Plant Biotechnology 28, 471–480 (2011)
DOI: 10.5511/plantbiotechnology.11.1027b

Original Paper

Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*

Hideki Hirakawa[1], Yasukazu Nakamura[1,2], Takakazu Kaneko[1,3], Sachiko Isobe[1], Hiroe Sakai[1], Tomohiko Kato[4], Takashi Hibino[4], Shigemi Sasamoto[1], Akiko Watanabe[1], Manabu Yamada[1], Shinobu Nakayama[1], Tsunakazu Fujishiro[1], Yoshie Kishida[1], Mitsuyo Kohara[1], Satoshi Tabata[1], Shusei Sato[1,*]

[1] Department of Plant Genome Research, Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Mishima, Shizuoka 411-8510, Japan; [3] Faculty of Life Sciences, Kyoto Sangyo University, Kita, Kyoto 603-8555, Japan; [4] Forestry Research Institute, Oji Paper Co. Ltd., Kameyama, Mie 519-0212 Japan
* E-mail: ssato@kazusa.or.jp Tel: +81-438-52-3923 Fax: +81-438-52-3924

Received October 7, 2011; accepted October 27, 2011 (Edited by T. Demura)

Abstract The genetic information in the genome of *Eucalyptus camaldulensis* was investigated by sequencing the genome and the cDNA using a combination of the conventional Sanger method and next-generation sequencing methods, followed by intensive bioinformatics analyses. The total length of the non-redundant genomic sequences thus obtained was 654,922,307 bp consisting of 81,246 scaffolds and 121,194 singlets. These sequences accounted for approximately 92% of the gene-containing regions with an average G+C content of 33.6%. A total of 77,121 complete and partial structures of protein-encoding genes have been deduced. Comparison of the genes mapped on the KEGG pathways or located in the KOG classification with those in other plant species revealed the characteristics of the *E. camaldulensis* genome, and it was found that 23 pathways contained enzymes present only in the *E. camaldulensis* genome. Polymorphism analysis using microsatellite markers developed from the genomic sequence data obtained was performed with six *Eucalyptus* species collected from various parts of the world to estimate their genetic diversity, and the usefulness of these markers was demonstrated. The genomic sequence and accompanying information presented here are expected to serve as valuable resources for the acceleration of fundamental and applied research with *Eucalyptus*, especially in the fields of paper production and industrial materials. Further information on the genomic and cDNA sequences and microsatellite markers is available at http://www.kazusa.or.jp/eucaly/.

Key words: *Eucalyptus camaldulensis,* cDNA sequencing, genome sequencing, microsatellite markers, genetic diversity.

Eucalyptus (Myrtaceae) is a large genus that includes about 700 species of hardwood trees and shrubs. *Eucalyptus* species are widely distributed in the temperate zone in the southern hemisphere, and are used as material for pulp and paper production around the world (Turnbull and Pryor 1984). The paper manufacturing industry has particularly high demand for *Eucalyptus* wood, because the trees have superior growth properties and pulp quality. There are two major aims for the development of *Eucalyptus* biotechnology. The first aim is to secure land areas for *Eucalyptus* tree plantations. It has become more and more difficult recently to secure areas for tree plantations for the paper pulp industry, since crops for food and bio-fuel production compete with trees for land. One possible way to sidestep these competing demands is to expand tree plantations into dry, saline and acidic areas by developing more stress-tolerant tree species. The development of *Eucalyptus* biotechnology is thus necessary for breeding stress-tolerant trees suitable for marginal lands.

The second aim is to improve the paper quality. Paper quality is greatly affected by the fiber properties of wood, such as fiber length, cell wall thickness, and cellulose and lignin content (Higgins 1984). However, little is known about the genes that affect these fiber properties. Thus, *Eucalyptus* biotechnology based on tree genomics is necessary for elucidating the genes that control the fiber properties.

In 2005, with the release of the whole genome sequence of *Populus* (Tuskan et al. 2006), forest tree biotechnology has entered a "post-genomic" research phase. Two *Eucalyptus* whole-genome sequencing projects have been launched: the present study, which

This article can be found at http://www.jspcmb.jp/
Published online December 10, 2011

began in 2004, and another by the DOE Joint Genome Institute (JGI), USA, which began in 2007. Additionally, there are a large number of *Eucalyptus* ESTs in private or semi-private databases, though these EST data are not always open to the public domain. For example, Arborgen Inc. (USA, http://www.arborgen.com/) has 218,000 ESTs (23,000 contigs) from *E. grandis* and other species, the ForEST consortium (*Eucalyptus* genome sequencing project consortium) supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil, http://www.fapesp.br/) has 123,000 ESTs from *E. grandis* (Bocca et al. 2005), and the Genolyptus consortium (Brazil, http://www.ieugc.up.ac.za/newsgenolyptus-25may2006.htm) has 120,000 ESTs from *E. grandis*, *E. globulus* and other species.

ESTs and genome sequencing data could be combined to produce contigs (unigenes or tentative consensus sequences) that would be useful for microarray design, for annotation of the genome, and for the development of molecular markers (possibly by using single nucleotide polymorphism (SNP) chips) to be used to anchor the genome sequence to a high-density genetic linkage map of the *Eucalyptus* genome (Grattapaglia et al. 2011). These data from different *Eucalyptus* species also could aid in the discovery of polymorphisms (e.g. simple sequence repeats (SSRs) which are highly informative as markers in mapping projects).

Among the *Eucalyptus* species, *E. camaldulensis,* known as river red gum, is naturally distributed in most of the Australian mainland, and is planted in many tropical and subtropical countries (Butcher et al. 2002). Because of its diploid nature (2n=22) and feasibility of Agrobacterium-medicated genetic transformation (Mullins et al. 1997), *E. camaldulensis* is suitable for molecular genetic analysis and application of genetic engineering. To quickly survey the genetic information carried by this plant and to accelerate the process of molecular breeding, we analyzed the structure of the whole genome of *E. camaldulensis*. For genome sequencing, we adopted a combination of shotgun sequencing by the conventional Sanger method and the next-generation sequencing method, which was followed by intensive bioinformatics analyses. In addition, microsatellite markers were developed using the sequence information, and polymorphism among six *Eucalyptus* species was examined. The information on the *E. camaldulensis* genome generated in this study will enhance both fundamental and applied research on *Eucalyptus* and related plants.

