BMC Genomics

BioMed Central
The Open Access Publisher

Advancing Eucalyptus genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries

BMC Genomics 2011, **12**:137 doi:10.1186/1471-2164-12-137

Jorge A P Paiva (jorgep@itqb.unl.pt)
Elisa Prat (elisa.prat@toulouse.inra.fr)
Sonia Vautrin (sonia.vautrin@toulouse.inra.fr)
Mauro D Santos (mdsa@kdbio.inesc-id.pt)
Helene San-Clemente (sancle@scsv.ups-tlse.fr)
Sergio Brommonschenkel (shbromo@ufv.br)
Paulo G S Fonseca (pgsf@kdbio.inesc-id.pt)
Dario Grattapaglia (dario@cenargen.embrapa.br)
Xiang Song (xsong@ag.arizona.edu)
Jetty S S Ammiraju (jettyr@ag.arizona.edu)
David Kudrna (dkudrna@ag.arizona.edu)
Rod A Wing (rwing@ag.arizone.edu)
Ana T Freitas (atf@kdbio.inesc-id.pt)
Helene Berges (hberges@toulouse.inra.fr)
Jacqueline Grima-Pettenati (grima@scsv.ups-tlse.fr)

ISSN 1471-2164

Article type Research article

Submission date 1 June 2010

Acceptance date 4 March 2011

Publication date 4 March 2011

Article URL http://www.biomedcentral.com/1471-2164/12/137

BMC Genomics

BioMed Central
The Open Access Publisher

http://www.biomedcentral.com/info/authors/

Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries

Jorge A.P. Paiva[1,2,§], Elisa Prat[3], Sonia Vautrin[3], Mauro D. Santos[4], Hélène San-Clemente[5,6], Sérgio Brommonschenkel[7], Paulo G.S. Fonseca[4], Dario Grattapaglia[8], Xiang Song[9], Jetty S.S. Ammiraju[9], David Kudrna[9], Rod A. Wing[9] , Ana T. Freitas[4], Hélène Bergès[3], Jacqueline Grima-Pettenati[5,6]

[1]Instituto de Investigação Científica Tropical (IICT), Centro de Florestas e dos Produtos Florestais, Tapada da Ajuda, 1349-018 Lisboa, Portugal

[2]Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2781-901 Oeiras, Portugal

[3]INRA-CNRGV, Chemin de Borde Rouge, 31326 Castanet-Tolosan, France

[4]Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID/IST), R. Alves Redol 9, 1000-029 Lisboa, Portugal

[5]Université de Toulouse ; UPS ; UMR 5546, Surfaces Cellulaires et Signalisation chez les Végétaux ; BP 42617, F-31326, Castanet-Tolosan, France

[6]CNRS ; UMR 5546 ; BP 42617, F-31326, Castanet-Tolosan, France

[7]BIOAGRO - Federal University of Viçosa, Av. P. H. Rolfs, s/n - 36570-000 - Viçosa, MG - Brasil

[8]EMBRAPA Genetic Resources and Biotechnology, EPqB Final W5 NOrte, 70770-910 Brasilia, DF, Brazil

[9]Arizona Genomics Institute, School of Plant Sciences and BIO5 Institute, The University of Arizona, Tucson AZ 85721 USA

[§]Corresponding author

Email addresses:

JAPP: jorgep@itqb.unl.pt

EP: elisa.prat@toulouse.inra.fr

SV:sonia.vautrin@toulouse.inra.fr

MDS: mdsa@kdbio.inesc-id.pt

HSC: sancle@scsv.ups-tlse.fr

SB: shbromo@ufv.br

PGSF: pgsf@kdbio.inesc-id.pt

DG:dario@cenargen.embrapa.br

XS: xsong@ag.arizona.edu

JA: jettyr@ag.arizona.edu

DK: dkudrna@ag.arizona.edu

RAW: rwing@ag.arizone.edu

ATF: atf@kdbio.inesc-id.pt

HB:hberges@toulouse.inra.fr

JGP:grima@scsv.ups-tlse.fr

Abstract

Background

Eucalyptus species are among the most planted hardwoods in the world because of their rapid growth, adaptability and valuable wood properties. The development and integration of genomic resources into breeding practice will be increasingly important in the decades to come. Bacterial artificial chromosome (BAC) libraries are key genomic tools that enable positional cloning of important traits, synteny evaluation, and the development of genome framework physical maps for genetic linkage and genome sequencing.

Results

We describe the construction and characterization of two deep-coverage BAC libraries EG_Ba and EG_Bb obtained from nuclear DNA fragments of *E. grandis* (clone BRASUZ1) digested with *Hind*III and *BstY*I, respectively. Genome coverages of 17 and 15 haploid genome equivalents were estimated for EG_Ba and EG_Bb, respectively. Both libraries contained large inserts, with average sizes ranging from 135Kb (Eg_Bb) to 157Kb (Eg_Ba), very low extra-nuclear genome contamination providing a probability of finding a single copy gene ≥ 99.99%. Libraries were screened for the presence of several genes of interest *via* hybridizations to high-density BAC filters followed by PCR validation. Five selected BAC clones were sequenced and assembled using the Roche GS FLX technology providing the whole sequence of the *E. grandis* chloroplast genome, and complete genomic sequences of important lignin biosynthesis genes.

Conclusions

The two *E. grandis* BAC libraries described in this study represent an important milestone for the advancement of *Eucalyptus* genomics and forest tree research. These

BAC resources have a highly redundant genome coverage (>15x), contain large average inserts and have a very low percentage of clones with organellar DNA or empty vectors. These publicly available BAC libraries are thus suitable for a broad range of applications in genetic and genomic research in *Eucalyptus* and possibly in related species of *Myrtaceae*, including genome sequencing, gene isolation, functional and comparative genomics. Because they have been constructed using the same tree (*E. grandis* BRASUZ1) whose full genome is being sequenced, they should prove instrumental for assembly and gap filling of the upcoming *Eucalyptus* reference genome sequence.

Background

Renowned for their fast growth, valuable wood properties and wide adaptability, *Eucalyptus* species are among the most planted hardwoods in the world. The genus *Eucalyptus* includes over 700 species [1], including the most planted species *E. grandis* and *E. urophylla* (section *Transversaria*), *E. globulus* (section *Maidenaria*) and *E. camaldulensis* (section *Exsertaria*), all belonging to the subgenus *Symphyomyrtus*. Native to Australia, these fast growing trees were rapidly introduced into India, France, Chile, Brazil, South Africa, and Portugal in the first quarter of the 1800s [2] and were promptly adopted for plantation forestry. Nowadays, *Eucalyptus* and their hybrids are among the world's leading sources of woody biomass and are the main hardwoods used for pulpwood and timber. In particular, *E. grandis* is grown in tropical and subtropical regions and *E. globulus* in limited temperate regions of the world while their hybrids are among the most widely used for industrial plantations because of their rapid growth rate, their adaptability to diverse ecological conditions and their good quality wood fiber.

Because of their high commercial value, *Eucalyptus* species are major targets for genetic improvement. Nevertheless from a genetics perspective, *Eucalyptus* are still in the early stages of domestication when compared to crop species and this fact has important implications when applying genomics approaches to understand the structural and functional biology of the genome [3-5].

In the last 15 years, several studies have led to a better understanding of the *Eucalyptus* genome and the development of an important set of genetic/genomic tools, which can be used to enhance future breeding efforts. *Eucalyptus* sp. are diploid plants with a

haploid chromosome number of 11 [6-8]. Grattapaglia and Bradshaw [8] estimated the haploid genome size of Eucalyptus species to range ranging from 370 to 700 million base pairs (Mbp), with the *Symphyomyrtus* species having on average a haploid genome size of 650 Mbp. The high level of genetic diversity, the ability to generate large progeny sets, the relatively small genome size and low proportion of repetitive DNA facilitated genetic linkage mapping in *Eucalyptus* [9]. Most *Eucalyptus* linkage maps [3-5], have been constructed from highly heterozygous parents and segregating half-sib of full-sib families, with progenies up to 200 individuals often from different species. The development of co-dominant markers such as microsatellites allowed comparative mapping studies among *Eucalyptus* species. Extensive map synteny and co linearity were detected between species of the subgenus *Symphomyrtus* [10]. QTL mapping in *Eucalyptus* has been applied to the identification of genetic loci associated with variation in biomass productivity, stem form, wood properties (wood density and composition, fiber traits, bark composition), vegetative propagation, biotic/abiotic stress responses, development, foliar chemistry, inbreeding depression, and transcript level [4, 5]. Expressed sequence tag (EST) catalogues are available [11-16] and private consortia generated thousands of other ESTs. The complete nucleotide sequence of the chloroplast genome from *E. globulus* is also available [17], and is similar to other angiosperms, with an inverted repeat (IR) separated by a large single copy (LCS) region, and a small single copy (SSC) region.

