The genome of *Eucalyptus grandis*

Alexander A. Myburg1,2, Dario Grattapaglia3,4, Gerald A. Tuskan5–6, Uffe Hellsten5, Richard D. Hayes7, Jane Grimwood2, Jerry Jenkins2, Erika Lindquist3, Hope Tice3, Diane Bauer5, David M. Goodstein3, Inna Dubchak5, Alexandre Poliakov5, Eschkar Mizrahi1,2, Anand K. R. Kullan1,2, Steven G. Hussey1,2, Desre Pinard1,2, Karen van der Merwe1,2, Pooja Singh1,2, Margaret Byrne1,2, Charles A. Hefer8,13, Erich Bornberg-Bauer9,14, Anna R. Kersting14,15, Kelly Vining16, Vindhya Amarasinghe16, Martin Ranik16, Rajani Raja17, Christopher Sullivan18, Elisson Romanel19,20,21, Marcio Alves-Ferreira21, Carsten Külheim22, William Foley22, Eshchar Mizrachi1,2, Anand R. K. Kullan1,2, Steven G. Hussey1,2, Desre Pinard1,2, Karen van der Merwe1,2, Pooja Singh1,2, Jacqueline Grima-Pettenati12, Jerome Salse11, Yves Van de Peer2,10, Daniel S. Rokhsar5 & Jeremy Schmutz5.

Eucalypts are the world’s most widely planted hardwood trees. Their outstanding diversity, adaptability and growth have made them a global renewable resource of fibre and energy. We sequenced and assembled >94% of the 640-megabase genome of *Eucalyptus grandis*. Of 36,376 predicted protein-coding genes, 34% occur in tandem duplications, the largest proportion thus far in plant genomes. Eucalyptus also shows the highest diversity of genes for specialized metabolites such as terpenes that act as chemical defences and provide unique pharmaceutical oils. Genome sequencing of the *E. grandis* sister species *E. globulus* and a set of inbred *E. grandis* tree genomes reveals dynamic genome evolution and hotspots of inbreeding depression. The *E. grandis* genome is the first reference for the eudicot order Myrtales and is placed here sister to the euroids. This resource expands our understanding of the unique biology of large woody perennials and provides a powerful tool to accelerate comparative biology, breeding and biotechnology.

A major opportunity for a sustainable energy and biomaterials economy in many parts of the world lies in a better understanding of the molecular basis of superior growth and adaptation in woody plants. Part of this opportunity involves species of *Eucalyptus* L’Her, a genus of woody perennials native to Australia.1 The remarkable adaptability of eucalypts coupled with their fast growth and superior wood properties has driven their rapid adoption for plantation forestry in more than 100 countries across six continents (>20 million ha), making eucalypts the most widely planted hardwood forest trees in the world. The subtropical *E. grandis* and the temperate *E. globulus* stand out as targets of breeding programmes worldwide. Planted eucalypts provide key resourceful properties for the production of pulp, paper, biomaterials and bioenergy, while mitigating human pressures on native forests. Eucalypts also have a large diversity and high concentration of essential oils (mixtures of mono- and sesquiterpenes), many of which have ecological functions as well as medicinal and industrial uses. Predominantly outcrossers with hermaphroditic animal-pollinated flowers, eucalypts are highly heterozygous and display pre- and postzygotic barriers to selfing to reduce inbreeding depression for fitness and survival.8

To mitigate the challenge of assembling a highly heterozygous genome, we sequenced the genome of ‘BRASUZ1’, a 17-year-old *E. grandis* genotype derived from one generation of selfing. The availability of annotated forest tree genomes from two separately evolving rosid lineages, *Eucalyptus* (order Myrtales) and *Populus* (order Malpighiales), in combination with genomes from domesticated woody plants (for example, *Vitis, Prunus, Citrus*), provides a comparative foundation for addressing...
fundamental evolutionary questions related to the biology of woody perennials. Moreover, the unique palaeogeographic evolution of Eucalyptus, that is, isolation from other members of the rosid clade, enables disentangling of the events that led to the modern members of the rosids by characterizing shared and unique whole-genome duplication events and syntenic gene space with other sequenced genomes. The draft genome of E. grandis suggests that the Eucalyptus genome has been shaped by an early lineage-specific genome duplication event and a subsequent high rate of tandem gene duplication.

Sequencing, assembly and annotation

We assembled a non-redundant chromosome-scale reference (V1.0) sequence for BRASUZ1 based on 6.7× whole-genome Sanger shotgun coverage, paired bacterial artificial chromosome (BAC)-end sequencing and a high-density genetic linkage map (see Methods and Supplementary Information section 1). An estimated 94% of the genome is organized into 11 pseudomolecules (605 megabases (Mb), Fig. 1). Anchoring the genome assembly to an independent linkage map revealed that the remaining 4,941 smaller unanchored scaffolds (totalling 85 Mb) correspond largely to repeat-rich sequences and segments of alternative haplotypes of the assembled chromosomes derived from regions of residual heterozygosity in the otherwise inbred BRASUZ1 genome.

The E. grandis genome encodes a large number of predicted protein-coding loci (36,376) of which 89% are expressed in vegetative and reproductive tissues (Extended Data Fig. 1) plus various classes of non-coding genes (Supplementary Information section 2). Of the 36,376 predicted proteins, 30,341 (84%) are included in gene clusters shared with other rosid lineages (Extended Data Fig. 2). Retrotransposons account for the major portion of the genome (44.5%), with long terminal repeat retrotransposons being the most pervasive class (21.9%). DNA transposons encompass only 5.6% of the genome. For this class, Helitron elements are the most abundant with an estimated 15,000 copies or 3.8% of the genome (Supplementary Information section 2).