Materials and methods

Plant materials

Eucalyptus camaldulensis, *E. dunnii, E. globules, E. grandis, E. nitens* and *E. urophylla* were grown in an experimental field at the Forestry Research Institute (Oji Paper Co. Ltd.). A clone of *E. camaldulensis*, named CPT1, was used as a material for EST collection and genome sequencing in this study.

EST collection

The cDNA libraries were constructed from xylem tissues of three-year-old stems, leaves and roots of ‘CPT1’. Total RNA was extracted from 5 g of each tissue using a Plant RNA Isolation Reagent (Invitrogen, USA) following the supplier’s protocol. Purification of polyadenylated RNA and conversion to cDNA was performed as described previously (Azamizu et al. 1999). Synthesis and cloning of cDNA were performed with a SMART™ cDNA Library Construction Kit (Clonetech, USA). Synthesized cDNA fragments were cloned into *Sfi*I-digested arms of a λTriplEx2 vector (Clonetech) and introduced into an *Escherichia coli* BM258 strain (Clonetech) according to the supplier’s protocol. For generation of ESTs, plasmid DNAs were prepared from the colonies and sequenced using the BigDye Terminator cycle sequencing ready reaction kit (Applied Biosystems, USA). The reaction mixtures were run on an automated DNA sequencer ABI PRISM 3730 (Applied Biosystems), and the collected data were processed as described below.

In addition, ESTs were also collected from the seedling plants grown from cutting propagation of ‘CPT1’. The total RNA was extracted from young leaves, cane tops of shoots and roots as described above with slight modification. Five cDNA libraries, ELS from shoots without dark treatment, EDS from shoots with dark treatment for 26 h, EDL from leaves with dark treatment for 26 h, EHR from roots and ENM from a mixture of shoots, leaves and roots, were constructed from their polyadenylated RNA purified using the PolyATtract mRNA Isolation System (Promega). Conversion to cDNA and size-selection of cDNA were performed as previously described (Asamizu et al. 1999). The cDNA fragments were cloned into *Eco*Rl–*Xho*I sites of pBluescript II SK-plasmid vectors (Agilent Technologies, USA) and transformed into *E. coli* ElectroTen-Blue cells (Agilent Technologies). Each average insert size in ELS, EDS, EDL and EHR was approximately 4 kb, and both ends of the inserts were sequenced. The ENM library was subjected to normalization as described (Asamizu et al. 1999), of which the average insert size was approximately 2 kb, and both ends of the inserts were sequenced. The sequencing was performed as described in the section “Shotgun sequencing of the genome”.

Construction of BAC libraries and BAC sequencing

BAC genomic libraries were constructed using the genomic DNA of *E. camaldulensis* partially digested with *Hin*dIII and Copy Control pCC1BAC as a cloning vector. The average insert size of these libraries was 86.5 kb. A total of 57,600 clones, which covers the haploid genome 7.7 times in total, were constructed and arrayed in 384-well microtiter plates (150 total).

To analyze end sequences, BAC DNAs were amplified using a TempliPhi large construction kit (GE Healthcare, UK), and the end sequences were analyzed according to the Sanger method using a cycle sequencing kit (Big Dye-terminator kit, Applied Biosystems, USA) with DNA sequencer type 3730xl

(Applied Biosystems). The sequencing process yielded 98,006 reads that were quality checked with PHRED (Ewing et al. 1998, Ewing and Green 1998), allowing the identification and removal of low-quality sequences.

High-quality BAC sequences were determined according to the shotgun strategy using the Sanger sequencing protocol with five times redundancy, as described previously (Sato et al. 2008).

Shotgun sequencing of the genome

For sequencing by the Sanger method, shotgun libraries with average insert sizes of 2.5 kb were generated using pUC118 as a cloning vector, and these were used to transform *E. coli* ElectroTen-Blue (Agilent Technologies, Santa Clara, CA, USA). The shotgun clones were propagated in microtiter plates, and the plasmid DNA was amplified using a TempliPhi kit (GE Healthcare). Sequencing was performed using a cycle sequencing kit (Big Dye-terminator Cycle Sequencing kit, Applied Biosystems) with DNA sequencer type 3730xl (Applied Biosystems) or DeNOVA-5000HT (Shimadzu Co., Japan) according to the protocols recommended by the manufacturers.

In parallel, a shotgun library with an average insert size of 2.5 kb was constructed from a pooled DNA sample isolated from 26,715 selected BAC clones (Selected BAC Mixture: SBM) in which neither end sequence showed homology with any of the highly repetitive sequences of the *E. camaldulensis* genome identified as contigs by the assembly of 0.5 million WGS reads. Template preparation and sequencing of this shotgun library, designated as a BAC mixture shotgun library, were carried out using the same strategy as described above. A total of 1,341,303 reads and 2,893,145 reads were accumulated from the WGS library and BAC mixture shotgun library, respectively.

454 paired-end sequence

For paired-end sequencing on the 454/Roche GS-FLX sequencer, two distinct shotgun libraries, namely 3-kb and 8-kb paired-end libraries, were constructed. Amplification and sequencing of these libraries were performed using GS FLX Titanium Sequencing Kits and 454 Genome Sequencer FLX Instruments following the manufacturer's protocols (Roche Applied Science, Mannheim, Germany). In total, 805.7 Mb of sequence data was generated from 1.91 million 3-kb paired-end reads and 0.56 million 8-kb paired-end reads.

Assembly of sequence data

EST sequences from the libraries of leaves, roots, stems and shoots were collected. Sequencing chromatograms were evaluated with PHRED and vector-derived sequences were trimmed with CROSS_MATCH (Ewing and Green 1998). The EST reads were quality-trimmed by the PHRED quality score at a position where five ambiguous bases (PHRED score under 16) were found within 15 contiguous bases. Reads comprising ≥100 contiguous bp of satisfactory quality were submitted to the DDBJ/EMBL/GenBank databases. Assembly was performed using MIRA 3.2.1 program (Chevreux et al. 2004) with the EST mode.

