAAGCCCAAAT	GF101884	1232965
EMBRA1451	tri	$(TCC)_{17}$	293	GCGTACTTGAAGATCCGCTC	AGACGCATCACTAGCGGAAG	GF101902	1232983
EMBRA2002	tri	$(CCA)_{26}$	263	CGTGATACCCGTGATGTACG	ATCAAACCCTGAAGCACCAC	GF101907	1232988

significantly larger than the expected one, suggesting the amplification across intronic sequences in all species but *E. camaldulensis*. Overall, the allele size ranges for each locus did not vary much across the 6 species for both dinucleotide and trinucleotide microsatellites (Supplementary Material Tables S1 and S2). This fact significantly facilitates the design of multiplexed amplification and analysis systems that would be usable across all 6 species.

A clear distinction in the rate of polymorphism and consequently the genetic information content was seen between the set of dinucleotide and trinucleotide microsatellites. As expected from the higher mutation rate of shorter sequence repeats (Chakraborty et al. 1997), the dinucleotide repeat microsatellites are more hypervariable than the trinucleotides, a commonly seen phenomenon in plant species (Vigouroux et al. 2002) with almost twice the overall average number of alleles across species and across loci ($t = 5.18$; $P = 0.00015$). When these *Eucalyptus* EST–derived microsatellites are compared with those derived from nongenic sequences from enriched genomic libraries, the average number of alleles and the observed and expected heterozygosities are significantly lower for the dinucleotides and even more so for the trinucleotide microsatellites. For example, although the overall average number of alleles for the 10 dinucleotide repeat microsatellites for all 6 species was 8.3 and 5.5 for the trinucleotide (Supplementary Material Tables S1 and S2), the average number of allele for dinucleotide microsatellites from enriched libraries was 14.3 (Brondani et al. 2006) or up to 16.3 for an earlier selected set of 20 loci (Brondani et al. 1998). The average observed and expected heterozygosities were higher in dinucleotide than trinucleotide microsatellites developed in this work. Interestingly, however, they were in the same range when comparing dinucleotides from ESTs (this work) ($H_o = 0.689$; $H_e = 0.803$) and a selected set of 20 microsatellites from enriched genomic libraries ($H_o = 0.570$; $H_e = 0.840$) (Brondani et al. 1998), indicating that the larger number of alleles found in the genomic-derived microsatellites is mostly due to rare alleles that have a very slight contribution to the total final observed heterozygosity and by consequence to relevant parameters for mapping purposes such as PIC or individual identification and parentage analysis such as PE and PI. Comparative analyses of rates of polymorphism are evidently always subject to ascertainment bias and should be taken with due caution. However, our results suggest that EST-derived microsatellites for the same type of repeat size are in fact as polymorphic as those derived from random genomic sequences with the added advantage of being far more transferable across species due to the conservation of the genic-flanking sequences onto which primers are designed.

To provide an overview of the genetic information content of the markers developed, average values of A, H_o, H_e, PIC, PE, and PI are presented for each marker across the 6 species and for each species averaging for all markers in each repeat category (Supplementary Material Tables S1 and S2). Looking across species for the dinucleotide microsatellites, a larger number of alleles and higher

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

information content was seen for *E. grandis* and *E. saligna* (averaging 9.8 and 9.6 alleles per locus) as compared with *E. urophylla* and *E. dunnii* (averaging 6.7 and 7.0 alleles/locus) ($t = 2.04$; $P = 0.0036$) in line with the significantly larger geographical distribution of the 2 former species (Doughty 2000) and possibly also derived from the fact that most ESTs mined for marker development were from *E. grandis,* with *E. saligna* being a sympatric and phylogenetically close species to *E. grandis*.

No significant ranking pattern emerged for the microsatellites in terms of the number of alleles and information content across species. An analysis of variance on the number of alleles revealed that no significant difference exists among species across dinucleotide microsatellites ($F = 2.21$; $P = 0.07$), whereas the difference among microsatellites across species was borderline significant ($F = 2.17$; $P = 0.042$). For the trinucleotide microsatellites again no significant difference exists among species ($F = 1.56$; $P = 0.194$) and also none among these microsatellites ($F = 0.73$; $P = 0.674$). To provide a guideline in terms of choosing a subset of most informative markers across all species, for dinucleotide microsatellites, we applied a threshold of a minimum of 6 alleles in every species and a minimum variance of allele number across species. Markers EMBRA1468, EMBRA1928, EMBRA1284, and EMBRA1990 are the 4 more consistently polymorphic across the 6 species. For the trinucleotides, with a threshold of a minimum of 4 alleles, EMBRA914, EMBRA975, EMBRA1135, and EMBRA1363 were identified as being the 4 best markers. If the objective, however, is to maximize resolution for population genetics studies or individual fingerprinting in any particular species, slightly different sets of microsatellites could be selected although these 8 markers are likely to always be a good choice.