Currently in progress, the first draft of a whole genome shotgun assembly of the *Eucalyptus* genome (*E. grandis*, clone BRASUZ1; www.eucagen.org) is expected to be publicly available by mid 2010. This effort will represent a major achievement and can be used as the first *Eucalyptus* reference sequence for future genomic undertakings [5].

Bacterial artificial chromosome (BAC) libraries have served as an essential genomic tool to enable positional cloning of important traits, synteny evaluation, and the development of genome framework physical maps for genetic linkage and genome sequencing [18]. A number of BAC libraries have been reported for woody plants, including forest trees such as *Liriodendron tulipifera* [19], *Pinus pinaster* [20, 21], *Populus trichocarpa* [22, 23] and *P. tremuloides* [23, 24]. The first *E. grandis* BAC library was built in Brazil in the context of the Genolyptus project, supplying full genomic clones of key genes as well as end-sequences used to derive microsatellite and STS markers to help anchoring selected BAC clones to the existing linkage maps [25]. Unfortunately, due to intellectual property restrictions, this library in not publicly available and no detailed report has been published yet about this specific work.

In this study, we report the construction and characterization of two deep-coverage BAC libraries (15-17x) from *E. grandis* (genotype "BRASUZ1") whose genome is currently being sequenced, becoming the public reference genome for the genus *Eucalyptus* (www.eucagen.org). The libraries described here i.e. EG_Ba (constructed from *Hind*III restricted nuclear DNA fragments) and EG_Bb library (constructed from *Bst*YI-digested DNA), are the first large-insert DNA libraries publicly available for genus *Eucalyptus* (http://www.genome.arizona.edu/orders/). The Roche GS FLX technology was successfully used to sequence five BAC clones allowing us to report here the whole sequence of the *E. grandis* chloroplast genome as well as the genomic sequences of important lignin biosynthetic genes.

Results and discussion

BAC library characterization

Two genomic BAC libraries (EG_Ba and EG_Bb) were constructed using partially digested (*Hind*III or *BstY*I) and size-selected nuclear DNA isolated from *E. grandis* (genotype BRASUZ1) and the pAGIBAC1 cloning vector, as described in Materials and Methods. Two libraries were constructed, using different restriction enzymes to avoid a biased distribution of clones along the *Eucalyptus* genome [26-29]. Each library, EG_Ba (*Hind*III) and EG_Bb (*BstY*I), contains 73,728 robotically picked clones arrayed into 192 384-well microtiter plates.

To evaluate the average insert size of each library, BAC DNA was isolated about 384 randomly selected clones from each library, restriction enzyme digested with the rare cutter *Not*I, and analyzed by Pulsed-Field Gel Electrophoresis (PFGE). All fragments generated by *Not*I digestion contained the 7.5 kb vector band and various insert fragments (see Figure 1a,b). Analysis of the insert sizes from the EG_Ba library showed that more than 87% of the library contained inserts >120kb while the average insert size was 157kb (Figure 2a). Analysis of the insert sizes from the EG_Bb library showed that more than 89% of the library contained inserts >111kb while the average insert size was 135kb (Figure 2b). Since the haploid genome of *E. grandis* is about 660 Mb, the library coverage is predicted to be 17 and 15 haploid genome equivalents for the EG_Ba and EG_Bb libraries, respectively: large large enough coverage to ensure these libraries will be useful for positional cloning, physical mapping and genome sequencing [30]. The estimated probability of finding any specific sequence is greater than 99.99% considering both libraries together [31].

BAC library screening

To characterize these BAC libraries and facilitate clone identification, we prepared high-density macroarrays on nylon filters from a subset of the libraries representing 8.5X and 7.5X of genome coverage for EG_Ba and EG_Bb libraries, respectively. Membranes were hybridized with a series of pooled probes representing the chloroplast and mitochondria genome as well as with probes derived from lignin biosynthesis related genes.

Contamination with extranuclear genomes was estimated at 0.55% of the total number of BAC clones in both libraries. BAC clones containing *E. grandis* chloroplast sequences represented about 0.48% of all BAC clones in both libraries, lower than the estimates for several other plant species libraries [32-37]. The mitochondrial genome was represented by 0.07% of our BAC clones; slightly higher as compared to the 0.03%, 0.012%, and 0.04% found in coffee [33], tomato [38], and banana [39] BAC libraries, respectively.

To evaluate the potential of these two BAC libraries to supply genomic sequences to contain candidate genes for cell wall biosynthesis, we screened the libraries with cDNA probes for lignin biosynthesis genes, [14, 40-42] and regulatory genes (*EgMyb1*, *EgMyb2*, *EgRAC1*) [11, 43, 44]. An average of 15.6 positive clones (Table 1) was obtained when the EG_Ba library was screened with probes derived from the following genes: *EguCAD2* (22 clones), *EguCCoAOMT* (19 clones), *EguCCR1* (13 clones), *EguF5H* (17 clones), *EguHCT* (13 clones), *EguMyb1* (10 clones) and *EguPAL* (15 clones). An average of 14.2 positive clones was obtained by probing the EG_Bb macroarray against *Egu4CL* (15 clones), *EguC3H* (5 clones), *EguC4H* (6 clones), *EguMyb2* (7 clones), *EguCOMT* (19 clones) and *EguRAC1* (33 clones) genes. The results of the macroarray

hybridization gene screening suggest an over-representation of positive clones for *CAD2, CCoAOMT, COMT* and *RAC* genes. However, the probes used in the macroarray hybridizations were relatively long allowing the possibility of cross-hybridization with other members of multigene families.

To remove false positives and target a single genetic locus, we performed an additional confirmation by PCR, using specific primer pairs designed from available *E. gunnii* available cDNA sequences. On average, the estimated proportions of 62% and 57% of the hybridization positive clones were confirmed by PCR screening for the EG_Ba and EG_Bb libraries, respectively (Table 1).

The results of the hybridization experiments compared to those obtained by PCR validation suggest that these genes may be present in duplicate or belong to multigene families present in the *Eucalyptus* genome, in agreement with the EST analysis of Rengel et al. [14] that found different unigene members for some of these genes.

Sequencing of selected BAC clones

Five BAC clones were selected, sequenced and assembled with Roche GS FLX sequencing, and Newbler assembly methodology, respectively (Table 2). They included one BAC clone containing the chloroplast genome (EG_Ba_35H24), one BAC clone randomly selected from the EG_Ba library (EG_Ba_18G23), and three BAC clones (EG_Ba_2B15, EG_Ba_11K15, EG_Bb_94G18) that were hybridization positive for three genes of interest - *EguCCR, EguCAD2 and EguRAC1* - respectively.

The shotgun sequencing of these BAC clones produced on average 29,491 high quality reads per clone sequenced with a mean read length of 261nt. These clones were sequenced to different levels of sequence coverage ranging from 44.6X to 62.1X.

On average 98% of these reads were used to assemble the full sequences of each clone into a minimum of five and a maximum of 36 contigs, for EG_Ba_35H24 and EG_Ba_18G23, respectively. The number of large contigs (>500bp) varied among clones from 4 (for clone EG_Ba_35H24) to 25 (for clone EG_Bb_94G18). These contigs were then reassembled into one (clone EG_Ba_2B15) to six (clone EG_Bb_94G18) scaffolds, allowing the reconstruction of the full sequence of all BAC clones except clone EG_Bb_94G18 (6 scaffolds). In this latter case, the presence of repetitive sequences prevented our ability to order and orient four of the six. Such a problem was already reported in barley by Wicker et al. [45]. Despite the relatively restricted number of clones sequenced, our results suggest the feasibility of using 454 sequencing for rapid and cost-effective sequencing and assembly of *Eucalyptus* BAC clones. The increased length of the 454 reads currently achievable with the Titanium chemistry (expected size ~400-550bp) should result in regions of high-quality finished genomic sequences. BAC clone sequences were deposited into GenBank (accession ID HM366540 to HM366544).