Reconstruction of the genome sequence of *E. camaldulensis* CPT1 was performed by assembly of sequence data generated by different types of DNA sequencers. The sequence data collected by the Sanger protocol using a 3730xl capillary sequencer with low quality and including the genome sequence of *E. coli* K12 MG1655 (GenBank: U00096) were excluded. The remaining sequences were subjected to the masking of cloning vectors (UniVec: http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) and trimming of the masked and low-quality bases (QV<15) by CROSS_MATCH and TRIM2 programs (Huang et al. 2006). These Sanger sequences and the pyrosequences with paired-end (3 kb and 8 kb) were directly assembled using the Celera assembler 6.1 program (generally called CABOG) (Koren et al. 2010). Scaffolds were constructed by adding N characters between the contigs by the WGS6.1 program. Scaffolds and singlets generated by the assembly were subjected to similarity searches against the bacterial genome sequences and non-redundant amino acid sequences (nr) published in the NCBI database, and the chloroplast of *E. grandis* (GenBank: NC_014570) and the mitochondria of *Arabidopsis thaliana* (GenBank: NC_001284) using BLASTN or BLASTX programs (Altschul et al. 1997). Matching sequences were then removed. Finally, sequences that were 499 bp and shorter were removed. The resulting scaffolds and singlets were designated as follows. The scaffolds were prefixed with 'EcC' followed by a seven-digit number, and the singlets were prefixed with 'EcS' followed by a seven-digit number.

Repetitive sequence analyses

Repetitive elements in the *E. camaldulensis* genome were identified by comparing all of the contig units using BLASTN and processing the outputs using the RECON program (Bao and Eddy 2002). A total of 99 consensus sequences of repetitive elements that appeared at least 20 times were identified. The consensus sequences of these elements were subjected to a similarity search against known repeat elements in the RepBase (http://www.girinst.org/). For the consensus sequences with features of class I or class II transposable elements (TE), full-length candidate sequences were identified by comparing the 10 kb upstream and downstream of the corresponding genomic regions to find long terminal repeats or terminal inverted repeats. For unclassified consensus sequences, the longest representative sequences were selected by comparing the corresponding genome sequences using the CLUSTALW (Thompson et al. 1994) multiple alignment program. Full-length elements of TE and representative sequences of unclassified repeats were collected into a repeat sequence library and used as references for RepeatMasker (http://www.repeatmasker.org) analysis to delineate the occurrence of these elements in the *E. camaldulensis* genome sequences.

Simple sequence repeats (SSRs) 15 nucleotides or longer in length, which contained all possible combinations of di-nucleotide (NN), tri-nucleotide (NNN) and tetra-nucleotide (NNNN) repeats, were identified from the *E. camaldulensis* genome sequences and the unigene set of EST sequences using the MISA program (Thiel et al. 2003).

Gene assignment

Gene prediction and modeling were performed by automatic gene assignment programs that employ *ab initio* gene finding and similarity searches. For *ab initio* gene finding, predictions of protein-coding regions were carried out using GENEMARK.HMM (Lukashin and Borodovsky 1998) and GENESCAN (Burge and Karlin 1997) programs with the matrix trained by an *A. thaliana* gene set, and predictions of exon–intron structure were performed using NETGENE2 (Hebsgaard et al. 1996) and SPLICEPREDICTOR (Brendel and Kleffe 1998) programs. Similarity searches for the potential protein-coding regions and for all contigs were performed respectively against the Uniref database (http://www. ebi.ac.uk/uniref/) by BLASTX or BLASTP programs with a cut-off E-value $\leq$1e-10. The exon–intron structure of the potential protein-coding regions and the contigs homologous to the Uniref database were predicted using the NAP program (Huang and Zhang 1996). Suitable exon–intron structures were determined by considering all the information above. The predicted gene structures were further confirmed by comparison to the cDNA sequences analyzed in this study. The protein-coding genes assigned in this manner were denoted by IDs with the contig names followed by sequential numbers from one end to another. They were classified into four categories based on sequence similarity to registered genes: genes with complete structure (intrinsic genes), pseudogenes, genes with partial structure (partial genes), and transposons/retrotranspons (TE). In addition, gene prediction was performed using the AUGUSTUS program (Stanke et al. 2006) with the *A. thaliana* training set. The genes which were not homologous to those predicted by GENEMARK.HMM and GENESCAN were then included. The amino acid sequences with lengths $>$50 amino acid residues were used for further analyses.

Functional assignment and classification of potential protein-coding genes

To assign gene families, functional domains, GO terms, and GO accession numbers (Ashburner et al. 2000), the predicted genes were searched against InterPro using InterProScan (Hunter et al. 2009) software. Genes with E-values $\leq$1.0 were selected. GO terms were grouped into plant GO slim categories using the map2slim program (http://www.geneontology.org/GO.slims.shtml).

The predicted protein-encoding genes were mapped onto KEGG metabolic pathways (Ogata et al. 1999) using the BLASTP program against the GENES database (Ogata et al. 1999). Thresholds of amino acid sequence identity $\geq$25% and of length coverage of the query sequence $\geq$50% with a cut-off E-value $\geq$1e-10 were applied.

The predicted genes were classified into eukaryotic clusters of orthologous groups (KOG) categories according to the results of BLASTX searches against amino acid sequences in the KOG database (Tatusov et al. 2003). These sequence similarities were considered to be significant when the E-value was less than 1e-10.

Phylogenetic analysis

The genes with two domains related to terpene synthase, namely IPR001906 (terpene synthase-like) and IPR005630 (terpene synthase, metal-binding domain) were identified using InterProScan from the total genes of *E. camaldulensis* and *A. thaliana* (TAIR10 (Garcia-Hernandez et al. 2002)). Evolutionary relationships of terpene synthase genes were inferred for their amino acid sequences using the bootstrap consensus tree in 1,000 replicates resulting from neighbor-joining analysis by CLUSTALW program (ver. 2.0.12). Phylogenetic trees were constructed with MEGA 5 software (Tamura et al. 2007).