In general, the estimates of expected heterozygosities (H_e) in all species were nominally larger than the observed heterozygosities (H_o). However of the 120 exact tests for HWE genotype proportions, 21 were significant at $\alpha = 0.05$ and 11 of these 21 at $\alpha = 0.01$ (Supplementary Material Tables S1 and S2). A concentration of markers deviating from HWE were seen for *E. urophylla* (8 loci), *E. saligna* (7 loci), and *E. camaldulensis* (4 loci) due to a deficiency of heterozygotes in all but one case (EMBRA928 in *E. saligna*). This could be due to actual inbreeding, sampling effects, or an increased incidence of null alleles in these species causing the observation of apparent homozygous genotypes due to the presence of frequent sequence polymorphisms (SNPs and/or indels) in the priming sites for these loci (Dakin and Avise 2004). Evidence in favor of this last hypothesis comes from the observation of a coincidence between significant deviation from HWE and an increased incidence of missing data (up to 40%) even after replicated genotyping attempts for some loci such as EMBRA949, EMBRA904, EMBRA975, and EMBRA1363. To provide additional evidence in favor of this hypothesis, we estimated the frequency of null alleles for all 20 loci in all 6 species under the IIM (Supplementary Material Tables S1 and S2). This model is adequate for species with mixed mating system as it provides 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). The estimated null allele frequency was considered significantly different from zero whenever $P < 0.01$ for the HWE test. We found that a relatively high frequency of null allele (between 0.1 and 0.3) was estimated in almost all cases where a significant deviation from HWE was observed due to an excess of homozygotes. Furthermore, the probability (β) of random amplification failure estimated for each locus × species combination (data not shown) was always lower than 1% indicating that the data are not consistent with random amplification failure but rather with the occurrence of true null alleles. Markers that displayed null allele frequency significantly different from zero were EMBRA979, EMBRA1924, and EMBRA975 in *E. camaldulensis*; EMBRA1284, EMBRA1924, EMBRA904, EMBRA975, and EMBRA1363 in *E. urophylla*; EMBRA1363 in *E. dunnii*; and EMBRA928, EMBRA949, and EMBRA904 in *E. saligna*. For these microsatellites, primers could be redesigned to try alternative primer-flanking sequences and possibly avoid frequently polymorphic sites. However, in Eucalyptus, such a task might be challenging due to the very high nucleotide diversity of 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 resequencing survey of 23 genes (Kulheim et al. 2009).

Microsatellite Performance for Genetic Analysis

Following the relatively homogenous distribution of the number of alleles across species for all markers, the average estimates for the PIC, PEs (PE_1 and PE_2) and PI were consistent across the 6 species and within the same range for most loci. The top estimates were obtained for *E. grandis* and the phylogenetically close species *E. saligna* followed by *E. globulus*. This might be due to the fact that most of the ESTs used for microsatellite discovery were from *E. grandis* (~50%) and *E. globulus* (~25%). For dinucleotide repeat microsatellites, these parameters were in the range of 0.678–0.828 for PIC, 0.322–0.536 for PE_1, 0.511–0.655 for PE_2, and 0.098–0.131 for PI, with *E. grandis* as the species with the highest values and *E. urophylla* with the lowest (Supplementary Material Table S1). For trinucleotide repeats, the range of these parameters were 0.571–0.672 for PIC, 0.262–0.340 for PE_1, 0.384–0.489 for PE_2, and 0.128–0.246 for PI (Supplementary Material Table S2). Interestingly, although *E. urophylla* ranked last for average information content for the dinucleotide repeat microsatellites, this was not the case for trinucleotides where *E. dunnii* was the species displaying the lowest average number of alleles and lowest average estimates of discrimination power. As expected from the estimated average number of alleles per locus, the dinucleotide microsatellites are more powerful when it comes to parentage and individual identification than trinucleotide repeat microsatellites. The combined probabilities of parentage exclusion provide a better assessment of the combined use of these

