RESEARCH ARTICLE

Eucalyptus microsatellites mined *in silico*: survey and evaluation

R. YASODHA*, R. SUMATHI, P. CHEZHIAN, S. KAVITHA and M. GHOSH

Division of Plant Biotechnology, Institute of Forest Genetics and Tree Breeding, P.O.Box 1061, Forest Campus, Coimbatore 641 002, India

Abstract

Eucalyptus is an important short rotation pulpy woody plant, grown widely in the tropics. Recently, many genomic programmes are underway leading to the accumulation of voluminous genomic and expressed sequence tag sequences in public databases. These sequences can be utilized for analysis of simple sequence repeats (SSRs) and single nucleotide polymorphism (SNPs) available in the transcribed genes. In this study, *in silico* analysis of 15,285 sequences representing partial and full-length mRNA from *Eucalyptus* species for their use in developing SSRs or microsatellites were carried out. A total of 875 EST-SSRs were identified from 772 SSR containing ESTs. Motif size of 6 for dinucleotide and 5 for trinucleotide, tetranucleotide, and pentanucleotides were considered in locating the microsatellites. The average frequency of identified SSRs was 12.9%. The dinucleotide repeats were the most abundant among the dinucleotide, trinucleotide and tetranucleotide motifs and accounted for 50.9% of the *Eucalyptus* genome. Primer designing analysis showed that 571 sequences with SSRs had sufficient flanking regions for polymerase chain reaction (PCR) primer synthesis. Evaluation of the usefulness of the SSRs showed that EST-derived SSRs can generate polymorphic markers as all the primers showed allelic diversity among the 16 provenances of *E. tereticornis*.

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

Introduction

Microsatellites or simple sequence repeats (SSRs) are short tandem repeats dispersed evenly throughout the genome. SSRs are preferentially applied as molecular markers in numerous plant species particularly regarding their potential to reveal high level of allelic polymorphisms, codominant mode of inheritance, transferability across species and high reproducibility. The major disadvantage of SSRs as markers has been their time consuming development in the laboratory (Zane *et al*. 2002). However, with the fast-paced increase of nucleic acid sequences in recent years, it became realistic to screen *in silico* for microsatellites in databases for several plant species.

SSRs have been widely applied in various areas of forestry including studies of genome variation, genome mapping, integration of physical and genetic maps, and comparative genome analyses (Echt *et al*. 1999). In important forestry species like *Eucalyptus*, about 450 SSR primer pairs targeting $(AG)_n$ and $(AC)_n$ repeats, designated as EMBRA, were developed by Brazilian scientists (Brondani *et al*. 1998, 2002, 2006). Five primer pairs (FMRSA) were developed through ISSR-enrichment technique and their transferability across species were documented at the University of Pretoria, South Africa. Twelve primer pairs, designated EMCRC, for the amplification of SSRs were isolated at University of Tasmania (Steane *et al*. 2001). Additionally, about 30 primer pairs were developed using $(CA)_n$ and $(CAG)_n$ genomic-enriched library (Ottewell *et al*. 2005). EMBRA primers have been widely used for individual identification, development of genetic linkage maps for interspecific and intraspecific crosses and also employed in QTL analysis of various commercially important traits (Brondani *et al*. 1998, 2002; Marques *et al*. 2002; Rungis *et al*. 2004; Kirst *et al*. 2005).

The transferability of SSR markers among species has made comparative genome analysis possible for *Eucalyptus* species, where SSRs constitute anchor points for specific regions of the genome of different species (Marques *et*

*For correspondence. E-mail: yasodha@icfre.org.

Keywords. *Eucalyptus*; EST-SSR; simple sequence repeats; allelic diversity.

al. 2002). Recently, a consensus linkage map for *Eucalyptus* was developed exclusively on interspecific transferable microsatellites (Brondani *et al*. 2006). Comparative mapping also facilitates cloning of orthologous genes by identifying crossreference genes (Varshney *et al*. 2005a). However, the number of SSR primer pairs developed so far in *Eucalyptus* is lesser (about 500 SSRs, including all species) than its counterparts like the *Pinus* and *Populus* (Kirst *et al*. 2005). More SSR markers are essential to have high-resolution saturated genetic map and to develop comparative mapping of other *Eucalyptus*. Previously, SSRs were developed from SSR-enriched libraries, however, recently an alternative source comes from data mining sequence information from the publicly available EST, genes and cDNA clonal sequence databases (Mahalakshmi *et al*. 2002; Varshney *et al*. 2002; Qureshi *et al*. 2004).

