doi:10.1038/nature13308

CONTENTS

SUPPLEMENTARY NOTES 2
S1. Additional notes on genome sequencing and assembly 2
S1.1 Construction of the 8X scaffold assembly 2
S1.2 Chromosome-scale pseudomolecule construction 2
S1.3 Identification of additional alternative haplotypes 3
S1.4 Screening and final assembly release 3
S1.5 Validation of assembly of tandem duplicate arrays 4
S2. Genome annotation 6
S2.1 Transposable element analysis 6
S2.2 Gene prediction 6
S2.3 High-throughput empirical curation of 5' untranslated regions 7
S2.4 Gene Ontology (GO) annotation 9
S2.5 ncRNA annotation 9
S2.6 miRNA annotation 10
S3. Genome evolution and phylogenetics 11
S3.1 Gene family evolution 11
S3.2 Protein domain evolution 12
S3.3 Reconstruction of green plant phylogeny based on protein coding nuclear genes 13
S3.4 Segmental and whole-genome duplication 14
S3.5 Genome structural evolution 15
S3.6 Collinearity analysis and dating of the whole-genome duplication 15
S3.7 Orthologous segments of conserved synteny between *Eucalyptus* and *Vitis* 17
S3.8 Genome synteny with *Populus* 17
S3.9 Tandem gene duplication 18
S3.10 Genome comparison to *E. globulus* 19
S4. Genetic load, inbreeding depression and heterozygosity 22
S5. Lignocellulosic biomass production 24
S5.1 SCW related TFs 24
S5.2 Lignin biosynthesis 25

S5.3 Carbohydrate active enzyme (CAZyme) diversity....26
S5.4 Cellulose and xylan biosynthesis....26
S6. Secondary metabolites and oils....27
S7. MADS Box gene family....28
S8. S-domain-Receptor-Like Kinase (SDRLK) gene family....30
SUPPLEMENTARY TABLES....31
EXTENDED DATA FIGURES....51
SUPPLEMENTARY DATA....52
LITERATURE CITED....53

SUPPLEMENTARY NOTES

S1. Additional notes on genome sequencing and assembly

S1.1 Construction of the 8X scaffold assembly

Tree G7J1, an elite *E. grandis* tree from the Suzano Pulp and Paper company breeding program, originally derived from seed lots collected in Coffs Harbor (Australia), was artificially self-pollinated by emasculation and flower isolation in 1987. An unrecorded number of putative S_1 seedlings were generated and planted in 1990. In 2007 the selfed origin of the most vigorous G7J1 S_1 offspring tree was verified by genotyping it and its mother tree with a set of 15 highly polymorphic microsatellites and observing the correct expected segregation of exclusively maternal alleles in the S_1. This S1 tree was named 'BRASUZ1' (for BRAzil SUZano S_1). High-molecular-weight DNA was prepared from young leaves of clonally propagated plants of BRASUZ1 grown in partial shade at the Federal University of Viçosa, Brazil, during 2007. 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 and used to prepare megabase-size DNA embedded in agarose plugs. Agarose plugs containing high-molecular-weight nucleic DNA were subsequently sent submerged in ethanol 95% to the Arizona Genomic Institute for the construction of two BAC libraries and to JGI for plasmid and fosmid library preparation.

A total of 7,738,368 Sanger reads (see **Supplementary Table 1** for clone size breakdowns) were assembled using a modified version of Arachne v.20071016[1] with parameters maxcliq1=100, correct1_passes=0, n_haplotypes=2 and BINGE_AND_PURGE=True. Subsequent directed Arachne modules were applied to collapse adjacent heterozygous contigs (Rebuilder max_overlap_to_accept=200000, max_overlap_score=200000, max_errors_in_align=10000, end_stretch_in_align=24, use_sw_gap=True max_gap_in_swgap=6000, max_indel_in_swgap=6000, check_coverage=False and SquashOverlaps TRIM_LAP_TO=700). The entire assembly was then run through another Arachne process starting at Stage 6 Rebuilder. This reduced the contig number and allowed the final repeat resolution steps to be more effective (see **Supplementary Table 2** for scaffold and contigs totals).

S1.2 Chromosome-scale pseudomolecule construction

A total of 2,588 DArT and microsatellite markers[2] were used to identify scaffold breaks in the initial assembly. A total of 19 breaks (16 in high coverage [> 6x], three in low coverage [≤ 6x]) were made in scaffolds based on linkage group discontiguity resulting in 6,062 scaffolds in the broken assembly. DArT

markers were placed on the broken assembly and a total of 174 scaffolds (525.9 Mb) were marked representing 78% of the assembled bases. A total of 5,888 scaffolds (166.8 Mb) contained no markers. Our typical goal of map integration was to reach 90% or higher of the assembled sequence, however the initial map integration was 97.5 Mb below the target. The primary problem was that the marker placements were biased toward the high-coverage (gene-rich) scaffolds as has previously been demonstrated for DArT markers[3,4]. **Supplementary Table 3** illustrates the bias, as only 15% of the marked bases are found in low coverage scaffolds.

To solve this problem, we employed a variety of techniques to attempt to order and orient the low coverage scaffolds into the assembly. Initially, we segregated the scaffolds based on coverage and placed the markers in the two scaffold sets individually with the goal of identifying significant, but lower quality marker alignments in low coverage scaffolds that could be used to order and orient additional scaffolds. Second, other available maps[4] were used for ordering and orienting scaffolds. Finally, BAC and fosmid pairs were placed on the broken assembly to identify additional scaffolds that could be reliably joined into the assembly and to orient scaffolds with indeterminate orientation.

Using the methods above, 94 additional scaffolds were identified that could be reliably oriented and ordered into pseudomolecule chromosomes. The final set of 268 broken scaffolds were combined using 257 joins to form the 11 pseudomolecule chromosomes. Map joins were denoted with 10,000 Ns. **Supplementary Table 4** lists the 605.9 Mb in the final set of joined scaffolds (4,941 smaller scaffolds comprising 85.4 Mb remained unmapped). Care was taken to ensure that all telomeric sequence was properly oriented at the ends of the resultant chromosomes. During the construction of the chromosomes, linkage groups 1, 2, 4, 6 and 8 contained a small subset of scaffolds that were fragmented, with interspersed sections of the scaffold placing at multiple locations in the map. This was observed in the DArT marker map, as well as the microsatellite map for the five aforementioned linkage groups. The remaining six linkage groups did not exhibit scaffold fragmentation.

S1.3 Identification of additional alternative haplotypes

The 4,941 scaffolds representing 85.4 Mb of unmapped *E. grandis* sequence were repeat masked using 24mers that occurred more than 50x in the release genome to mask repetitive content. Then, a set of 21mers was obtained for each linkage group that was not shared with any other linkage group. The unshared set was subsequently used to identify shared content between a given linkage group (scaffolds A-K) and the 4,941 unmapped scaffolds. A set of 2,619 unmapped scaffolds between 5 kb and 75 kb, totaling 40.5 Mb, were aligned to the linkage group that indicated the strongest kmer association. A BLAT parameter set (-minMatch=1 -tileSize=10 -repMatch=10000 -minIdentity=85 –extendThroughN) was selected to find lower quality alignments, as the alignment between mapped and unmapped scaffolds was expected to resemble cross-species comparison. The analysis revealed that the majority of alignments with >80% identity represented fragments of alternative haplotypes of between 5 kb and 20 kb. In total, only 443 scaffolds (5.3 Mb) out of 2,619 possible scaffolds had an alignment of at least 80% identity suggesting that 6% of the 85.4 Mb of unmapped sequence represent alternative haplotypes of sequences in the 11 chromosome assemblies.

S1.4 Screening and final assembly release

We classified the remaining scaffolds in various bins depending on sequence content. We identified contamination using megablast against GenBank NR and blastp against a set of known microbial proteins. A total of 51 scaffolds were identified as prokaryotic contamination. We classified additional scaffolds as unanchored rDNA (12), mitochondrion (6), chloroplast (140) and repetitive (218). We also removed 426 scaffolds that were less than 1 kb in sequence length. The resulting final statistics are shown in **Supplementary Table 5.**

S1.5 Validation of assembly of tandem duplicate arrays

Tandem duplicated genes with high nucleotide identity can present a challenge for genome assembly, especially in highly heterozygous genomes where allele diversity and tandem gene differences are within the same range. This may lead to the separate assembly of allelic forms and generation of artificially duplicated gene loci where both allelic forms are included in the local assembly. This was anticipated from the outset of the study and mitigating steps were therefore incorporated into the design of the project. We present several lines of evidence in support of our finding that the *Eucalyptus grandis* genome has a large number of tandemly duplicated genes detected in the V1.0 assembly (**Supplementary Information S3.9)** and that these are largely correctly assembled in the current draft:

First, we selected a confirmed selfed (S_1) individual (BRASUZ1) for genome sequencing. Evidently, the BRASUZ1 genome is far less homozygous than expected of an individual produced by selfing (see **Supplementary Information S4**.) This is most likely due to genome-wide selection in the S_1 progeny against regions carrying genetic load resulting in BRASUZ1 and its live siblings being on the low end of homozygosity distribution in the family. Nevertheless, large segments of the genome (approx. 146 Mb, 24%, regions marked green in **Figure 1g** and **h, Supplementary Data 12**) are in fact completely homozygous and would therefore have presented a single haplotype for assembly. If erroneous assembly in highly heterozygous regions (i.e., failure to collapse heterozygous haplotypes) resulted in inflated estimates of tandem duplication, one would expect an elevated frequency of tandem clusters and tandem genes in such regions compared to the homozygous regions of the BRASUZ1 genome. To test this, we examined the frequency of tandem clusters and tandem genes (as defined in **Supplementary Information S3.9**) in homozygous and heterozygous regions of the genome. We show that the observed rate of tandem duplication in heterozygous regions (4.4 clusters/Mbp and 15.2 tandem duplicate genes/Mb) is not higher than in homozygous regions (4.9 clusters/Mbp and 17.0 tandem genes/Mbp, **Supplementary Data 12**). Both are in fact slightly lower in heterozygous regions than homozygous regions presumably due to the generally higher quality and completeness of assembly expected in homozygous regions.

We also show from mapping reads (Illumina PE100) generated from the BRASUZ1 genotype itself to the reference genome assembly (for the same genotype) and calling SNP genotypes from such data (see **Supplementary Information S4** and **Supplementary Data 12)**, that heterozygous regions of the genome (**Figure 1g** - purple blocks) are not interspersed with genes that appear homozygous (**Figure 1h** - green lines) as would be expected if there was widespread incorporation of alternative haplotypes as pseudo-tandem repeats in heterozygous regions. Only 88 (0.75%) of 11,656 genotyped genes of BRASUZ1 have homozygous genotypes in otherwise heterozygous regions (**Supplementary Data 12)**.

To further verify the correct assembly of tandem duplicates, we performed a genome-wide analysis of contiguity of assembly in tandem gene regions (**Extended Data Figure 5**). Pairs of genes in previously identified tandem gene regions were interrogated for possible misassembly. For each gene pair (<50kb separation) in the tandem gene regions, the number of contig breaks between the genes and the ancestry (inbred, outbred, or unknown region) was determined. A total of 7,279 gene pairs met this criteria and of these 2,624 (36%) were of undetermined ancestry, 1,331 (18%) fell in inbred regions and 3,323 (46%) fell in outbred regions. The contiguity of the tandem gene regions is high. **Extended Data Figure 5a** shows that 65% of the gene pairs contain no contig breaks between them indicating that they are on one assembled contig. Furthermore, the number of contig breaks separating tandem duplicates is proportional to the distance between the genes (**Extended Data Figure 5b)**, as would be expected for any pair of genes (duplicated or not).

Widespread artificial creation of tandem duplicates from alterative haplotypes in heterozygous regions would lead to an excess of tandem gene pairs with high sequence similarity. We evaluated Ks divergence

for all tandem clusters with exactly two members (most relevant when considering possible misassembly, **Extended Data Figure 5c**). Only about 8% of these exhibit less than 2% divergence, suggesting a strong upper limit on the proportion that could be due to misassembly of alleles. Furthermore, a gradient of similarity is observed consistent with the expected age distribution of authentic tandem duplicates added at an approximately constant rate since divergence from other rosid ancestors (**Extended Data Figure 5d** and **Supplementary Information S4**).

To further validate the accuracy of the Sanger sequence assembly across loci with highly similar tandem gene copies and confirm the absence of putative alternative allelic forms, we analysed BAC clones (from a library of the BRASUZ1 genome) covering two arrays of duplicated genes. First, we analysed an array of nine predicted R2R3MYB genes (**Supplementary Data 3**, Eucgr.J01553.1, Eucgr.J01555.1, Eucgr.J01569.1, Eucgr.J01573.1, Eucgr.J01574.1, Eucgr.J01576.1, Eucgr.J01578.1, Eucgr.J01580.1, Eucgr.J01586.1) spanning a genomic region of 760 kb on chromosome 10. The nine predicted CDSs show higher than 95% pairwise identity (**Supplementary Data 3)**. We selected from the EG_Ba BAC library constructed from *Hind*III restricted nuclear DNA fragments of the BRASUZ1 genome[5], four BAC clones (EG_Ba337C19, EG_Ba086k13, EG_Ba068F10 and EG_Ba029N23) which cover the region of the genomic R2R3 MYB cluster. We designed primers to amplify specific regions of each of these nine genes in the corresponding BAC clones. Specific amplicons were obtained and used for Sanger sequencing using the corresponding forward and reverse primers (**Supplementary Data 3**). Obtained sequences were aligned against the *E. grandis* (BRASUZ1) genome assembly using CLUSTALW[6], and in the case of the CDS, keeping the bases in codons when opening gaps. Identity values were calculated using the SIAS webpage (http://imed.med.ucm.es/Tools/sias.html) considering only the region comprised in the shortest sequence in the alignment and without considering gaps. Visualization of the position of the primers, amplicons and gene models in the chromosome and in the BAC clones was performed using Fancygene[7]. The analysis of the MYB gene cluster revealed that all of the amplified BAC sequences showed perfect identity and contiguity with the genome sequence assembly confirming that the *E. grandis* genome was properly assembled at least in this region of 760 kb (**Supplementary Data 3)**.

Similarly, we investigated an array of eight caffeic acid O-methyltransferase (COMT) genes (Eucgr.K00949, Eucgr.K00950, Eucgr.K00951, Eucgr.K00953, Eucgr.K00954, Eucgr.K00955 Eucgr.K00956 and Eucgr.K00957) covering a genomic region of 185 kb on chromosome 11 **(Supplementary Data 3).** One BAC clone (EG_Ba70107) spanning all eight COMT genes was analysed. Amplicons were designed and analysed for the eight COMT genes and, in all cases, the orientation and alignment of the amplified BAC sequences to the genome sequence validated the accuracy of assembly over the highly duplicated genomic region spanning 185 kb.

Finally, to validate the correct assembly of these highly duplicated gene regions, we examined the alignment of 235 million Illumina PE100 reads derived from BRASUZ1 (see **Supplementary Information S3.10**) to the Sanger sequence assembly and specifically extracted those mapping to the assembled R2R3 MYB gene cluster. We used Nuclear and Vision software (Gydle Inc.) to: i) correct a small number of local assembly errors (likely due to low coverage, Sanger sequencing artefacts in homopolymer/SSR regions and artifactual sequences in contig edges), ii) bridge gaps between contigs and iii) validate sequence connectivity (metric of reads connecting adjacent segments) to create a near-finished sequence of the region (two gaps and two weak connectivity points remain). The analysis showed that the ~760 kb region, spanning Contigs J-0291 to J-0301 in the *E. grandis* assembly, to be largely correctly assembled with consistent 1X coverage across the entire region (**Extended Data Figure 6a**). We similarly mapped X46 (*E. globulus*) derived Illumina PE100 reads (See **Supplementary Information 3.10**) to the ~760 kb BRASUZ1 region and show that about 350 kb is present in *E. globulus*, with a consistent 1X coverage (**Extended Data Figure 6b**) for five of the nine MYB genes and complete absence of coverage of the others, suggesting that much of the tandem duplication is specific to the BRASUZ1 (*E. grandis*) lineage.

S2. Genome annotation

S2.1 Transposable element analysis

The transposable element (TE) annotation was carried out by a combination of several specialized tools. The initial annotation step was performed using the REPET pipeline[8], a comprehensive collection of programs for *de novo* identification and classification of repetitive elements in the genome. The starting point is an all-versus-all similarity search of the genome (Phytozome version 8.0; assembly 201) against itself using overlapping windows of 200 kb. The pair-wise matches were then clustered and, from the ensuing groups, multiple alignments and corresponding consensi were derived. This procedure permits the construction of a TE database used to rescan the genome and annotate according to an established hierarchical TE classification system[9].

In addition to the general annotation outlined, an *in silico* prediction was performed to refine the LTR retrotransposon superfamily using the combination of LTRharvest[10] and LTRdigest[11]. The former was used to scan the genome for full-length LTR retrotransposons based on structural features and the latter to refine the prediction by taking into consideration protein domain information and characteristic internal features, such as the primer binding site and polypurine tract. There was considerable prediction overlap, but elements found by only one of these tools were still retained given the advantages to find both canonical elements (LTRdigest) and degraded remnants of ancient LTR retrotransposons. Homology-based TE identification was performed using RepeatMasker[12] against plant-specific repeat families in RepBase (v17.11)[13].

About half of the *E. grandis* genome (11 chromosome scaffolds) is composed of transposable elements (TE), reinforcing the correlation between genome size and TE abundance when compared to other angiosperms[14]. As in the majority of plant species, the retrotransposons (class I TEs) account for the major portion of the genome (44.5%), with long terminal repeat (LTR) retrotransposons being the most pervasive class I elements (21.9%). The proportion of the genome occupied by DNA transposons (class II TEs) is lower (5.6%) likely due to their transposition strategy. The complete dissection of TE distribution and categorization per chromosome can be found in **Supplementary Data 13.**

S2.2 Gene prediction

A *de novo* repeat library was made by running RepeatModeler (Arian Smit, Robert Hubley) on the genome to produce a putative library of repeat sequences. Sequences with Pfam domains associated with non-transposable element functions were removed from the library of repeat sequences and the library was then used to mask ~38% of the genome with RepeatMasker.