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

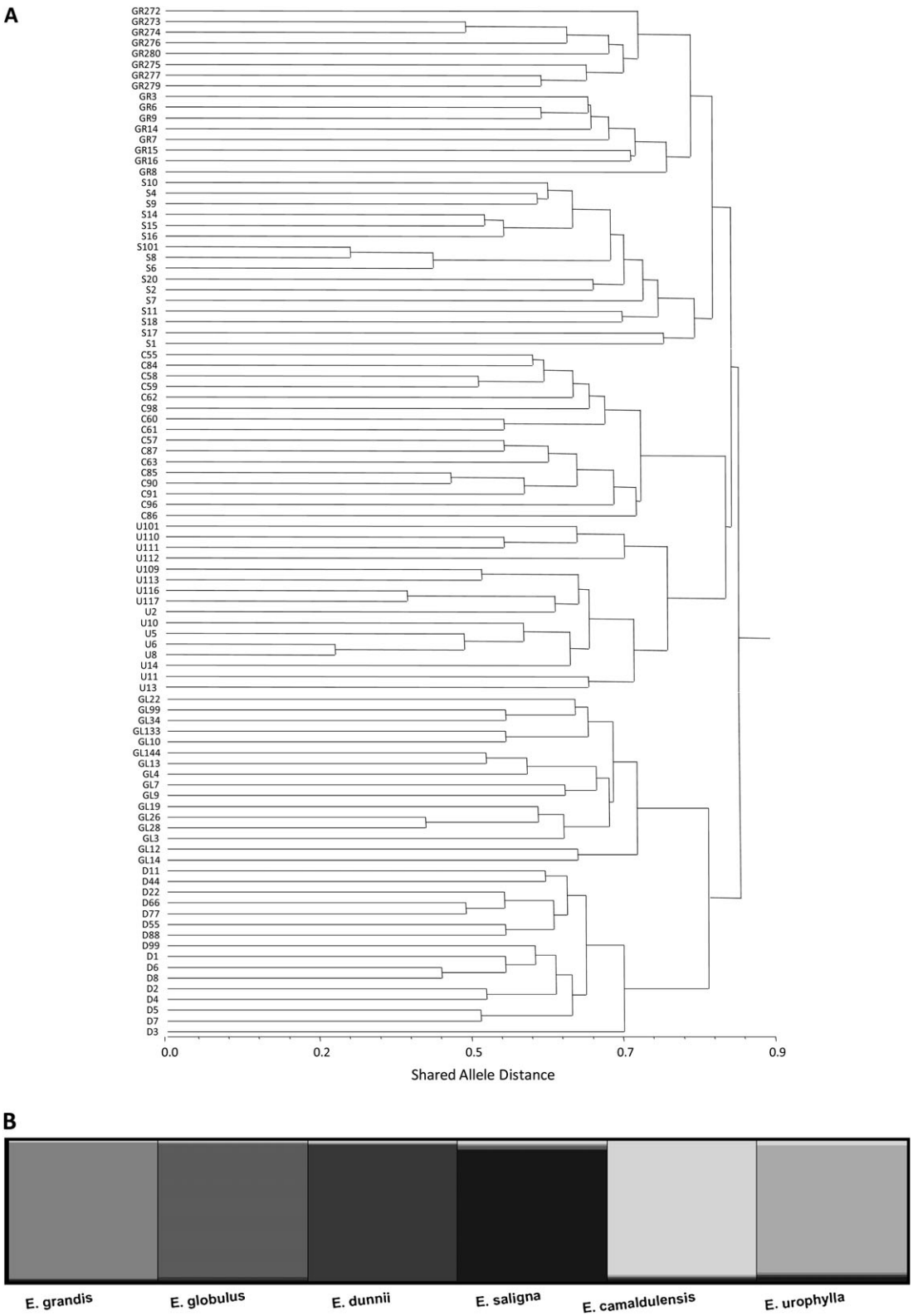

Figure 1. (**A**) UPGMA phenogram based on pairwise estimates of shared allele distances between all 96 individual trees (GR = *Eucalyptus grandis*; S = *Eucalyptus saligna*; C = *Eucalyptus camaldulensis*; U = *Eucalyptus urophylla*; GL = *Eucalyptus globulus*; and D = *Eucalyptus dunnii*). (**B**) Population analysis using STRUCTURE showing the correct assignment of all individuals to their respective species.