Genome evolution and phylogeny

To address the phylogenetic position of Eucalyptus, we performed genome-wide analysis of 17 sequenced plant genomes, generating a matrix of 697,423 aligned amino acid positions from 3,268 orthologue gene clusters (Methods and Supplementary Information section 3). Studies employing broad taxon sampling but a modest number of genes have consistently recovered two very well-supported clades of eurosids—the fabids and malvids—and grouped Eucalyptus and other Myrtales with the malvids. Our analysis alternatively places Eucalyptus as a sister taxon to the eurosids (Extended Data Fig. 3) and supports the grouping of Populus and Jatropha (order Malpighiales) with malvids rather than fabids, in agreement with other recent whole-genome studies.

The discrepancy between our genome-wide analyses and the angiosperm phylogeny group (APG) consensus highlights important methodological trade-offs between sampling more characters (as in our genome-wide study) versus more taxa (as per APG)11,12.

The evolutionary history of the Eucalyptus genome is marked by a lineage-specific palaeotetraploidy event newly revealed by our genomic analysis, superimposed on the earlier palaeohexaploidy event shared by the Eucalyptus and other Myrtales with the malvids. This WGD event is considerably older, having occurred 109.9 (105.9–113.9) million years (Myr) ago (Supplementary Data 1).

Figure 1 | Eucalyptus grandis genome overview. Genome features in 1-Mb intervals across the 11 chromosomes. Units on the circumference show megabase values and chromosomes. a, Gene density (number per Mb, range 6–131). b, Repeat coverage (22–88% per Mb). c, Average expression state (fragments per kilobase of exon per million sequences mapped, FPKM, per gene per Mb, 6–41 per Mb). d, Heterozygosity in inbred siblings (proportion of 28 S1 offspring heterozygous at position, 0.39–0.93). e, Telomeric repeats. f, Tandem duplication density (2–50). g, h, Single nucleotide polymorphisms (SNPs) identified by resequencing BRASUZ1 in 1-Mb bins (g) and per gene (h, 11,656 genes); homozygous regions (~24%) and genes in green and heterozygous regions and genes in purple. Central blue lines connect gene pairs from the most recent whole-genome duplication event (Supplementary Data 1).
that the WGD event could be directly related to the origin of the clade. More precise timing will require genomic analysis of other families and genera from the Myrtales.

The Eucalyptus genome exhibits substantial conservation of synteny with other rosids as has been demonstrated for the basal rosid lineage represented by Vitis vinifera\(^5\). Extending the method previously described\(^5\) we identified 480 pairwise segments of conserved synteny between Eucalyptus and Vitis (Supplementary Information section 3). These segments include 68% of Eucalyptus genes and 76% of Vitis genes used in the analysis. The WGD in the Eucalyptus lineage relative to Vitis is clearly revealed by the 2:1 pattern in which two different Eucalyptus regions are typically collinear with a single region in Vitis. However, the gene content of these segments varies, as more than 95% of the paralogues in Eucalyptus have been lost subsequent to the WGD (a total of 5,896 genes and 76% of Eucalyptus genome). The loss of redundant genes after the WGD in Eucalyptus is accompanied by a compaction of those parts of the genome.

Eucalyptus chromosome 3, the largest single chromosome in the Eucalyptus genome, is the only chromosome that does not contain inter-chromosomal segmental duplications (Fig. 1), having fused with its WGD homologue. A similar situation occurs in Populus chromosome XVIII. Interestingly, Eucalyptus chromosome 3 and Populus chromosome XVIII nearly exclusively contain the ancestral eudicot chromosome 2 (Fig. 2c), despite their independent WGDs. There are no other examples among the currently sequenced dicotyledon genomes that contain a sole single copy of an ancestral chromosome. Moreover, in Eucalyptus and Populus, all other ancestral chromosomes appear to be dispersed and rearranged among the extant chromosomes (Fig. 2c). The conserved gene content and order (Supplementary Information section 3) on these chromosomes in two distantly related species could be due to: (1) convergent selection and positional stoichiometry of genes related to long-lived perennial woody habit that favours preservation of certain genes on these chromosomes in two distantly related species could be due to: (1) convergent selection and positional stoichiometry of genes related to long-lived perennial woody habit that favours preservation of certain genes related to perennial habit, including homologues of NAM (no apical meristem, PF02365) and senescence-associated protein (PF02365), several syntetic sets of disease-resistance genes, as well as genes related to cell-wall formation (Supplementary Data 2).
We also find that *E. grandis* has the largest number of genes in tandem repeats (12,570, 34% of the total) reported among sequenced plant genomes (Table 1 and Supplementary Information section 3). The low frequency of contig breaks separating tandem gene pairs (Extended Data Fig. 5) and conserved gene order on independent BAC clones spanning two large tandem gene arrays (Supplementary Data 3 and Supplementary Information 1) support the accuracy of the assembly across highly similar tandem copies. Tandem duplication often involves stress-response genes that are retained in a lineage-specific fashion, suggesting that tandem duplication is important for adaptive evolution in dynamically changing environments. For example, more than 80% of the S-domain receptor-like kinase (SDRLK) subfamily occurs in tandem arrays (Supplementary Data 4). There also seems to be a bias in gene retention following tandem duplication in comparison to segmental and whole-genome duplication.