Allele frequency of the microsatellite markers of Eucalyptus germplasms

A total of six *Eucalyptus* species, *E. camaldulensis, E. dunnii, E. globules, E. grandis, E. nitens and E. urophylla,* were used for polymorphic analysis of genome and EST-derived microsatellite markers. The numbers of tested genome and EST-derived microsatellite markers were 512 and 500, respectively. PCR amplifications (5 μl) were performed on 0.7 ng of *Eucalyptus* genomic DNA in 1×PCR buffer (BIOLINE, London, UK), 3 mM $MgCl_2$, 0.04 U BIOTAQ^{TM} DNA Polymerase (BIOLINE, USA), 0.8 mM dNTPs, and 0.4 mM of each primer, using the modified 'Touchdown PCR' protocol described by Sato et al. (2008). PCR products were separated by 10% polyacrylamide gel electrophoresis using TBE buffer, and data were collected as described previously (Sato et al. 2008). Allele detection and genotype code typing were performed using the Polyans program (ver.1.1; http://www.kazusa.or.jp/polyans/). The presence or absence of amplification and the number of different-sized fragments, which was regarded as the number of alleles, were recorded using Polyans ver.1.1 (http://www.kazusa.or.jp/polyans/) software. Loci where no amplification was observed were regarded as null alleles. Single bands were regarded as homozygous loci. In the presence of multiple loci, the alleles of the clearest locus were qualified as the observed data of the marker. The HZ (heterozygosity) of the markers was estimated by the following equation:

$$HZ_i = 1 - \sum_{j=1}^{i} P_{ij}^2$$

Here P_{ij} is the frequency of the jth allele for the ith marker.

Results and discussion

Collection and clustering of ESTs

A total of 70,683 ESTs were collected from several tissues of *E. camaldulensis* CPT1 by the Sanger method as shown in Materials and methods. After trimming of the vector sequence and the low-quality regions, 58,584 ESTs were selected. These include 11,436 sequences from leaves, 13,128 from roots, 21,847 from stems, 3,009 from the ELS library, 2,265 from the EDS library, 2,844 from the EDL library, 1,182 from the EHR library, and 2,873 from the ENM library.

The ESTs were subjected to assembly using the MIRA ver. 3.2.1 (Chevreux et al. 2004) program, and 20,020 non-redundant sequences (unigenes) consisting of 9,022 contigs and 10,998 singlets were generated. The total length of the contigs and singlets was 12,792,464 bp. The average contig length was 781 bp, with the longest being 3,366 bp.

Genome sequencing

Details of the sequencing of the genome of *E. camaldulensis* CPT1 are summarized in Supplemental Figure 1. A total of 4,376,079 sequence reads were obtained by the conventional Sanger sequencing method. After removal of reads with low quality and those derived from the *E. coli* genome, the remaining 3,676,829 reads, consisting of 2,479,289 SBM sequences, 122,893 BAC end sequences, and 1,074,647 WGS sequences, were subjected to the trimming of vector sequences and low-quality bases. These Sanger sequences were combined with 1,909,515 and 555,045 reads of 3-kb and 8-kb paired-end 454 FLX pyrosequencing, and were assembled using the CABOG (Celera assembler 6.1) program (Koren et al. 2010) as described in Materials and methods. As a result, 81,246 scaffolds and 121,194 singlets were obtained. This dataset was denoted as EUC_r1.0. The total length of the genomic sequences excluding the scaffold gaps was 654,922,307 bp. The average genome size of the subgenus *Symphyomyrtus*, which includes the majority of the most widely planted and bred species of *Eucalyptu*s (e.g. *E. camaldulensis* and *E. grandis*), has been estimated to be 650 Mb (Grattapaglia and Bradshaw 1994); thus, the genome sequences obtained in this study represent 96% of the presumptive genome size. Statistics of the assembly of EUC_r1.0 are summarized in Table 1.

The coverage of gene space in EUC_r1.0 sequences was estimated by matching the EST unigene sequences described above. Among 20,020 non-redundant sequences, 18,336 matched the EUC_r1.0 sequences, suggesting that 91.6% of the gene space in the *E. camaldulensis* genome was covered by EUC_r1.0 sequences.

To investigate the structural features of the *E. camaldulensis* genome, high-quality sequences of 64 BAC clones were obtained by manual finishing. The total length of the sequences of the BAC clones was 4,641,094 bp, of which 4,040,869 bp (87.1%) was covered by EUC_r1.0 sequences. For the 723,718 bp (15.6%) regions covered by the BAC sequences, two corresponding but different EUC_r1.0 sequences were

Table 1. Statistics of the assembly of EUC_r1.0

Scaffolds	
Total length (bp)	624,468,648
(Total length excluding scaffold gaps (bp))	(535,166,177)
Total number	81,246
Average length (bp)	7,686
Maximum length (bp)	708,721
N50	18,024
G+C content (%)	33.6
Singlets	
Total length (bp)	119,756,130
Total number	121,194
Average length (bp)	988
Maximum length (bp)	34,761
G+C content (%)	39.4