Recently, Varshney *et al*. (2005b) reviewed the importance of EST-SSRs including various applications and potential uses in plant genetics and breeding. Analysis of *Eucalyptus* EST databases revealed one SSR for every four sequences (Ceresini *et al*. 2005) or one SSR per 2.5-kb sequence (Rabello *et al*. 2005). SSRs primers developed from such databases were used for genome analysis of barley (Thiel *et al*. 2003), wheat (Gupta *et al*. 2003) and cotton (Blenda *et al*. 2006) however, there is no such information available for *Eucalyptus*. In this study, the expressed sequences available in GenBank were used to investigate the type and distribution of repeat motifs in *Eucalyptus*.

Materials and methods

Plant material and DNA isolation

Seeds of 16 provenances were obtained from Australian Tree Seed Centre, CSIRO, Australia. The seedlings were raised and planted in Panampally Research Station, Kerala, India, and was used for the present study. One tree from each provenance was selected for the present study. Methods of DNA isolation has been described elsewhere (Balasaravanan *et al*. 2005).

Source of sequences

All the *Eucalyptus* ESTs, partial and full-length mRNA sequences (as on 9 August 2006) were downloaded from NCBI\PubMed\Nucleotide (http://www.ncbi.nlm.nih.gov) using limits of '*Eucalyptus*' for organism and 'mRNA' for molecules. The sequences were downloaded for five species of *Eucalyptus* (*E. camaldulensis, E. globulus, E. grandis, E. gunnii* and *E. tereicornis*) in FASTA format and saved as individual files. A total of 15,285 sequences (3341 EST sequences and 11,944 partial and full length 5′ mRNA sequences) representing various species and tissues were downloaded from the database and the details for each *Eucalyptus* species are given in table 1.

Detection of SSRs and designing of primers

The perfect SSRs were identified using SSRIT (simple sequence repeats identification tool), (Temnykh *et al*. 2001) with the criteria of selecting dinucleotide motifs with a minimum repeat length of 6 and trinucleotide, tetranucleotide, and pentanucleotide motifs with a minimum repeat length of 5. In the present study, data were assembled up to pentanucleotide repeats only.

After the detection of SSRs, flanking primers of size 18–27 nucleotides in length were designed using PRIMER3 (Rozen and Skaletsky 1998) based on the criteria of optimum 55% GC content, optimum melting temperature of 60°C, and absence of secondary structure.

Species wise information for the new microsatellite primers including the GenBank accession number of the sequences from which the microsatellite primer pairs were designed are given in tables 1–5 of electronic supplementary material at (http://www.ias.ac.in/jgenet/). Information includes: repeat motif, number of repeats, forward and reverse primer sequences, melting temperature (°C) and size of expected product in base pairs.

PCR conditions and electrophoresis

In the present study primers were synthesized (Sigma Aldrich, Bangalore) for *E. gunnii* (AJ627770), and *E. globules* (AJ697753, AF046122) to examine the cross species transferability.

PCR was carried out in 10 μl reactions consisting of a 1× PCR buffer (Genei, Bangalore) including 1.5 μM $MgCl_2$, 200 μM dNTPs, 250 nM of each primer, 0.5 U of *Taq*

Table 1. Frequency and distribution of SSR motifs in *Eucalyptus* species.

Species	No. of sequences	No. of SSR -EST	% of SSR -EST	No. of SSR	No. of motifs DN	TN	TTN	PN	% of motifs DN	TN
E. camaldulensis	8	3	37.5	3	1	2	-	-	33.3	66.7
E. globulus	4005	96	2.4	100	67	21	1	11	67	21
E. grandis	1591	187	11.8	224	121	97	4	2	54.0	43.3
E. gunnii	8550	392	4.6	445	269	167	9	-	60.4	37.6
E. tereticornis	1131	94	8.3	103	41	58	3	1	39.8	56.3
Total	15285	772	12.9[a]	875	499	345	17	14	50.9[a]	45.0[a]

[a]Average of percentages; DN, dinucleotide; TN, trinucleotide; TTN, tetranucleotide and PN, pentanucleotide.

polymerase (Genei, Bangalore), and 30 ng of genomic DNA. All fragments were amplified using MJ Research PTC-200 Thermal Cycler (MJ Research, USA) with the following temperature profile of initial denaturation of 3 min at 94°C followed by 35 cycles with denaturation at 94°C for 30 s, annealing temperature 55°C for 30 s and extension at 72°C for 30 s. After 35 cycles final extension step was performed at 72°C for 5 min.