Even within the genus *Eucalyptus*, tandem duplication appears to be dynamic, for example, a cluster of MYB transcription factor genes in *E. globulus* lacks four of the nine tandem duplicates found in *E. grandis* (Extended Data Fig. 6).

Despite having the same number of chromosomes \((n = 11)\) and highly co-linear genomes, *Eucalyptus* vary considerably in genome size. *E. grandis* (640 Mb) and *E. globulus* (530 Mb) represent different sections (Latuangsulatae and Maidenaria) within the subgenus Symphomyrtus, estimated to have diverged in the past 36 million years. Resequencing of the subtropical *E. grandis* (BRASUZ1) and a representative of the temperate *E. globulus* (X46, Supplementary Information section 3) revealed that many small, non-transposable element (TE)-derived changes distributed throughout the genome (164,813 regions; mean length 538 bp, median 230 bp, maximum 30,610 bp, total 88.7 Mb) account for nearly all of the genome size difference between the two species. Recent TE activity accounts for only 2 Mb of the size difference. This is in contrast to other studies in closely related plant species that report a predominant role for TEs in genome size evolution. Using sequence data from other *Eucalyptus* species taxonomically positioned around the *E. grandis–* *E. globulus* split (J. Tibbits, unpublished data), we estimate that since divergence, *E. grandis* has gained 58 Mb and lost 12 Mb, while *E. globulus* has gained 15 Mb and lost 24 Mb, suggesting more active genome size evolution than was apparent from previous estimates.

### Genetic load and heterozygosity

Eucalyptus are preferentially outcrossing with late-acting post-zygotic self-incompatibility resulting in outcrossing rates that can exceed 90%, high levels of nucleotide variation, and accumulation of genetic load and expression of inbreeding depression. A microsatellite survey of BRASUZ1 and its inbred siblings indicated putative hotspots of genetic load (Supplementary Information section 4). To investigate the distribution of preserved heterozygosity further, we resequenced an unrelated (outbred) *E. grandis* parental genotype M35D2 and 28 of its S1 offspring. The offspring were genotyped using 308,784 high-confidence heterozygous sites (within 22,619 genes) identified in M35D2 (Methods and Supplementary Information section 4). Contrary to Mendelian expectation of 50% retained heterozygosity after selfing, we observed 52% to 79% heterozygosity in the 28 S1 offspring (average of 66%). In all chromosomes except 5 and 11, heterozygosity was high (>80%) in long chromosome segments with peaks at >90% on chromosomes 6, 7, and 9 (Fig. 1d). Despite the strong bias towards heterozygosity in these regions, a small proportion of either homozygous haplotype was always present, suggesting that there are genetic backgrounds in which homozygosity of any particular gene is not lethal. One exception is on chromosome 4, where a 25-Mb region is completely devoid of one homozygous class across all surveyed genotypes (Extended Data Fig. 7 and Supplementary Information section 4).

The genetic architecture of genetic load and contribution of individual loci to inbreeding depression are largely unknown for woody perennials and present a barrier to rapid domestication via recurrent inbred mating. Our results suggest that a model of genome-wide cumulative effects of many small recessive alleles affecting overall fitness and survival best explains the architecture of inbreeding depression in *Eucalyptus*. This result is consistent with recent genome-wide selection experiments in *Eucalyptus* showing that a multifactorial model of a few hundred small effects throughout the genome contribute additively to height growth in contrast to earlier suggestions of the existence of a relatively small number of loci of larger effect as reported in several biparental QTL mapping studies.

### Lignocellulosic biomass production

Whereas woody growth habit (the ability to produce radial secondary tissues from a vascular cambium) is polyphyletic, having appeared and disappeared multiple times across more than 30 diverse taxa, secondary cell wall formation itself is highly conserved across vascular plants. Large woody plants produce secondary cell walls on a vastly different scale from that of herbaceous plants. Approximately 80% of woody biomass comprises cellulose and hemicellulose, with the remaining biomass primarily composed of lignin. A major determinant of industrial processing efficiency lies in secondary cell wall ultrastructure, which is dependent on interactions among these biopolymers. We identified putative functional homologues of genes encoding 18 enzymatic steps of cellulose and hemicellulose biosynthesis (Supplementary Information section 5). Despite the lineage-specific WGD event and the high number of genes in tandem duplications, relative and absolute expression levels suggest that most of the key enzymatic steps involve only one or two functional homologues (Fig. 3), which are highly and specifically expressed in xylem tissue. The xylem expression pattern of genes involved in sucrose catabolism suggests that *Eucalyptus* uses both direct (SUSY) and indirect (INV) pathways for the production of UDP-glucose. Notably, the two *sucrose synthase* 4 homologues (Eucgr.C00769 and Eucgr.C000769) are expressed at high levels in xylem tissue and account for 70% and 18% of total sucrose synthase expression, respectively. These genes, found on chromosome 3 in *Eucalyptus*, are part of a syntenic set of genes found on *Populus* chromosome XVIII, indicating that these genes pre-date the speciation events that separate these genera. There are 10 multigene families encoding phenylpropanoid biosynthesis genes that have expanded, mostly through tandem duplication, to include 174 genes in *E. grandis* (Supplementary Information section 5). Phylogenetic analysis and expression profiling have allowed us to define a core set of 24 genes, as well as five novel lignification candidates, preferentially and highly expressed in developing xylem (Extended Data Fig. 8 and Supplementary Information section 5). These results highlight the central role of tandem gene duplication in shaping functional diversity in *Eucalyptus* and suggest that subfunctionalization within these expanded gene families has prioritized specific genes for woody formation.