We aligned ~2.9M *E. grandis* EST sequences and ~2.2M EST sequences from sister *Eucalyptus* species (**Supplementary Table 6**) using PASA (Brian Haas[15]), which aligns ESTs to the genome assembly sequence via gmap, then filters hits to ensure proper splice boundaries. This produced datasets of 152,659 and 109,198 candidate gene loci, respectively. A dataset of 218,939 cDNAs from diverse *Eucalyptus* species were aligned separately by PASA to produce 27,690 candidate loci also included in the full gene modeling pipeline. However, due to the proprietary nature of some of these sequences (ArborGen Dataset), they have been removed from downstream results, including all shared data files and display in the Phytozome *Eucalyptus* Gbrowse tool.

To produce the current gene annotation set, the Phytozome gene prediction pipeline included the following components: protein sequences from diverse angiosperms (*Arabidopsis thaliana*, *Oryza*, soybean, *Populus trichocarpa* and *Vitus vinifera*) were aligned to the genome by BLASTX and extended by the EXONERATE algorithm[16]. Similarly, the full set of ~290,000 predicted EST assemblies from PASA analysis were run through the FgenesH+ and GenomeScan algorithms[17,18], with parameters

allowing up to 2 kb wiggle room. The best of these predictions at each gene locus was selected based on a weighted scoring of metrics including its predicted peptide to seed peptide BLASTP score divided by the MBH seed score (c-score), EST coverage (**Supplementary Table 7**), protein homolog coverage and degree of support for predicted exon-intron boundaries. This predicted dataset was extended by a second run through the PASA algorithm to model UTR and alternative splice variants based on EST support and the final results were filtered to remove genes identified as transposon-related or overlapping the RepeatModeler *de novo* repetitive region predictions by greater than 20%.

Further filtering of this initial V1.0 annotation dataset was done, employing stricter c-score and peptide homology coverage thresholds, especially when considering partial transcripts missing a modeled start or stop codon. EST support was also examined to check that aligned coverage followed the same intron splicing pattern as the gene model. In total, 8620 low-confidence gene models were removed from the original V1.0 annotation to produce a high-confidence gene set of 37,376 gene loci (V1.1 annotation, **Supplementary Table 8, Supplementary Table 9**) representing 17,883 gene family clusters in the Phytozome V9.1 release (http://www.phytozome.net/). Of these, a core set of 6,738 family clusters have membership from rosid genomes represented by *Eucalyptus*, *Populus*, *Vitis* and *Arabidopsis* (**Supplementary Table 10**).

We further validated the expression of the predicted genes using deep Illumina RNA-Seq expression profiling (PE50, 240 million) of vegetative and reproductive tissues of *E. grandis.* RNA-seq reads obtained from Illumina sequencing of six *Eucalyptus* tissues (i.e., young leaf, mature leaf, shoot tips, flower, immature xylem and phloem) were mapped to the *Eucalyptus* genome (See **Methods** in main text. The aligned read files were processed by Cufflinks[19], with RNA-Seq fragment counts (i.e., Fragments Per Kilobase of exon per Million fragments mapped - FPKM) to measure the relative abundances of transcripts (**Extended Data Figure 1**). Expression was detected (FPKM>1) in all six tissues profiled for a total of 23,485 gene models (64.6%) and 32,697 (89.9%) were detected in at least one of the six tissues.

S2.3 High-throughput empirical curation of 5' untranslated regions

The *E. grandis* 5' untranslated regions (UTRs) were predicted simultaneously with protein-coding loci using FGenesH[20]. *In silico* 5' UTR annotation uses *Arabidopsis* derived genomic features such as the TATA-Box and oligomer frequencies to computationally assign a putative transcription start site (TSS) and thus delimit the 5' UTR[20,21]. False positive estimates for this approach range from 1/700 to 1/1000 bp and there is a bias toward TATA promoter sensitivity[21]. In an attempt to circumvent these biases, particularly in cases of multiple possible TSSs per locus, the most 3' annotation is preferred, possibly truncating the 5' UTR model. Due to the low positive predictive value, sensitivity is evidently low with only ~57% of the current gene models assigned a 5' UTR.

5' UTRs are empirically distinguishable by transcript support such as ESTs[15,22], full length cDNA[23-25] and next-generation sequencing (NGS) RNA-Seq data. We hypothesize that a subset of the predicted 5' UTR annotations are truncated and that RNA-Seq and EST expression evidence better delimit the start of transcription at these loci. High confidence TSS distribution (TSSD)[26] annotation is still rare in plants and requires an exact protocol such as DeepCAGE[27]. However, the most 5' position of continuous high-quality, full-length expressed transcripts empirically refers to a position within this distribution[15,23].

We aimed to empirically curate a maximum number of *E. grandis* 5' UTRs using available transcript evidence, specifically EST and high-throughput RNA-Seq data. This includes ~2.9 million *E. grandis* ESTs (454 reads produced by the DOE-JGI from the BRASUZ1 genotype) and 240 million paired-end RNA-Seq data from seven diverse tissues of *E. grandis*[28]. We analysed the current Phytozome[29] *E. grandis* 5' UTR models and assessed the concurrence of empirical evidence and FGenesH[20] predictions. We selected a prioritized set of 5' UTR models from the above three sources (see Methods below). We

expect that including and prioritizing empirically substantiated models will enrich the total number and quality of *E. grandis* 5' UTR models. In lieu of TSSD annotations, the TSSs inferred by these models will be referred to as distal TSSs (dTSS). We provide **Supplementary Data 14** defining each of the *in silico* predictions and two empirical sources of 5' UTR annotations.

Methods

The Phytozome[29] v8.0 ("Egrandis_201") unmasked genome assembly and gene-exon gff3 file were downloaded from the Phytozome *E. grandis* ftp repository (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Egrandis/). Approximately 2.9 million *E. grandis* ESTs, assembled by *PASA*[15] into 90,266 locus-associated EST assemblies, were downloaded per request from the ORCAE curation platform[30]. 5' UTRs predicted by FGenesH and defined by the Phytozome gene models are referred to as "*FGH*" 5' UTRs in the remainder of the text; those inferred by the PASA EST assembly data are referred to as "*PASA*" 5' UTRs; those inferred by RNA-Seq[28], as "*NGS*" 5' UTRs; and the most 3' initiation codon per locus is referred to as "*INIT*".

Paired-end Illumina RNA-Seq data for immature xylem, shoot tip, phloem, mature leaf, young leaf, flower and root tissues were mapped individually to the genome[28]. The resulting SAM alignments of the ~700 million reads were merged into a bulk alignment file using SAMtools[31], which was subsequently filtered for a mapping quality of at least one (less than ten possible hits in the genome) and for introns less than 6500 bp, as extracted by the extended CIGAR[31] string. The remaining reads were reconverted to BAM format. Igvtools[32] was then used to convert the BAM to the per-base coverage wiggle format. All possible splicing events were extracted from the CIGAR strings, which were submitted in unison with the bulk BAM file to bam2ssj[33] for splice junction quantification. Those splice-events with a splicing index and PSI value of at least 0.4 and raw splice-junction coverage of at least 5, were retained. The wiggle represented coverage and high confidence splice junctions were then used to delimit the dTSSs. The following exclusionary criteria were applied to *NGS* coverage data: i) overlapping gene models; ii) continuous coverage spanning more than 9 kb; iii) continuous 5' coverage into an upstream gene model; iv) coverage of less than 5 at *INIT*; and v) a 5' UTR splice junction spanning to an adjacent gene model. If the above criteria at a given locus were permissive, the *NGS* 5' UTR regions were defined by: $NGS \in \{kc_1, kc_2, kc_3 \dots kc_n\}$ where $k = {}^{1}/_{r}$; r is the raw coverage at *INIT*; and c is the raw coverage at positions $\{1,2,3 \dots n\}$ upstream of *INIT*, so that $c \geq 1$. Thus, a normalised array was built from *INIT* upstream until there was a discontinuation in coverage. Areas of zero coverage were only permitted if they were bounded by a significantly supported splice junction unique to that respective gene model. The *FGH* 5'UTR regions were extracted from the gff3 as the region from the most 5' exon position to and including the most 3' initiation codon of a gene model. The *PASA* 5' UTRs were extracted from the tabular *PASA* file[30] if a given locus-associated EST assembly overlapped the most 3' initiation site of the respective locus in the correct orientation. Again, the 5' UTR is defined from the most 5' position of the EST assembly, to and excluding *INIT*.

At each locus, the *FGH*, *PASA* and *NGS* 5' UTR lengths were compared. From these lengths, 5' UTR annotation sources were prioritized per locus (n). An empirical annotation (*PASAn*; *NGSn*) is prioritized over *FGHn*. The longest empirical transcript is preferred, i.e. *PASAn* > *NGSn* or *NGSn* > *PASAn*. Those loci which have a 5' UTR reported by only *FGH* retain their *FGH* annotation as the best current annotation. Distal transcription start sites may be inferred by the first position of the prioritized 5' UTR, with greater confidence associated with those that are empirically substantiated and, if by *NGS*, those with higher coverage at *INIT*, the scores of which are available in **Supplementary Data 14.**

Results and Discussion

Of the 20,760 *E. grandis* loci annotated with a 5' UTR, 35% (7,306) were represented by *PASA* data, whereas 75% (15,569) were represented by *NGS*. The co-occurrence of 5' UTR annotation per locus (**Supplementary Data 14**) shows that 16% (5,790) of all loci are supported by all three sources. Despite

the co-evidence at loci, the respective annotations differ significantly (pairwise comparison provided in **Supplementary Table 11**). Only 519 loci had annotations of a similar length (standard deviation < 5 nt). The Pearson Correlation Coefficient between *FGH* and *PASA* 5' UTR lengths is 0.58, 0.28 between *NGS* and *FGH* and only 0.20 between *PASA* and *NGS*. The mean and median for *NGS* 5' UTRs are 539 and 283 nt respectively, differing from both *FGH* (394; 154) and *PASA* (533; 151). As hypothesized, the *FGH* models are significantly shorter (KS test, $p<0.05$), indicating possible truncation of *in silico* models. This was expected as FGenesH predictions favor those most proximal to the translation start site. *NGS* data is suited to splice junction detection and identified 2,095 high confidence 5' UTR introns, with unfiltered alternative splicing occurring prolifically at ~1.27 times per intron. This exemplifies the ability for *NGS* data to detect longer transcript isoforms, which are in minor abundance.

Using high-throughput empirical curation, we were able to substantiate or improve 17,085 5' UTRs and added 7,596 which had no prior annotation. A prioritized composite set of annotations comprises the longest empirical 5' UTR and in cases of no empirical evidence and only a *FGH* prediction, the *FGH* predicted 5' UTR, proportions of which are detailed in **Supplementary Data 14**. We have augmented the total percentage of annotated 5' UTRs from 57% of all gene models to 78%, achieving the desired increase in both quality and number of 5' UTR annotations. Empirically annotated 5' UTRs allow greater confidence in the inference of dTSSs, yet *in silico* annotations remain invaluable *in lieu* of transcriptome data for all loci. The total 28,358 5' UTRs annotations are provided in **Supplementary Data 14.**

S2.4 Gene Ontology (GO) annotation

Interproscan[34] provided the domain annotations for proteins encoded by 21,066 genes, while SignalP[35] was used to predict localization to the mitochondrial or plastid or secretion pathway, besides providing signal peptide cleavage sites. Further high confidence predicted localization to either or both the mitochondrion or plastid was performed with Predotar[36]. Transmembrane domains in the peptide sequences were annotated with TMHMM[37]. After collecting these annotations, standardized protocols for assigning the GO annotations were adopted. Mapping files provided by the GO consortium were employed to annotate the genes with the three GO categories (gene_ontology.obo file dated May-11-2011), namely Molecular Function, Biological Process and Cellular Component. The majority of these annotations were inferred by electronic annotation (IEA) evidence codes except the cases where the predicted scores for SignalP and Predotar were reviewed after computational analysis (RCA). Besides these, the *Arabidopsis thaliana* GO annotations (dated July 12, 2011) provided by the TAIR database[38] were respectively imported to enrich the *Eucalyptus* gene annotations by way of gene based orthology suggested by the gene family clustering methods described elsewhere in this communication.

Annotation coverage in the *Eucalyptus* genome is equivalent to that of *Arabidopsis*, a five-fold smaller genome. Preliminary annotation of ~29,841 genes suggested that compared to *Arabidopsis* the *E. grandis* genome maintains: 1) more genes for 'molecular function' categories defined for transport, electron carrier and catalytic activities and binding to protein and nucleotides (**Supplementary Data 15**), 2) more assignments for genes with biological processes such as transport, protein, lipid and carbohydrate metabolism, signaling and response to biotic stimulus and 3) more gene counts for those assigned to cellular localizations such as mitochondrion, plastid and membrane. Lower counts were recorded for ribosome, vacuole, cytoskeleton and chromosome.

S2.5 ncRNA annotation

Methods

To predict ncRNAs in *Eucalyptus*, the genome sequence was scanned by using Infernal[39] with the covariance models (i.e., a combination of sequence consensus and RNA secondary structure consensus) of 1,973 RNA families in the RFam database version 10.1[40,41]. The bit score cutoff of 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 e-value cutoff of 0.01. To examine the ncRNA conservation between *Eucalyptus* and other plant genomes, the *Eucalyptus* ncRNA candidate sequences obtained from Infernal search were used as queries to search against the genome sequences of *Arabidopsis lyrata* (Phytozome; www.phytozome.net), *A. thaliana* (www.arabidopsis.org), *Brachypodium distachyon* (Phytozome), *Brassica rapa* (brassicadb.org/brad/), *Carica papaya* (Phytozome), *Chlamydomonas reinhardtii* (Phytozome), *Cucumis sativus* (Phytozome), *Fragaria vesca* (www.rosaceae.org/projects/strawberry_genome), *Glycine max* (Phytozome), *Malus* x *domestica* (www.rosaceae.org/projects/apple_genome), *Oryza sativa* (Phytozome), *Phoenix dactylifera* (qatar-weill.cornell.edu/research/datepalmGenome/), *Physcomitrella patens* (Phytozome), *Populus trichocarpa* (Phytozome), *Ricinus communis* (Phytozome), *Selaginella moellendorffii* (Phytozome), *Solanum tuberosum* (potatogenomics.plantbiology.msu.edu), *Sorghum bicolor* (Phytozome), *Theobroma cacao* (www.cacaogenomedb.org/main), *Vitis vinifera* (Phytozome), *Volvox carteri* (Phytozome) and *Zea mays* (Phytozome) using blat[42] with a minimum coverage (i.e., minimum fraction of query that must be aligned) of 80% and a minimum identity of 60%.

Results and Discussion

Noncoding RNAs (ncRNAs) are functional RNA molecules without being translated into proteins, including microRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), signal recognition particle RNAs (SRP RNAs), tRNAs, rRNAs, regulatory RNAs, small RNAs (sRNAs) and other non-coding transcripts[43-46]. ncRNAs play important roles in a great variety of processes, such as the rRNAs and tRNAs involved in mRNA translation, snRNAs involved in splicing and snoRNAs involved in the modification of rRNAs[44]. To date, many ncRNAs have been identified in other sequenced woody plants such as *Populus*, *Vitis* and *Malus*[47-50].

From Rfam-based *in silico* prediction, consolidated with cross-species conservation (i.e., a minimum coverage of 80% and a minimum identity of 60%) and/or RNA expression evidence (i.e., >30% RNA-seq coverage), we identified 1,265 *Eucalyptus* ncRNAs (**Supplementary Data 16**, not including miRNAs), including 508 tRNAs, 269 rRNAs, 125 spliceosomal snRNAs, 19 antisense RNAs, 22 signal recognition particle (SRP) RNAs, 67 cis regulatory (Cis-reg) RNA elements, 175 snoRNAs and 80 sRNAs (**Supplementary Table 9**). Four types of ncRNAs (i.e., tRNA, rRNA, snRNA and antisense RNA) were generally conserved between *Eucalyptus* and other species. However, a large proportion of the other four types of ncRNAs (i.e., SRP RNA, Cis-reg RNA, snoRNA and sRNA) appear to be *Eucalyptus*-specific indicating that these ncRNA families have experienced significant lineage-specific expansion.

S2.6 miRNA annotation

E. grandis microRNAs (miRNAs) were identified by deep sequencing small RNA libraries (smRNA-Seq) in one lane of an Illumina GAII sequencer using a 36 bp single-end sequencing kit, using barcoded RNA samples from leaves and xylem extracted from a BRASUZ1 adult tree. The resulting small RNA reads were processed by a custom computational pipeline starting with a sequence cleaning step involving quality trimming and adapter removal. Cleaned reads were size sorted, quantified (tag counting) and used to create a non-redundant set of sequences using the program UCLUST[51]. The non-redundant reads were mapped against the reference *E. grandis* (BRASUZ1) genome using BWA[52].

In order to identify conserved miRNA sequences, a similarity search using PatMan[53] against the mirBase database (version 19)[54] was performed using plant-specific mature sequences therein and allowing up to four mismatches. Novel miRNAs were predicted using the miRDeep2 pipeline[55], which performs a genome-wide search for potential miRNA precursors based on the extension of regions with mapped reads followed by secondary structure prediction and stability evaluation[56]. Messenger RNA target

prediction of the potential miRNA sequences was performed using psRNATarget[57], with default settings, against the *E. grandis* transcript database taken from Phytozome (v8.0)[29].

Results

Reads resulting from both libraries (leaves and xylem) of the BRASUZ1 smRNA-Seq experiment displayed the characteristic two-peak profile at 21 and 24 nt found in other angiosperms, representing potential miRNAs and small interfering (siRNAs) respectively. Similarity searches against the miRBase database resulted in 470 BRASUZ1 smRNA sequences with at most four mismatches. Further processing, taking into consideration the secondary structure the potential precursor, identified 60 compatible loci, likely genuine miRNA genes. Complete annotation information on these sequences is available in **Supplementary Data 17.** Additionally, a genome-wide *in silico* prediction of potentially novel miRNA genes was performed using the miRDeep2 pipeline. A total of 178 loci were found matching the following criteria: expression as probed in smRNA-Seq, mapping to the genome and compatible precursor secondary structure. Additionally, computational mRNA target prediction for these putative miRNAs indicates that 163 candidates target at least one annotated gene model. From these, 153 loci do not have significant similarity to miRBase sequences and are potentially new *E. grandis* miRNA genes (**Supplementary Data 17**). This new set of potentially lineage-specific miRNAs awaits further experimental validation. However, the enrichment of specific target mRNA functional classes indicates that gene families subject to rapid expansion, notably disease resistance genes[58] and penta tricopeptide repeat (PPR) proteins[59] may evolve in concert with novel regulatory modules.