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

markers. For dinucleotide microsatellites, they ranged from a highest PE_1 = 99.962% and PE_2 = 99.998% for *E. grandis* to the lowest PE_1 = 98.312% and PE_2 = 99.956% for *E. urophylla*. For trinucleotides from a highest PE_1 = 98.633% and PE_2 = 99.860% for *E. globulus* to the lowest PE_1 = 96.807% and PE_2 = 99.892% for *E. urophylla*. By combining all 20 loci, these combined estimates always exceed 99.999% for both parentage-testing situations. For fingerprinting purposes, the combined estimates of PI are less than 10^{-12} for dinucleotides, except for *E. urophylla* and less than 10^{-9} for trinucleotides, except for *E. dunnii*. In practical terms, a PI of 10^{-9} means that the likelihood of 2 genetically unrelated trees displaying the same multilocus genotype is 1 in 1 billion. These results show that these microsatellites can be used with high efficiency in applications that require individual identification or parentage testing. Although dinucleotides provide a higher discrimination power, the trinucleotides have an intrinsic advantage when it comes to the precision of the allele size determination due to a larger allele size difference of 3 base pairs. It should be pointed out, however, that all these EST-derived microsatellites were also selected based on the fact that they supply higher quality allelic profiles with minimal stuttering and rare duplication when compared with loci that we previously developed from random genomic sequences (Brondani et al. 2006). In this respect, we also noted that, particularly for the trinucleotides, the relatively smaller number of alleles per locus has an advantage when it comes to the ability of multiplexing the loci in single electrophoretic runs. Narrower allele size ranges allow fitting more markers in the same fluorescence detection spectrum when designing and testing multiplexed amplification and detection systems.

A preliminary evaluation of the power of these microsatellites for genetic distance and population structure analyses showed that this set of microsatellites provides very high resolution for individual and species discrimination. The clustering observed is consistent with the known phylogenetic relationships and can be explained by the strong association typically found in *Eucalyptus* between genetic similarity and geographic proximity of species and populations (Butcher et al. 2002; Steane et al. 2006; Payn et al. 2008). *Eucalyptus grandis* and *E. saligna* belong to the same section and series (*Latoangulatae*, *Transversae*), and the sampled populations are geographically close in New South Wales (Australia) (see Material and Methods). *Eucalyptus urophylla*, on the other hand, although in the same section, belongs to a different series (*Annulares*) (Brooker 2000) and is located several thousand kilometers away from NSW, in Timor Island. This probably explains why *E. grandis* and *E. saligna* group together, whereas *E. urophylla* forms a separate cluster together with *E. camaldulensis*, whose sampled population in Northern Queensland is geographically closer to *E. urophylla*. These in turn are separate from the temperate species *E. dunnii* and *E. globulus* that belong to the more distantly related section *Maidenaria* forming another cluster (Figure 1). Estimates of the average shared allele distances (D_{SA}s) within species ranged from 0.637 (*E. dunnii*) to 0.727 (*E. grandis*). When individuals of all species were analyzed jointly, the average pairwise distance increased to a highest estimate of 0.853 (*E. globulus* × *E. urophylla*) to a lowest 0.801 (*E. globulus* × *E. dunnii*) and an overall estimate of 0.810. These within-species estimates are consistent with higher levels of genetic variation in species with more widespread geographical distribution. The between-species estimates are in turn coherent with the phylogenetic relationships among them. Model based clustering with STRUCTURE at $\Delta K = 6$ resulted in the grouping of the 6 species (Figure 1). All individuals were correctly assigned to their respective species indicating that these microsatellite provide an effective tool to differentiate the tested species. Individuals of each species were sampled from the same population so that no structure is expected within species. Given the resolution of these markers, it is reasonable to say, however, that they may also prove useful to infer genetic structure in *Eucalyptus* populations, a necessary procedure to avoid spurious results when carrying out association genetics studies in potentially structured populations. Furthermore, these results suggest that these microsatellites may be effectively used for assignment tests to infer species composition of spontaneous hybrids.