### Secondary metabolites and oils

It is generally thought that the extremely diverse array of secondary metabolites observed within *Eucalyptus* defends against a comparably diverse array of biotic pests, pathogens and herbivores encountered across its natural range. Many of the defence compounds are terpenoid based, including the commercially valuable eucalyptus oil, which is composed largely of 1,8-cineole. The conjugation of terpenes with phloroglucinol derivatives, as well as the formation of monoterpane glucose esters, leads to the myriad of defence compounds that vary across the genus. *E. grandis* has the largest observed number of terpene synthase genes among all sequenced plant genomes \((n = 113)\) compared to a range of...
Figure 3 | Genes involved in cellulose and xylan biosynthesis in wood-forming tissues of *Eucalyptus*. Relative (yellow–blue scale) and absolute (white–red scale) expression profiles of secondary cell-wall-related genes implicated in cellulose and xylan biosynthesis are shown. Sugar and polymer intermediates are shown in blue. Detailed protein names, annotation and mRNA-seq expression data are provided in Supplementary Data 5. ST, shoot tips; YL, young leaves; ML, mature leaves; FL, floral buds; RT, roots; PH, phloem, IX, immature xylem. Absolute expression level (FPKM50) is only shown for immature xylem, the target secondary cell-wall-producing tissue. DUF, domain of unknown function; GATL, galacturonosyl transferase-like; GUX, glucuronic acid substitution of xylan; HEX, hexokinase; INV, invertase; IRX, irregular xylan; PGM, phosphoglucomutase; SUSY, sucrose synthase; RWA, reduced wall acetylation; UGD, UDP-glucose dehydrogenase; UGP, UDP-glucose pyrophosphorylase; UXS, UDP-glucose synthase.

### Reproductive biology

The genus *Eucalyptus* is named for its unusual floral structure derived from the Greek κυτη, well, and καλλιφτος, covered, which refers to the operculum that covers the floral buds before anthesis. The ability to produce large amounts of pollen and seed over long generation times increases the reproductive success of woody perennials and impacts on adaptation and population genetics. Interestingly, the evolution and genetic control of the unique floral structure in *Eucalyptus* may be reflected in the expansion and deletion of genes typically associated with floral structure (for example, the APETALA1/FRUITFUL-like clade, Supplementary Information section 7). *SOC1*, a type II MADS-box gene that integrates multiple signals related to initiation of flowering, including long days, vernalization and pathways related to gibberellin signalling, has been markedly expanded in *E. grandis* compared to other angiosperms (Extended Data Fig. 10). *Eucalyptus* is a diverse genus of over 700 species distributed in a wide range of environments ranging from tropical, sub-tropical and temperate forests. This environmental heterogeneity encompasses extensive variation in the onset, season and intensity of flowering. Because of *SOC1’s* diverse roles in environmental control of flowering, the expansion and subfunctionalization of the *SOC1* subfamily may have contributed to the evolutionary diversification of *Eucalyptus* by integrating multiple signals into flowering responses relevant to different geographical zones. *Eucalyptus* may thus provide a model for the evolution of responses to divergent sets of flowering cues required for wide colonization and speciation.

### Conclusions and future directions

The availability of a high-quality reference represents a timely step forward in fundamental studies of adaptation across the diversity of habitats...
occupied by eucalypt species. The unique biology and evolutionary history of Eucalyptus are reflected in its genome, for example, the expansion of terpene synthesis genes and the large number of tandem repeats, respectively. The coincidence of a lineage-specific WGD with the origin of the Myrttales reinforces the proposed role of genome duplication in angiosperm evolution and underscores the value of additional genome sequencing of families and genera in this important rosid lineage. Future studies of variation in functional genes will provide insights into the relative influences of drift and selection on Eucalyptus evolution and identify mechanisms of speciation and adaptive divergence. Such insight will lead to improved understanding of the response of eucalypts to environmental change. Comparative analysis of the E. grandis genome with those of other large perennial species will add crucial insights into the evolutionary innovations that have made eucalypts a keystone species that shape biodiversity in diverse ecosystems. The prospect of accelerating breeding cycles for productivity and wood quality via genomic prediction of complex traits will provide increasing valuable insights into the complex connections between individual genomic elements and the extraordinary phenotypic variation in Eucalyptus.

METHODS SUMMARY

We used whole-genome shotgun sequencing (6.73x final sequence coverage from 7.7 million Sanger reads) followed by assembly in Arachne v.20071016 (ref. 36) and Illumina PE100 DNA sequencing. Lignin, cellulose, xylan, terpene and flowering-related genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10, 203–209 (2005).


Supplementary Information is available in the online version of the paper.