S3. Genome evolution and phylogenetics

S3.1 Gene family evolution

Gene Homology Analysis

An integrated analysis[60] based on Inparanoid algorithm[61] was used to find orthologous genes and paralogous genes that arise by duplication events. Ortholog clusters were seeded with a reciprocal best pair match, after which an algorithm for adding in-paralogs was applied. First all-vs.-all (reciprocal) BLAST search was run using sequences files from any two given species pairs (e.g., X and Y). Sequence pairs with mutual best hits were detected. Sequences from out-group species were used to detect cases of selective loss of orthologs. The X-Y sequence pairs were eliminated if either sequence X or sequence Y scored higher to an out-group sequence than to each other. Additionally, orthologs based on in-paralogs were clustered together with each remaining pair of potential orthologs. Overlapping clusters were resolved by a set of rules adopted from Inparanoid. The parameters used were the defaults of a cutoff bit score of 50, in-paralog confidence cutoff of 0.05 and group overlap of 0.5, where homologues were grouped if they were similar enough to each other. Other parameters used were a cutoff of 0.5 for total sequence overlaps and 0.25 for segment overlaps. The peptide sequences used were from twenty-nine species, including *Arabidopsis lyrata, A. thaliana, Brachypodium distachyon, Caenorhabditis elegans, Chlamydomonas reinhardtii, Danio rerio, Ectocarpus siliculosus, Escherichia coli, Eucalyptus grandis, Fragaria vesca, Glycine max, Homo sapiens, Jatropha curcas, Mus musculus, Neurospora crassa, Nostoc punctiforme, Oryza sativa, Phoenix dactylifera, Physcomitrella patens, Populus trichocarpa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Selaginella moellendorffii, Solanum tuberosum, Sorghum bicolor, Synechocystis pcc6803, Theobroma cacao, Vitis vinifera, Zea mays.* The sequences were downloaded from Gramene[62,63], Phytozome (www.phytozome.net) and Ensembl (www.ensembl.org) databases.

Eucalyptus unique gene clusters

A total of 113,954 gene sequences from four major rosid lineages represented by *Arabidopsis, Vitis, Eucalyptus* and *Populus* clustered together in 16,048 gene families (**Extended Data Figure 2a**). Of the 36,354 protein coding genes in the *E. grandis* genome (V1.1 annotation, Phytozome), 30,341 were

included in 10,049 of the identified family clusters, with 851 gene clusters (8.5%) containing 1,357 genes that were unique to *Eucalyptus* in the four-way rosid comparison. Of these 1,357 only 939 had GO assignments and 968 had some InterPro domain assignments (**Supplementary Data 18**). Most InterPro domains found in this subset of 968 genes belong to domains with associations to unknown function DUF594/247 (IPR007658, IPR016158), pentatricopeptide repeat (IPR002885), transferase (IPR003480), Cytochrome P450 (IPR001128), plant disease response protein (IPR004265), domains associated with the ubiquitination and various types of transcription regulators. The remaining 418 of the 1,357 peptides are novel predicted genes of unknown function.

A core set of 6,926 gene clusters comprising 86,723 genes (average 3.13 genes per cluster per species) had membership in all four species. We also found 6,086 *E. grandis* genes that did not cluster with any of the genes from the 29 species (**Supplementary Data 15**). These numbers are consistent with relative proportions from other sequenced genomes (**Extended Data Figure 2a**). In the larger set of 6,086 unique *E. grandis* genes, 3,211 had assignments to GO (160 GO terms, **Supplementary Data 15**) and only 1,043 genes had assignments to 516 InterPro domains. Prominent GO assignments include transcription regulation (GO:0006355), oxidoreductase activity (GO:0016491), transmembrane signaling receptor activity (GO:0006355), innate immunity (GO:0045087) and cellular locations such as mitochondrion, plastid and endomembrane system. These annotations generally consistent with the genome level GO analysis (**Supplementary Data 15**). The InterPro assignments of 1,043 unique *E. grandis* genes show high representation for domains associated with disease resistance, innate immunity and signal transduction such as the protein kinase core (IPR000719), leucine-rich repeat (IPR001611), serine/threonine protein kinase domains (IPR017442, IPR002290 and IPR008271), in addition to those already mentioned for *E. grandis* unique genes identified among the four genomes.

S3.2 Protein domain evolution

The domain and domain arrangement content of *E. grandis* was compared to those of predicted proteomes of *E. grandis*, *Populus trichocarpa*[50], *Arabidopsis thaliana*[64] and *Vitis vinifera*[65] (**Extended Data Figure 2b)**. For proteins with different transcripts, all but the longest sequence were removed. The proteins were scanned with the utility hmm_scan from hmmer3.0 against PfamA[66]. Predicted domains with a length < 0.3 of the hmm model length were removed. Overlapping domains were removed by keeping the domain with the lowest e-value. Domain arrangements were created by concatenating domains in their order on the sequence. Domains belonging to the category 'Repeat' or 'Motif' were collapsed if occurring in series[67]. Domains and domain arrangements were compared within the rosids to distinguish between a core set of domains and arrangements 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 1.5x higher frequency in *Eucalyptus* compared to the average abundance in the rosids were defined as overrepresented.

Eucalyptus has 13 PfamA domains which are not found in the three other major rosid lineages (represented by *Arabidopsis*, *Populus* and *Vitis*, **Extended Data Figure 2b)** and 236 domains and 992 arrangements which are (1.5x) overrepresented in the *Eucalyptus* proteome compared to the other three rosid lineages (**Supplementary Data 8)**. Some of the most over-represented domain and domain arrangements include domains involved in the cellulose biosynthesis such as members of the CBM and Glyco_hydro family as well as domains involved in terpene synthesis such as Terpene_synth_C (PF03936) and Terpene_synth (PF01397). Besides the overrepresented domains, 392 domain arrangements are unique to *Eucalyptus*, such as 'PF03936;PF03936'.

S3.3 Reconstruction of green plant phylogeny based on protein coding nuclear genes

Methods

A total of 174,020 peptides encoded by single copy protein coding nuclear genes were identified from 17 plant genomes *A.s lyrata, A. thaliana, B. distachyon, E. grandis, F. vesca, G. max, J. curcas, O. sativa, P. dactylifera, P. patens, P. trichocarpa , S. moellendorffii, S. tuberosum, S. bicolor, T. cacao, V. vinifera* and *Z. mays* by using previously described gene orthology clustering methods [68] and described in the gene clustering section of this communication. These were further analysed by Hal[69] which is an automated workflow for phylogenomic analyses. The Hal workflow was initiated with the filtered set of orthologous genes available from InParanoid[70] analysis carried out by us and described earlier in this communication. These InParanoid orthologs were subjected to a second all-vs.-all BLASTP analysis, which further reduced the orthologous gene set to 162,431. This step was followed by the MCL clustering across a range of inflation parameters (1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 and 5.0). Resulting orthologous clusters (OCs) were filtered to remove any redundant clusters and to retain only OCs that contain no more than one protein per genome and that had at least 50% of the species represented in the cluster[69]. Using this approach, a total of 3268 single copy clusters were identified across the 17 genomes. Alignments were generated for each cluster with MUSCLE [71] and low quality regions of the alignments were identified and trimmed with GBlocks[72,73] using two settings: liberal ("maximum number of contiguous nonconserved positions allowed is 8; minimum length of a block allowed is 5") and conservative ("a maximum number of contiguous nonconserved positions allowed is 4; minimum length of a block allowed is 10"). Individual trimmed protein alignments were concatenated into conservative and liberal superalignments that retained 42.26% (697,423 aa) and 46.35% (764,978 aa) of the original 1,650,340 aa concatenated alignment, respectively. Amino acid positions in both the conservative and liberal alignments were further categorized into eight rate categories using the program PAML[74]. Maximum likelihood phylogenetic analyses using RAxML[75] and the PROTGAMMAWAG model of evolution were performed on both conservative and liberal alignments followed by three sets of analyses that involved the removal of the three (8th, 8th+7th, 8th-6th) fastest rate categories. The latter sets of analyses were performed to assess the sensitivity of tree topology and branch support to sites characterized by higher variability and potentially greater levels of homoplasy. The RAxML tree shown in **Figure 2** (main text) is from the conservative supermatrix; trees produced from the liberal supermatrix are provided as **Supplementary Data 7.**

Results and discussion

Since much of plant adaptation and manifestation of morphology, anatomy and physiology affecting plant fitness and speciation is determined by the nuclear-encode genes, we began with the hypothesis that green plant phylogeny and the position of *E. grandis* within that can be reconstructed by analyzing the nuclear genome unlike conventional methods often using genes/markers from the organellar genomes. Comparative phylogenetic analysis of 17 sequenced plant genomes surprisingly revealed that the currently accepted phylogenetic placement of *Eucalyptus*, in the Malvids cluster may be incorrect. The maximum likelihood phylogenetic analysis produced a well-supported tree (**Figure 2**, **main text**) that differed from earlier studies showing placement of *Eucalyptus (Myrtales)* with *Arabidopsis* in the Malvidae instead of sister/basal to the core rosids. The liberal and conservative super alignments concatenated by harboring 764,978 and 697,423 amino acid characters resulted in similar tree topology. After generating the trees we reanalysed the MCL protein clusters put together after the all-vs-all BLASTP to find the number of unique and shared peptides at each of the species nodes (**Supplementary Data 7**). Of the 9,836 *E. grandis* single copy peptides initially filtered by the Inparanoid analysis prior to running Hal, 476 peptides were found to be unique to the *E. grandis* node. The ancestral node shared between *E. grandis* and core rosids contains 179 peptides. *E. grandis* and core rosids share an ancestral node with *Vitis* which contains 308 peptides. The root of the tree with all 17 plant species analysed shared 5,518 peptides.

Widely used phylogenetic analyses of plants and angiosperms in particular, based exclusively on chloroplast genes have consistently resolved the two large rosid clades, Fabidae and Malvidae[76] with strong support. In contrast, several recent studies involving the mitochondrial, nuclear and plastid genes[77,78] resulted in the placement of *Populus* in the Malvidae and *Fragaria* in the Fabids[68]. Our analysis is consistent with the placement of *Fragaria* (strawberry) and *Populus* reported by Shulaev *et al*.[68], which was carried out using ~250 peptides from sequenced plant genomes. Recent phylogenomics studies with large scale datasets from the genome and transcriptome were successfully applied to find whole-genome duplication events[79-81] and to reconstruct plant phylogeny[82,83]. This is also the second such report on placing *Phoenix dactylifera* a non-grass monocot in the angiosperm phylogeny based on the genome-scale gene family analysis[81]. In a similar analysis, we show for the first time that *Eucalyptus* (representative of the order Myrtales) is placed as a sibling/sister to the core rosids rather than in the Malvids clade of the core rosids. In previous studies based on the smaller set of plastid and nuclear genome markers the Myrtales clade was placed either tightly with the Malvids or as an extended member of the clade[84-88]. The novel placement of *Eucalyptus* (Myrtales) as sister to the core rosids may be due to evolutionary differences between the markers selected in previous reports that are often based in a small number of genes or markers. We hypothesize that though chloroplast genome markers have been successfully used in resolving the plant phylogeny, these markers of organellar genomes may not be congruent with a large number of nuclear variations observed for metabolic pathways from carbohydrate, amino acid, lipid and secondary metabolite metabolism and gene regulation leading to anatomical differences and speciation events. Therefore, the subtlety of the physiology and adaptation must be resolved wherever possible by looking at the whole-genome level. Also, selection of nuclear genes is important because the nuclear genome often undergoes more frequent changes due to adaptation and selection pressure from the native environment involving frequent genome/gene duplication and losses[89] leading to selective mutations in genes determining the biology of the plants. Therefore, a large-scale analysis of all the genes in the genomes, such as the one reported in this communication, is capable of resolving the subtle differences that are crucial for physiological and genetic adaptation. Analysis such as this is also important for projecting genome annotations based on genome alignments and gene orthology. Though the *Arabidopsis* genome is the most favored and enriched genome supported by the sheer amount of empirical evidence available for assigning functions to genes, the taxon itself may appear distant in terms of comparing the plant's biology and physiology. This may be compounded by the number of gene functions resulting from local adaptation and selection pressures that the *E. grandis* genome may have undergone. In this context, if empirical evidence is available for gene function and expression from closely placed asteroids, such as potato, tomato and rosids, such as *Vitis* and *Populus*, these genomes would be as valuable as *Arabidopsis* for genome annotation by functional projection onto the *E. grandis* genome.

S3.4 Segmental and whole-genome duplication

For analysis of whole-genome duplications, tandem regions were first ‘collapsed’ by retaining only the longest gene within a tandem duplicated region for further analysis. Likewise, genes were eliminated with more than 15 hits to other genes with blast e-values at least 1e-20 from analysis after collapsing the tandems ("promiscuous" genes). Finally, only pairwise hits with a blast score of at least 30% of the best blast hit within the genome by any of the two genes in the pair (‘c-score $\geq$ 0.3’) were considered for further analysis. All of the remaining genes still under consideration, i.e. moderately unique genes, were ordered in increasing order along the scaffolds and analysed using a ‘segmentator’ algorithm to identify stretches of intragenomic synteny. A maximum of 18 non-intervening genes between genes in runs of conserved synteny was allowed. This relatively lenient threshold was necessary to detect many of the duplicated genes, since the exact gene order has been scrambled and the vast majority of duplicated genes have lost the second copy. In order to estimate the false positive rate, the gene IDs were randomly scrambled while retaining the position data. Of the real data set, 2,118 duplicated gene pairs were found in regions of conserved synteny, but only 31 in the scrambled data set, i.e., a false positive rate of ~1.5%.

4TDV analysis of the duplicated genes (**Supplementary Data 19**) suggests three age classes of duplicated genes: **i) 4DTV < 0.25.** A total of 143 gene pairs are presumably segmental duplications from a variety of relatively recent epochs. They are present on all chromosomes except 4, 6, 9 and 10. **ii) 0.25 < 4DTV < 0.36.** The majority of these 947 gene pairs (located on all 11 chromosomes) are from a whole-genome duplication event (0.27 corrected 4DTV) that happened soon after the divergence of the Myrtales lineage from other rosid lineages. **iii) 4DTV ≥ 0.36.** The majority of these 1,028 gene pairs (present on all 11 chromosomes) arose in the more ancient hexaploidization event that happened at corrected 4DTV = 0.425 and is shared among all rosids. It is important to note that the separation based on 4DTV is not perfect and some individual segments may be incorrectly binned, but the analysis is nevertheless reliable overall.

S3.5 Genome structural evolution

Seven major duplications were identified and characterized in *Eucalyptus* covering 47% of the genome and involving the following chromosome to chromosome relationships: e4-e5 (blue), e3-e10-e6 (turquoise), e2-e8-e11 (purple), e7-e10-e2-e11-e1 (orange), e8-e6-e7-e9-e10 (green), e2-e11-e7-e9-e10 (yellow), e6-e5-e1-e11 (red) (*cf*. **Figure 3a, main text**). We identified with precise dating[90] two duplication events corresponding to a lineage-specific WGD (involving 393 paralogous relationships) dating back to ~90 mya, as well as the well-known shared paleohexaploidization event (referred to as γ, involving 427 paralogous relationships), dating back ~ 130 mya (**Figure 3b, main text**).

Using *Vitis* as the reference genome, known to be the closest relative to the eudicot ancestor structured in seven protochromosomes (colour code used for chromosome painting), 4,106 orthologous relationships were identified between *Eucalyptus* and *Vitis* (**Figure 3c, main text**) covering 72% of the *Eucalyptus* genome. The following chromosome to chromosome relationships were established (**e** for *Eucalyptus* and **g** for *Vitis*/grape as chromosome nomenclature): e1/g8-g12-g5-g16-g7-g18, e2/g17-g6-g1-g5-g14-g8-g7-g4-g18, e3/g11-g4, e4/g2-g10-g15-g1-g6, e5/g2-g15-g10-g1, e6/g18-g19-g9-g14, e7/g13-g1-g3-g14-g8-g12, e8/g5-g7-g14-g4-g1-g17-g18-g3, e9/g14-g3-g4-g7-g12-g9, e10/g6-g13-g14-g9-g4-g1, e11/g8-g12-g17-g7-g16.

Integration of independent analyses of duplications within and synteny between seven eudicot genomes led to the precise characterization in *Eucalyptus* of seven paleoduplications identified recently as the basis of the definition of seven ancestral chromosomal groups in eudicots[91,92]. These ancestral shared duplications were found on the following chromosome pair combinations in *Eucalyptus* compared to the seven ancestral paleoduplications reported in *Vitis*: g1-g14-g17/ e2-e11-e7-e9-e10, g2-g15-g12-g16/ e4-e5, g3-g4-g7-g18/ e8-e6-e7-e9-e10, g4-g9-g11/ e3-e10-e6, g5-g7-g14/ e2-e8-e11, g6-g8-g13/ e7-e10-e2-e11-e1, g10-g12-g19/ e6-e5-e1-e11.

Based on the ancestral paleohexaploidization event reported for the eudicots, it becomes possible to propose an evolutionary scenario that has shaped the 11 *Eucalyptus* chromosomes from the seven-chromosome eudicot ancestor and more precisely to the 21-chromosome paleohexaploid intermediate (**Figure 3c**). We suggest from the 21-chromosome intermediate ancestor at least 69 fissions and 79 fusions to reach the modern 11 *Eucalyptus* chromosome structure (**Figure 3c)**. Finally, we suggest that the modern *Eucalyptus* genome went through one additional whole genome duplication (*cf* **Figure 3b-c**).