In conclusion, we have exploited existing resources of *Eucalyptus* ESTs to develop, select, and carry out a detailed characterization of a new set of 20 microsatellite markers that are polymorphic and transferable across 6 of the major planted *Eucalyptus* species. Considering the existing literature and unpublished data in our laboratory regarding the assessment of transferability and polymorphism across species of *Eucalyptus*, together with the fact that they are derived from genic regions, we expect that these new microsatelites should perform well for other species within the subgenus *Symphyomyrtus*. Outside *Symphyomyrtus*, however, microsatellite transferability tends to drop, although rates between 30% and 60% have been reported (Kirst et al. 1997; Steane et al. 2001; Ottewell et al. 2005; Nevill et al. 2008). The specific advantage of this set of markers over previously developed sets is their higher quality electrophoretic profiles, their higher and confirmed interspecific transferability, and the homogeneously high rate of polymorphism across these 6 species. We showed that these markers provide very good resolution for genetic population studies in phylogenetically contrasting species within the subgenus *Symphyomyrtus*. As they were developed from more conserved transcribed regions, the transferability and polymorphism of these microsatellites will most likely extend to the other 300 or more species within the same subgenus *Symphyomyrtus* further highlighting their applied value for *Eucalyptus* genetics and breeding.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

Funding

Brazilian Ministry of Science and Technology through FINEP grant 1755-01 and CNPq grant 520489/02-0 (part of the Genolyptus project); doctoral fellowships from CAPES (to E.M.C.M. and D.A.F.); research fellowships from CNPq (to G.J.P.Jr and D.G.).

References

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.

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 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.

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.

Brooker MIH. 2000. A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). Aust Syst Bot. 13:79–148.

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.

Byrne M, Marquezgarcia MI, Uren T, Smith DS, Moran GF. 1996. Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust J Bot. 44:331–341.

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, Jin L. 1993. Determination of relatedness between individuals using DNA-fingerprinting. Human Biol. 65:875–895.

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 U S A. 94:1041–1046.

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

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

Doughty RW. 2000. The eucalyptus. A natural and commercial history of the gum tree. Baltimore (MD) and London: The Johns Hopkins University Press.

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.

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.

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. 2004. Integrating genomics into *Eucalyptus* breeding. Genet Mol Res. 3:369–379.

Grattapaglia D, Alfenas AC, Coelho ASG, Bearzoti E, Pappas GJ, Pasquali G, Pereira G, Colodette J, Gomide JL, Bueno J, et al. 2004. Building resources for molecular breeding of *Eucalyptus*. In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M, Tomé M, editors. International IUFRO Conference: Eucalyptus in a changing world. Oct 11-15 Aveiro (Portugal): RAIZ, Instituto Investigação da Floresta e Papel. p. 20–32.

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

Grattapaglia D, Ribeiro VJ, Rezende GD. 2004. 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.

Jones ME, Shepherd M, Henry R, Delves A. 2008. Pollen flow in *Eucalyptus grandis* determined by paternity analysis using microsatellite markers. Tree Genet Genomes. 4:37–47.

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, Brondani RPV, Brondani C, Grattapaglia D. 1997. Screening of designed primer pairs for recovery of microsatellite markers and their transferability among species of *Eucalyptus*. In: EMBRAPA, editor. Proceedings of the IUFRO Conference on Eucalyptus Genetics and Silviculture. 1997 August 24–29; Salvador, BA (Brazil). EMBRAPA CNPF. p. 167–171.

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.

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.

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

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.

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

Nevill PG, Reed A, Bossinger G, Vaillancourt RE, Larcombe M, Ades PK. 2008. Cross-species amplification of *Eucalyptus* microsatellite loci. Mol Ecol Resour. 8:1277–1280.

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr., Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics. 9: 312.

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.

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.

Poke FS, Vaillancourt RE, Elliott RC, Reid JB. 2003. Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). Mol Breeding. 12:107–118.

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

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018

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

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

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

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.

Varshney RK, Graner A, Sorrells ME. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23: 48–55.

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.

Received October 13, 2009; Revised February 5, 2010; Accepted February 8, 2010

Corresponding Editor: David B. Wagner

Downloaded from https://academic.oup.com/jhered/article-abstract/101/4/512/906641
by Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) user
on 23 August 2018