Acknowledgements The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. The research and writing of the manuscript was supported, in part, by the Office of Biological and Environmental Research in the US Department of Energy Office of Science under contract DE-AC05-00OR22725 as part of the US DOE Bioenergy Center. Funding for additional components of the study was provided by the Brazilian Ministry of Science, Technology and Innovation (MCTI) through its research funding agencies (CNPq and FINEP), the Brazilian Federal District Research Foundation (FAP-DF), the public-private Genolyptus network of Brazilian forestry companies, the Tree Biosafety and Genomics Research Cooperative (TBGRC, Oregon State University), South African forestry companies Sappi and Mondi, the Technology and Human Resources for Industry Programme (THIRP, UID 80118), the South African National Department of Science and Technology (DST) and National Research Foundation (NRF, UID 18312 and 86936), the Laboratoire d’Excellence (LABEX TULIP ANR-10-LABX-41), the Agence Nationale pour la Recherche (Project Tree For Joules ANR-2010-B2BE-007-01), Fundação para a Ciência e Tecnologia (FCT, P-ABBE_AQR_00001/2010), the Centre National pour la Recherche Scientifique (CNRS), the University Paul Sabatier Toulouse III (UPS). Part of this work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by the Hercules Foundation, the Flemish Government–department EWI. We also acknowledge S. Oda and E. Gonzalez of Suzano Paper and Pulp for providing genetic material of E. grandis genotypes BRASUZ1, GJ1, M5D52 and their progeny used for genome sequencing and resequencing. Forestal Mininco (Chile) for genetic material of X46, the E. globulus genotype used for genome resequencing. M. Hinchee and W. Rottmann of ArborGen for EST sequences used to support gene annotation, Sappi (South Africa) for genetic material of the population used for genetic linkage mapping, and Sappi and Mondi (South Africa) for E. grandis tissues used for RNA sequencing. We acknowledge Neil of the University of Pretoria for technical assistance with E. grandis RNA sequencing.


Author Information The E. grandis whole-genome sequences are deposited in GenBank under accession number AUSX00000000. A genome browser and further information on the project are available at http://www.phytozome.net/eucalyptus.php. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A.M. (zander.myburg@up.ac.za).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0/
METHODS

Whole-genome shotgun sequencing and assembly. All sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL automated sequencers at the Joint Genome Institute, Walnut Creek, CA. Three different sized libraries were used for the plasmid subclone sequencing process and paired-end sequencing. A total of 3,446,208 reads from the 2.6-kb sized libraries, 3,479,232 reads from the 6.0-kb sized libraries and 518,016 reads from a 32.6–40.6-kb library were sequenced. Two BAC libraries (EG_Ba, 127.5-kb insert and EG_Bb, 155.0-kb insert) were end sequenced to add an additional 294,912 reads for long-range linking.

The sequence reads were assembled using a modified version of Arachne v.20071016 (ref. 36) with parameters maxclq = 100, correct1_passes = 0, n_haplotypes = 2 and RINGE AND PURGE = True. The resulting output was then passed through Rebuilder and SquashOverlaps with parameters to merge adjacent assembled alternative haplotypes and subsequently run through another complete Arachne assembly process to finalize the assembly. This produced 6,043 scaffold sequences, with a scaffold L50 of 4.9 Mb and total scaffold size of 692.7 Mb. Scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and were removed if found to be a contaminant. Additional scaffolds were removed if they (1) consisted of >95% of base pairs that occurred as 24mers four other times in the scaffolds larger than 50 kb; (2) contained a majority of unanchored RNA sequences; or (3) were less than 1 kb in length.

For chromosome-scale pseudomolecule construction, markers from the genetic map were placed using two methods. SSR-based markers were placed using three successive rounds of e-PCR with N = 0, N = 1 and N = 3. Markers that had sequence associated with them, including SNP markers, were placed with BLAT\(^51\) and blastn\(^52\). A total of 19 breaks (16 in high coverage (≥6x), 3 in low coverage (5x≤6x)) were made in scaffolds based on linkage group discontinuity; a subset of the broken scaffolds were combined using 257 joins to form the 11 pseudomolecule chromosomes. Map joins were denoted with 10,000 repeats of the letter N (Ns). The pseudomolecules contained 605.9 Mb out of 691.3 Mb (88%) of the assembled sequence. The final assembly contains 4,952 scaffolds with a contig L50 of 67.2 kb and a scaffold L50 of 4.9 Mb and total scaffold size of 692.7 Mb. Scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and were removed if found to be a contaminant. Additional scaffolds were removed if they (1) consisted of >95% of base pairs that occurred as 24mers four other times in the scaffolds larger than 50 kb; (2) contained a majority of unanchored RNA sequences; or (3) were less than 1 kb in length.

Gene prediction. To produce the current gene set, we used the homology-based FgenesH and GenomeScan predictions. The best gene prediction at each locus was selected and integrated with EST assemblies using the PASA program\(^37\). The gene set shown in the browser was generated from the input gene models at JGI. The gene prediction pipeline was structured as follows: peptides from diverse angiosperms and ∼260,000 EST assemblies (from ∼2.9 M filtered E. grandis ESTs and ∼2.4 M EST sequences from other closely related (‘sister’) Eucalyptus species, assembled with PASA) were aligned to the genome and their overlaps used to define putative protein-coding gene loci. The corresponding genomic regions were extended by 1 kb in each direction and submitted to FgenesH and GenomeScan, along with related angio- sperm peptides and/or ORFs from the overlapping EST assemblies. These two sets of predictions were integrated with expressed sequence information using PASA\(^37\) against ∼260,000 Eucalyptus EST assemblies. The results were filtered to remove genes identified as transposon-related.

Gene family cluster and gene ontology analysis. The Inparanoid algorithm\(^38,39\) was used to identify orthologous and paralogous genes that arose through duplication events. Clusters were determined using a reciprocal best pair match and then an algorithm for adding in-paralogues was applied. The peptide sequences used were from Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Caenorhabditis elegans, C. elegans, Drosophila melanogaster, D. melanogaster, D. melanogaster, E. coli, E. coli, E. coli. These sequences were translated into integrated approach involving Interproscan\(^40\), SignalP\(^41\), Predotar\(^42\), TMHMM\(^43\) and orthology-based projections from Arabidopsis.