S3.6 Collinearity analysis and dating of the whole-genome duplication

Homologous gene pairs were determined through an all-vs.-all blastp[93] analysis using the proteomes of *Vitis* and *Eucalyptus* (e-value cutoff was set to e^{-5} and the number of hits to 500). Next, gene families were constructed using the blast results as input for Tribe-MCL[94], parameter *I* was set to 2 and all other parameters were left as default. Collinear and duplicated regions in the *Eucalyptus* genome were

identified by running iADHoRe3.0[95,96] with the previous described gene families and the ordered gene list from *Vitis* and *Eucalyptus*. iADHoRe parameters were set as follows: *table type=family*, *alignment method*=gg2, *gap size*=20, *cluster gap*=25, *q value*=0.9, prob *cutoff*=0.001, *anchor points*=4 and *level 2 only*=true. In total, we were able to uncover 214 duplicated segments in *Eucalyptus* and 760 collinear regions with *Vitis* containing 1911 and 13,010 homologous genes, respectively. For almost every *Vitis* chromosome two collinear regions can be identified in the *Eucalyptus* genome. This 1-to-2 relationship between *Vitis* and *Eucalyptus* genomic regions is a clear indication of a duplication event in the *Eucalyptus* lineage after the split with *Vitis.*

The K_S age distribution of the *E. grandis* genome (**Extended Data Figure 4a**) was reconstructed as previously described[97]. Paralogous gene pairs located in duplicated segments (anchors) were detected using i-ADHoRe 3.0[95,96]. Anchors with a K_S of between 0.8 and 1.5, corresponding to the most recent WGD (see **Extended Data Figure 4b**) were used for phylogenetic dating as described in Fawcett *et al*.[98], with some small modifications. For each gene pair, an orthogroup was created that included the two paralogous anchors and several orthologous genes identified by Inparanoid[61] using a broad taxonomic sampling with one representative from the Fabales, Malpighiales, Brassicales, Malvales and Monocots[99]. In total, 104 such orthogroups could be collected. The node joining the two *Eucalyptus* paralogs was dated using the BEAST v1.7 package[100] using an uncorrelated relaxed clock model and a GTR+G (4 rate categories) evolutionary model partitioned on each codon position. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the topology discussed in the main text (see **Figure 2**): (((Fabales,((Brassicales,Malvales),Malpighiales)),(Eucalyptus1,Eucalyptus2)),Monocots). To account for the fact this topology differs from the consensus of APGIII in the placement of the Malpighiales and Myrtales[101], the MCMC for each orthogroup was run accommodating for phylogenetic uncertainty by estimating each topology independently. The following fossil calibrations were implemented: a lognormal calibration prior on the node uniting the Brassicales and Malvales based on the fossil *Dressiantha bicarpellata*[102] with offset=82.8, mean=3.8528 and SD=0.5[103]; and a lognormal calibration prior on the root with offset=124, mean=4.0786 and SD=0.5[104]. The offset of these calibrations represent conservative hard minimum boundaries, while their mean was placed according to the most recent and taxonomic complete dating studies available for these specific clades because the location of the calibration peak mass has been shown to effect the posterior of the age estimate to some extent[105]. A run without data was performed to ensure proper placement of the marginal calibration prior distributions[106]. The MCMC for each orthogroup was run for 100 million generations, sampling every 10,000 generations resulting in a sample size of 10,000. The resulting trace files of all 104 orthogroups were evaluated manually using Tracer v1.5[100] with a burn-in of 1000 samples to ensure proper convergence for each dated anchor pair (minimum ESS for all statistics at least 200). All 104 orthogroups were accepted and estimated ages were then processed as described in Fawcett *et al*[98]. The resulting absolute age distribution of these WGD estimates in presented in **Extended Data Figure 4c.**

Together these results support a WGD pattern (**Extended Data Figure 4d)** that includes a lineage-specific WGD occurring around 109.93 mya in the lineage (order Myrtales) leading to *Eucalyptus*, which followed the older paleohexaploidy event shared by rosids and asterids. This is considerably earlier than the WGD events typically detected in other rosid genomes such as *Arabidopsis*, *Populus* and *Glycine*. An age of 109.93 mya places the WGD firmly in the Gondwanan ancestor of the lineage at about the period when Australia and Antarctica began to separate from East Gondwana. The WGD could therefore have played a pivotal role in the early evolution of the lineage and its subsequent differentiation from other rosid ancestors. Genome sequencing of other Myrtaceae genera and other families in the Myrtales will reveal the extent to which this event has shaped the evolution of this important order of flowering plants.

S3.7 Orthologous segments of conserved synteny between Eucalyptus and Vitis

Segments of conserved synteny between the *Eucalyptus* and *Vitis vinifera* genomes were identified by gene similarity at the peptide level, using the *segmentator* algorithm previously described[50]. Tandem duplicated genes were identified and excluded prior to the analysis in both genomes, as were genes showing sequence similarity to more than 10 genes in the other genome with BLAST e-value < 1e-20. We also excluded pairwise hits between genes in the two genomes with scores less than 90% of the best reciprocal score between any of the two genes and any other genes in the opposite genome. This step drastically reduces the finding of paralogous segments from the common ancient hexaploidy events: orthologous genes are empirically close to being reciprocal best hits, whereas older paralogs show comparatively weaker similarity. From the rungs of conserved synteny identified by this analysis, we identified cases in which more than a single region of *Eucalyptus* or two regions of *Vitis* were part of conserved synteny segment, in which case we kept only the longer (or longest two for *Vitis*) regions. This last step eliminates a few more, small, ancient paralogous segments. We finally merged adjacent segments in *Eucalyptus* showing colinearity to adjacent regions in *Vitis* into single segments, provided distances between segments were less than 3 Mb of sequence and that there were no intervening segments suggesting alternative blocks of synteny between the two. The resulting 480 orthologous segments of conserved synteny cover some 68% of the *Eucalyptus* genes and 76% of *Vitis* genes included in the analysis.

S3.8 Genome synteny with Populus

Methods

We used the VISTA pipeline infrastructure[107,108] for the construction of genome-wide pairwise DNA alignments between *Eucalyptus grandis* and *Populus trichocarpa*. 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[109] to local alignments produced by translated BLAT[42]. 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.

Syntenic regions between *Eucalyptus* chromosome 3 and *Populus* chromosome XVIII were defined as segments of contiguous sequence based on the VISTA procedures outlined above. Each contiguous block of DNA was annotated and cross-compared between the two species. Gene models within the syntenic 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 single locus in both cases.

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 the VISTA browser at http://pipeline.lbl.gov/and Phytozome http://phytozome.org. We have also extended this approach to compare duplicated segments in *E. grandis*. In the self-alignment of *E. grandis*, we detected a significant number of tandem repeats, defined as sequences separated by no more than 100 bp. They range in size from 25 bp to 36,225 bp and cover 4,331,798 bp or 0.7% of the genome.

Results

After applying the whole-genome DNA alignment procedures[110], 75.51% of the length in coding exons, 15.74% of UTRs, 13.6% of 100-bp-long upstream regions, 10.56% of 200-bp-long upstream regions, 6.6% of 500-bp-long upstream regions and 7.3% of 200-bp-long downstream regions of the *E. grandis* chromosome 3 are covered by significant pair-wise alignments with *P. trichocarpa* chromosome XVIII.

At the gene level, there are 104 syntenic blocks of genes shared between *Eucalyptus* chromosome 3 and *Populus* chromosome XVIII that contain 778 genes in 522 gene families (**Supplementary Data 2**). The most common family found in this conserved gene space is represented by 33 disease resistance genes; the next most common gene family was a protein kinase superfamily protein (KOG0192, PFam7714). Using the syntenic blocks conserved in *Eucalyptus* and *Populus* as reference, there are also 85 syntenic blocks that are evident in *Arabidopsis*; these shared syntenic blocks contain 196 gene models with a maximum number of syntenic genes equal to six. There are also 10 syntenic blocks that share the same break/rearrangement points among the three species.

Co-regulation and co-expression have been proposed as possible mechanisms for maintaining gene order and leading to high levels of synteny across different taxa[111,112]. The uniform expression pattern across the entire chromosome 3 of the *Eucalyptus* genome lends support to the co-expression hypothesis and may provide insight into preservation of gene order in *Eucalyptus* and high level of synteny shared by this chromosome with *Populus* chromosome XVIII. Deeper characterization, not only of functional gene classes, but also regulatory elements that lie within intergenic regions and their putative targets, will be required to provide a definitive answer to the conserved synteny between chromosome 3 in *Eucalyptus* and chromosome XVIII in *Populus*.

Chromosome 3 exhibits the highest level of intra-chromosomal duplication relative to other chromosomes in the *Eucalyptus* genome. Out of the 824 conserved syntenic genes, 429 had at least one duplicate copy. The protein kinase superfamily had the highest number of duplicates with a total of 19 copies. Fischer *et al*.[113] proposed segmental gene duplication on alternate chromosomes, followed by loss of the ancestral gene, either via deletion or accumulation of mutations, as a mechanism by which synteny is disrupted. Thus, intra-chromosomal duplication may be viewed as means of maintaining ancestral gene signatures on the same chromosome, leading to preserved ancestral chromosome structure. When duplication is not followed by gene loss, the tendency for segmental duplications to occur within the same chromosome could lead to enrichment of genes with related functions that also share the same regulatory features and result in the occurrence of co-regulated and co-expressed genes in conserved chromosome segments. In this instance, both the co-regulation/co-expression and duplication/loss hypotheses appear to be complementary and support the observation in *Eucalyptus* and *Populus*.

It is also noteworthy that there appears to be high fidelity in whether particular *Eucalyptus* chromosomal regions favor intra- versus inter-chromosomal segmental duplication, with chromosome 3 and 6 exhibiting the highest level of intra-chromosomal and lowest level of inter-chromosomal duplication and chromosomes 1 and 11 have the opposite tendency. Additionally, the notion that intra-chromosomal duplication seems to be correlated with low inter-chromosomal duplication is supported by the observation that regions harboring tandem duplication appear to share little or no synteny with other chromosomes, even though parts of the same chromosome might share high synteny with other chromosome. Notably, tandem duplications on chromosome 5 and 8 provide clear examples of this phenomenon (**Figure 1f, Main text).** The tendency for highly specialized genes, such as those mediating plant-microbe or plant-insect interactions, to tandemly duplicate has been well documented[114] and in this case may offer insight into the molecular mechanisms that lead to distinct duplication behavior and consequently differing levels of gene order conservation between chromosomes in the same genome.

S3.9 Tandem gene duplication

The criterion used for tandem gene detection was that tandem duplicates were genes with blast hits with e-values at least 1e-20, with a maximum of five intervening genes. Clusters were build up by single linkage, i.e., if a third gene showed similarity to and were within five genes from another gene already member of a tandem cluster, it was included in the cluster, etc. A total of 3,185 tandem expanded regions (clusters, **Supplementary Data 19**) were detected on all chromosomes (**Figure 1f, main text)**. These

clusters contain a total of 12,570 genes (34.5% of the 36,376 genes in the V1.1 annotation of the genome). This is more than have been reported for any of the sequenced plant genomes (**Table 1, main text**). More than 97% of the duplicated genes are separated by fewer than five non-participating genes and most are next to each other. A total of 23,806 genes (65.4%) do not have tandem duplicates. We validated the accuracy of assembly across clusters of highly similar tandem duplicates by analyzing the contiguity of tandem assemblies and analyzing independent BAC clones spanning two large gene clusters (**Supplementary Information 1.5).**

The tandem duplicates in the *E. grandis* genome are not the result of a recent burst of duplication, but of a steady (and unusually high) rate of tandem duplicate gain. Applying an analysis to the tandem-expanded regions in which pairwise 4DTV distances among members in each region are first sorted in increasing order and the list then traversed and the epochs of the various tandem duplication events recorded (in essence constructing a UPGMA tree and recording the time of the nodes), the distribution of distances and divergence times can be determined (**Extended Data Figure 5**). Except for *Arabidopsis* which clearly has a different evolutionary pattern, the rosid genomes analysed all follow nearly exponential distributions, consistent with a simple, Lynch-Connery style model[115], in which tandem duplications occur at a constant rate per 4DTV unit (or per million years) and loss also has a constant probability per time. *Vitis*, *Populus* and *Eucalyptus* have similar loss rates, but the tandem duplication rate in *Eucalyptus* is three times as high (1352 duplications per million years) as in *Populus* and five times as high as in *Vitis*, while the rate of gene loss is more similar in the three genomes (**Extended Data Figure 5**). Tandem duplications therefore appear to have happened at a constant rate in the three rosid genomes, without any specific "bursts". Some gene families may deviate from this pattern.

S3.10 Genome comparison to E. globulus

We generated 266.6 Million (M) Illumina-HiSeq paired reads (2 x 100 bp) from DNA of *E. globulus* clone X46 (*E. globulus* X46) extracted using the method of Tibbits *et al*[116]. Filtering the raw data using NUCLEAR software (Gydle Inc.) produced 191 M high-quality (HQ) reads (164 M paired-end, 27 M single-end) representing 32.8 Gb of total sequence suitable for alignment. The filtering process trimmed reads based on quality, characterized read complexity and removed sequencing controls. We identified 4.9 M chloroplastic reads (2.6% total) and used the VISION software (Gydle Inc.) to assemble a complete sequence of the *E. globulus* X46 chloroplast genome (160,268 bp). Among the remaining HQ reads we used the NUCLEAR software (Gydle Inc.) to map 179.5 M (96.4%) to the *E. grandis* genome scaffolded assembly v.1.0 (569.71 Mb in 21,378 scaffolded contigs, hereafter referred to as *E. grandis* scaffolds). To ensure a proper coverage of repeat regions, the NUCLEAR parameters were set to use its multi-HSP, paired-end-aware adaptive scoring system that guarantees the best hit from all hits above a threshold, set to allow for up to 20% local allelic variability. Mapped reads were then locally realigned and subsequently analysed in a multiple alignment context in downstream analyses. Among the paired-end reads mapped to the scaffolds, 88% had both ends aligned to their scaffold sequence. The DNA library size was assessed from paired-end reads mapped to the X46 chloroplast and gave a major peak around 430 bp (80% inserts within 360-500 bp).

The coverage of *E. grandis* scaffolds by *E. globulus* was computed using the aligned portions of the mapped X46 reads (84.8% of the bases in 179.5 M reads). This coverage distribution can be split into distinct *additive* components. First, the uncovered regions, with sequence coverage of 5 reads or fewer (0x-5x), amount to a total of 100.3 Mb (17.6% scaffold length). The bell-shaped distribution centered on a coverage of 58x represents the homologous single copy portion of the two genomes (1:1) and accounts for 328.9 Mb (57.7% of scaffold length). A second bell-shaped distribution, centered around 29x and comprising 73.8 Mb (13% of scaffold length), corresponds to half covered regions (0.5:1). The high coverage regions (minimum 151x, mean 284x) amount to 15.4 Mb (2.7% of the scaffold length) and represent, in addition to their single copy (58x) equivalent, an extra 60.1 Mb of sequence with respect to

the *E. grandis* scaffolds. The remaining component, all low coverage (6x-28x), amounts to 51.3 Mb (9% scaffold length) and represents 13 Mb at single-copy (58x) equivalent.

To compare our *E. globulus* analysis with similar metrics generated from *E. grandis*, we performed the same analysis on the Illumina reads generated at JGI from BRASUZ1. A total of 287.8 M paired reads (2 x 100 bp) were filtered to produce 253.6 M HQ reads (199.6 M paired-end, 54 M single-end), representing 38.2 Gb of usable sequence. We identified 10.6 M chloroplastic reads (4.2% total) and mapped 235.1 M reads (96.8% of the 243 M remaining reads) to the *E. grandis* scaffolds, similarly locally realigning and integrating them downstream into multiple sequence alignments. The DNA library insert size) was centered on 160 bp (95% inserts within 100-250 bp). Among the reads mapped to the scaffolds, 68.7% were aligned at 100% identity over their complete sequence and among paired-end reads mapped to the scaffolds, 97.3% have both ends aligned to their scaffold sequence. This mapping result highlights the quality of the *E. grandis* scaffolds in representing locally accurate consensus sequence.

Similarly the coverage of *E. grandis* scaffolds by BRASUZ1 reads **b**), was computed using the aligned portions of the mapped BRASUZ1 Illumina reads (91.1% bases in 235.1 M reads). Here, the uncovered portion represents only 2 Mb (0.3% of the scaffold length) and is mostly composed of very high GC% regions, which are consistently underrepresented in PCR amplified Illumina libraries. The 1:1 portion is centered on 54x and covers 479.7 Mb (84.2% scaffold length). Similar to X46, there is 80.3 Mb (14.1% scaffold length) covered at about half depth (32x; 0.5:1). High coverage regions (minimum 151x, mean 417x) amount to 6.4 Mb (1.1% scaffold length) and represent, in addition to their single copy (54x) equivalent, an extra 42.9 Mb of the *E. grandis* genome outside of the scaffolds. The remaining component represents 1.3 Mb (0.2%) and 0.4 Mb at single-copy (54x) equivalent.

Coverage analysis of the *E. grandis* scaffolds by X46 and BRASUZ1 reads provides the basis for genome size estimations for *E. globulus* and *E. grandis* that are not dependent on the reference assembly being either complete or non-redundant. We estimate the total genome sizes of *E. globulus* and *E. grandis* by first adding the 1:1 equivalent lengths from each of their distribution classes. We then correct that sum based on the aligned fraction of the reads used in the coverage computation (respectively 84.76% and 91.1%). Among the remaining unassigned ~3% HQ reads (non-chloroplastic and unmapped to scaffolds) of each genome, 15% could be mapped to the unscaffolded contigs of the v1.0 assembly (70.9 Mb in 11,384 contigs) adding about 2.5 Mb to the genome size estimate for each species. This low mapping rate supports the hypothesis that these unscaffolded contigs represent (besides a large proportion of repetitive sequences - **Supplementary Information 1)** nearly exclusively allelic variants of the scaffolded genome. The remaining HQ reads (unmapped to any contig in the v1.0 assembly) containing high-complexity sequence were excluded from the final estimations as i) they account for ~12 Mb which is within the margin of error of our estimate and ii) they possibly represent contamination. As a result, we estimate the total genome size of *E. globulus* at 535 Mb $\pm$ 15 Mb, consistent with the estimates (498-545 Mb) obtained from flow cytometry[117,118] and the total genome size of *E. grandis* at 625 Mb $\pm$ 15 Mb consistent with a previous estimate of 640 Mb[117]. This gives an apparent genome size difference between *E. grandis* and *E. globulus* of 90 Mb.