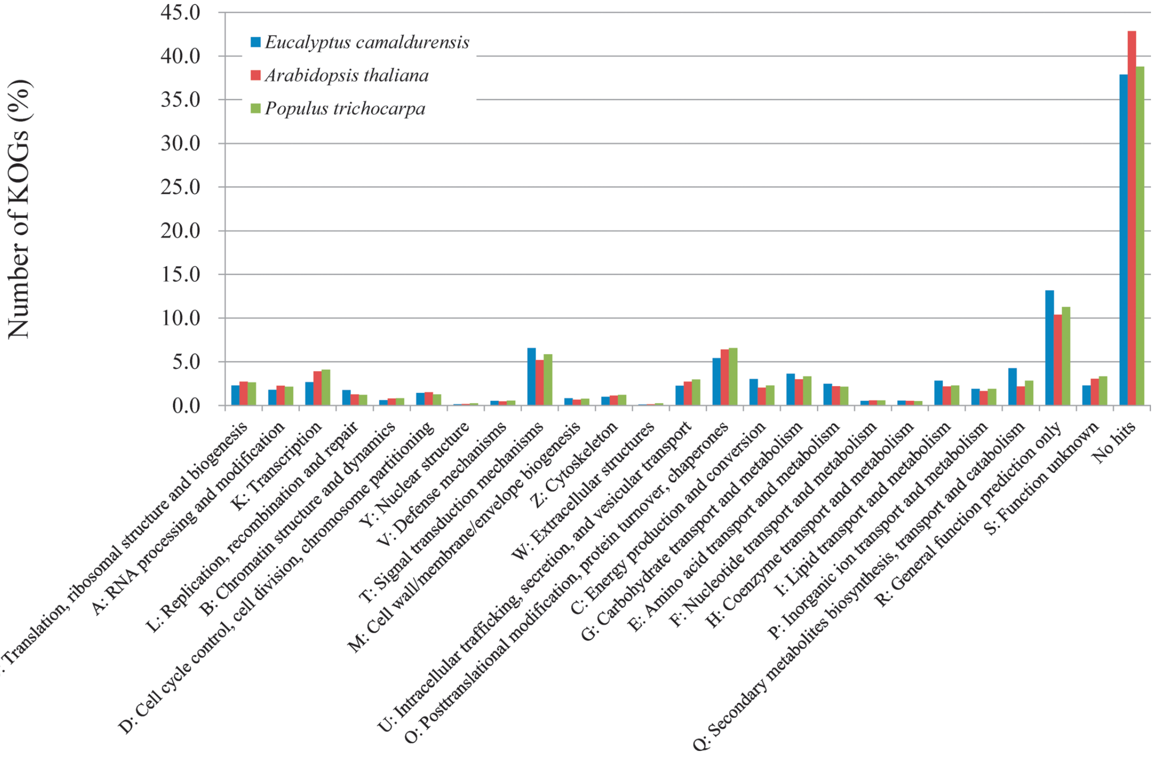

Figure 1. Gene assignment to KOG functional categories.

identified. These regions could be derived from highly heterozygous regions, and thus were difficult to assemble into a single contig in EUC_r1.0.

Repetitive sequences

A total of 91,570 di-, tri-, and tetra-nucleotide simple sequence repeats (SSRs) equal to or longer than 15 bp were identified in the *E. camaldulensis* genomic sequences (Supplemental Table 1). Provided that the size of the *Eucalyptus* genome is 650 Mb (Grattapaglia and Bradshaw 1994), the frequency of occurrence of the above SSRs was estimated to be one SSR in every 7.1 kb. Di-, tri-, and tetra-nucleotide SSRs accounted for 53.4%, 30.3%, and 16.3% of the identified SSRs, respectively. The SSR patterns that appeared frequently were (AG)n, (AAG)n, and (AAAT)n, each representing 81.9% of di-, 35.3% of tri-, and 47.3% of tetra-nucleotide repeat units.

In the cDNA fraction, on the other hand, a total of 1,746 SSRs were identified in the unigene sequences of 12.8 Mb in total. Thus, the average frequency of occurrence of SSRs is one SSR in every 7.3 kb slightly higher rate than 1 per entire genome. The frequencies of trinucleotide SSRs, such as (GGC)n, (GGA)n and so forth, were high in the cDNA unigene sequences. The ratio of (AG)n was also high, presumably corresponding to the UTR regions.

A search using the repeat-sequence-finding program RECON against the *E. camaldulensis* genomic sequences revealed the presence of a variety of repeat elements including class I and class II transposable element (TE) subfamilies and those that are difficult to classify into known subfamilies. The composition of these repeat sequences was analyzed with the RepeatMasker program (http://repeatmasker.org/); the results are summarized in Supplemental Table 2. Among the highly repetitive elements identified by the RECON program, EcRE02, a *copia*-type retrotransposon, was found in 3.9% of EUC_r1.0 sequences (data not shown).

A large number of short repetitive sequences including miniature inverted repeat transposable elements (MITEs), terminal-repeat retrotransposons in miniature (TRIM) and short interspersed nuclear elements (SINEs) were found in the *E. camaldulensis* genome. Over 40,000 copies of MITEs were identified (Supplemental Table 2) and classified into 9 types (data not shown). A substantial portion of these short repetitive sequences was found in introns and 3′ regions (putative 3′ UTRs) of the predicted genes.

Gene prediction and annotation

The *E. camaldulensis* genomic sequences were subjected to the automatic assignment of protein-encoding genes, and a total of 118,501 genes including 38,227 transposon related genes, were assigned. Complete structures with lengths ≥50 aa were predicted for 77,121 genes, but only partial structures were predicted for 32,729 genes. In addition, 2,869 genes and 26,997 genes with complete and truncated structures, respectively, were likely to be pseudogenes. It should be noted that the gene number may be overestimated by multiply counting the genes separated in multiple contigs and singlets. Of the 77,121 presumptive protein-encoding genes, 22,537 (29.2%) carried ESTs with a sequence identity of 90% or more for a stretch of 50 nucleotides.

Structural features of the protein-encoding genes in *E. camaldulensis* were investigated in detail for 278 genes predicted on the 64 BAC clones (4.60 Mb in total) for which high-quality sequences were obtained by manual finishing and annotation. As shown in Table 2, the basic structures of the protein-encoding genes in *E. camaldulensis* were similar to those of *A. thaliana* except for the average lengths of genes and introns: 2,863 bp versus 1,918 bp for genes and 383 bp versus 157 bp for introns in *E. camaldulensis* and *A. thaliana*, respectively.

A similarity search of the translated amino acid sequences of the 77,121 presumptive protein-encoding genes was performed using the TrEMBL database as a protein sequence library (Bairoch and Apweiler 1996). The results indicated that 51,642 (67.0%) genes had significant (E-value ≤1e-20) sequence similarity to genes in this database. Of these genes, 33,530 (64.9%) genes showed sequence similarities to the unigene constructed in this study with a cutoff E-value ≤1e-20 using TBLASTN.

The 77,121 presumptive protein-encoding genes assigned in *E. camaldulensis*, the 35,386 genes in *A. thaliana* released by TAIR10 (Garcia-Hernandez et al. 2002) and the 45,033 genes in populus (*Populus trichocarpa* released by Phytozome (http://www.phytozome.net/)), which is a model for the molecular biology of a fast-growing tree, were classified into plant GO slim categories (Carbon et al. 2009). The numbers of

Table 2. Structural features of the genes in *E. camaldulensis*

	E. camaldulensis			*A. thaliana*		
	Average	Min.	Max.	Average	Min.	Max.
Length of gene including intron (nt)	2,863	243	22,097	1,918	78	17,203
Length of gene (aa)	436	55	2,351	427	25	4,706
Number of introns per gene	4	0	36	4	0	48
Length of exon (bp)	260	3	4,035	256	2	5,966
Length of intron (bp)	383	25	4,630	157	23	2,989

genes classified into the various GO slim categories (i.e., Biological Process [BP], Cellular Component [CC] and Molecular Function [MF]) are listed in Supplemental Table 3.