Green plant phylogeny. We used an integrated approach of gene ontology clustering\(^44\) and an automated workflow for phylogenetic analyses\(^45\) to reconstruct land plant phylogeny of peptide sequences. A total of 174,020 peptides encoded by single-copy protein coding orthologous nuclear genes from 17 plant genomes (Supplementary Data 7) were identified, aligned and assembled into a supermatrix resulting from conservative and liberal superalignments that retained 42.26% (697,423 amino acids) and 46.35% (764,978 amino acids) of the original 1,650,340 amino acid concatenated alignment for maximum likelihood phylogenetic reconstruction\(^46\), respectively (Supplementary Data 7). These alignments come from 3268 orthologous gene clusters with each cluster carrying single copy genes from at least 9 (50%) species.

Protein domain analysis. Domains and domain arrangements were compared within the rosids to distinguish a core set of domains and conserved trichomes present in all rosids and those shared by one or more of the four rosid lineages included in the analysis (Supplementary Data 8). Domains occurring at twice the frequency in Eucalyptus compared to the average abundance in the rosids were defined as overrepresented. If several splice variants were present for one protein, we excluded all but the longest transcript. All proteomes were scanned for domains with the Pfam\(_\text{scan}\) utility and HMMER 3.0 against the Pfam-A and Pfam-B databases\(^47,48\). For the annotation of Pfam-A domains, we used the model-defined gathering threshold and query sequences were required to match at least 30% of the defining model\(^49,50\). Pfam-B domains were annotated using an e-value cutoff of 10^{-5}. When possible, Pfam-A domains were mapped to clans and consecutive sections of the same domains were collapsed into one large pseudo-domain\(^51\). We defined domain arrangements as ordered sets of domains for each protein. For the analysis of arrangements, only Pfam-A domains were used.

Genome-wide mRNA expression profiling. To study the expression of predicted protein-coding and ncRNA genes, RNA-seq reads obtained from Illumina sequencing of seven Eucalyptus tissues (that is, shoot tips, young leaf, mature leaf, flower, roots, phloem and immature xylem, http://www.eucuegenie.org, Hefer et al., unpublished data) were mapped to the Eucalyptus genome using TopHat\(^52\) with the Bowtie algorithm\(^53\) for performing the alignment. The aligned read files were processed by Cufflinks\(^54\), with RNA-seq fragment counts (that is, fragments per kilobase of exon per million fragments mapped (FPKM)) to measure the relative abundance of transcripts. Differential ncRNA expression between the seven Eucalyptus tissues was determined using Cuffdiff\(^55\).

ncRNA analysis. To predict ncRNAs in Eucalyptus, the genome sequence was scanned using Infernal\(^56\) with the covariance models (that is, a combination of sequence consensus and RNA secondary structure consensus) of 1,973 RNA families in the RFam database v10.1 (refs 64, 65). The bit score cutoff of the Infernal search was set as the TC cutoff value that was used by RFam curators as the trusted cutoff. The Infernal search result was further filtered by an e-value cutoff of 0.01. To examine the ncRNA conservation between Eucalyptus and other plant genomes, the Eucalyptus ncRNA candidate sequences obtained from the Infernal search were used as queries to search against the genome sequences listed above using BLAT\(^51\) with a minimum coverage (that is, minimum fraction of query that must be aligned) of 80% and a minimum identity of 60%.

5′ UTR empirical curation. Approximately 2.9 million E. grandis ESTs and ∼700 million RNA-seq reads from seven diverse tissues were used to empirically curate 5′ UTR annotations. At each locus, the predicted, EST and RNA-seq derived 5′ UTR lengths were compared. An empirical annotation was prioritized over an in silico prediction and the longest empirical transcript was preferred. Those loci which had a 5′ UTR reported by only FgenesH retained their annotation as the best current empirical annotation.

Genome evolution. We used the E. grandis genome sequence information (http://www.phytozome.net/eucalyptus.php) to unravel the Myrtaceae evolutionary palaeo-history leading to the modern Eucalyptus genome structure of 11 chromosomes. Independent intraspecific (that is, parologue) and interspecific (that is, orthologue) comparisons were necessary to infer gene relationships between Eucalyptus and the other rosids genomes. We applied a robust and direct approach\(^45,46\) allowing the characterization of genome duplications by aligning the available genes (36,376) on themselves with stringent alignment criteria and statistical validation.

We used the VISTA pipeline infrastructure\(^46\) for the construction of genome-wide aligned mRNA alignments between E. grandis and other related genomes. To align genomes we used a combination of global and local alignment methods. First, we obtained an alignment of large blocks of conserved synteny between the two species by applying Shuffle-LAGAN global chaining algorithm\(^57\) to local alignments produced by translated BLAT\(^51\). After that we used Supermap, the fully symmetric whole-genome extension to the Shuffle-LAGAN. Then, in each syntenic block we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions.

Syntic regions between Eucalyptus chromosome 3 and Populus chromosome XVIII were defined as segments of contiguous sequence. Each contiguous block of scaffolds was annotated and cross-compared between the two species. Gene models within in the syntinic blocks were compared based on a sliding window representing 10 gene models with an allowance of two intercalated gene models. Genes occurring in tandem repeats on either the Eucalyptus or Populus chromosomes were counted as a single locus in either case. The constructed genome-wide pair-wise alignments can
be downloaded from http://pipeline.lbl.gov/downloads.shtml and are accessible for browsing and various types of analysis through Phytozome (http://phytozome.org).