PCR products were separated on denaturing polyacrylamide gels consisting of 5% polyacrylamide (AA : BIS = 9.5 : 0.5) and 7 M urea in 1× TBE buffer. To this end PCR reactions were mixed with equal volume of loading buffer (formamide containing 0.8 M EDTA and traces of xylene cyanol and bromophenol blue), denature at 95°C for 5 min and snap cooled on ice. Later the samples were loaded on preheated Sequi-Gen GT sequencing gel (Bio-Rad, USA), which were run at 1800 V for 2 to 3 h. After electrophoresis, the fragments were visualized by silver staining. The gel was documented using digital camera and scoring was done by visual inspection.

Results and discussion

Single-pass sequencing of cDNA clones is considered as the best approach for gene identification in many economically important species including forest trees. These (ESTs), which target the transcribed genes can be utilized in various ways including the development of SSR and SNP markers. EST-SSR markers are widely used as an anchor points in comparative genome mapping among species (Yu *et al*. 2004; La Rota *et al*. 2005). In industrially important genus like *Eucalyptus*, genome sequencing is in progress for few species (Nehra *et al*. 2005) hence, the SSRs polymorphism generated from ESTs could be extended to other less characterized species. In *Eucalyptus*, large research projects like FORESTs and Genolyptus have been established specifically for EST sequencing and identification (Poke *et al*. 2005) however, these EST databases are not publicly accessible. In this study, 15,285 ESTs and partial or full-length mRNA sequences from various species of *Eucalyptus* were downloaded from GenBank were used for the analysis of microsatellites.

The SSRIT detected 772 SSRs containing sequences (12.9%) from a total of 15,285 sequences, of which 571 SSRs had sufficient flanking regions to design primers. The average frequency of SSR motif in the *Eucalyptus* sequence collection was 12.9% (table 1). Similar studies conducted in *Eucalyptus* (Ceresini *et al*. 2005; Rabello *et al*. 2005) detected 25.5% and 29% of SSR containing ESTs, however these studies included mononuclotides as well as compound repeats also. Cereal genome analysis showed an average of 3.2% ESTs consisted of SSRs (Varshney *et al*. 2002). The list of SSRs available in the EST sequences are summarized as additional Excel file with separate work sheets for each species. It provides the information on the GenBank accession number, type of repeat, frequency of the motif, forward and reverse primers, annealing temperatures and product size.

The occurrence of individual types of repeat motifs in ESTs also varied depending upon the species. The dinucleotide and trinucleotide motifs in *Eucalyptus* were 50.9% and 45% respectively, with few representations of tetranucleotide and pentanucleotide motifs, similar observations were reported in *Eucalyptus* (Ceresini *et al*. 2005; Rabello *et al*. 2005) however, trinucleotide repeats were found to be the most common, followed by either dinucleotide repeats or tetranucleotide repeats (Varshney *et al*. 2005a).

Among the dinucleotide motifs the most abundant was AG/TC (89.8%) and other types of dimers were rarely found in all the species (table 2). Similar distribution was noticed in most of the plant species (Mahalakshmi *et al*. 2002; Varshney *et al*. 2002; Ceresini *et al*. 2005; Rabello *et al*. 2005) but in case of the animals AC/TG was found to be the most abundant (Ju *et al*. 2005). Among the trinucleotide repeats GGC/CCG (16.2%) and AAG/TTC (15.0%) were the most common ones. In many plant species similar occurrence was reported (Thiel *et al*. 2003; Ceresini *et al*. 2005; Rabello *et al*. 2005) except for species like *Medicago* (Mahalakshmi *et al*. 2002) and wheat (Varshney *et al*. 2002)

The variation in length and location of an SSR motif was also considered for its usefulness as a marker (Varshney *et al*. 2002; Ju *et al*. 2005). The average length of the *Eucalyptus* microsatellite was 16.0 bases. However, the number of SSRs with length variations ranged from 12 to 64 bases and maximum number of SSRs contained 12–20 bases followed by 22–30 bases with few occurrences in more than 30 bases. However in the FORESTs database the average size of the SSRs was found to be 22.42 bp (Ceresini *et al*. 2005).