Characterization of genomic nuclear BAC clone sequences

RepeatMasker (http://www.repeatmasker.org/) was used to estimate GC content and search for interspersed repeats and low complexity DNA sequences in the four BAC sequences (588,509bp, 13 scaffolds). The four BAC clone sequences revealed a low number of transposable elements, and low percentage of low complexity sequences (Additional file 1) when compared to 10 to 35% in other plant genomes analyzed (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). However, these estimates cannot be generalized to the entire *E. grandis* genome

due to the small number of BAC clones analyzed as well as the ascertainment bias resulting from the selection of gene-containing BACs. Furthermore, it is known that the distribution of these repetitive elements vary along the genomes. Indeed, clone EG_Ba_2B15 presented the lowest repeat content (1.2% of 152,083bp) while the highest levels were found for clones EG_Bb_94G18 (5.5% of 129,018bp) and clone EG_Ba_11K15 (5.4% of 137,697bp). Six putative retroelements were found within the scaffold sequences of clone EG_Ba_18G23 (1.2% total length, 3 LINES and 3 LTRs) whereas none were identified in clone EG_Ba_2B15. Four LTR class retroelements were found within the scaffold sequences of clone EG_Bb_94G18 (2.5% total length, 3 Ty1/Copia and one Gypsy/DIRS1) and EG_Ba_11K15 (3,6% total length, 4 Gypsy/DIRS1). Low complexity sequences covered 0.8%, to 1.9% of the BAC clones sequence scaffolds analysed

Using Sputnik (Abajian, 1994, http://www.cbib.u-bordeaux2.fr/pise/sputnik.html) a total of 88 microsatellites also called simple sequence repeats (SSR) were found within the BAC scaffold sequences and these can be developed as new genetic markers. SSR markers have been extensively used in linkage analysis and comparative mapping of *Eucalyptus* species[10, 46-48], genetic fingerprinting [49], population genetics [50], and for clonal fidelity assessment [51]. One possible application of these new BAC-derived genetic markers could be the anchoring of a physical map to the available *Eucalyptus* genetic maps. Within the BAC clone sequence scaffolds we found 57 perfect Class I SSRs, that are more likely to be polymorphic [52] as SSR mutations tend to be positively correlated with SSR length [53]. This class of SSRs was found to occur on average every 10.3 kb within the sequences of the four nuclear BAC clones sequenced, from a minimum of 7.2 kb [EG_Bb_94G18] to 15.3kb [EG_Ba_11K15]. These differences could

reflect the non-random distribution of SSR in plant genomes [52, 54, 55]. The frequency of ClassI SSRs observed in this study was similar to that observed by Mun et al. [54] for selected *Medicago truncatula* BAC clone sequences, but very low when compared to that observed in another tree species *Populus,* where SSRs occurred on average every 2.5kb [55]. However, one cannot discard the possibility that the existence of gaps generated by the presence of repetitive sequences, where 454 sequencing has trouble going through, and/or due to a low sequence coverage in the region, might potentially underestimate the SSR frequency. Furthermore, the frequency of SSRs within the BAC clones is also very low when compared to that observed by Ceresini et al. [56]) and Rabello et al. [57] that reported a frequency of one Class I SSR every 2.5-2.7 kb in *Eucalyptus* cDNA libraries.

Comparison of the structure and sequence of the CCR and CAD2 genes from *E. grandis* and *E. gunnii*

Cinnamoyl CoA reductase (CCR; EC 1.2.1.44), that catalyzes the conversion of cinnamoyl CoA esters to their corresponding cinnamaldehydes and Cinnamyl alcohol dehydrogenase (CAD; EC 1.1.195) that catalyzes the conversion of these aldehydes to the corresponding alcohols are considered key enzymes in lignin biosynthesis. For instance, it was previously found that CCR down-regulated plants had lower lignin levels than controls [58, 59] and the extractability of the lignin polymer was improved in CAD down-regulated plants [60, 61]. Sequences of positive BAC clones for *CAD* and *CCR* probes were analyzed for gene identification via homology searches. Searches with BLASTN performed against a non-redundant databases (e-value cut off of $1e^{-5}$) allowed us to easily identify the *E. grandis* homologous genes to the *E. gunnii CCR* gene (within

scaffold #1 of EG_Ba_2B15) and the CAD2 gene (within scaffold#2 of clone EG_Ba_11K15).

Global sequence alignments performed using the Needle algorithm included within the Emboss package [62] allowed for the calculation of the percent identity between *E. grandis* and *E. gunnii CCR* and *CAD* sequences and to compare their intron/exon structure. Results of these structural comparisons are schematically presented in Figure 3 (a, b), and revealed that the number of exons, the intron/exon structure and junction boundaries were strictly conserved for both genes in both species. Moreover, the sequences are highly conserved particularly in the exons, with identity percentages varying between 98 and 100%. It seems that for both genes, the three first exons are slightly more conserved between the two species than the fourth and fifth exons.

Non-coding regions were also very well conserved between *E. gunnii and E. grandis CAD and CCR* sequences, respectively. The *CAD* promoter regions exhibited sequence identity of 96% for the first 2.5kb. Whereas the sequence conservation was lower between the *CCR* promoter sequences, showing 88% sequence identity within the first 500pb upstream the transcription start. The alignment was interrupted by an insertion of 444 bp in the *E. grandis* sequence and the identity level in the remaining 5' sequence dropped to 85%.

Concerning introns, identities between *E. grandis* and *E. gunnii* increased from 90% to 97%, the less conserved being intron 4, showing 90% and 93% identity for *CAD2* and *CCR* respectively. Intron 1 in the *E. grandis CAD2* gene exhibited a deletion of 71nt after position 122 as compared to the corresponding intron in *E. gunnii*.

The indels, reported here, between the *CCR* promoters and between introns 1 of *CAD2* could be used to develop markers to discriminate these two species, as they seem highly conserved within each species (Additional file 2).

The *E. grandis* chloroplast genome characterization

The large size of *E. grandis* BAC clone inserts coupled with the library screening strategy (arrays hybridized with a pool of probes selected along the *E. globulus* chloroplast genome, NCBI accession number NC_008115) allowed for the identification of BAC clones with inserts that potentially contained the entire sequence of the *E. grandis* choloroplast genome. Clone EG_Ba_35H24 was selected for sequencing since this clone was shown to be positive for several chloroplast probes and also presented an insert size close to that of the previously sequence of *E. globulus* chloroplast genome (~160kbp). The The Roche GS FLX reads were assembled into four long contigs sharing more than 99% sequence identity with the *E. globulus* chloroplast genome sequence. Due to the presence of inverted repeats (IRs) in the *E. globulus* chloroplast, a manual rearrangement of the sequences was performed based on this reference chloroplast genome [17] which allowed us to obtain a unique continuous fragment with 160,137bp (EMBL HM347959), a length that is close to that observed for *E. globulus* (160,268 bp) [17] and *Vitis* (160,928 bp) [63] but larger than the one reported earlier for *E. nitens* (151Kbp) based on restriction enzyme mapping [64]. The chloroplast genome of *E. grandis* includes a pair of inverted repeats 26,390 bp long, separated by a small, and large single copy regions of 18,478 bp and 88,879 bp, respectively. The GC-content of the *E. grandis* chloroplast genome is 36.9%, which is comparable to that of the *E.*

globulus chloroplast genome and to other tree plant plastids (e.g. 38.5% in *Pinus thunbergii* [65], 36.7% in *Populus trichocarpa* [66], 37.4% in *Vitis vinifera* [63]). Figure 4 illustrates a very high conservation of the chloroplast genomes between the *E. grandis* and the *E. globulus*. Moreover, the annotation of the *E. grandis* chloroplast genome sequences reveals that gene order is conserved in the two species. The large size of the inserts in the BAC libraries allowed us to obtain the chloroplast genome sequence in one single BAC clone. The sequencing of this BAC clone with Roche GS FLX technology was as efficient as the bridging shotgun library strategy used for *E. globulus* [17] or *Vitis* [63] chloroplast genome sequencing. Such an approach could be readily applied to study other non-nuclear genomes, such as mitochondrial genomes. The complete and annotated *E. grandis* chloroplast genome sequence was deposited into GenBank (accession ID NC_014570).

Conclusions

The two *Eucalyptus* BAC libraries described in this study represent an important milestone for the advancement of *Eucalyptus* genomics and forest tree research. These BAC resources have a highly redundant genome coverage (>15x), contain large average inserts (157 and 135kb) and have a very low percentage of clones with organellar DNA or empty vectors. This indicates that these publicly available BAC libraries are suitable for a broad range of applications in genetic and genomic research in *Eucalyptus* and possibly in related species of *Myrtaceae*, including genome sequencing, gene isolation, functional and comparative genomics. The analysis of ~0.6Mb of BAC clone sequences generated by Roche GS FLX sequencing technology provided an overview of the

composition of the *Eucalyptus* nuclear genome and the feasibility of using this high-throughput technology for low-cost and efficient sequencing and assembly of the targeted *Eucalyptus* sequences. SSRs identified within the BAC clone sequences could be used to develop new genetic markers for multiple genotyping purposes. In addition, we report the full chloroplast genome sequence of *E. grandis* (160,137 bp) allowing comparison of this genome with *E. globulus* and other plant species. Comparative analysis of the *CAD2* and *CCR* genes between *E. grandis* and *E. gunnii* showed a high conservation of the structure of genes as well as a high identity both in the coding and non coding sequences.