To investigate the components underlying the genome size difference between the two species, we analyzed regions representative of each genome's specific content. To represent *E. grandis* specific content, we defined X46 uncovered regions as those with 100 bp or greater length with a consistent coverage lower than 5x. This yielded 88.7 Mb in 164,813 individual regions (mean 538 bp, median 230 bp, max 30,610 bp), with 2,076 regions of 4 kbp or larger (12.9 Mb), 18,753 regions of 1 kbp to 4 kbp (33.1 Mb) and 143,984 regions smaller than 1 kbp (42.7 Mb; Figure 3). The fraction of these regions artificially duplicated in the scaffolds (representative 0.5:1 BRASUZ1 regions as defined above) was the same (13.5%) as that of the whole assembly, indicating that i) the X46 uncovered regions are unbiased in representing *E. grandis* specific sequence and ii) they represent 82.7 Mb of unique sequence (91.9% of

the genome size difference). The distribution of these regions along chromosomes showed small but significant differences in density, with chromosomes 1, 6, 9, 10 and 11 (Scaffold A, F, I, J and K) having less uncovered bases per Mb than the other chromosomes. Chromosome 3 (Scaffold C) had significantly more uncovered bases per Mb than any other chromosome. These differences hold if large uncovered regions (> 1 kbp) are omitted from the analysis indicating these chromosome level differences do not arise from large outlier regions.

To represent *E. globulus* specific content, we defined X46 overrepresented regions as those 100 bp or greater in length with a read coverage high in X46 (151x or higher) but not in BRASUZ1 (less than 108x, which is twice the center value of the 1:1 distribution). This identified 9 Mb of scaffolds in 29,991 regions (mean 300 bp, median 150 bp, max 18,200 bp) representing single-copy equivalents of 39.1 Mb in *E. globulus* and 12 Mb in *E. grandis*, thus accounting for 27.1 Mb genome size difference. Investigation of the distribution of these regions shows they evenly distributed across chromosomes, without any chromosome showing a significant difference in density. These regions are also unbiased as they do not overlap (< 2%) with the duplicated assembly regions in the scaffolds. Overall this analysis shows the apparent genome size difference to be almost entirely attributable to thousands of broadly distributed small changes rather than any single large change such as major chromosomal or chromosomal arm duplication.

Genome comparisons between closely related species indicate that the most significant contributor to genome size variability is changes in the mobile LTR- (Long Terminal Repeat) retrotransposons (TEs)[119]. To investigate the possible contribution of transposable elements (TEs) to the genome size difference between *E. grandis* and *E. globulus*, we identified 214 Mb of TE sequences (~38% of scaffold sequence) in 275,000 intervals in the *E. grandis* scaffolds. We then studied, for both uncovered and high-coverage representative regions, their overlap with these TE sequences as well as their possible duplications since divergence (presence of highly similar regions elsewhere in the genome). Among the uncovered regions, 40% contained TEs over 43% of their sequence, with more than half of the increase (2 Mb) over the average genome content (38%) attributable to 311 regions of complete autonomous LTR-TEs elements. Hence, the uncovered regions do not appear especially enriched in TE derived sequences and, contrary to the findings of other plant genome comparisons[119,120], recent TE expansion in *E. grandis* is only a minor contributor to the genome size difference. We also looked for specific motifs and similarities in the uncovered regions and found no evidence of MITE expansion and no significant association with SSR/LINE/SINE elements. To investigate if the distribution of uncovered regions is more heavily concentrated in TE-rich genomic regions, we compared the total length of TE annotated regions per Mb with that of uncovered regions per Mb. We found that the local uncovered region per Mb is correlated with the local level of TE annotation per Mb, but with the correlation varying by chromosome (P-value < 2e-16, R^2 = 0.299; **Supplementary Table 12**). Hence, while recent TE activity does not explain much of the overall genome size difference, heavily TE-enriched regions appear to preferentially accumulate small genomic changes, despite these not being particularly enriched for TE derived sequence.

To attribute genome specific sequence to gain or loss in *E. grandis* and *E. globulus*, we further analysed representative regions using reads (Tibbits, J. unpublished data) from species that are taxonomically positioned around the *E. grandis*/*E. globulus* split (*E. nitens* and *E. aromaphloia* [ingroup; Combined 37.9x coverage]; *E. cladocalyx*, *E. polyanthemos* and *E. melliodora* [outgroup; Combined 61.9x coverage]). We assigned a most probable estimate of 1:1 coverage in each genome and then used it to obtain a scaled coverage estimate of the uncovered and high-coverage regions in these genomes, using gene regions as a control. Both the ingroup and outgroup were consistent in the number of single copy *E. grandis* genes detected as 1:1 ([0.66 1.33]:1 scaled) coverage (ingroup – 88.2% s.e. 0.6%; outgroup – 85.5% s.e. 0.6%). For X46 uncovered regions (82.7 Mb of unique sequence), the fraction also uncovered in other genomes (< 0.25:1 scaled coverage) was 78.1% for the ingroup (s.e. 2.9%) and 63% for the outgroup (s.e. 3.4%). From this, we estimate that, since divergence, about 58 Mb has been gained in *E.*

grandis and 24 Mb has been lost in *E. globulus*. For X46 overrepresented regions (27.1 Mb of unique sequence), the fraction also overrepresented in other genomes (> 2:1 scaled coverage) was 55.4% for the ingroup (s.e. 4.7%) and 34.4% for the outgroup (s.e. 1.0%). From this, we estimate that, since divergence, about 12 Mb has been lost in *E. grandis* and 15 Mb has been gained in *E. globulus*.

S4. Genetic load, inbreeding depression and heterozygosity

Plant Material

G7J1 family. Tree G7J1, an elite *E. grandis* tree from the Suzano Pulp and Paper company breeding program was artificially self-pollinated by emasculation and flower isolation in 1987. An unrecorded number of putative S_1 seedlings were generated and planted in 1990. In 2007 the selfed origin of the most vigorous G7J1 S_1 offspring tree was verified by genotyping it and its mother tree G7J1 with a set of 15 highly polymorphic microsatellites and observing the correct expected segregation of exclusively maternal alleles in the S_1. Additionally, when selecting BRASUZ1 for genome sequencing, an expanded microsatellite survey found that only 33 (22%) of 147 genotyped loci heterozygous in the selfed parent tree (G7J1) went to homozygosity in BRASUZ1. This is significantly (p = 1.2e-11, two-tailed binomial distribution) below the expected Mendelian proportion of 50% for products of a self-cross indicating strong selection against homozygosity in the BRASUZ1 genome. This analysis also revealed large variation in the level of preserved heterozygosity across the 11 chromosomes, suggesting putative hotspots of genetic load. Leaf, cambium and xylem samples were taken from this S1 tree which was named BRASUZ1, for BRAzil SUZano S_1. In 2011 cambium tissue of an additional set of six S_1 sibs of BRASUZ1 were sampled in the field for a resequencing experiment.

M35D2 S1 family. Tree M35D2, an elite *E. grandis* tree from the Suzano Pulp and Paper company breeding program was artificially self-pollinated using a one-stop pollination method. Approximately 120 putative S_1 seedlings were generated. Selfed origin was checked by genotyping a set of 15 highly polymorphic microsatellites and observing the correct expected segregation of exclusively maternal alleles in the S_1 offspring. Seedlings were planted in single tree plots in 2004 in the vicinity of Itapetininga, SP, Brazil (23°35'30" S/ 48°03'11" W). In July 2011, 33 S_1 sibling tree were randomly sampled, 18 of them were still live and 15 had died within the previous year. Cambium samples were taken from the 33 S1 siblings and the mother tree M35D2 for DNA extraction.

Leaf or cambium tissue from all samples described above were ground to a fine powder in liquid nitrogen and used for DNA extraction using either a Qiagen Plant DNA kit or a CTAB protocol as described earlier. DNA samples were quantified by a picogreen method and sent to JGI for sequencing.

Results

M35D2 Parent resequencing and genotyping. A total of 204.7 million pairs of 2x100 Illumina reads were trimmed for poor quality at the ends and aligned to the reference sequence using *bwa (-q 15)*. From the perfectly mapped pairs with mapping quality q29 or better, a pileup was generated using *samtools mpileup* with BAQ scores disabled. An initial set of heterozygous SNPs were called at sites with 20-100x coverage of bases of quality Q30 or higher, with at least 4 bases confirming both variants. This set had 3.18 million SNPs (0.92%) out of more than 346 million sites with eligible coverage. A much more stringent subset of these was selected for genotyping of the selfed offspring: first, to reduce effects from tandem duplicated regions; we restricted ourselves to sites with coverage in the relative narrow window between 40 and 65 (depth distribution peaks at 52). Furthermore, at each site we required the count of the least observed variant be at least 35.8% of the total, which corresponds to the 95% confidence interval of binomially distributed variants with p = 50% and a total depth of 52. Finally, we chose to genotype only at annotated genes (introns and exons) to avoid effects from mis-aligned intergenic repeat families. The

genotype of intergenic region can usually be inferred by that of the adjacent genes. In total, 308,784 high-confidence heterozygous SNPs within 22,619 genes were selected for genotyping of the selfed offspring.

M35D2 S_1 siblings resequencing and genotyping. DNA from the 33 siblings, 18 live and 15 dead trees, were barcoded, pooled and sequenced as 11 lanes of 2x100 paired-end Illumina reads. The yield varied substantially for the individual barcodes, with four of the dead trees having too little sequence for proper genotyping. Of the remaining trees, the lowest average depth was ~2.4, whereas the highest was ~24.8. The individuals with lower coverage have a fairly large probability of getting mis-genotyped as homozygous in a heterozygous region at any single SNP. To mitigate this effect we genotyped only genes with at least four detectable SNPs of coverage depth > 3. Based on the distribution of the fraction *f* of SNPs per gene genotyped as heterozygous we chose to genotype genes with $f > 40\%$ as heterozygous, $f < 5\%$ as homozygous. Genes with *f* between 5% and 40% where considered ambiguous and not genotyped. By this method, one of the 18 live trees appeared to be heterozygous at all sites (a possible well contamination or barcoding issue) and was excluded from the sample. The genotypes of the remaining 28 siblings, of which 17 were live and 11 had died at the time of DNA collection, were compiled into ordered gene lists, retaining only genes for which at least 20 siblings could be genotyped. Finally, missing genotypes were imputed and genotyping errors were corrected, by a PERL script that first assigned all occurring patterns of 28 genotypes into clusters. All clusters with five or more members were regarded as confirmed genotype patterns, whereas the small, overwhelmingly single-instance clusters contained genotyping errors. Next, patterns with missing genotyping data were assigned to one of the confirmed clusters, if the 20 or more genotyped siblings were 100% compatible with that and only that, cluster. Finally, remaining unassigned patterns were assigned to the closest matching confirmed cluster, provided that the pattern differed from the cluster in at most two siblings. The resulting output file was manually edited to correct a few additional out-of-context genotyping errors, as well as eliminate genes that were locally genotyped as heterozygous in nearly all siblings, in contrast to the genotype patterns of surrounding genes. Such genes are likely tandem-duplicated genes that have been assembled and/or aligned incorrectly. In total, 10,543 genes covering the 11 chromosomes were genotyped for all 28 siblings (and the genotypes of most of the remaining genes can be inferred from that of the surrounding genes). The resulting data file is available as **Supplementary Data 20**. From this data it can be seen that the siblings range from 52% to 79% heterozygosity, with an average of 66%. This deviates grossly from the 50% heterozygosity one would expect under the assumption of random segregation of alleles: even for the sibling with 52% heterozygosity, 5512 heterozygous out of 10,543 genotyped genes, the null hypothesis of 50% heterozygosity is rejected with $p = 1.3\times10^{-6}$ based on a cumulative binomial distribution.

To further examine whether this is due to homozygous versions of certain alleles being lethal, or strongly selected against, we examined, for each bin, the fraction of siblings that were heterozygous as a function of position in the genome (**Extended Data Figure 7**). On most chromosomes peaks are present where heterozygosity is significantly higher than expected for an inbred family. An extreme example is on scaffold_6 where a single peak is present around 44 Mb where 26 of 28 siblings are heterozygous. In most of these cases, both homozygous classes are severely underrepresented suggesting the presence of genetic load on both parental homologs. However, on scaffold_4 there is a region of 25 Mb (11-36 Mb) that is completely devoid of one of the homozygous classes (**Extended Data Figure 7**) suggesting the presence of one or more genetic lesion on one of the parental homologs exhibiting near complete lethality when homozygous. Similar regions devoid of one homozygous class are apparent on scaffold_6, 7, 8, 9 and 10, but these mostly coincide with regions of high heterozygosity, which has both homozygote classes underrepresented. These results suggest a genome-wide distribution of genetic load with some regions exhibiting genotypic ratios consistent with semi-lethality to complete lethality. However, it is not clear whether single, large-effect mutations or the cumulative effect of many small-effect mutations underlie such regions. High-throughput genome resequencing in larger families or populations will be required to unravel the nature of the genetic load underlying each peak.

Predicted functional effects of SNPs segregating in M35D2. As a first step towards understanding the genetic load that may be segregating in this inbred family, we investigated the potential functional effects of the 308,784 SNPs detected in 22,619 protein-coding genes of M35D2 using the SnpEff software[121]. Some SNPs have more than one predicted effect as some genes have more than one transcript annotated and therefore a total of 316,294 "effects" were detected by SnpEff (**Supplementary Table 13**), the majority (74.4%) of which are in non-coding (UTR and intron) regions. Of the 81,061 SNPs in coding sequence, 38,849 (47.9%) are predicted to be non-synonymous mutations and of these, 618 are predicted loss of function (LOF)[122,123] mutations potentially impacting the function of 584 genes (**Supplementary Data 21**). Approximately 40% of the predicted LOF SNPs inactivate genes via premature stop codons and a similar proportion by splice site disruptions. LOF mutations in protein coding genes are likely to contribute to the observed inbreeding depression in M35D2, in addition to other mutations and structural variants (not detected here) in intergenic regions. We further investigated the rate and distribution of silent, missense, nonsense and predicted LOF SNPs, in the region around the peak of heterozygosity on scaffold_6 at ~44 Mb, where 26 out of 28 siblings are heterozygous (**Extended Data Figure 7**). We recorded the frequency (in 1 Mb and 200 kb bins, **Supplementary Table 13**) and distribution (**Supplementary Data 21**) of the different categories of SNP effects relative to the peak (at ~44 Mbp) of heterozygosity (region between ~40.3 Mb and 49.7 Mb) on scaffold_6. Surprisingly, the highest frequency of synonymous (silent) SNPs on the scaffold is found within 2 Mbp of the peak of heterozygosity (**Supplementary Data 21**). Synonymous mutations are traditionally considered neutral, but can have functional effects through codon-usage bias[124]. However, the increased frequency of synonymous mutations in the region could simply reflect increased heterozygosity and nucleotide diversity in the region. Indeed, all SNP effect categories occur at higher frequency in this region compared to the average for scaffold_6 and for the genome (**Supplementary Table 14**). We further tested for enrichment of genes with tandem duplicates in the 584 genes with predicted LOF SNPs based on the hypothesis that inactivation of redundant genes can be tolerated by organisms[122]. We find that 45.4% of the LOF genes have tandem duplicates (**Supplementary Data 21**), which is indeed higher than the genomic average of 34.6%. However, of the 16 genes with LOF SNPs within the region of heterozygosity on scaffold_6, only five had tandem duplicates. One or more of the 11 LOF genes with no tandem duplicates could explain the extreme levels of heterozygosity in this region if any of them have essential functions in *Eucalyptus* trees. For example, the *Arabidopsis* homolog of one of these genes, Eucgr.F03739 (At3G19770.1 encoding VACUOLAR PROTEIN SORTING 9A, VPS9A) was previously shown to be essential for embryo development[125]. Embryogenesis was arrested at torpedo stage in an *Arabidopsis* mutant for the gene. A LOF mutation in the *Eucalyptus* ortholog could result in strong selection against homozygotes. Local accumulation of such recessive mutations on both parental homologs would explain the near absence of both homozygous classes in inbred progeny of M35D2.

S5. Lignocellulosic biomass production

S5.1 SCW related TFs

Over the past five years, studies in *Arabidopsis* and other plants have revealed that secondary cell wall biosynthesis is regulated by a complex hierarchical network of transcription factors most of which belong to two large families namely R2R3-MYB and NAC (NAM/ATAF/CUC)[126-130]. Members of the auxin-response transcription factor families, Aux/IAA and ARF have also been associated with wood formation[131,132]. The structural characteristics and reconstruction of the evolutionary history of these protein families are fully described in companion papers (Hussey *et al.;* Soler *et al.*; Yu *et al.*, manuscripts in preparation). The number of proteins comprising the NAC, R2R3-MYB, Aux/IAA and ARF families in *E. grandis* was compared to those described for the dicots *A. thaliana, P. trichocarpa, V. vinifera* and the monocot model *O. sativa* (**Supplementary Table 15**). Of these species, *E. grandis* has the largest NAC domain family (189 members) in part due to the tandem duplication in several clades).

For the MYB gene family (141 members in *E. grandis*), only *Populus* has a larger R2R3-MYB family possibly as a result of gene retention subsequent to the more recent genome-wide duplication in the Salicaceae lineage. In contrast, the *E. grandis* repertoires of Aux/IAA and ARF genes (26 and 17 members, respectively) do not show expansion relative to other plants suggesting a different evolutionary scenario for auxin-response transcription factors.

The vast majority of the R2R3-MYB proteins described in *Arabidopsis* were also found in *E. grandis*, but some subgroups show a notable expansion in woody species (*E. grandis*, *P. trichocarpa* and *V. vinifera*) such as those known to be involved in the control of the phenylpropanoid and anthocyanin pathways in *Arabidopsis* - subgroups 5 (tt2-related) and 6 (PAP-related) (**Extended Data Figure 9**). The former exhibited a notable specific expansion in *E. grandis* in agreement with the large range of phenylpropanoid-derived compounds in eucalypts. Remarkably, five subgroups seem to be 'woody-preferential' since they only contain members from the three woody genera. Some of these genes show preferential expression in the cambial region of *Eucalyptus* suggesting a role in the regulation of wood development (fully discussed in companion paper, Soler *et al*., manuscript in preparation). Another interesting feature is that the frequency of tandem gene duplications is dramatically higher in the 'woody-preferential' subgroups and in those more expanded in woody species than in the rest of the R2R3-MYB family.