To evaluate the usefulness of EST-SSRs across species as polymorphic markers primers were synthesized and amplified with the genomic DNA of 16 provenances of *E. tereticornis* (figure 1). All the three primers showed PCR products in expected size range in the individuals analysed and the average number of amplified products were 5.7. All the bands produced were polymorphic. Interspecific transferability was high for the primers developed showing that significant portion of the markers can be transferred to related species. EST-SSR markers have been proven to be associated or responsible for a target trait and recently wheat EST-SSR primers were used for candidate gene mapping (Gao *et al*. 2004).

In *Eucalyptus*, the occurrence and characteristics of genomic microsatellites was first reported in *E. nitens* by Byrne *et al*. (1996) and presently, construction of comprehensive microsattelite based linkage map for commercial species of *Eucalyptus* was well advanced (Brondani *et al*. 2006). *Eucalyptus* being an important source for pulp-wood worldwide, many genomic research programmes are under way for QTL mapping, and physical mapping, however it is restricted to

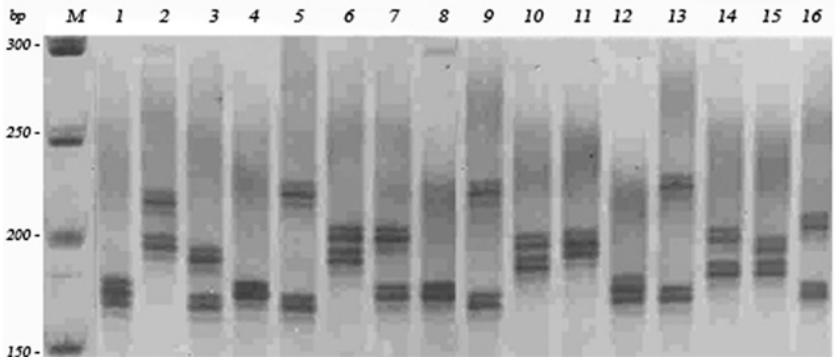

Figure 1. EST-SSR amplification in *E. tereticornis* showing allelic diversity among the 16 individuals. Primers: forward - 5′ CTAAGGACGACAAATTCGAGGA 3′ reverse - 5′ CTAAGGACGACAAATTCGAGGA 3′ (AJ697753).

Table 2. Types of SSR motifs in *Eucalyptus* species.

Nucleotides	Motif	*E. camaldulensis*	*E. globulus*	*E. grandis*	*E. gunnii*	*E. tereticornis*
Dinucleotide	AG/TC	1	53	118	236	40
	CG/GC	0	0	1	1	1
	AC/TG	0	2	2	13	0
	AT/TA	0	12	0	19	0
Trinucleotide						
	AAT/TTA	0	0	0	4	1
	CTA/GAT	0	3	0	7	1
	ATG/TAC	0	2	1	0	0
	ACT/TGA	0	1	2	4	0
	CTC/GAG	0	0	11	22	4
	AGG/TCC	0	0	23	12	8
	CAG/GTC	0	0	10	12	4
	AAG/TTC	2	2	3	37	8
	CAA/GTT	0	3	1	6	0
	AGC/TCG	0	1	5	14	3
	ACG/TGC	0	0	7	4	2
	AGA/TCT	0	2	2	11	3
	ACC/TGG	0	1	2	4	1
	GGC/CCG	0	5	17	18	16
	CGC/GCG	0	0	11	5	7
	CAC/GTG	0	0	2	4	0
	Others	0	1	0	3	0

few temperate species. Hence, the SSR primers identified will be useful for the scientific community to address various issues of *Eucalyptus* genetics.

Acknowledgements

This work was supported by a grant from Department of Biotechnology, Government of India, New Delhi.

References

Balasaravanan T., Chezhian P., Kamalakannan R., Ghosh M., Yasodha R., Varghese M. and Gurumurthi K. 2005 Determination of inter- and intra-species genetic relationships among six *Eucalyptus* species based on inter-simple sequence repeats (ISSR). *Tree Physiol*. **25**, 1295–1302.

Blenda A. V., Scheffler J. A., Scheffler B. E., Palmer M., Lacape J., Jesudurai C. *et al*. 2006 Cmo: a cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics*. **7**, 132 doi: 10.1186/1471-2164-7-132.

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

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

Brondani R. P. V., Williams E. R., Brondani C. and 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. doi:10.1186/1471-2229-6-20.

Byrne M., Marques-Garcia M. I., Uren T., Smith D. S. and Moran G. F. 1996 Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus. Aust. J. Bot*. **44**, 331–341.