Methods

Plant material and DNA preparation

High Molecular Weight (HMW) DNA was prepared from young leaves of clonally propagated plants of tree BRASUZ1 grown in partial shade at the Federal University of Viçosa, Brazil. BRASUZ1 is a S1 individual (one generation of selfing) of an elite *E. grandis* tree originally derived from seed lots collected in Coffs Harbor (Australia). BRASUZ1 was developed by Suzano Papel e Celulose Co. in 1987 and confirmed as S1 by microsatellite genotyping (D. Grattapaglia, unpublished). It is a fast growing tree that flowers normally and does not exhibit any sign of inbreeding depression. This tree is being used by the *Eucalyptus* Genome project currently under production at JGI (Joint Genome Institute) for its lower genomic heterozygosity as compared to a regular outcrossed tree to facilitate assembly. Tender, expanded leaves were collected during a

time period of two months and kept frozen at -80ºC. For each extraction, approximately 50 grams of frozen leaf tissue was ground to powder in liquid nitrogen with a mortar and pestle used to prepare megabase-size DNA embedded in agarose plugs as described by Zhang et al. [67] Agarose plugs containing high molecular weight (HMW) nuclei DNA were subsequently sent submerged in ethanol 95% to the Arizona Genomic Institute (AGI) for library construction.

BAC library construction and BAC clone characterization

BAC libraries were constructed [68] using modifications recently described for *Oryza* [30]. DNA digestions were performed with 0.5 Unit *Hind*III in 100µL reaction volume (EG_Ba library) and 0.8 Unit *Bst*YI also in 100µL reaction volumeFor (EG_Bb libary) to obtain the appropriate partial digestion conditions. the EG_Ba BAC library, pulsed-field gel electrophoresis (PFGE) size-selected restriction fragments were ligated to the pAGIBAC1 vector (a modified pIndigoBAC536Blue with an additional *Swa*I site), while for the EG_Bb library, size-selected *BstY*I digested fragments were ligated to BamHI digested vector. Ligation products were transformed into DH10B T1 phage resistant *Escherichia coli* cells (Invitrogen, Carlsbad, CA) and plated on LB agar that contained chloramphenicol (12.5µg.mL^{-1}), X-gal (20 mg.mL^{-1}) and IPTG (0.1M). Clones were robotically picked into 384-well plates containing LB freezing media. Plates were incubated for 18h, replicated and then frozen at -80°C for long term storage.

To estimate insert sizes, 5µL aliquots of extracted BAC DNA were digested with 5U of *Not*I enzyme for 3hrs at 37°C. The digestion products were resolved by PFGE (CHEF-DRIII system, Bio-Rad) in a 1% agarose gel in TBE buffer 0.5×. Electrophoresis was carried out for 16 hours at 14°C with an initial switch time of 5sec, a final switch time of

15sec, in a voltage gradient of 6V.cm^{-1}. Insert sizes were compared to those of the MidRange I PFG Marker (New England Biolabs).

BAC library screening

a) High Density Filter production and hybridization

A subset (plates 1-96) of libraries EG_Ba and EG_Bb was used for screening. High density colony filters for both libraries were prepared using a Genetix Q-bot (Genetix, New Milton, Hampshire, United Kingdom). Each 22.5×22.5 cm filter (Hybond-N+; Amersham Bioscience,Piscataway, NJ, USA) contained 36,864 independent clones arrayed in a double spotted 6×6 pattern. Spotted nylon membranes were transferred to solid LB agar plates supplemented with 12.5µg/mL chloramphenicol, and bacterial clones allowed to grow overnight at 37ºC, followed by transferred to 4ºC just until colonies became confluent. Membranes were kept for 4 min onto a Whatmann 3 MM paper saturated with denaturation buffer (0.5M NaOH, 1.5N NaCl), treated for 10min at 100ºC, and neutralized for 10min on Whatmann 3 MM paper saturated with neutralization buffer (1.5M Tris-HCL pH7.4, 1.5N NaCl). Immediately membranes were incubated at 37ºC for 45 min with 250 mg/L proteinase K in 100mM Tris-HCl pH 8, 50mM EDTA, 0.5N NaCl. Finally, membranes were dried for 45min at 80ºC and UV-crosslinked (120,000 µJ.cm^{-2} for 50sec).

Labelling of organellar or nuclear probes was performed with [α-^{33}P] dCTP (Perkin-Elmer, Waltham, Massachusetts, USA) on DNA (150ng amplified DNA fragment) using the Ready-To-GO DNA Labeling kit. The dCTP (GE Healthcare, Waukesha, WI) and the non-incorporated nucleotides were removed using Illustra-ProbeQuant G50 (GE Healthcare, Waukesha, WI) according to manufacture's instructions.

Before performing hybridization, membranes were incubated for 30min at 50ºC in 6x SSC. Filters were pre-hybridized for 2 hours in 50mL of hybridization buffer (6XSSC, 5X Denhardt, 0.5% SDS, 100µg.mL^{-1} denatured salmon sperm DNA) at 68ºC. Hybridizations were carried out in high stringency conditions at 68°C overnight using 50mL of fresh hybridization buffer supplemented with a minimum of 10^7cpm purified and denatured probe per mL of buffer. Membranes were washed for 15min at 50ºC in SSC2x and SDS 0.1% buffer, followed by a second wash at 50ºC for 30 min in SSC 0.5X and SDS 0.1% buffer. Finally, they were wrapped in a plastic film, exposed to the General Purpose PhosphorImager screen (Amersham Bioscience, Piscataway, NJ, USA) for a period of three days, and finally scanned using a Storm System (Amersham Bioscience (Amersham Bioscience, Piscataway, NJ, USA), set to a resolution of 50µm.

b) Organellar DNA content estimation

To estimate the percentage of chloroplast and mitochondrial DNA content in each library, one high-density filter from each library was screened separately with a pool of five *E. globulus* probes for chloroplast genes *psbA, psbB, psbD, rbcL, ndhB* and with a pool of probes for mitochondrial genes *ccb256, ccb452, cox3* [69]. Chloroplast probes were obtained by amplification of *E. globulus* DNA using specific chloroplast primers (Additional file 3). Positive clones for chloroplast and mitochondria probes were identified and 10% of these clones were used in the PCR validation step. BAC clone (EG_Ba_35H24] was selected to be full sequenced, and was subsequently sequenced by Roche GS FLX technology (454 Life Sciences, Branford, CT, USA) by Cogenics company(Grenoble, France).

c) Screening for selected genes

Each library was screened with different probes for genes of lignification biosynthesis pathway and for three regulatory genes (*EguRAC1; EguMyb1, EguMyb2*). Characteristics of each gene probe are described in Table 1. Clone inserts were amplified using universal primers M13 (5'CAGGAAACAGCTATGACC3') and M13 reverse (5'TGTAAAACGACGGCCAGT3'), and the identity of the amplicons checked by sequencing before use in the hybridization. Hybridization positive clones were then validated by PCR using specific gene primers (Additional file 4).

BAC clones sequencing

Once selected, BAC clones were isolated, and the presence of the gene of interest was re-confirmed by PCR on isolated colony. BAC insert size was determined as described above. BAC DNA extractions, sequencing library generation, and sequencing were performed by Cogenics (Grenoble, France). BAC insert were sequenced by pyrosequencing using a Roche GS FLX Life Sciences instrument (Branford, CT, USA) [70, 71]. The Newbler software (454 Life Sciences, Branford, CT, USA) was used to perform *de novo* assembly of the Roche GS FLX. 454 These assemblies may still contain indel errors even at high coverage.

***E. grandis* chloroplast genome sequence annotation**

The *E. grandis* chloroplast genome was annotated using the *E. globulus* chloroplast genome (NCBI accession id NC_008115) as the reference genome and following the annotation pipeline detailed in Addicional file 5.

Since these two genomes correspond to circular genomes, it was not guaranteed that the linear sequence obtained from sequencing was cut at the same genome position. An important step before using Multiple Sequence Alignment (MSA) algorithms on genome annotation of circular genomes includes the circularization and rotation of the genomes that are to be aligned. MSA algorithms were developed to deal with linear genomic sequences and in this sense they are very sensitive to the location where the genomic sequence begins. To improve the MSA alignment of both *E. grandis* and *E. globulus*, these two genomes were first processed using the CSA algorithm [72]. This algorithm identified the largest chain of non-repeated longest subsequences common to these two circular genomes. The genomes were then rotated and made linear for MSA purposes.

After the pre-processing step, the Dual Organellar GenoMe Annotator (DOGMA) software package [73] was used to perform BLAST searches against a custom database of plant chloroplast genomes.

To conclude the annotation procedure, the Artemis Comparison Tool (ACT) [74] was used. With this tool it was possible to visualize and solve any the remaining annotation inconsistencies, by comparing the annotation obtained from DOGMA with the *E. globulus* chloroplast genome sequence. Each gene structure was manually checked to define putative exons.