The NAC protein family in *E. grandis* has expanded considerably relative to other woody (*Populus*, *Vitis*) and non-woody (*Arabidopsis*, *Oryza*) plants (described fully in companion paper, Hussey *et al*, manuscript in preparation). Most *E. grandis* NAC (*EgrNAC*) genes occur in tandem duplicate arrays, with many paralogs exhibiting evidence of purifying selection and diverged expression patterns. Twenty-two subfamilies were identified, with EgrNAC proteins showing significant expansions in subfamilies associated with stress response, but this did not include those linked to SCW biosynthesis.

S5.2 Lignin biosynthesis

The biosynthesis of lignin involves the general phenylpropanoid pathway leading to the hydroxycinnamoyl-CoA esters, which are also the precursors of a wide range of end products that vary according to species, cell-types and environmental cues. A survey of the ten gene families involved in the biosynthesis of lignin precursors led to the identification of 174 genes distributed over all 11 scaffolds of the *E. grandis* genome (fully described in companion paper Carocha *et al*., manuscript in preparation). Comparative phylogenetic analysis with *Arabidopsis*, *Populus* and *Oryza* genomes revealed significant expansion of five families in *Eucalyptus* (*i.e.*, PAL, 4CL, CCoAOMT, COMT and CAD). With the notable exceptions of the C4H, F5H and CCR families, the phenylpropanoid gene families generally exhibit multiple tandem gene clusters in short, discrete chromosomal regions. This highlights a preeminent role for tandem gene duplication shaping phenylpropanoid gene families and associated functional diversity in eucalypts. However, it is worth noting that genes encoding *bona fide* lignin biosynthetic enzymes (those highly expressed in stem vasculature tissue) were not impacted by tandem duplication. Protein phylogenetic analyses using *bona fide* landmark genes, coupled to expression profiling of a wide range of tissues/organs, restricted the *E. grandis* core lignin toolbox to a group of 29 genes exhibiting high expression in wood-related, highly lignified tissues (**Extended Data Figure 8** and Carocha *et al*., manuscript in preparation).

The polymerization of monolignols has been attributed to members of the peroxidase and/or laccase protein families. Analysis of the class III peroxidases family known to contain members involved in the lignin oxidative polymerization step[133], revealed a dramatic expansion with more than 190 members detected *versus* 73 in *Arabidopsis* and 89 in *Populus*. Ninety percent of these genes, mainly located in scaffold 1, result from tandem and/or segmental duplications leading to large clusters specific to *E. grandis*, but few of which are expressed in xylem tissue. However, 16 of the 188 have expression

evidence in phloem or immature xylem tissues (**Supplementary Data 22**) collected from stems of actively growing *E. grandis* trees (Li *et al*., manuscript in preparation). Similar gene family expansion is observed for the laccases, with 88 members detected in the *E. grandis* genome versus 17 genes in *Arabidopsis*. Of the 88 *E. grandis* genes, 17 have gene expression evidence with strong phloem or xylem preferential expression, mainly orthologs of *Arabidopsis* LAC4 and LAC17 (**Supplementary Data 22)**, which have previously been shown to be involved in lignification[139]. Identification of these genes in the *E. grandis* genome together with xylem-specific expression provides one of the first genome-wide assessments of lignification genes in a fast-growing forest tree species and a starting place for identifying candidates for functional characterization as well as population-wide association analysis of these genes in *Eucalyptus*.

S5.3 Carbohydrate active enzyme (CAZyme) diversity

Carbohydrate biopolymers, including those deposited in woody biomass, are produced, modified and degraded by the concerted action of carbohydrate active enzymes (CAZymes). CAZyme protein domains fall into five classes based on their biochemical activity[134]. Glycosyl transferases (GTs) catalyze the formation of glycosidic bonds between an acceptor molecule and a donor sugar. Glycosyl hydrolases hydrolyze glycosidic bonds and polysaccharide lyases (PLs) and carbohydrate esterases (CEs) modify carbohydrate biopolymers by the removal of certain chemical bonds. Carbohydrate binding modules (CBMs) bind to carbohydrates, increasing the specificity and action of the 4 enzymatic CAZyme domain classes. CAZyme domains are classified into families based on amino-acid sequence alignments (www.cazy.org). CAZymes are responsible for carbohydrate metabolism and as such, they are of interest to wood scientists. We surveyed the genomes of sequenced rosids for CAZyme genes in order to identify CAZyme domains that are unique or over-represented in the genome of *E. grandis* and which may contribute to the unique biology of *E. grandis*.

CAZyme domain frequency in *E. grandis* (**Supplementary Table 16**) is higher in all classes than that of other rosid species (*P. trichocarpa*, *A. thaliana* and *V. vinifera*). *E. grandis* also has the highest percentage of CAZyme domain containing genes relative to all annotated genes in the genome at 6.9% compared to 5.0% in *P. trichocarpa*, 5.5% in *A. thaliana* and 5.4% in *V. vinifera*. In the *E. grandis* genome, 8 CAZyme families are unique compared to 10, 4 and 2 in the *P. trichocarpa*, *V. vinifera* and *A. thaliana* genomes, respectively, **Supplementary Table 17**). CBM family 54, which is present as a single member in *E. grandis*, has been shown to bind to xylan, yeast cell wall glucan and chitin as demonstrated by affinity assays[135]. The gene containing this domain is expressed in all *Eucalyptus* tissues and is relatively (~2x) highly expressed in immature xylem, young leaf and flower tissues (results not shown). The only GT domain family unique to *E. grandis* is GT11 the putative function of which, according to CAZy, is fucosyl transferase (http://www.cazy.org/GT11.html). The gene containing this domain is expressed in all tissues (results not shown) and has no assigned function in the JGI annotation available on Phytozome. The remaining unique domains in the *E. grandis* genome are GHs. 45 CAZyme domain families are >1.5X over represented in the *E. grandis* genome as compared to the genomes of *P. trichocarpa*, *A. thaliana* and *V. vinifera*. There are 788 genes that contain these domains that are expressed in at least one tissue in *E. grandis*. Of the CAZy domains that are 1.5X over-represented in the *E. grandis* genome, 14 are GTs, 18 are GHs, 6 are CBMs, 5 are CEs and 1 is a PL.

S5.4 Cellulose and xylan biosynthesis

Using homology to *Arabidopsis* genes and Pfam domain analysis (companion paper, Kersting *et al*., manuscript in preparation) we identified candidate homologs encoding 18 enzymatic reactions involved in four major processes in cellulose and xylan biosynthesis (**Supplementary Data 5, Figure 4 main text**) including i) the production of UDP-glucose, ii) UDP-glucose utilization directly into cellulose (CESA), iii) conversion into UDP-glucuronate (UGD [EC: 1.1.1.22]), followed by conversion to UDP-xylose and

iv) the biosynthesis of xylan backbone. All possible family members were identified. Next, gene expression was considered in terms of xylem specific expression relative to other tissues/organs in *E. grandis*. We also took into account each gene's expression relative to other family members/isoforms in xylem, as well as relative to the median (~90,000 FPKM), 90th (1.35 million FPKM), 95th (2.57 million FPKM) and 99th (7.78 million FPKM) percentiles of xylem expression in the entire transcriptome. Considering each gene's relative and absolute expression levels, all members expressed in xylem over median expression were noted (**Supplementary Data 5**). The functional importance of these genes in xylem is highlighted by the fact that all 18 enzymatic steps contained at least one gene member expressed in the 90th percentile of xylem gene expression and most (15 steps) contained at least one member expressed in the 95th and 99th percentile (11 steps). All steps, with the exception of the alternative pathway to UDP-glucose production from sucrose, contained at least one member showing highly xylem-specific expression (> 50% of total transcript abundance in xylem compared to other tissues).

S6. Secondary metabolites and oils

Methods

Genome data bases from eight plant species, for which a complete public genome sequence is available, were mined for the presence of two conserved protein domains from the terpene synthase gene family (TPS), PF01397 and PF03936[136]. The plant taxa included four eudicots (*E. grandis*, *P. trichocarpa*, *V. vinifera* and *A. thaliana*), two monocots (*S. bicolor* and *O. sativa*), as well as one Lycophyte (*S. moellendorffii*) and one moss (*P. patens*). Protein sequences for all TPS genes were extracted and aligned in Genious Pro using 'muscle' with standard parameters. The sequences were then manually adjusted with focus on various conserved regions of terpene synthases such as the 'RLLR', DDXXD and NSE/DTE motifs[137]. The alignment was then truncated to ensure that sites were homologous; resulting in a 462 amino acid alignment with 308 sequences. To create a phylogeny, we first tested which amino acid substitution model provided the maximum likelihood tree with the best AICc (Akaike's information criterion value, corrected for samples size) value and further tested whether gamma distribution estimation and/or proportion of invariable sites estimation improved the AICc value. The amino acid substitution models that were tested were: EHO, EX2, EX3, JTT, LG, UL2, UL3 and WAG. The tree with the highest AICc value was obtained with the JTT model with estimation of invariable sites and estimation of gamma distribution. The phylogenetic tree was then rooted at the split between type I (TPS –c, -e, -f and –h) and type III (TPS –a, -b and -g). Clades were labeled by their TPS sub family category and nodes were colored by species.

Results and discussion

One of the most distinctive features of eucalypts is their high concentration of terpene-dominated essential oils (mono- and sesquiterpenes and their adducts)[138], which are contained mainly in schizogenous secretory cavities in leaves, buds and bark. These terpenoids influence many biotic interactions such as feeding by iconic herbivores like the koala[139] and are also important contributors to atmospheric hydrocarbons[140]. *Eucalyptus* and related genera of Myrtaceae show striking variation in both the profile and the concentration of terpenoids between and within species, individual trees and even amongst different branches on a tree[141]. These differences do not originate from inducible changes as they are stable over many years.

Some of the quantitative variation can be explained by allelic variants in upstream genes of the terpenoid biosynthetic pathways. Quantitative variation of monoterpenes is associated with allelic variants in genes from the chloroplastic methylerythritol phosphate pathway and for sequiterpenes in genes from the cytosolic mevalonate pathway[142]. *E. grandis* leaves contain up to 100 mono- and sesquiterpenes with distinct chemotypic variation in terpenes wherein there are markedly different terpene profiles[138] amongst different individuals. Terpene chemotypes are widespread and common amongst *Eucalyptus* and related

genera. This diversity and chemotypic variation arises from differential regulation of the large number of terpene synthase genes (TPS) in the genome (**Supplementary Data 6**), e.g., Padovan *et al.*[143], because these genes can code for either a single end product or an array of products[144]. Subsequent modifications (e.g., by P450 enzymes) can further add to the diversity of products.

Our analysis of the TPS gene family shows that *E. grandis* has by far the largest family of TPS genes of any plant sequenced to date. We found 113 putative TPS genes in *E. grandis* compared to 83 in *V. vinifera*, 59 in *P. trichocarpa*, 51 and 47 in *O. sativa* and *S. bicolor*, respectively (**Figure 5,** main text). The subfamilies in which *Eucalyptus* is over-represented are subfamily –f, which contains an *Arabidopsis* diterpene synthase making geranyllinalool[145], subfamilies –g and –b, which form cyclic and acyclic monoterpenes, as well as isoprene and subfamily –a, which makes sesquiterpenes[146]. *Eucalyptus* is only over-represented in those TPS clades that that contain proteins for the formation of specialized metabolites; those that contain proteins responsible for the formation of primary metabolites (e.g. gibberellins via ent-kaurene) (subfamilies –c and –e) are represented in similar numbers to other species investigated here. Of the *Eucalyptus* TPS genes the great majority was found to be expressed, with highest levels of expression found in young leaves, followed by mature leaves, flowers, root, shoot tips, phloem and xylem (described in companion paper, Külheim *et al.*, manuscript in preparation). Several *E. grandis* TPS genes appear to be tissue-specific in expression, with profiles that are dominated by either green tissue (leaf, flowers and shoot tips), root or stem (stem, phloem and xylem; Külheim *et al.* manuscript in preparation).

In *Eucalyptus*, the TPS-a1, f, g and b sub-families have undergone a major expansion and duplication following speciation. A similar expansion was observed for *Vitis* TPS-a1, g and b and in *Arabidopsis* for TPS-a1 and b. TPS-h as reported earlier is unique to lycopod *Selaginella*, a species basal to vascular plants. TPS genes were presumably first introduced to the plant genomes via bryophyte mosses, such as *Physcomitrella*. All the angiosperm genomes in our analysis harbor between 30-60 TPS genes except for the major expansions observed in *Vitis* and *Eucalyptus* genomes. This may be indicative of the plant's development and adaptation to certain environments including selective pressure either by humans for grape domestication purposes or in response to biotic and abiotic stress in the Australian continent.

S7. MADS Box gene family

Methods

MADS related genes in the *E. grandis* genome were identified using InterProScan protein motif designations for *E. grandis* V1.1 annotation gene models. Any protein carrying the PFAM motif identifiers PF00319 (SRF-type transcription factor), PF01486 (K-box region), panther identifier PTHR11945 (MADS Box protein) were designated as a MADS Box related protein.

For *A. thaliana*, the 109 annotated MADS box sequences were downloaded from The *Arabidopsis* Information Resource (TAIR10, http://www.arabidopsis.org) and evaluated using the Pfam (http://pfam.sanger.ac.uk/)[147] and SMART databases (http://smart.embl-heidelberg.de/)[148]. Two genes (At1g22590, *AGL87*; At2g26880, *AGL41*) were excluded from the analysis because they did not contain the MADS domain. Six *Arabidopsis* MADS-box genes exhibited polymorphisms in the corresponding AtCode genes (*AGL34*, *AGL35*, *AGL36*, *AGL51*, *AGL64* and *AGL105)*. A new search using a local blastn against the downloaded file TAIR10_cDNA_20102114 recovered three more genes (At1g33070, At5g27810 and At5g27944) containing the MADS domain. Local blastn searches were conducted using a final list of *Arabidopsis* genes (109) against sets of all *Populus trichocarpa* and *Vitis vinifera* coding sequences downloaded from Phytozome v9.0 (Ptrichocarpa_210_cds.fa[50], Vvinifera_145_cds.fa[65] (http://www.phytozome.net/). Presence of a MADS-box domain in the putative MADS-box genes was confirmed using the Pfam and SMART databases by searching against the respective genome sequences.

In all these BLASTN analyses, we used a low stringency e-value cut-off (0.1) to avoid the loss of false negative genes.

For phylogenetic analyses, multiple alignments with complete protein sequences were conducted using the MUSCLE database (http://www.ebi.ac.uk/Tools/msa/muscle/)[149] with default parameters. Phylogenetic trees were constructed with a Neighbor-joining method[150], including pairwise deletion and *p*-distance calculation, using the MEGA 4.1 program[151].

Identification, annotation and classification of *Eucalyptus* MADS-box genes and other rosid plants: We revised the total number of *Arabidopsis* genes annotated as MADS-box transcription factors to 109, where 47 are type II and 63 are type I (**Supplementary Table 18, Supplementary Data 11)**. For *V. vinifera*, our analysis identified a total of 58 MADS-box genes, where 10 were type I MADS-box (not identified previously)[152] and 48 were type II MADS-box. We identified 15 new MADS-box genes not present in previous work[152]. We also found four putative MIKC subfamily genes in *Vitis*, which contained only the K-box domain (*VvAP3.2*, *VvAGL17.5*, *VvSVP-L.4* and *VvFLC.2*). The nomenclature of the *V. vinifera* MADS-box genes was adjusted to include the new genes. In *Populus*, the two available datasets have distinct classifications for the same genes in the MIKC clade[153,154]. The phylogenetic analysis was performed with these two datasets and our dataset. Since there were several ambiguities, it was necessary to reorganize the data[153,154]. We renamed all *P. trichocarpa* MADS-box genes based on the *Arabidopsis* and *Vitis* annotation rules. The search in the *Populus* genome resulted in 102 MADS-box genes, where 62 were type II and 39 were type I (**Supplementary Data 11**). Our analysis identified two putative MADS-box genes containing the K-box domain only (*PtFLC.6* and *PtAGL4*).

Phylogenetic analysis of *Eucalyptus grandis* MADS-box genes: The phylogenetic classification of all *E. grandis* MADS-box genes revealed 70 type II and 35 type I MADS-box genes. Among the MIKC clade, 19 genes were Mα type, eight were Mβ, eight were Mγ and one was Mδ type (**Supplementary Table 18**). To further investigate the phylogenetic relationship of *E. grandis* MADS-box genes and those of other eudicots plants, we generated two phylogenetic trees: one containing only the MIKC clade with 11 subfamilies (**Extended Data Figure 10**) and another with the Mα, Mβ, Mγ and Mδ clades with 12 subfamilies (**Supplementary Data 11)**. The high number of type II MADS-box genes in *E. grandis* is mainly due to the expansion of the *SOC1* subfamily, which accounts for 30 genes. The large number of genes and their chromosomal positions suggest several gene duplication events. The SOC subfamily is encompassed by members of three species (*Arabidopsis*, *Populus* and *Vitis*) (**Extended Data Figure 10**). In *Eucalyptus*, two additional groups are present: one with 10 loci (*EgSOC4*) and the other with three loci (*EgSOC5.1-3*), with no close homologue in the eudicots evaluated in this work. The MIKC phylogenetic tree revealed that the *Eucalyptus SOC2-L* subfamily (closely related to the AGL42-like group) consists of 16 species-specific loci, which underwent many tandem duplications, one locus closely related to AGL19/AGL14 (*EgSOC3*) and no putative homologue for the *SOC1* gene. Another unexpected result was the absence of putative homologues for the *AP1* gene in the *Eucalyptus* genome. On the other hand, we found duplicated *FUL* genes (*EgFUL.1* and *EgFUL.2*) as well as an *EgFUL-Like* gene (*EgFUL-L*). The phylogenetic analysis revealed several other gene duplication events: *EgSEP3* (*EgSEP3.1* and *EgSEP3.2*), *EgFLC* (*EgFLC.1* and *EgFLC.2*), *EgAP3* (*EgAP3.1* and *EgAP3.2*), *EgSVP* (*EgSVP.1* and *EgSVP.2*), *EgANR1-L* (*EgANR1-L.1* and *EgANR1-L.2*). Additional duplications also took place for *EgPI* (*EgPI.1*, *EgPI.2* and *EgPI.3*) and *EgANR1-like* (*EgANR1-L.3*, *EgANR1-L.4* and *EgANR1-L.5*). Moreover, we found species-specific genes in the SVP subclade. Based on our results, we decided to re-name these genes *SVP-like* (*SVP-L*) (**Supplementary Data 11**). The *E. grandis* genome has three *SVP* loci, named *EgSVP-L*, while *Populus* has five (*PtSVP-L*).