Ceresini P. C., Silva C. L. S. P., Missio R. F., Souza E. C., Fischer C. N., Guillherme I. R. et al. 2005 Satellypus: Analysis and database of microsatellites from ESTs of *Eucalyptus*. *Gen. Mol. Biol*. **28**, 589–600.

Echt C. S., Vendramim G. G., Nelson C. D. and Marquardt P. 1999 Microsatellite DNA as shared genetic markers among conifer species. *Can. J. For. Res*. **29**, 365–371.

Gao L. F., Jing R. L., Huo N.-X., Li Y., Li X.-P., Zhou R.-H., Chang X.-P., Tang J.-F.,. Ma Z.-Y and Jia J.-Z. 2004 One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat. *Theor. Appl. Genet*. **108**, 1392–1400.

Gupta P. K., Rustgi S., Sharma S., Singh R., Kumar N. and Balyan H. S. 2003 Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics*. **270**, 315–323.

Ju Z., Wells M. C., Kazianis S., Rains J. D. and Walter R. B. 2005 an in silico mining for simple sequence repeats from expressed sequence tags of *Zebrafish, Medaka, Fundulus*, and *Xiphophorus. In Silico Biol*. **5**, 439–463.

Kirst M., Cordeiro G. D., Rezende S. P. and Grattapaglia D. 2005 Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations. *J. Hered*. **96**, 1–6.

La Rota M., Varshney R. V., Yu J.-K. and Sorrells M. E. 2005 Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* **6**, 23.

Mahalakshmi V., Aparana P., Ramadevi S. and Ortiz R. 2002 Genomic sequence derived simple sequence repeat markers - Case study with *Medicago* spp. *Electron. J. Biotechnol*. **5**, 233–242.

Marques C. M., Brondani R. P. V., Grattapaglia D. and Sederoff R. 2002 Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species. *Theor. Appl. Genet*. **105**, 474–478.

Nehra N. S., Becwar M. R., Rottmann W. H., Pearson L., Chowdhury K., Chang S. *et al*. 2005 Forest biotechnology: Innovative methods, emerging opportunities, *In Vitro Cell Develop. Biol. Plant* **41**, 701–717.

Ottewell K. M., Donnellan S. C., Moran G. F. and Paton D. C. 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.

Poke F. S., Vaillancourt R. E., Potts B. M. and Reid J. B. 2005 Genomic research in *Eucalyptus. Genetica* **125**, 79–101.

Qureshi S. N., Saha S., Varshney R. V. and Jenkins J. N. 2004 EST-SSR: A New class of genetic markers in cotton. *J. Cotton Sci*. **8**, 112–123.

Rabello E., de Souza A. N., Saito D. and Tsai S.-M. 2005 *In silico* characterization of microsatellites in *Eucalyptus* spp.: Abundance, length variation and transposan associations. *Gen. Mol. Biol*. **28**, 582–588.

Rozen S. and Skaletsky H. J. 1998 Primer 3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Rungis D., Berube Y., Zhang J., Ralph S., Ritland C. E., Ellis B. E. *et al*. 2004 Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. *Theor. Appl. Genet*. **109**, 1283–1294.

Steane D. A., Vaillancourt R. E., Russell J., Powell W., Marshall D. and Potts B. M. 2001 Development and characterisation of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). *Silvae Genet*. **50**, 89–91.

Temnykh S., DeClerck G., Lukashova A., Lipovich L., Cartinhour S. and McCouch S. 2001 Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genet. Res*. **11**, 1441–1452.

Thiel T., Michalek V. and Graner A. 2003 Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet*. **106**, 411–422

Varshney R. K., Sigmund R., Börner A., Korzun V., Stein N., Sorrells M. E., Langridge P. and Graner A. 2005a Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Pl. Sci*. **168**, 195–202.

Varshney R. K., Graner A. and Sorrells M. E. 2005b Genic microsatellite markers in plants. *Trends Biotechnol*. **23**, 48–55.

Varshney R. K., Thiel T., Stein N., Langridge P. and Graner A. 2002 In Silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol. Biol. Lett*. **7**, 537–546.

Yu J. K., La Rota M., Varshney R. V. and Sorrells M. E. 2004 EST derived SSR markers for comparative mapping in wheat and rice. *Mol. Genet. Genomics* **271**, 742–751.

Zane L., Bargelloni L. and Patarnello T. 2002 Strategies for microsatellite isolation: a review. *Mol. Ecol*. **11**, 1–16.

Received 21 June 2007, in revised form 11 September 2007; accepted 12 September 2007
Published on the Web: 1 April 2008