Authors' contributions

JAPP participated in the conception of the study, in its design, and carried out all the experimental studies listed, and prepare the original draft of the manuscript. EP and SV participated in BAC macroarray preparation, hybridization and analysis. SB prepared HMW DNA agarose plugs. XS, JA, DK, RAW, produced the BAC libraries; MDS, PGSF, HSC

and ATF participate in bioinformatics analysis and annotation of BAC clones. DG check the identity of BRASUZ1 and participated in the definition of BAC libraries. HB and JGP participated in the conception of the study, in its design. All authors read and approved the final manuscript.

Acknowledgements

This work was partially supported by Fundação para a Ciência e Tecnologia (Portugal) (GenEglobwq project - PTDC/AGR-GPL/66564/2006). Authors also acknowledge the fundings from the ERA-PG EUCANET project nº ANR-06-ERAPG-10-03, from the Brazilian Ministry of Science and Technology (GENOLYPTUS project - FINEP grant 1755-01 and CNPq grant 520489/02-0) and a CNPq research fellowship to DG. The authors gratefully acknowledge Nathalie Ladouce for help on cDNA clones selection and amplification for probe preparation.

References

1. Brooker MIH: **A new classification of the genus *Eucalyptus* L'Her.** (*Myrtaceae*). *Australian Systematic Botany* 2000, 13(1):79-148.
2. Doughty RW: **The *Eucalyptus*. A natural and commercial history of the gum tree.** Baltimore USA, London UK: Johns Hopkins Universitary Press; 2000.
3. Poke FS, Vaillancourt RE, Potts BM, Reid JB: **Genomic research in *Eucalyptus*.** *Genetica* 2005, 125(1):79-101.
4. Myburg A, Potts B, Marques C, Kirst M, Gion J, Grattapaglia D, Grima-Pettenati J: ***Eucalyptus.*** In: *Genome mapping and molecular breeding in plants.* Edited by C K, vol. 7. New York, NY, USA: Springer; 2007: 115–160.
5. Grattapaglia D, Kirst M: ***Eucalyptus* applied genomics: from gene sequences to breeding tools.** *New Phytologist* 2008, 179(4):911-929.
6. Eldridge K, Davidson J, Harwood C, van Wyk G: **Eucalypt domestication and breeding**. In. Oxford, UK: Clarendon Press; 1993.
7. Potts BM, Wiltshire RJE: **Eucalypt genetics and genecology.** In: *Eucalypt ecology: individuals to ecosystems.* Edited by Williams J, Woinarski J. Cambridge, UK; 1997: 56-91.
8. Grattapaglia D, Bradshaw HD: **Nuclear-DNA content of commercially important *Eucalyptus* species and hybrids.** *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 1994, 24(5):1074-1078.
9. Grattapaglia D, Sederoff R: **Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using pseudo testcross mapping strategy and RAPD markers.** *Genetics* 1994, 137(4):1121-1137.

10. Marques CM, Brondani RPV, Grattapaglia D, Sederoff R: **Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species.** *Theoretical and Applied Genetics* 2002, 105(2-3):474-478.

11. Foucart C, Paux E, Ladouce N, San-Clemente H, Grima-Pettenati J, Sivadon P: **Transcript profiling of a xylem vs phloem cDNA subtractive library identifies new genes expressed during xylogenesis in *Eucalyptus*.** *New Phytologist* 2006, 170(4):739-752.

12. Paux E, Tamasloukht M, Ladouce N, Sivadon P, Grima-Pettenati J: **Identification of genes preferentially expressed during wood formation in *Eucalyptus*.** *Plant Molecular Biology* 2004, 55(2):263-280.

13. Paux E, Carocha V, Marques C, de Sousa AM, Borralho N, Sivadon P, Grima-Pettenati J: **Transcript profiling of *Eucalyptus* xylem genes during tension wood formation**. *New Phytologist* 2005, 167(1):89-100.

14. Rengel D, Clemente HS, Servant F, Ladouce N, Paux E, Wincker P, Couloux A, Sivadon P, Grima-Pettenati J: **A new genomic resource dedicated to wood formation in *Eucalyptus*.** *BMC Plant Biology* 2009, 9.

15. Rasmussen-Poblete S, Valdes J, Gamboa MC, Valenzuela PDT, Krauskopf E: **Generation and analysis of an *Eucalyptus globulus* cDNA library constructed from seedlings subjected to low temperature conditions.** *Electronic Journal of Biotechnology* 2008, 11(2).

16. Keller G, Marchal T, SanClemente H, Navarro M, Ladouce N, Wincker P, Couloux A, Teulieres C, Marque C: **Development and functional annotation of an 11,303-EST collection from *Eucalyptus* for studies of cold tolerance.** *Tree Genetics & Genomes* 2009, 5(2):317-327.

17. Steane DA: **Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (*Myrtaceae*).** *DNA Research* 2005, 12(3):215-220.

18. Monaco AP, Larin Z: **YACS, BACS, PACS and MACS - artificial chromosomes**. *Trends in Biotechnology* 1994, 12(7):280-286.

19. Liang HY, Fang EG, Tomkins JP, Luo MZ, Kudrna D, Kim HR, Arumuganathan K, Zhao SY, Leebens-Mack J, Schlarbaum SE *et al*: **Development of a BAC library for yellow-poplar (*Liriodendron tulipifera*) and the identification of genes associated with flower development and lignin biosynthesis.** *Tree Genetics & Genomes* 2007, 3(3):215-225.

20. Bautista R, Villalobos DP, Diaz-Moreno S, Canton FR, Canovas FM, Claros MG: **Toward a *Pinus pinaster* bacterial artificial chromosome library**. *Annals of Forest Science* 2007, 64(8):855-864.

21. Bautista R, Villalobos DP, Diaz-Moreno S, Canton ER, Canovas EM, Claros MG: **New strategy for *Pinus pinaster* genomic library construction in bacterial artificial chromosomes.** *Investigacion Agraria-Sistemas Y Recursos Forestales* 2008, 17(3):238-249.

22. Stirling B, Newcombe G, Vrebalov J, Bosdet I, Bradshaw HD: **Suppressed recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust.** *Theoretical and Applied Genetics* 2001, 103(8):1129-1137.

23. Tuskan GA, Gunter LE, Yang ZMK, Yin TM, Sewell MM, DiFazio SP: **Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa.*** *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 2004, 34(1):85-93.

24. Fladung M, Kaufmann H, Markussen T, Hoenicka H: **Construction of a *Populus tremuloides* Michx. BAC library.** *Silvae Genetica* 2008, 57(2):65-69.

25. Grattapaglia D, Alfenas A, Coelho A, Bearzoti E, Pappas G, Pasquali G, Pereira G, Colodette J, Gomide J, Bueno J *et al*: **Building resources for molecular breeding of *Eucalyptus*.** In: *International IUFRO Conference - Eucalyptus in a changing world: 2004: 2004; Aveiro, Portugal*; 2004: 20-32.

26. Zhang HB, Choi SD, Woo SS, Li ZK, Wing RA: **Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping population.** *Molecular Breeding* 1996, 2(1):11-24.

27. Chen Q, Sun S, Ye Q, McCuine S, Huff E, Zhang HB: **Construction of two BAC libraries from the wild Mexican diploid potato, *Solanum pinnatisectum*, and the identification of clones near the late blight and Colorado potato beetle resistance loci.** *Theoretical and Applied Genetics* 2004, 108(6):1002-1009.

28. Tao Q, Wang A, Zhang HB: **One large-insert plant-transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses.** *Theoretical and Applied Genetics* 2002, 105(6-7):1058-1066.

29. Wu CC, Nimmakayala P, Santos FA, Springman R, Scheuring C, Meksem K, Lightfoot DA, Zhang HB: **Construction and characterization of a soybean bacterial artificial chromosome library and use of multiple complementary libraries for genome physical mapping.** *Theoretical and Applied Genetics* 2004, 109(5):1041-1050.

30. Ammiraju JSS, Luo MZ, Goicoechea JL, Wang WM, Kudrna D, Mueller C, Talag J, Kim H, Sisneros NB, Blackmon B *et al*: **The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*.** *Genome Research* 2006, 16(1):140-147.

31. Clarke L, Carbon J: **Colony bank containing synthetic plasmids representative of of entire *Escherichia coli* genome.** *Cell* 1976, 9(1):91-99.

32. Luo MZ, Wang YH, Frisch D, Joobeur T, Wing RA, Dean RA: **Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon *Fusarium* wilt (Fom-2).** *Genome* 2001, 44(2):154-162.

33. Leroy T, Marraccini P, Dufour M, Montagnon C, Lashermes P, Sabau X, Ferreira LP, Jourdan I, Pot D, Andrade AC *et al*: **Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes.** *Theoretical and Applied Genetics* 2005, 111(6):1032-1041.