Of the 77,121 presumptive genes in the *E. camaldulensis* genomic sequences, 3,920 genes could be mapped onto 128 of the 152 metabolic pathways in the KEGG database (Ogata et al. 1999), whereas the 4,371 and 1,300 genes of *A. thaliana* and *P. trichocarpa* were mapped onto 130 and 128 pathways, respectively. The enzymes mapped by the genes in the *E. camaldulensis* genome alone were identified in the twenty-three pathways including 'pentose and glucuronate interconversion' and 'fructose and mannose metabolism' in carbohydrate metabolism, 'steroid hormone biosynthesis' in lipid metabolism, 'tyrosine metabolism' in amino acid metabolism, 'cyanoamino acid metabolism' in metabolism of other amino acids and 'betalain biosynthesis' in biosynthesis of other secondary metabolites (Supplemental Table 4).

Among 77,121 presumptive genes, 64,046 (83.0%) showed similarity to the amino acid sequences in the KOG database. With the aim of comparing the features of the genes in the other organisms, KOG classification was carried out against the genes of *A. thaliana* and *P. trichocarpa*. The distribution of the genes assigned to various KOG functional categories is shown in Figure 1. Slightly high proportions of presumptive genes were classified into the metabolism categories of Signal transduction mechanisms (KOG T), Energy production and conversion (KOG C), Carbohydrate transport and metabolism (KOG G), Lipid transport and metabolism (KOG I) and Secondary metabolites biosynthesis, transport and catabolism (KOG Q).

Gene features in the genome of E. camaldulensis

Genes for cellulose biosynthesis

Eucalyptus retains major lignocellulosic biomass generated by the accumulation of their cellulose microfibers over a long period. These celluloses are synthesized with the cellulose synthases (CesAs) using UDP-glucose as a substrate. Genes encoding CesAs form a common gene family. A total of 10, 18, and 9 genes for CesA have been identified in the genomes of *A. thaliana*, *P. trichocarpa*, and *Oriza sativa*. For *E. grandis*, 6 different cDNAs for CesA have been cloned (Ranik and Myburg, 2006). Based on the sequence similarity, orthologous genes of each of these six cDNAs were identified in the *E. camaldulensis* genomic sequences. In addition to these, 5 more candidate genes for CesA (EcS729250.10, EcC074048.10/EcS733245.10, EcC001269.10, EcC047978.10, EcC050269.10) were found, suggesting that the CesA gene family in *E. camaldulensis* consists of at least 11 genes.

UDP-glucose, one of the substrates for cellulose biosynthesis, is produced by the cleavage of sucrose by sucrose synthase. *E. camaldulensis* has 27 genes encoding sucrose synthase, while only 6 genes have been identified in *A. thaliana*. Another pathway for production of UDP-glucose involves phosphorylation of glucose 1-phosphate by UDP-glucose pyrophosphorylase. Three copies of the genes for UDP-glucose pyrophosphorylase, EcC053822.20, EcC054676.20 and EcC054864.40, were found in the genome of *E. camaldulensis*.

Genes for terpenoid synthesis

The genus *Eucalyptus* is known to produce essential oils, whose major oil compounds are monoterpenes and sesquiterpenes (Chen et al. 2004; Kampranis et al. 2007). It has been reported that immature flowers of *E. camaldulensis* mainly contain 1,8-cineol, beta-pinene and spathulenol (Giamakis et al. 2001). The genes related to terpenoid biosynthesis tend to be clustered in the genome of *A. thaliana*: Tholl et al. (2005) assigned 33 genes to the gene family related to terpenoid synthases (TPSs); each bore two protein domains, "Terpene synthase, metal-binding" (IPR005630) and "Terpene synthase-like" (IPR001906). The *E. camaldulensis* genome contains at least 146 presumptive genes for the TPSs, which is significantly more than those in other trees, *P. trichocarpa* (42 genes) and *Vitis vinifera* (74 genes). Fifty-seven and 40 of the TPS genes in *E. camaldulensis* formed sub-families with 6 monoterpene synthases and 20 sesquiterpene synthases in *A. thaliana*, respectively, suggesting that the genes for oil production in *E. camaldulensis* are diverse and many of them are unique to the species (Figure 2).

Genes for P450

The genes for Cyt P450 constitute a relatively large gene family, which is further classified into sub-families on the basis of nomenclature files on the Cyt P450 Homepage (http://drnelson.utmem.edu/cytochromeP450.html). Investigation of the EUC_r1.0 sequences revealed 1,021 putative cytochrome P450 genes, including partially predicted genes. This gene number is larger than those previously identified in the plant genomes, such as *A. thaliana* (275 genes), *P. trichocarpa*, (566 genes), *V. vinifera* (553 genes) and *Oryza sativa* (417 genes). The composition of some of the sub-families of cytochrome P450 genes in the *E. camaldulensis* genome was remarkably different from that of *A. thaliana*. Significant amplification of the members was observed in the sub-families CYP75B, CYP76G, CYP82C, CYP704A, CYP716A, and CYP734A, when compared with cytochrome P450 genes in the *A. thaliana* genome (Supplemental Table 5). While the detailed functions of these subgroups remain to be analyzed, the CYP75B subfamily is reportedly involved in the flavonoid synthesis pathway, while CYP716A and CYP734A are