For comparative analysis of the *E. grandis* and *E. globulus* genomes, enriched nuclei were extracted using a modified BAC library preparation protocol and DNA extracted following Tibbits et al. DNA was prepared for sequencing using Illumina TruSeq kits and 100-bp paired-end sequencing was performed on a HiSeq2000. NUCLEAR software (Gydle Inc.) was used to filter for high-quality reads that were then mapped to the *E. grandis* genome scaffold assembly. The VISION software was used to visualize assemblies and assembly metrics were computed using custom Perl, R and Shell scripts.

**Genome function analysis.** Using homology to *Arabidopsis* genes and Pfam domain analysis we identified candidate homologues for lignin, cellulose and xylan biosynthetic genes. All possible family members were identified and their gene expression evaluated in seven developing tissues of *E. grandis* using Illumina RNA-seq analysis. In particular, we analysed each gene’s expression relative to other family members/isofoms in xylem, as well as relative to the median (Supplementary Data 5 and Supplementary Data 9). Similarly, a search for conserved protein motifs for the terpene synthase gene family was conducted in eight plant genomes, including *E. grandis* (Supplementary Information section 6 and Supplementary Data 6). The amino acid sequences were aligned and truncated to compare homologous sites. A maximum likelihood tree was created, rooted by the split between two major types of terpene synthase genes, and nodes were coloured by species.

Extended Data Figure 1 | RNA-seq-based expression evidence for predicted *Eucalyptus grandis* gene models. Gene expression was assessed with Illumina RNA-seq analysis (240 million RNA sequences from six tissues, mapped to 36,376 *E. grandis* genes, V1.1 annotation). Genes were counted as expressed in a tissue if a minimum of FPKM = 1.0 was observed in the tissue. A total of 23,485 gene models (64.6%) were detected in all six tissues compared here and 32,697 (89.9%) in at least one of the six tissues. Expression profiles for individual genes are accessible in the *Eucalyptus* Genome Integrative Explorer (EucGenIE, http://www.eucgenie.org/).
Extended Data Figure 2 | Sharing of protein-coding gene families, protein domains and domain arrangements in *Eucalyptus*, *Arabidopsis*, *Populus* and *Vitis*. a, The four rosid lineages have a total of 16,048 protein coding gene clusters (from a total of 35,118 identified in 29 sequenced genomes; see Methods and Supplementary Information section 3) of which a core set of 6,926 clusters are shared among all four lineages. Of the 36,376 high-confidence annotated gene models in *E. grandis*, 30,341 (84%) are included in 10,049 clusters. *E. grandis* has 851 unique gene clusters (that is, not shared with any of the three other rosid genomes, but shared with at least one other of the 29 genomes). b, A total of 3,160 Pfam A domains are shared among the four rosid lineages, the majority of which are single-domain arrangements (3,138 shared among the four lineages). Thirteen PfamA domains were only detected in *Eucalyptus* and 392 domain arrangements are specific to *Eucalyptus* in this four-way comparison.
Extended Data Figure 3 | Green plant phylogeny based on shared gene clusters from 17 sequenced plant genomes. The phylogenetic tree was generated by RAxML analysis including at least one protein from at least half of the species per protein cluster in a concatenated MUSCLE alignment adjusted by Gblocks with liberal settings (Supplementary Data 7). The corresponding bootstrap partitions are provided at each node. The tree was rooted with Physcomitrella (a moss) as outgroup. The Myrtales lineage represented by Eucalyptus grandis is supported as sister to fabids and malvids (core rosid) clades together with the basal rosid lineage Vitales, whereas Populus trichocarpa (Malpighiales) is grouped with malvids.
Extended Data Figure 4 | Dating of the Eucalyptus lineage-specific whole-genome duplication event. 

a, Eucalyptus Ks whole-paranome (the set of all duplicate genes in the genome) age distribution. On the x axis the Ks is plotted (bin size of 0.1); on the y axis the number of retained duplicate paralogous gene pairs is plotted. 

b, Eucalyptus Ks anchor age distribution. On the x axis the Ks is plotted (bin size of 0.1); on the y axis the number of retained duplicate anchors is plotted. Anchors falling within the Ks range of 0.8–1.5 were used for absolute dating. 

c, Eucalyptus absolute dated anchors from the most recent WGD. The smooth green curve represents the maximum likelihood normal fit of dated anchors derived from the most recent WGD in Eucalyptus, whereas the blue dots represent a histogram of the raw data. The dashed line indicates the ML estimate of the distribution mode, whereas the dotted lines delimit the corresponding 95% confidence intervals. The mode of dated anchors is estimated at 109.93 Myr ago with its lower and upper 95% boundaries at 105.96 and 113.91 Myr ago, respectively. 