34. Stevens MR, Coleman CE, Parkinson SE, Maughan PJ, Zhang HB, Balzotti MR, Kooyman DL, Arumuganathan K, Bonifacio A, Fairbanks DJ *et al*: **Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins.** *Theoretical and Applied Genetics* 2006, 112(8):1593-1600.

35. Cavagnaro PF, Chung SM, Szklarczyk M, Grzebelus D, Senalik D, Atkins AE, Simon PW: **Characterization of a deep-coverage carrot (*Daucus carota* L.) BAC library and initial analysis of BAC-end sequences.** *Molecular Genetics and Genomics* 2009, 281(3):273-288.

36. Terol J, Naranjo MA, Ollitrault P, Talon M: **Development of genomic resources for *Citrus clementina*: Characterization of three deep-coverage BAC libraries and analysis of 46,000 BAC end sequences.** *BMC Genomics* 2008, 9.

37. Nilmalgoda SD, Cloutier S, Walichnowski AZ: **Construction and characterization of a bacterial artificial chromosome (BAC) library of hexaploid wheat (*Triticum aestivum* L.) and validation of genome coverage using locus-specific primers.** *Genome* 2003, 46(5):870-878.

38. Hamilton CM, Frary A, Xu YM, Tanksley SD, Zhang HB: **Construction of tomato genomic DNA libraries in a binary-BAC (BIBAC) vector.** *Plant Journal* 1999, 18(2):223-229.

39. Safar J, Noa-Carrazana JC, Vrana J, Bartos J, Alkhimova O, Sabau X, Simkova H, Lheureux F, Caruana ML, Dolezel J *et al*: **Creation of a BAC resource to study the structure and evolution of the banana (*Musa balbisiana*) genome**. *Genome* 2004, 47(6):1182-1191.

40. Lacombe E, Hawkins S, VanDoorsselaere J, Piquemal J, Goffner D, Poeydomenge O, Boudet AM, GrimaPettenati J: **Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: Cloning, expression and phylogenetic relationships.** *Plant Journal* 1997, 11(3):429-441.

41. Poeydomenge O, Marolda M, Boudet AM, Grimapettenati J: **Nucleotide sequence of cDNA-enconding mitochondrial malate-dehydrogenase from *Eucalyptus*.** *Plant Physiology* 1995, 107(4):1455-1456.

42. Gion JM, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C: **Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes**. *Molecular Breeding* 2000, **6**(5):441-449.

43. Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N *et al*: ***Eg*MYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis.** *Plant Journal* 2005, 43(4):553-567.

44. Legay S, Lacombe E, Goicoechea M, Briere C, Seguin A, Mackay J, Grima-Pettenati J: **Molecular characterization of *Eg*MYB1, a putative transcriptional repressor of the lignin biosynthetic pathway.** *Plant Science* 2007, 173(5):542-549.

45. Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N: **454 sequencing put to the test using the complex genome of barley.** *BMC Genomics* 2006, 7.

46. Brondani RPV, Brondani C, Tarchini R, Grattapaglia D: **Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*.** *Theoretical and Applied Genetics* 1998, 97(5-6):816-827.

47. Brondani RPV, Brondani C, Grattapaglia D: **Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers.** *Molecular Genetics and Genomics* 2002, 267(3):338-347.

48. Brondani RPV, Williams ER, Brondani C, Grattapaglia D: **A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus.** *BMC Plant Biology* 2006, 6.

49. Kirst M, Cordeiro CM, Rezende G, Grattapaglia D: **Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations.** *Journal of Heredity* 2005, 96(2):161-166.

50. Steane DA, Jones RC, Vaillancourt RE: **A set of chloroplast microsatellite primers for *Eucalyptus* (*Myrtaceae*).** *Molecular Ecology Notes* 2005, 5(3):538-541.

51. Tripathi SB, Mathish NV, Gurumurthi K: **Use of genetic markers in the management of micropropagated Eucalyptus germplasm.** *New Forests* 2006, 31(3):361-372.

52. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S: **Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential.** *Genome Research* 2001, 11(8):1441-1452.

53. Ellegren H: **Microsatellites: Simple sequences with complex evolution.** *Nature Reviews Genetics* 2004, 5(6):435-445.

54. Mun JH, Kim DJ, Choi HK, Gish J, Debelle F, Mudge J, Denny R, Endre G, Saurat O, Dudez AM *et al*: **Distribution of microsatellites in the genome of *Medicago truncatula*: A resource of genetic markers that integrate genetic and physical maps**. *Genetics* 2006, 172(4):2541-2555.

55. Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR, Tuskan GA: **Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome.** *New Phytologist* 2009, 181(2):498-503.

56. Ceresini PC, Silva C, Missio RF, Souza EC, Fischer CN, Guillherme IR, Gregorio I, da Silva EHT, Cicarelli RMB, da Silva MTA *et al*: **Satellyptus: Analysis and database of microsatellites from ESTs of *Eucalyptus*.** *Genetics and Molecular Biology* 2005, 28(3):589-600.

57. Rabello E, de Souza AN, Saito D, Tsai SM: ***In silico* characterization of microsatellites in *Eucalyptus* spp.: Abundance, length variation and transposon associations.** *Genetics and Molecular Biology* 2005, 28(3):582-588.

58. Piquemal J, Chamayou S, Nadaud I, Beckert M, Barriere Y, Mila I, Lapierre C, Rigau J, Puigdomenech P, Jauneau A *et al*: **Down-regulation of caffeic acid O-methyltransf erase in maize revisited using a transgenic approach.** *Plant Physiology* 2002, 130(4):1675-1685.

59. Piquemal J, Lapierre C, Myton K, O'Connell A, Schuch W, Grima-Pettenati J, Boudet AM: **Down-regulation of cinnamoyl-CoA reductase induces significant changes of lignin profiles in transgenic tobacco plants.** *Plant Journal* 1998, 13(1):71-83.

60. Halpin C, Knight ME, Grimapettenati J, Goffner D, Boudet A, Schuch W: **Purification and characterization of cinnamyl alcohol dehydrogenase from tobacco stems.** *Plant Physiology* 1992, 98(1):12-16.

61. Halpin C, Knight ME, Foxon GA, Campbell MM, Boudet AM, Boon JJ, Chabbert B, Tollier MT, Schuch W: **Manipulation of lignin quality by down regulation of cunnamyl alcohol dehydrogenase.** *Plant Journal* 1994, 6(3):339-350.

62. Rice P, Longden I, Bleasby A: **EMBOSS: The European molecular biology open software suite.** *Trends in Genetics* 2000, 16(6):276-277.

63. Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H: **Phylogenetic analyses of *Vitis* (*Vitaceae*) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids.** *BMC Evolutionary Biology* 2006, 6.

64. Byrne M, Moran GF, Tibbits WN: **Restriction map and material inherithance of chloroplast DNA in *Eucalyptus nitens*.** *Journal of Heredity* 1993, 84(3):218-220.

65. Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M: **Loss of all NDH genes as determined by sequencing the entire chloroplast genome of the blackpine *Pinus thunbergii*.** *Proceedings of the National Academy of Sciences of the United States of America* 1994, 91(21):9794-9798.

66. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, 313(5793):1596-1604.

67. Zhang HB, Zhao XP, Ding XL, Paterson AH, Wing RA: **Preparation of megabase-size DNA from plant nuclei.** *Plant Journal* 1995, 7(1):175-184.

68. Luo M, R.A. W: **An improved method for plant BAC library construction.** In: *Plant Functional Genomics.* Edited by E. G. Totowa, NJ.: Human Press Inc.; 2003: pp. 3-20.

69. Duminil J, Pemonge MH, Petit RJ: **A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA.** *Molecular Ecology Notes* 2002, 2(4):428-430.

70. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, 437(7057):376-380.

71. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT *et al*: **Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005).** *Nature* 2006, 441(7089):120-120.

72. Fernandes F, Pereira L, Freitas AT: C**SA: An efficient algorithm to improve circular DNA multiple alignment**. *BMC Bioinformatics* 2009, 10: 230.

73. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA.** *Bioinformatics* 2004, 20(17):3252-3255.

74. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: the Artemis comparison tool.** *Bioinformatics* 2005, 21(16):3422-3423.

Figures

Figure 1. NotI digest of random *E. grandis* BAC clones. PFGE random selected BAC clones from the a) EG__Ba and b) EG__Bb libraries. Size standards and cloning vector are indicated.

Figure 2. Size distribution of the inserts in the two BAC libraries. Histogram shows the distribution of insert sizes from random selected BAC clones from the a) EG__Ba BAC library and the b) EG__Bb BAC library.

Figure 3. Genomic structure comparison of the CCR (a) and the CAD2 (b) genomic clones between *E. grandis* and *E. gunnii*. Global alignement was performed by the Needle software (EMBOSS package). (a) The predicted *E. grandis* CCR genomic sequence found in scaffold #1 (108,677 to 111,886bp in Eg__Ba_2B15) was compared to the *E. gunnii* CCR promoter (EMBL AJ132750) linked to the genomic sequence (EMBL X97433). (b) The predicted *E. grandis* CAD genomic sequence found in scaffold #2 (3,492 to 8553bp in Eg__Ba_2B15) was compared to the *E. gunnii* CAD genomic sequence (EMBL X75480).