The phylogenetic analysis of the Mα/Mβ/Mγ/Mδ group in *Eucalyptus*, *Arabidopsis*, *Populus* and *Vitis* identified many duplication events and species-specific duplication events (**Supplementary Data 11**). *Eucalyptus* shows duplication in *EgAGL80* (four loci), *EgAGL47*/*AGL82* (eight loci), *EgAGL57*/*AGL88*

(eight loci) and *EgAGL29*/*AGL91* (three loci) subfamilies. The phylogenetic tree showed that *Eucalyptus* have many AGL-like specific genes which received the AGLS name, such as *EgAGLS1* (Mα) (three loci), *EgAGLS2* (Mα) (one locus), *EgAGLS3* (Mγ) (one locus), *EgAGLS4* (Mβ/Mγ) (one locus) and *EgAGLS5* (Mγ) (one locus) (**Supplementary Data 11**).

S8. S-domain-Receptor-Like Kinase (SDRLK) gene family

Introduction

Receptor kinases are major players in perceiving and transducing extracellular signals into appropriate cellular responses in living organisms. Fully sequenced plant genomes suggest that receptors-like kinases (RLKs) form one of the largest gene families in plants (e.g., ~600 in *Arabidopsis* ~1100 in *Oryza* and ~1200 in *Populus*[155]. A typical receptor kinase is anchored in the plasma membrane by a single transmembrane (TM) spanning domain that connects the ligand-binding extracellular region (ECR) with the cytoplasmic kinase domain. The binding of a ligand to the ECR of its cognate receptor induces activation (e.g. auto-phosphorylation) of the receptor kinase, which in turn triggers a signaling pathway by phosphorylation of one or more cytoplasmic targets. The sequence and structural diversity in the ECR of RLKs is consistent with involvement of its members in diverse cellular plant processes including the regulation of meristem proliferation[156], organ specification [157-159], hormone signal transduction[157,160,161], defense[162,163] and reproduction[164,165].

The high diversity in the sequence and structure of ECRs of RLKs has been used to subgroup RLKs. The largest and widely studied class of RLKs (~216 members in *Arabidopsis*) is characterized by the presence of leucine-rich repeats (LRRs) in the ECR[166]. The second largest subgroup of RLKs is the S-domain (SD)-RLK group[167], so named because its members share sequence and structural similarity with "S-Domain", the ECR of the *Brassica* S-locus receptor kinase (SRK)[168]. SRK is an integral plasma membrane protein of the stigma epidermis[169] and acts as the female determinant of self-incompatibility (SI), a post-pollination genetic mechanism that prevents germination and entry of genetically related pollen into the pistil and eventually, fertilization[164,170]. "ECR" of Brassica SRK6[168] is predicted to contain four structural domains; the two amino terminal lectin-like domains (LLDs), followed by an epidermal growth factor (EGF)-like domain and a PAN_Apple domain. The SD-RLK family appears to be ubiquitous in plants, as genes exhibiting similarity to the SRK-ECR are found in dicots monocots, conifers, liverworts and algae[166,171].

Results

We identified 253 SD-RLK genes in the *Eucalyptus* genome through domain searches and ortholog finding methods. The analysis of domain structure of 253 *Eucalyptus* SD-RLK family members suggest that *Eucalyptus* genome encodes 101 typical SD-RLKs with one TM, 20 full-length SD-RLKs containing two TMs, 3 full-length SD-RLKs containing 3 TMs and 129 partial duplicates of a typical SD-RLK. Among the partial duplicates include 44 SD-RLKs lacking one or more ECR domains with single TM, 3 SDRLKs containing partial ECR with 2 TMs, 5 proteins containing partial ECR with 1 TM but lacking KD. In addition, 77 proteins that lack TM but retained parts of ECR and/or KD were found including 6 receptor like-proteins and 71 proteins containing full or partial ECR fused to KD. Phylogenetic analysis indicates the expansion of SD-RLKs in *Eucalyptus* through numerous tandem duplications and fusion of one or more structural domains. A total of 222 genes (88%) are present in close proximity of two or more SD-RLK genes (**Supplementary Data 4**). Tandem duplications of RLKs are a consistent feature in plant genomes, although, the pattern of tandem duplication differs among RLK subfamilies in lineage specific manner[167,172].

SUPPLEMENTARY TABLES

Supplementary Table 1 | Genomic libraries included in the *Eucalyptus grandis* genome assembly. 32
Supplementary Table 2 | Summary statistics of the output of the whole-genome shotgun assembly. 33
Supplementary Table 3 | Statistics of the marker placements in high and low coverage scaffolds in the broken *Eucalyptus* assembly. 34
Supplementary Table 4 | Statistics of the final set of joined scaffolds in high and low coverage scaffolds in the broken *Eucalyptus* assembly. 34
Supplementary Table 5 | Final summary assembly statistics for the chromosome-scale assembly. 34
Supplementary Table 6 | EST resources included in the genome annotation pipeline. 35
Supplementary Table 7 | Expressed sequence tag coverage metrics. 36
Supplementary Table 8 | Annotation summary metrics. 38
Supplementary Table 9 | Summary of *Eucalyptus grandis* gene annotation. 39
Supplementary Table 10 | Phytozome v9.1 gene family cluster membership (Eudicot node). 39
Supplementary Table 11 | Pairwise comparison of 5' UTR lengths from FGH, PASA and NGS sources. 40
Supplementary Table 12 | Correlation of local uncovered bases (bp per Mb) with local TE annotation (bp per Mb) by chromosome from mapping of *Eucalyptus globulus* X46 Illumina reads against the *E. grandis* BRASUZ1 v1.0 genome scaffolds. 41
Supplementary Table 13 | Summary of SNP effects predicted genome-wide in MD35D2 and in a region of interest on scaffold_6. 42
Supplementary Table 14 | Per site rate of various SNP effect categories on scaffold_6 and in the region with increased heterozygosity in S1 progeny. 43
Supplementary Table 15 | Comparison of the number of NAC, R2R3-MYB, Aux/IAA and ARF transcription factors in *Arabidopsis thaliana*, *Populus trichocarpa*, *Eucalyptus grandis*, *Vitis vinifera* and *Oryza sativa*. 44
Supplementary Table 16 | Frequency of CAZyme domain families in the genomes of *Eucalyptus grandis*, *Populus trichocarpa*, *Arabidopsis thaliana* and *Vitis vinifera*. 45
Supplementary Table 17 | Unique CAZyme domain families in the *Arabidopsis thaliana*, *Populus trichocarpa*, *Eucalyptus grandis* and *Vitis vinifera* genomes (in this four way comparison). 49
Supplementary Table 18 | MADS-Box gene clade sizes for *Eucalyptus grandis*, *Arabidopsis thaliana*, *Populus trichocarpa* and *Vitis vinifera*. 50

Supplementary Table 1 | Genomic libraries included in the *Eucalyptus grandis* genome assembly.

Library ID	Average Insert Size	Read Number	Assembled Sequence Coverage (X)
GAXU	2,576	1,616,640	1.44
FUNX	2,579	1,829,568	1.66
GGHX	6,028	65,664	0.04
GFTB	6,035	1,544,064	1.59
FUNW	6,048	1,869,504	1.29
GITC	36,348	154,080	0.13
FUNC	38,529	205,056	0.19
GNFS	39,417	80,640	0.07
GGZH	40,182	39,072	0.03
GFUO	41,185	39,168	0.03
CAK	127,458	147,456	0.13
CAL	154,948	147,456	0.12
Total		**7,738,368**	**6.73**

The assembled sequence coverage is shown for each library in the final release.

Supplementary Table 2 | Summary statistics of the output of the whole-genome shotgun assembly.

Statistics are before screening, removal of organelles and contaminating scaffolds and chromosome-scale pseudomolecule construction. The table shows total contigs and total assembled base pairs for each set of scaffolds greater than the given size.

Size	Number	Contigs	Scaffold Size	Base pairs	% Non-gap Base pairs
5,000,000	41	9,568	330,370,723	319,047,727	96.57%
2,500,000	80	13,708	468,110,999	451,527,098	96.46%
1,000,000	131	18,157	556,142,625	531,520,158	95.57%
500,000	190	20,940	598,718,562	567,344,231	94.76%
250,000	246	22,393	619,419,373	582,653,481	94.06%
100,000	363	23,779	638,716,902	595,280,356	93.20%
50,000	531	24,805	649,711,081	603,681,175	92.92%
25,000	790	25,686	658,925,415	611,647,370	92.82%
10,000	1,823	27,915	674,393,605	626,617,165	92.92%
5,000	3,146	30,142	684,026,576	635,804,116	92.95%
2,500	5,231	33,088	691,721,087	643,220,187	92.99%
1,000	5,620	33,527	692,367,534	643,834,880	92.99%
0	6,043	33,950	692,678,021	644,145,367	92.99%

Supplementary Table 3 | Statistics of the marker placements in high and low coverage scaffolds in the broken *Eucalyptus* assembly.

Coverage	Total Scaffolds	Total Bases (Mb)	Marked Scaffolds	Marked Bases (Mb)
>6x	520	478.5 (69%)	108	446.2 (85%)
<= 6x	5,542	214.2 (31%)	66	79.7 (15%)
Total	6,062	692.7	174	525.9

Supplementary Table 4 | Statistics of the final set of joined scaffolds in high and low coverage scaffolds in the broken *Eucalyptus* assembly.

Coverage	Total Scaffolds	Total Bases (Mb)	Joined Scaffolds	Marked Bases (Mb)
>6x	520	478.5 (69%)	125	469.1 (77%)
<= 6x	5,542	214.2 (31%)	143	136.8 (23%)
Total	6,062	692.7	268	605.9

Supplementary Table 5 | Final summary assembly statistics for the chromosome-scale assembly.

Scaffold total	4,952
Contig total	32,762
Scaffold sequence total	691.3 Mb
Contig sequence total	640.6 Mb (7.3% gap)
Scaffold N/L50	5/53.9 Mb
Contig N/L50	2261/67.2 kb

Supplementary Table 6 | EST resources included in the genome annotation pipeline.

Summary of EST resources used to support gene annotation in the BRASUZ1 genome

A) *E. grandis* sequences

Source	Sequence Type	Count
JGI	454 sequencing	1,977,495
Genolyptus (Brazil)	ABI 3100/Megabase	36,554
Kirst[173]	454 sequencing	1,041,876
NCBI	various cDNAs from NR database	1,939

Total : 3,057,864 Filtered for redundancy : 2,916,546

B) Sibling species sequences

Source	Species	Sequence Type	Count
JGI	*E. globulus*	454 sequencing	2,108,985
University of Pretoria (Myburg)	*E. grandis* x *E. urophylla*	~72 million RNA-Seq paired-end reads assembled by BWA/TopHat (Eucpresso dataset[174])	18,894 assembled contigs
Genolyptus	*E. globulus*, *E. pellita*, *E. urophylla*, or mixture	ABI 3100/Megabase	47,452
NCBI	Various	cDNAs from NR database	35,042
ArborGen Proprietary Dataset	Various	cDNAs	218,939[a]
Total : 2,210,373	**Filtered for redundancy : 2,139,444**		

[a] Proprietary dataset not included in table subtotal and downstream analyses

Supplementary Table 7 | Expressed sequence tag coverage metrics.

(A) EST coverage with strict enforcement of intron-exon splice boundaries for matching EST sequences

Chromosome	Total Gene Count	Genes With EST Strict	Genes With Strict Coverage ≥ 30%	Average Strict EST Coverage	Genes With Genomic EST Matches	Genes With Sibling EST Matches	Genes With ArborGen[a] EST Matches	Genomic EST Seq. Count	Sibling EST Seq. Count	ArborGen[a] EST Seq. Count
scaffold_1	3332	2896 (86.91%)	2707 (81.24%)	95.40%	2563	1951	1289	3585	2660	1654
scaffold_2	4445	3860 (86.84%)	3630 (81.66%)	98.38%	3382	2654	1691	4870	3540	2122
scaffold_3	4291	3547 (82.66%)	3222 (75.09%)	22.61%	3000	2260	1319	4336	3040	1627
scaffold_4	2832	2375 (83.86%)	2229 (78.71%)	35.48%	2084	1578	987	2976	2118	1232
scaffold_5	4034	3243 (80.39%)	2950 (73.13%)	72.70%	2802	2078	1108	3971	2807	1345
scaffold_6	5125	4471 (87.24%)	4194 (81.83%)	74.91%	3916	3195	1944	5764	4367	2424
scaffold_7	3537	2985 (84.39%)	2779 (78.57%)	76.19%	2581	1994	1150	3817	2701	1421
scaffold_8	5205	4302 (82.65%)	3979 (76.45%)	93.26%	3764	2821	1665	5299	3881	2100
scaffold_9	3108	2661 (85.62%)	2510 (80.76%)	41.00%	2343	1777	1213	3330	2434	1549
scaffold_10	3750	3297 (87.92%)	3110 (82.93%)	30.34%	2869	2276	1472	4320	3227	1880
scaffold_11	3999	3463 (86.60%)	3272 (81.82%)	92.15%	3047	2443	1575	4409	3356	1927
all chromosomes	43658	37100 (84.98%)	34582 (79.21%)	69.06%	32351	25027	15413	46677	34131	19281
full genome	**46315**	**38977 (84.16%)**	**36267 (78.31%)**	**39.41%**	**33881**	**26023**	**15772**	**48666**	**35358**	**19685**

[a] Proprietary dataset not included in downstream analyses

(B) EST coverage by less strict examination of CDS with any EST overlap

Chromosome	**Total Gene Count**	**Genes With EST Overlap**	**Genes With Overlap Coverage ≥ 30%**	**Average Overlap Coverage**
scaffold_1	3332	2995 (89.89%)	2860 (85.83%)	100.00%
scaffold_2	4445	4023 (90.51%)	3837 (86.32%)	98.38%
scaffold_3	4291	3703 (86.30%)	3452 (80.45%)	33.52%
scaffold_4	2832	2485 (87.75%)	2346 (82.84%)	35.66%
scaffold_5	4034	3408 (84.48%)	3174 (78.68%)	75.62%
scaffold_6	5125	4634 (90.42%)	4431 (86.46%)	74.93%
scaffold_7	3537	3091 (87.39%)	2933 (82.92%)	92.89%
scaffold_8	5205	4501 (86.47%)	4256 (81.77%)	93.86%
scaffold_9	3108	2760 (88.80%)	2648 (85.20%)	64.25%
scaffold_10	3750	3422 (91.25%)	3294 (87.84%)	65.35%
scaffold_11	3999	3592 (89.82%)	3454 (86.37%)	99.52%
all chromosomes	43658	38614 (88.45%)	36685 (84.03%)	85.01%
full genome	**46315**	**40604 (87.67%)**	**38503 (83.13%)**	**39.41%**

EST coverage was calculated for the three separate EST assembly datasets built from different EST input sequence sources (Genomic: any sequence known to be from *E. grandis* BRASUZ1; Sibling: all other sequences from other *Eucalyptus* species; ArborGen: the proprietary 218,939 cDNA sequence dataset from ArborGen, Inc.)

A) EST coverage based on overlap of EST assemblies, enforcing strict splicing of the EST assembly sequence to match intron-exon splice junctions in the corresponding CDS sequences. Gene counts are included as a percentage of the total genes per chromosome in column 2.

B) EST coverage based on overlap of EST assemblies, based on relaxed rules of any overlap of EST assembly sequence to corresponding CDS sequences. We detected that short 5' sequences of some EST assemblies, particular from 454 data, would overhang the CDS 3' splice junction in some cases where the EST was not full length, resulting in an under reporting of EST coverage percentage in table A.

Supplementary Table 8 | Annotation summary metrics.

(A) Primary Transcripts only:

Chromosome	Total Gene Count	Avg. CDS Length (per gene)	Avg. Exon Count (per gene)	Avg. Intron Count (per gene)	Avg. Intron Length (per gene)
scaffold_1	2561	1513.81	4.83	3.85	1655.62
scaffold_2	3408	1513.29	5.14	4.15	1677.64
scaffold_3	3431	1456.22	4.64	3.66	1537.54
scaffold_4	2226	1478.26	4.64	3.66	1696.72
scaffold_5	3314	1394.47	4.32	3.34	1497.06
scaffold_6	3853	1615.03	5.18	4.20	1737.00
scaffold_7	2764	1495.36	4.76	3.79	1665.45
scaffold_8	4169	1446.17	4.70	3.71	1600.25
scaffold_9	2380	1506.13	5.00	4.01	1681.87
scaffold_10	2746	1562.55	5.22	4.24	1749.27
scaffold_11	3065	1526.50	4.95	3.97	1729.51
full genome	**36376**	**1473.95**	**4.76**	**3.78**	**1608.58**

(B) Full annotation:

Chromosome	Total Gene Count	Avg. CDS Length (per gene)	Avg. Exon Count (per gene)	Avg. Intron Count (per gene)	Avg. Intron Length (per gene)
scaffold_1	3332	1613.65	5.54	4.56	1995.83
scaffold_2	4445	1603.73	5.82	4.84	1993.18
scaffold_3	4291	1552.17	5.36	4.38	1893.09
scaffold_4	2832	1563.34	5.29	4.32	2018.95
scaffold_5	4034	1487.37	4.93	3.95	1824.10
scaffold_6	5125	1699.80	5.96	4.98	2085.27
scaffold_7	3537	1572.44	5.45	4.47	1995.26
scaffold_8	5205	1530.20	5.33	4.34	1888.57
scaffold_9	3108	1616.00	5.76	4.77	2015.84
scaffold_10	3750	1675.38	6.03	5.05	2101.47
scaffold_11	3999	1617.47	5.64	4.66	2078.55
full genome	**46315**	**1568.17**	**5.46**	**4.47**	**1936.45**

Supplementary Table 9 | Summary of *Eucalyptus grandis* gene annotation.