d, Genome duplication pattern in the core eudicot (rosid and asterid) ancestor and lineages leading to Solanum (asterid), Vitis and Eucalyptus (basal rosids) and the core rosids. The three Eucalyptus (E1–E3), Vitis (V1–V3) and Solanum (S1–S3) orthologues were generated by the shared hexaploidy event (purple box, ~130 to 150 Myr ago) and an additional set of Eucalyptus orthologues (E1’–E3’) were created in the lineage-specific WGD (orange boxes, ~110 Myr ago).
Extended Data Figure 5 | Genome-wide analysis of tandem gene assemblies. The number and distribution of contig breaks was evaluated for pairs of tandem genes (located within 50 kb of each other). a, Distribution of the number of contig breaks between gene pairs (blue bars) and cumulative proportion of gene pairs separated by contig breaks (black line). b, Distribution of the number of contig breaks per separation distance showing that the number of breaks is positively correlated with separation distance. The red line shows the distribution of distance between gene pairs with three or more contig breaks. c, Distribution of $K_s$ divergence of tandem gene pairs in clusters with exactly two tandem genes showing a gradient of similarity (that is, age of duplication) expected for authentic tandem gene pairs. d, Rate of tandem gene duplication (TD) and gene loss in *Eucalyptus grandis* (Eg), *Populus trichocarpa* (Pt), *Vitis vinifera* (Vv) and *Arabidopsis thaliana* (At). All of the rosid genomes (except *Arabidopsis*) exhibit constant rates of tandem duplication and loss. The rate of tandem gene duplication in *Eucalyptus* has been stable and consistently higher than in *Populus* and *Vitis*. 1 Myr $\sim$ 0.0026 transversions at fourfold degenerate sites, consistent with *Populus* and *Eucalyptus* having diverged $\sim$100 Myr ago.
Extended Data Figure 6 | Illumina PE100 read coverage of the ~760-kb region containing a R2R3–MYB tandem gene array. Illumina PE100 reads generated from BRASUZ1 (E. grandis) and X46 (E. globulus) were aligned to the E. grandis (BRASUZ1, V1.0) genome assembly, and insert (green bars) and sequence (blue line) coverage investigated for the ~760-kb region including a R2R3–MYB tandem array (details in Supplementary Data 3) in the E. grandis genome assembly. a, Read coverage profile of the BRASUZ1 reads mapped to the region showing 1× relative coverage across all nine of the tandem duplicates (red blocks) in the region, and b, X46 (E. globulus) reads mapped to the region showing 1× relative coverage on approximately half of the region with some tandem duplicates apparently absent from the E. globulus genome. Note that insert coverage (green bars) is relatively higher for E. globulus (X46, panel b) due to the larger insert size of the genomic library sequenced for X46 (~300 bp) than for BRASUZ1 (~150 bp).
Extended Data Figure 7 | Alternative homozygous classes observed in the 28 M35D2 siblings as a function of position on chromosomes 1–11. Several peaks of conserved heterozygosity (peaks >80%) are seen on all chromosomes except 5 and 11. A region of 25 Mb on chromosome 4 from 11 to 36 Mb is completely devoid of homozygous versions of one of the alleles (red line), but has roughly 25–32% of the siblings homozygous for the other allele (green line) and the rest heterozygous in a roughly 1:4 ratio. The blue line is the total proportion of siblings out of 28 that are heterozygous in the region. One would expect 50% under the null model, but almost the entire chromosome is biased towards heterozygosity. In several other regions (for example, chromosomes 6, 7, 9 and 10) both homozygous classes are depleted, suggesting the presence of genetic load at different loci along the two parental homologues and explaining the strong selection for heterozygosity in such regions.
Extended Data Figure 8 | Genes involved in lignin biosynthesis in woody tissues of *Eucalyptus*. Relative (yellow–blue scale) and absolute (white–red scale) expression profiles of secondary cell-wall-related genes implicated in lignin biosynthesis. Detailed gene annotation and mRNA-seq expression data are provided in Supplementary Data 9. Five novel *Eucalyptus* candidates that have not previously been associated with lignification are indicated by asterisks (Carocha *et al.*, unpublished data). ST, shoot tips; YL, young leaves; ML, mature leaves; FL, floral buds; RT, roots; PH, phloem, IX, immature xylem. Absolute expression level (FPKM<sup>50</sup>) is only shown for immature xylem.
Extended Data Figure 9 | Phylogenetic tree of R2R3 MYB sequences from subgroups expanded and/or preferentially found in woody species. A total of 133 amino acid sequences from Eucalyptus grandis (50), Vitis vinifera (34), Populus trichocarpa (40), Arabidopsis thaliana (6) and Oryza sativa (3) corresponding to three woody-expanded (subgroups 5, 6 and AtMYB5 based on Arabidopsis classification) and five woody-preferential subgroups (I through V). The latter do not contain any Arabidopsis nor Oryza sequences. Sequences were aligned using MAFFT with the FFT-NS-i algorithm (Supplementary Data 10). Evolutionary history was inferred constructing a Neighbour-joining tree with 1,000 bootstrap replicates (bootstrap support is shown next to branches) using MEGA5 (ref. 70). The evolutionary distances were computed using the Jones-Taylor-Thornton substitution model and the rate variation among sites was modelled with a gamma distribution of 1. Positions containing gaps and missing data were not considered in the analysis. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. RNA-seq-based relative transcript abundance data for six different tissues, expressed in FPKM values (fragments per kilobase of exon per million fragments mapped), are shown for each Eucalyptus gene next to each subgroup. ST, shoot tips; YL, young leaves; ML, mature leaves; FL, flowers; PH, phloem; and IX, immature xylem.
Extended Data Figure 10 | Phylogenetic tree of type II MIKC MADS box proteins. Neighbour-joining consensus tree of the type II MIKC sub-clade using protein sequences from *Eucalyptus grandis*, *Arabidopsis thaliana*, *Populus trichocarpa* and *Vitis vinifera* (Supplementary Data 11). Bootstrap values from 1,000 replicates were used to assess the robustness of the tree. Bootstrap values lower than 40% were removed from the tree. *Eucalyptus* genes are denoted with green dots, *Arabidopsis* genes with red dots, *Populus* genes with yellow dots and *Vitis* genes with blue dots. The gene model numbers from *Populus* and *Vitis* were abbreviated to better fit in the figure (P. trichocarpa, Pt; V. vinifera, Vv).