Figure 4. CSA algorithm output showing the conserved sequences identified between the *E. grandis* (EMBL HM347959) and E*. globulus* (NC_008115) chloroplast genomes.

TABLES

Table 1 – BAC libraries screening. cDNA probes used were either involved in the lignin biosynthetic pathway (*PAL, C4H, HCT, C3H, CCoAOMT, CCR, F5H/CAld5H, CAD*) or regulating this pathway (*MYB1* and *2*, *RAC1*).

Gene family	Gene	EMBL accession	library	Positive clones - Array hybridization (a)	PCR (b)	Ratio (b/a) %
Phenylalanine ammonia lyase (PAL)	*EguPAL_*	CT987001	EG__Ba	15	11	63
Cinnamate 4-hydroxylase (C4H)	*EguC4H*	CT988030	EG__Bb	6	2	33
4 coumarate CoA ligase (4CL)	*Egu4CL*	AJ244010	EG__Bb	15	12	80
Hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyl transferase (HTC)	*EguHTC*	CT980202	EG_Ba	13	13	100
p-coumarate 3-hydroxylase (C3H)	*EguC3H*	CT986440	EG__Bb	5	3	60
Caffeoyl-CoA *O*-methyltransferase (CCoaOMT)	*EguCCoaOMT*	AF168778	EG_Ba	19	7	37
Cinnamoyl CoA reductase (CCR)	*EguCCR*	X79566	EG_Ba	13	10	77
Ferulate 5 hydroxylase/ coniferaldehyde 5 hydroxylase (F5H/CAld5H)	*EguF5H*	CT987560	Eg_Ba	17	5	29
Caffeic acid/5-hydroxyconiferaldehyde *O*-methyltransferase (COMT)	*EguCOMT*	X74814	EG__Bb	19	12	63
Cinnamyl Alcohol dehydrogenase	*EguCAD2*	X65631	EG__Ba	22	13	59
Rho-related small GTP-binding protein	*EguRAC1*	DR410036	EG__Bb	33	3	9
R2R3 MYB transcription factor	*EguMyb1*	AJ576024	EB_Ba	10	9	90
R2R3 MYB transcription factor	*EguMyb2*	AJ576023	EG__Bb	7	6	86

Screening was made on subsets of EG_Ba and EG_Bb BAC libraries with redundant genome coverage of 8.5X and 7.5X, respectively. Number of positive clones obtained (a) and validated by PCR (b).

Table 2 –Sequencing (454FLX) and assembly of selected BAC clones.

BAC	Gene/Probe	Reads	% in the assembly	% repeat	Scaffolds	Large Contigs (>500Kb)	Observed lenght	Assembly length
EG__Ba_2B15	***EguCCR***	27,155	97.3	4.1	1	19	~145kb	147,199
EG__Ba_11K15	***EguCAD2***	30,804	97.8	1.6	2	21	~130kb	136,759
EG__Bb_94G18	***EguRAC1***	28,580	98	2.9	6	25	~120kb	-
EG__Ba_18G23	***Randomly chosen***	29,837	97.7	0.9	4	24	~160kb	174,430
EG__Ba_35H24	**Chloroplast**	31,081	98.3	1.72	1	4	~170kb	160,137

Additional files

Additional file 1 -–BAC sequences characteristics.

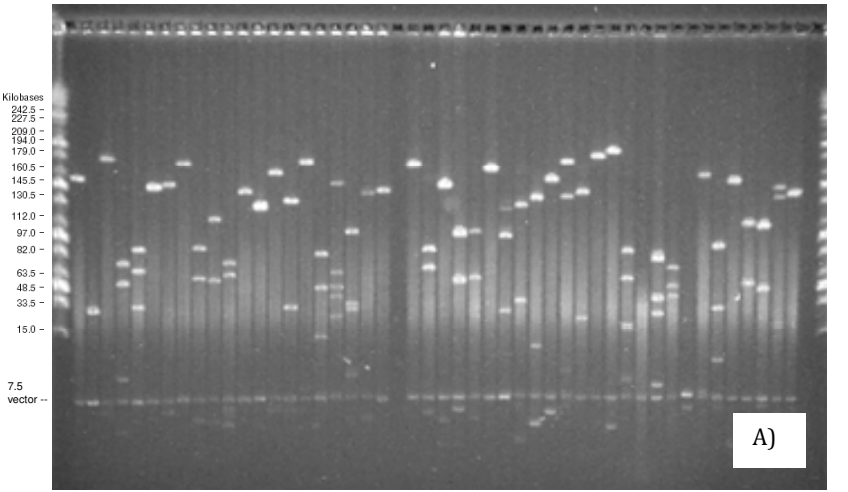
Additional file 2 - Analysis of polymorphism found at promoter of CCR gene and intron of CAD gene beetween the *E.* grandis (Transversaria section) and *E. gunnii* (Maidenaria section).

Additional file 3 - Screening of the *E. grandis* BAC libraries for chloroplast and mitochondria genomes. Primers used for PCR amplification of organelles specific probes.

Additional file 4 - Table of primers used to validate hybridization positive clones

Additional file 5 - *E. grandis* chloroplast genome annotation pipeline.