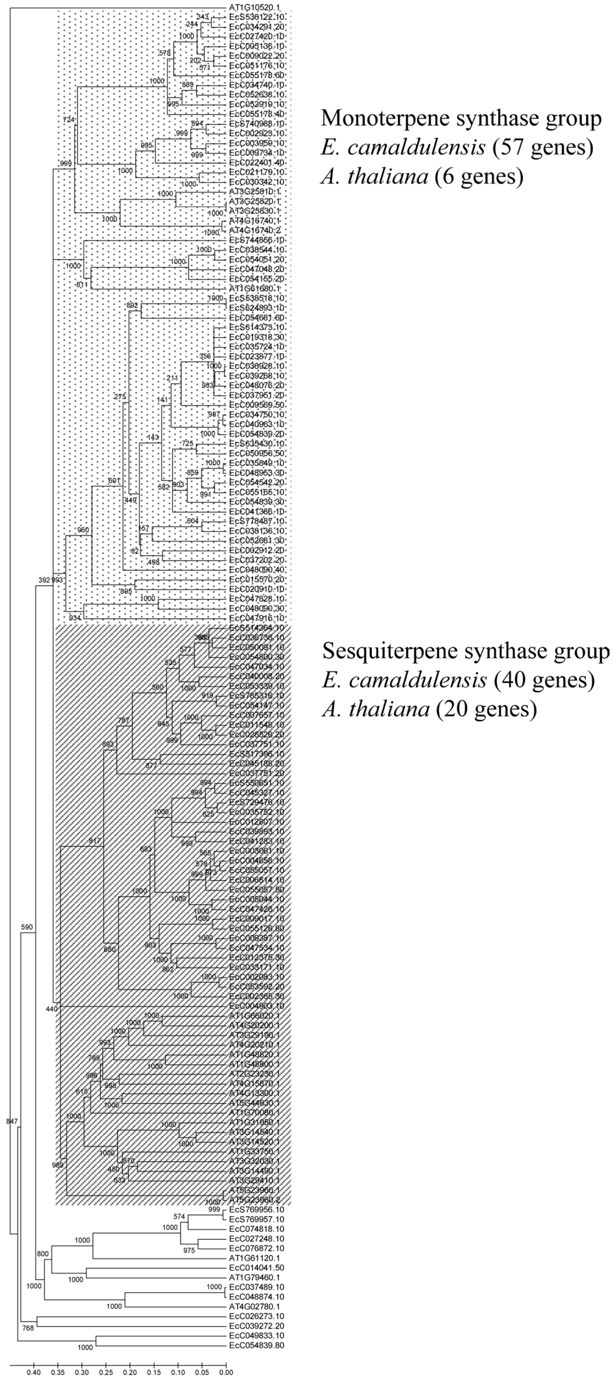
Monoterpene synthase group
E. camaldulensis (57 genes)
A. thaliana (6 genes)
Sesquiterpene synthase group
E. camaldulensis (40 genes)
A. thaliana (20 genes)
0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.00

Figure 2. Phylogeny of the genes for terpenoid synthases (TPS).

believed to be respectively involved in the monoterpene biosynthesis pathway and the triterpene, sterol, and brassinosteroid metabolism pathway, based on the co-expression analysis (Ehlting et al. 2008).

Genetic diversity

A total of 512 and 500 primer pairs to amplify microsatellites were designed based on the genomic sequences and ESTs of *E. camaldulensis*, respectively, and PCR amplification was examined for six *Eucalyptus* species: *E. camaldulensis, E. dunnii, E. globules, E. grandis, E. nitens* and *E. urophylla*. As a result, 477 and 449 primer pairs from genomic sequences and ESTs, respectively, successfully amplified polymorphic bands (Supplemental Table 6). Of these, null alleles were detected for 46–68 of genome-derived and 44–53 of EST-derived microsatellite markers for each of the six *Eucalyptus* species (Supplemental Table 7). The number of genome- and EST-derived microsatellite markers for which multiple loci were detected multiple loci was 257 (53.5%) and 138 (30.7%), respectively. The number of alleles per locus ranged from 1 to 9 with a mean value of 4.1 in genome-derived microsatellite markers, while those of EST-derived markers ranged 1 to 11 with a mean value of 3.8 (Figure 3). The HZ value ranged from zero to 0.88 for both genome- and EST-derived markers. No marker was observed for those with HZ values in the range 0≤HZ≤0.1. The mean HZ values of genome- and EST-derived microsatellite markers were 0.61 and 0.53, respectively.

Generally, high heterozygosity of DNA markers leads to high transferability of the marker, because polymorphism between haplotypes is essential to transfer the information of the DNA marker. Up to the present, genome-derived microsatellite markers showing high heterozygosity and transferability within intraspecies genomes have been reported in *E. leucoxylon* (Ottewell et al. 2005), *E. grandis* (Kirst et al. 2005), *E. sieberi* (Glaubitz et al. 2001) and *E. nitens* (Byrne et al. 1996). In this study, we investigated the interspecies polymorphisms. The mean HZ value of 0.61 and 0.53 for genome- and EST-derived microsatellite markers, respectively, suggested a high transferability of the microsatellite markers found in this study across species.

The diversity of genus *Eucalyptus* is large, containing more than 700 species (Brooker 2000). According to the phylogenetic groups defined by Brooker et al. (200), the six *Eucalyptus* species used in this study belong to subgenus *Symphyomyrtus*, and further into three sections, *Exseriaria,* (*E. camaldulensis*), *Latoangulatae* (*E. grandis, E. urophylla*) and *Maidenarla* (*E. globules, E. nitens* and *E. dunnii*). Despite the fact that the primer pairs for the microsatellite markers were designed based on the genomic information in *E. camaldulensis*, the result obtained in this study indicated the high conservation of microsatellite regions across the subgenus *Symphyomyrtus*.

In *Eucalyptus*, microsatellite markers have been used for ecological studies such as the estimation of gene flow in orchards (Chaix et al. 2003) and effects of native forest regeneration on genetic diversity (Glaubitz et al. 2003), as well as breeding studies such as the identification of QTLs (Thamarus et al. 2004) and development of a breeding method (Grattapaglia et al. 2004). To date, more than 300 genome-derived microsatellite markers have been reported for several

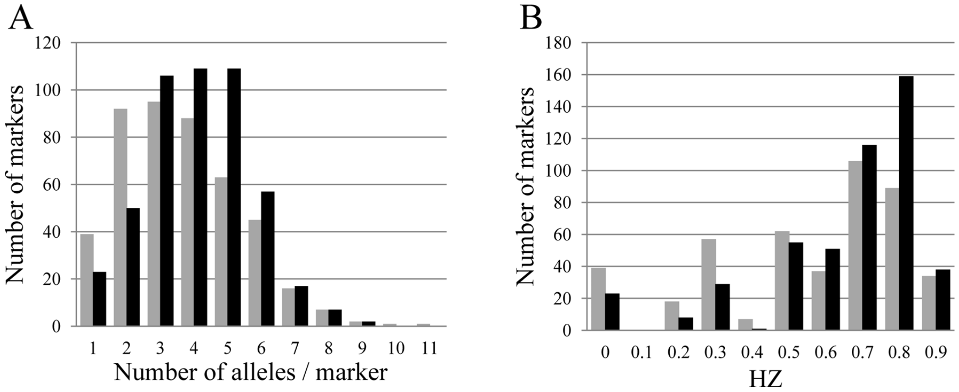

Figure 3. Allele frequency of the genome and EST microsatellite markers. a: Distribution of the number of alleles per locus; b: Distribution of HZ values. Black and gray bars represent genome and EST microsatellite markers, respectively.

species of *Eucalyptus*, including *E. grandis* and *E. upophylla* (Brondani et al. 1998, 2002, 2006), *E. globulus* (Steane 2002), *E. nitens* (Byrne et al. 1996), *E. sieberi* (Glaubitz et al. 2001) and *E. leucoxyon* (Ottwell et al. 2005). The genome- and EST-derived DNA markers developed in this study will be a useful resource to accelerate genetic analyses not only in *E. camaldulensis* but in all *Symphyomyrtus* plants.