EG__Ba



EG__Bb

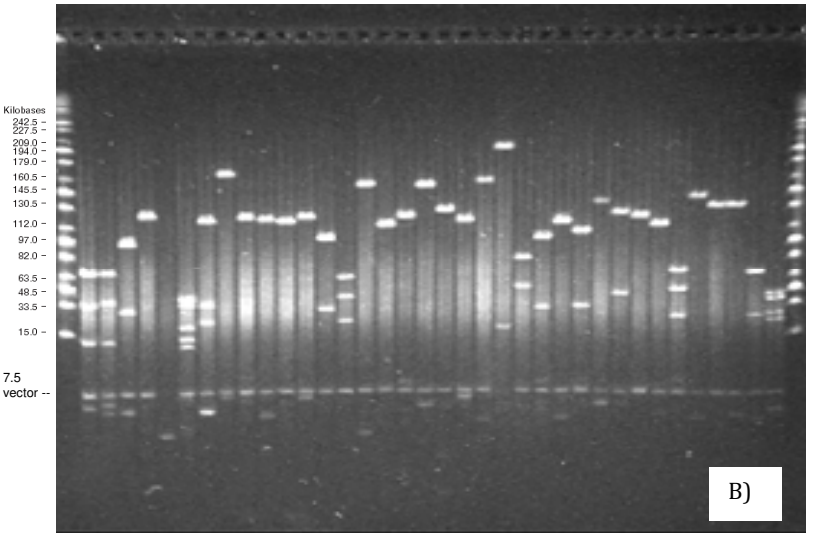


Figure 1

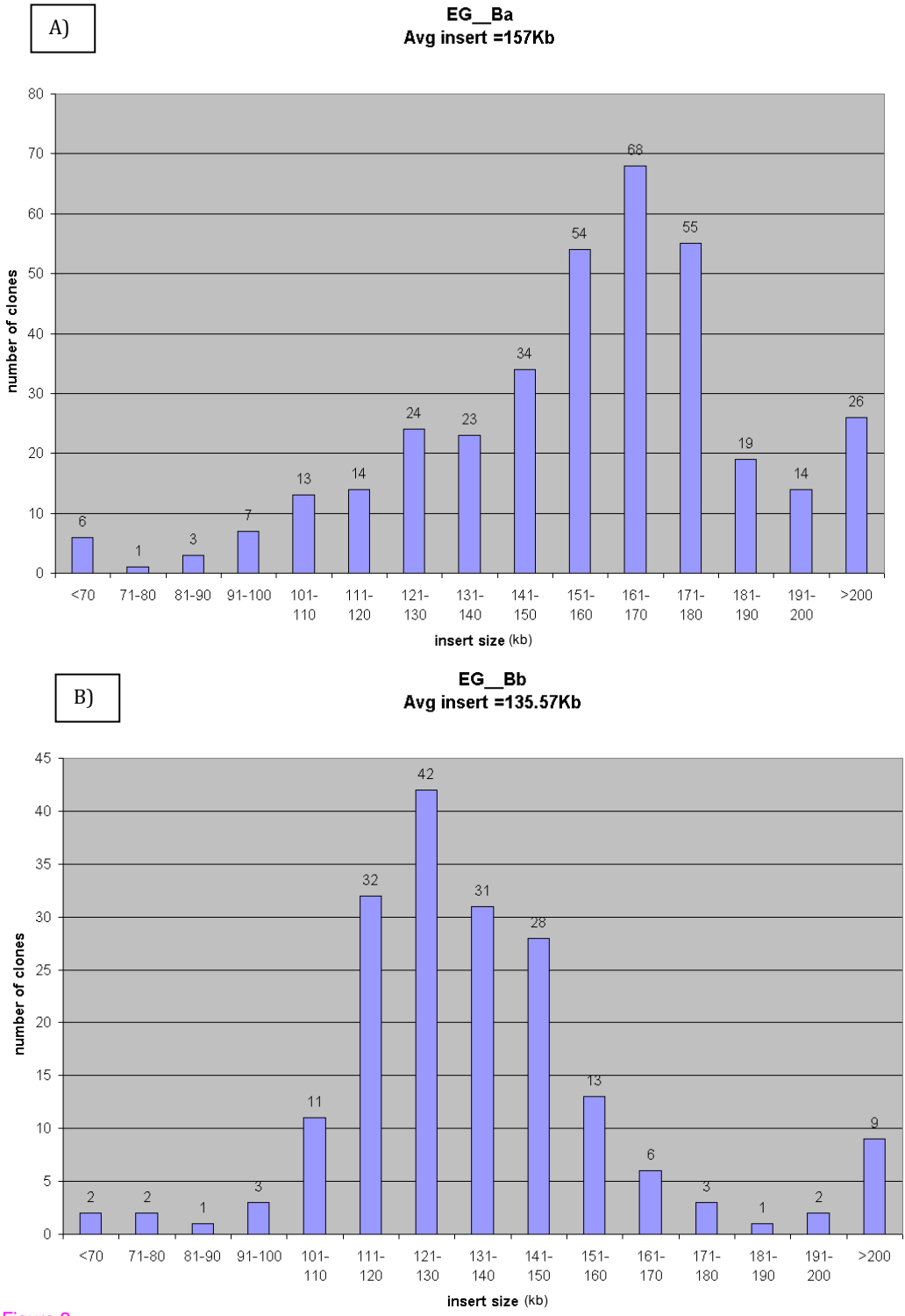
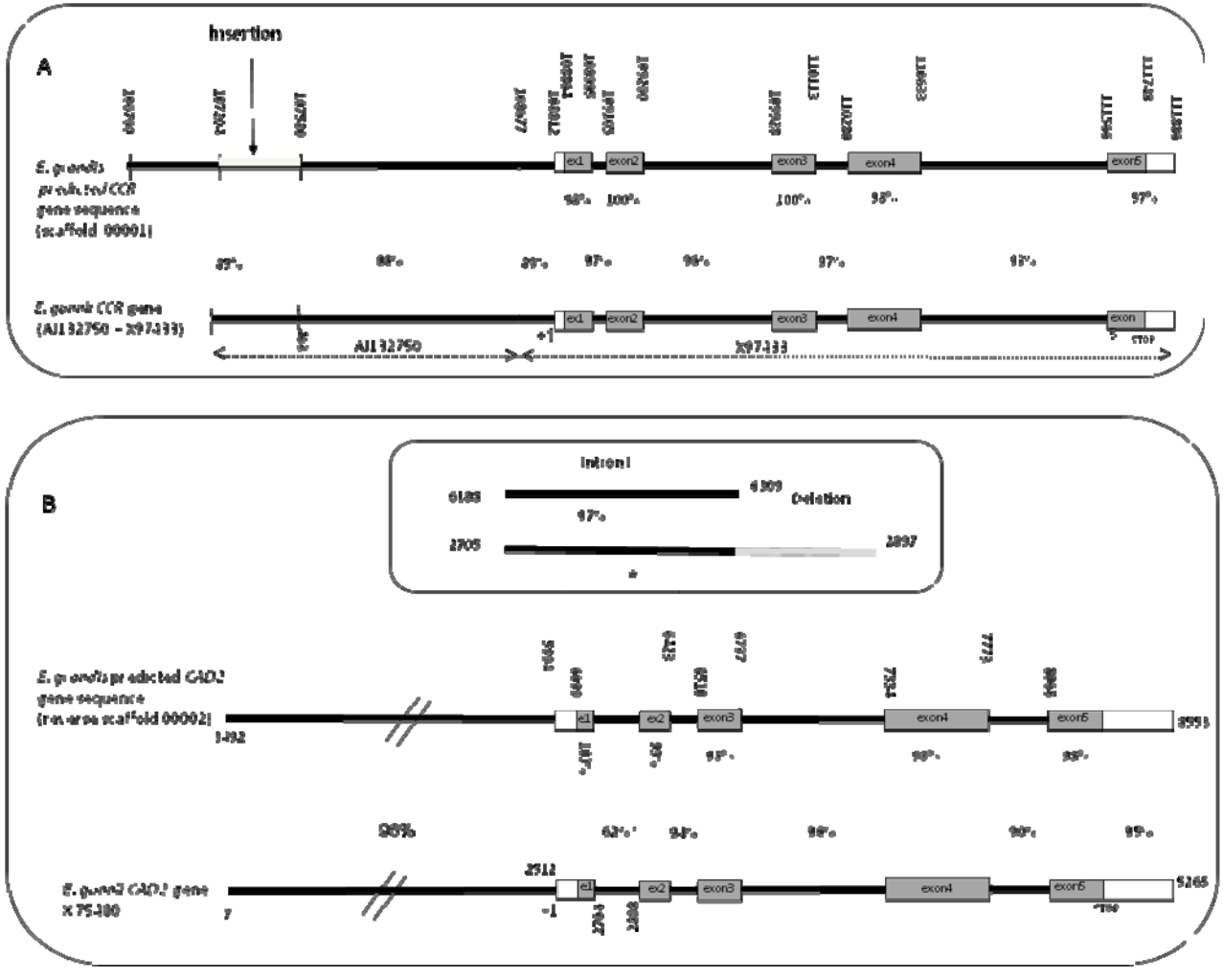

A)
EG__Ba
Avg insert =157Kb
number of clones
insert size (kb)
<70
71-80
81-90
91-100
101-110
111-120
121-130
131-140
141-150
151-160
161-170
171-180
181-190
191-200
>200
6
1
3
7
13
14
24
23
34
54
68
55
19
14
26
B)
EG__Bb
Avg insert =135.57Kb
number of clones
insert size (kb)
2
2
1
3
11
32
42
31
28
13
6
3
1
2
9

Figure 2

Figure 3


A
Insertion
E. grandis predicted CCR gene sequence (scaffold 000011
ex1
exon2
exon3
exon4
exon5
E. grandis CCR gene (AJ132750 – X97333)
AJ132750
X97333
B
Intron1
6188
6309
Deletion
97%
2705
2897
E. grandis predicted CAD2 gene sequence (reverse scaffold 00002)
1492
ex1
ex2
exon3
exon4
exon5
98%
E. grandis CAD2 gene X75180
2512
5265
-1

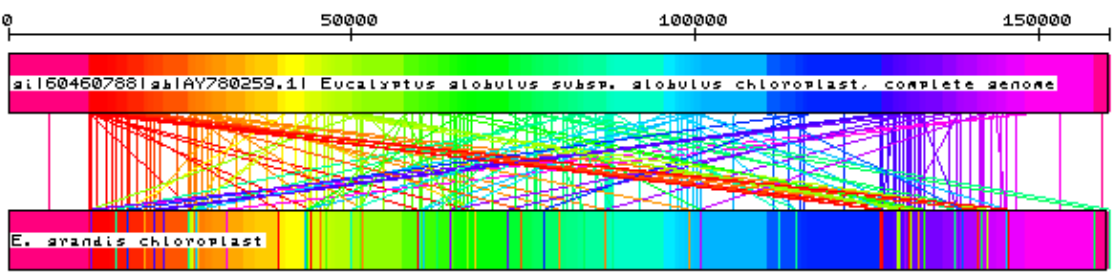

0
50000
100000
150000
gi|60460788|gb|AY780259.1| Eucalyptus globulus subsp. globulus chloroplast, complete genome
E. grandis chloroplast
202 block chains

Figure 4

Additional files provided with this submission:

Additional file 1: BAC sequences characteristics..pdf, 30K
http://www.biomedcentral.com/imedia/1083027450520143/supp1.pdf
Additional file 2: Analysis of polymorphism found at promoter of CCR gene and intro, 75K
http://www.biomedcentral.com/imedia/1998881415520143/supp2.pdf
Additional file 3: Screening of the E. grandis BAC libraries for chloroplast and mi, 26K
http://www.biomedcentral.com/imedia/1902172702520144/supp3.pdf
Additional file 4: Additional file 4.pdf, 64K
http://www.biomedcentral.com/imedia/1387710911524936/supp4.pdf
Additional file 5: E. grandis chloroplast genome annotation pipeline..pdf, 64K
http://www.biomedcentral.com/imedia/8132125305201453/supp5.pdf