Database and accession numbers

A web-based database that provides the nucleotide sequences and the predicted genes is available at http://www.kazusa.or.jp/eucaly/.

The ESTs were deposited in the DDBJ/EMBL/GenBank databases with the accession numbers FY782538 to FY841121 (58,584 entries). Information about the genomic sequences (contigs and singlets) and BAC clone sequences is available through international databases (DDBJ/Genbank/EMBL) under accession numbers BADO01000001 to BADO01274001 (27,4001 entries). Paired-end sequences for 3 kb and 8 kb libraries by the GS FLX sequencer are available through DDBJ Sequence Read Archive under accession numbers DRA000466 and DRA000467, respectively.

Acknowledgements

This work was supported by the Kazusa DNA Research Institute Foundation.

References

Altschul SF, Madden TL, Schäffer AA (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402

Asamizu E, Nakamura Y, Sato S, Fukuzawa H, Tabata S (1999) A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res* 6: 369–373

Ashburner M, Ball CA, Blake JA, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29

Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24: 21–25

Bao Z, Eddy SR (2002) Automated de novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res* 12: 1269–1276

Bocca SN, Magioli C, Mangeon A, Junqueira RM, Cardeal V, Margis R, Sachetto-Martins G (2005) Survey of glycine-rich proteins (GRPs) in the Eucalyptus expressed sequence tag database (ForEST). *Genet Mol Biol* 28: 608–624

Brendel V, Kleffe J (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. *Nucleic Acids Res* 26: 4748–4757

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

Brondani RPV, 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

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268: 78–94

Butcher PA, Otero A, Mcdocald 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

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* 15: 288–289

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

Chen F, Ro D, Petri J, Gershenzon J, Bohlmann J, Pichersky E, Tholl D (2004) Characterization of a Root-Specific Arabidopsis Terpene Synthase Responsible for the Formation of the Volatile Monoterpene 1,8-Cineole1. *Plant Physiol* 135: 1956–1966

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14: 1147–1159

Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175-185

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194

Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, Huala E, Knee E, Lambrecht M, Miller N, Mueller LA, et al. (2002) TAIR: a resource for integrated Arabidopsis data. *Funct Integr Genomics* 2: 239–253

Giamakis A, Kretsi O, Chinou I, Spyropoulos CG (2001) Eucalyptus camaldulensis: volatiles from immature flowers and high production of 1,8-cineole and beta-pinene by in vitro cultures. *Phytochemistry* 58: 351–355

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

Glaubitz JC, Murrell JC, Moran GF (2003) Effects of native forest regeneration practices on genetic diversity in *Eucalyptus consideniana*. *Theor Appl Genet* 107: 422–431

Grattapaglia D and Bradshaw HD (1994) Nuclear DNA content of commercially important Eucalyptus species and hybrids. *Can J For Res* 24: 1074–1078

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

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

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* 24: 3439–3452

Higgins HG (1984) Pulp and Paper. In: Hillis WE and Brown AG (eds) *Eucalyptus for wood production.* Academic Press Inc, London, pp 290–316

Huang X, Zhang J (1996) Methods for comparing a DNA sequence with a protein sequence. *Comput Appli in the Biosci* 12: 497–506

Huang X, Yang SP, Chinwalla AT, Hillier LW, Minx P, Mardis ER, Wilson RK (2006) Application of a superword array in genome assembly. *Nucleic Acids Res* 34: 201–205

Hunter S, Apweiler R, Attwood TK, et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37: D211–215

Kampranis SC, Ioannidis D, Purvis A, Mahrez W, Ninga E, Katerelos NA, Anssour S, Dunwell JM, Degenhardt J, Makris AM, et al. (2007) Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: structural insights into the evolution of terpene synthase function. *Plant Cell* 19: 1994–2005

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

Koren S, Miller JR, Walenz BP, Sutton G (2010) An algorithm for automated closure during assembly. *BMC Bioinform* 11: 457

Lukashin A, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 26: 1107–1115

Mullins KV, Llewellyn DJ, Hartney VJ, Strauss S, Dennis ES (1997) Regeneration and transformation of *Eucalyptus camaldulensis*. *Plant Cell Rep* 16: 787–791

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27: 29–34

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

Ranik M, Myburg AA (2006) Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. *Tree Physiol* 26: 545–556

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, et al. (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res* 15: 227–239

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34(Web Server issue): W435–439

Steane DA, Nicolle D, McKinnon GE, Vaillancourt RE, Potts BM (2002) Higher-level relationships among the eucalypts are resolved by ITS-sequence data. *Aust Syst Bot* 15: 49–62

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24: 1596–1599

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinform* 4: 41–54

Thamarus K, Groom K, Bradley A, Raymond CA, Schimleck LR, Williams ER, Moran GF (2004) Identification of quantitative trait loci for wood and fibre properties in two full-sib properties of *Eucalyptus globulus*. *Theor Appl Genet* 109: 856–864

Tholl D, Chen F, Petri J, Gershenzon J, Pichersky E (2005) Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from Arabidopsis flowers. *Plant J* 42: 757–771

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 122: 4673–4680

Turnbull JW, Pryor LD (1984) Choice of specific and seed sources. In: Hillis WE and Brown AG (eds) *Eucalyptus for wood production.* Academic Press Inc, London, pp 6–65

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) The Genome of Black Cottonwood, *Populus trichocarpa* (Torr & Gray). *Science* 15: 1596–1604