Annotation	Count	Proportion of Total Genome Assembly
Protein coding loci[a]	36,796	7.8% (16.2%) [b]
Alternative transcripts[c]	9,939	2.8% (7.3%) [b]
ncRNA	1478	< 0.1 %
tRNA[d]	508	< 0.1%
miRNA[e]	213	< 0.1%
rRNA[d]	269	< 0.1%
snoRNA[d]	175	< 0.1%
Spliceosomal snRNA[d]	125	< 0.1%
sRNA[d]	80	< 0.1%
Cis-regulatory RNA[d] elements[c]	67	< 0.1%
SRP RNA[d]	22	< 0.1%
Antisense RNA[d]	19	< 0.1%
Transposable elements [f]		~50%
Retrotransposons (class I)		44.5%
DNA transposons (class II)		5.6%

[a] Each gene locus represents a single (longest) primary transcript and zero or more alternative spliceforms
[b] Percentage of bases in spliced transcripts, including UTR (unspliced, with introns, in parentheses)
[c] Alternative transcripts are distributed between 6110 gene loci (16.6% of total)
[d] Predictions via RFam
[e] From deep sequencing of BRASUZ1 leaves and xylem small RNA libraries
[f] See Supplementary Note for methods

Supplementary Table 10 | Phytozome v9.1 gene family cluster membership (Eudicot node).

	Eucalyptus gene family cluster count[a]
Any cluster with *Eucalyptus* membership > 0	17883
Eucalyptus > 0 and *A. thaliana* > 0	8469
Eucalyptus > 0 and *P. trichocarpa* > 0	8644
Eucalyptus > 0 and *V. vinifera* > 0	8146
Membership from all four species	6738

[a]Counts from Phytozome v9.1 biomart

Supplementary Table 11 | Pairwise comparison of 5' UTR lengths from FGH, PASA and NGS sources.

The most prominent result per comparison is highlighted.

	FGH vs PASA	
Comparison	**Absolute frequency**	**Percentage**
FGH > PASA	2,550	11.86
PASA > FGH	675	3.14
FGH = PASA	4,081	18.99
FGH only	13,454	62.59
PASA only	734	3.41
Total	**21,494**	
	PASA vs NGS	
PASA > NGS	844	3.42
NGS > PASA	5,463	22.13
PASA = NGS	35	0.14
PASA only	1,698	6.88
NGS only	16,641	67.42
Total	**24,681**	
	FGH vs NGS	
FGH > NGS	2,051	7.28
NGS > FGH	13,431	47.67
FGH = NGS	87	0.31
FGH only	5,191	18.42
NGS only	7,414	26.32
Total	**28,174**	

FGH - FgenesH 5' UTR annotation; PASA - EST-based 5' UTR annotation; NGS - next-generation sequence (RNA-seq) based 5' UTR annotation

Supplementary Table 12 | Correlation of local uncovered bases (bp per Mb) with local TE annotation (bp per Mb) by chromosome from mapping of *Eucalyptus globulus* X46 Illumina reads against the *E. grandis* BRASUZ1 v1.0 genome scaffolds.

Chromosome	R^2	*P*-value
1	0.202	0.0036
2	0.389	3.8e-08
3	0.0947	0.0055
4	0.585	3.8e-09
5	0.210	3.6e-05
6	0.329	5.7e-06
7	0.440	8.5e-08
8	0.0654	0.028
9	0.477	1.2e-06
10	0.268	0.00073
11	0.204	0.0017
Overall	**0.2998**	**<2e-16**

Supplementary Table 13 | Summary of SNP effects predicted genome-wide in MD35D2 and in a region of interest on scaffold_6.

		Genome-wide summary				Scaffold_6 (region around 44 Mb) summary			
Type of SNP	**Type of SNP effects**	**Number of SNP effects**	**%**	**No. of genes impacted**	**Average no. of mutations per gene**	**Number**	**%**	**No. of genes impacted**	**Average no. of mutations per gene**
All SNPs	Coding	81061	25.6	16507	4.85	2603	27.0	614	4.19
	Non-coding	235233	74.4	18203	12.49	7022	73.0	694	10.43
	Total	**316294**				**9625**			
All SNPs	5' UTR	8985	2.8	4186	2.14	226	2.3	151	1.90
	3' UTR	32492	10.3	9418	3.44	935	9.7	366	2.97
	Intron	193501	61.2	15505	12.48	5858	60.9	600	10.10
	Silent	42212	13.3	13306	3.17	1345	14.0	507	2.65
	Missense	38269	12.1	12768	2.99	1239	12.9	465	2.66
	Nonsense	580	0.2	546	1.06	19	0.2	18	1.06
	Splice site disrupter	255	0.1	245	1.04	3	0.0	3	1.00
	Total	**316294**				**9625**			
LOF SNPs	Nonsense	254	41.1	238	1.07	9	52.9	8	1.13
	Splice site donor disrupted	129	20.9	126	1.02	2	11.8	2	1.00
	Splice site acceptor disrupted	126	20.4	123	1.02	1	5.9	1	1.00
	Start lost	109	17.6	108	1.01	5	29.4	5	1.00
	Total	**618**				**17**			

Supplementary Table 14 | Per site rate of various SNP effect categories on scaffold_6 and in the region with increased heterozygosity in S1 progeny.

Mutation	Genome-wide	Entire scaffold_6	Scaffold_6 (region around 44 Mbp)
LOF	1.01E-06	1.37E-06	2.10E-06
Nonsense	9.52E-07	8.89E-07	1.00E-06
Missense	6.28E-05	8.91E-05	1.43E-04
Silent	6.93E-05	9.81E-05	1.59E-04

Supplementary Table 15 | Comparison of the number of NAC, R2R3-MYB, Aux/IAA and ARF transcription factors in *Arabidopsis thaliana*, *Populus trichocarpa*, *Eucalyptus grandis*, *Vitis vinifera* and *Oryza sativa.*

	Arabidopsis	*Populus*	*Eucalyptus*	*Vitis*	*Oryza*
NAC	107	163	189	79	151
R2R3-MYB	126	180	141	123	106
Aux/IAA	29	35	26	23	31
ARF	23	39	17	19	25

Supplementary Table 16 | Frequency of CAZyme domain families in the genomes of *Eucalyptus grandis*, *Populus trichocarpa*, *Arabidopsis thaliana* and *Vitis vinifera*.

CAZyme domain	*Arabidopsis*	*Eucalyptus*	*Populus*	*Vitis*
CBM13	5	5	13	4
CBM16	8	8	12	4
CBM18	10	17	23	6
CBM20	7	9	13	6
CBM22	18	34	11	7
CBM23	1	4	1	1
CBM32	4	15	9	4
CBM4	1	1	0	0
CBM42	2	2	1	0
CBM43	65	56	73	33
CBM45	6	8	6	4
CBM48	9	13	12	9
CBM49	3	5	3	2
CBM50	13	53	39	17
CBM53	4	13	4	4
CBM54	0	1	0	0
CBM57	59	178	152	76
CBM61	2	0	1	0
CE1	63	151	73	51
CE10	49	144	104	65
CE11	5	2	1	1
CE12	4	10	8	4
CE13	12	24	11	7
CE14	2	3	1	1
CE15	1	2	0	1
CE16	104	144	137	94
CE2	0	0	1	0
CE3	0	0	1	1
CE5	4	0	2	0
CE6	2	8	5	3
CE7	19	36	26	16
CE8	70	54	101	55
CE9	6	13	8	6
GH1	66	80	76	48
GH10	11	22	10	6
GH100	9	15	16	11
GH101	0	0	0	1
GH109	5	6	4	1

GH110	0	0	3	0
GH113	2	1	1	1
GH116	4	6	5	3
GH117	2	4	3	2
GH119	0	0	1	0
GH123	1	2	1	2
GH13	11	23	18	16
GH14	10	12	14	10
GH16	33	45	41	35
GH17	51	71	89	52
GH18	12	58	32	22
GH19	14	38	24	14
GH2	8	11	15	5
GH20	3	9	5	5
GH23	1	2	4	2
GH27	10	52	20	16
GH28	71	104	87	70
GH29	1	5	3	3
GH3	15	28	25	26
GH31	5	37	11	8
GH32	10	12	10	14
GH33	1	4	1	2
GH35	25	48	31	32
GH36	5	18	14	8
GH37	2	2	8	1
GH38	4	17	7	6
GH4	1	0	1	1
GH42	0	1	0	0
GH43	4	7	5	5
GH45	2	1	0	0
GH47	5	10	6	9
GH5	17	47	25	26
GH50	0	3	0	0
GH51	2	0	2	1
GH55	17	22	20	19
GH57	0	2	0	0
GH62	0	1	0	0
GH63	2	1	1	1
GH64	0	1	3	1
GH68	0	0	1	0
GH70	0	0	0	1
GH72	0	0	1	0
GH74	1	1	3	1
GH76	1	1	1	1

GH77	2	2	3	2
GH78	0	1	0	0
GH79	4	17	6	7
GH80	0	1	0	0
GH81	2	15	1	1
GH85	3	2	5	1
GH86	0	0	1	0
GH87	0	1	1	1
GH89	1	11	2	2
GH9	26	33	35	22
GH90	0	0	1	0
GH93	0	0	2	0
GH95	1	5	8	1
GT1	124	511	270	126
GT10	4	7	3	2
GT11	0	1	0	0
GT13	1	10	4	5
GT14	33	28	38	27
GT16	1	1	1	0
GT17	7	9	5	3
GT18	1	2	1	1
GT19	1	1	1	1
GT2	37	57	47	45
GT20	12	13	13	10
GT21	1	1	2	1
GT22	3	5	4	3
GT23	3	5	5	3
GT24	1	4	1	1
GT28	5	7	11	7
GT29	3	6	5	3
GT30	1	1	2	1
GT31	34	44	46	28
GT32	6	6	6	3
GT33	1	2	1	1
GT34	8	5	7	5
GT35	2	8	6	4
GT37	14	5	17	5
GT38	0	0	1	0
GT4	28	45	44	32
GT41	167	243	241	159
GT43	4	13	7	4
GT47	39	48	72	37
GT48	12	27	11	11
GT5	9	26	15	14

GT50	1	2	1	1
GT51	0	0	0	1
GT57	2	3	3	3
GT58	2	1	1	1
GT59	1	5	1	1
GT61	8	18	11	6
GT64	3	6	4	2
GT65	12	12	16	12
GT66	2	3	6	3
GT68	15	20	12	10
GT70	0	0	1	0
GT75	7	7	12	7
GT76	1	1	2	1
GT77	18	19	21	12
GT8	43	79	61	44
GT83	0	4	3	1
GT9	1	1	1	1
GT90	9	17	7	11
GT92	10	5	7	5
PL1	32	30	34	22
PL10	2	8	9	7
PL12	5	0	0	0
PL14	0	0	0	1
PL22	0	9	2	2
PL4	0	12	10	6
PL6	9	0	0	0
PL7	1	0	0	1

Supplementary Table 17 | Unique CAZyme domain families in the *Arabidopsis thaliana*, *Populus trichocarpa, Eucalyptus grandis* and *Vitis vinifera* genomes (in this four way comparison).

Arabidopsis Only	*Populus* Only	*Eucalyptus* Only	*Vitis* Only
PL12	CE2	CBM54	GH101
PL6	GH110	GH42	GH70
	GH119	GH50	GT51
	GH68	GH57	PL14
	GH72	GH62	
	GH86	GH78	
	GH90	GH80	
	GH93	GT11	
	GT38		
	GT70		

Supplementary Table 18 | MADS-Box gene clade sizes for *Eucalyptus grandis, Arabidopsis thaliana, Populus trichocarpa* and *Vitis vinifera*.

Species	MIKC	Mα	Mβ	Mγ	Mδ	Non-group	Total
Eucalyptus grandis	69	19	8	8	1	1	107
Arabidopsis thaliana	39	25	22	16	7	-	109
Populus trichocarpa	56	23	10	6	6	1	102
Vitis vinifera	44	9	-	1	4	-	58

EXTENDED DATA FIGURES

Extended Data Figure 1 | RNA-Seq based expression evidence for predicted *Eucalyptus grandis* gene models.

Extended Data Figure 2 | Sharing of protein-coding gene families, protein domains and domain arrangements in *Eucalyptus*, *Arabidopsis*, *Populus* and *Vitis.*

Extended Data Figure 3 | Green plant phylogeny based on shared gene clusters from 17 sequenced plant genomes.

Extended Data Figure 4 | Dating of the *Eucalyptus* lineage-specific whole-genome duplication event.

Extended Data Figure 5 | Genome-wide analysis of tandem gene assemblies.

Extended Data Figure 6 | Illumina PE100 read coverage of the ~760 kb region containing a R2R3-MYB tandem gene array.

Extended Data Figure 7 | Alternative homozygous classes observed in the 28 M35D2 siblings as a function of position on scaffolds 1-11.

Extended Data Figure 8 | Genes involved in lignin biosynthesis in woody tissues of *Eucalyptus.*

Extended Data Figure 9 | Phylogenetic tree of R2R3 MYB sequences from subgroups expanded and/or preferentially found in woody species.

Extended Data Figure 10 | Phylogenetic tree of type II MIKC MADS box proteins.

SUPPLEMENTARY DATA

Supplementary Data 1 | Input scripts for Circos genome overview graph

Supplementary Data 2 | Syntenic genes on *Eucalyptus chr3* and *Populus chr18*

Supplementary Data 3 | Validation of assembly of MYB and COMT tandem gene cluster

Supplementary Data 4 | SD-RLK genes identified in the *Eucalyptus grandis* genome

Supplementary Data 5 | Predicted cellulose and xylan biosynthetic genes and their expression profiles

Supplementary Data 6 | Predicted terpene synthase gene loci

Supplementary Data 7 | Data for green plant phylogeny

Supplementary Data 8 | Shared, unique and over-represented protein domains and domain arrangements

Supplementary Data 9 | Predicted phenylpropanoid biosynthesis genes and their expression profiles

Supplementary Data 10 | Data for woody expanded MYB genes

Supplementary Data 11 | MADS and K-Box genes identified in *Eucalyptus grandis*

Supplementary Data 12 | Heterozygosity analysis of tandem genes and clusters

Supplementary Data 13 | Transposable element analysis

Supplementary Data 14 | *Eucalyptus grandis* 5'UTR annotations

Supplementary Data 15 | GO annotation of *Eucalyptus grandis* genes

Supplementary Data 16 | Predicted ncRNA loci

Supplementary Data 17 | Known and predicted novel miRNAs detected in the *Eucalyptus grandis* genome

Supplementary Data 18 | InterPro domains found in a subset of 968 *Eucalyptus* specific genes

Supplementary Data 19 | Tandem segmental and whole-genome duplications in the *Eucalyptus grandis* genome

Supplementary Data 20 | SNP genotypes for genes segregating in the M35D2 S1 siblings

Supplementary Data 21 | Genes with predicted loss-of-function mutations in M35D2

Supplementary Data 22 | Peroxidase and laccase genes in *Eucalyptus grandis*

LITERATURE CITED

1 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-96, doi:10.1101/gr.828403 (2003).

2 Kullan, A. R. K. *et al.* High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* × *E. urophylla*. *Tree Genet. Genomes* **8**, 163-175, doi:10.1007/s11295-011-0430-2 (2012).

3 Wittenberg, A. H. *et al.* Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Mol. Genet. Genomics* **274**, 30-39 (2005).

4 Petroli, C. D. *et al.* Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. *PLoS One* **7**, e44684, doi:10.1371/journal.pone.0044684 (2012).

5 Paiva, J. A. *et al.* Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. *BMC Genomics* **12**, 137, doi:10.1186/1471-2164-12-137 (2011).

6 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

7 Rambaldi, D. & Ciccarelli, F. D. FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* **25**, 2281-2282, doi:10.1093/bioinformatics/btp381 (2009).

8 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526, doi:10.1371/journal.pone.0016526 (2011).

9 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982, doi:10.1038/nrg2165 (2007).

10 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, doi:10.1186/1471-2105-9-18 (2008).

11 Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002-7013, doi:10.1093/nar/gkp759 (2009).

12 Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open - version 3.0.*

13 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467, doi:10.1159/000084979 (2005).

14 Tenaillon, M. I., Hollister, J. D. & Gaut, B. S. A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* **15**, 471-478, doi:10.1016/j.tplants.2010.05.003 (2010).

15 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

16 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31 (2005).

17 Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516-522 (2000).

18 Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803-816, doi:10.1101/gr.175701 (2001).

19 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

20 Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7 Suppl 1**, S10.11-12, doi:10.1186/gb-2006-7-s1-s10 (2006).

21 Shahmuradov, I. a., Solovyev, V. V. & Gammerman, a. J. Plant promoter prediction with confidence estimation. *Nucleic Acids Res.* **33**, 1069-1076, doi:10.1093/nar/gki247 (2005).

22 Kan, Z., Gish, W., Rouchka, E., Glasscock, J. & States, D. UTR reconstruction and analysis using genomically aligned EST sequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8**, 218-227 (2000).

23 Molina, C. & Grotewold, E. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6**, 25, doi:10.1186/1471-2164-6-25 (2005).

24 Soderlund, C. *et al.* Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS Genet.* **5**, e1000740, doi:10.1371/journal.pgen.1000740 (2009).

25 Tanaka, T., Koyanagi, K. O. & Itoh, T. Highly diversified molecular evolution of downstream transcription start sites in rice and *Arabidopsis*. *Plant Physiol.* **149**, 1316-1324, doi:10.1104/pp.108.131656 (2009).

26 Zhao, X., Valen, E., Parker, B. J. & Sandelin, A. Systematic clustering of transcription start site landscapes. *PloS One* **6**, e23409, doi:10.1371/journal.pone.0023409 (2011).

27 Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255-265, doi:10.1101/gr.084541.108 (2009).

28 Hefer, C. A., Van der Merwe, K., Mizrachi, E. & Myburg, A. A. The *Eucalyptus* Genome Integrative Explorer (EucGenIE, http://eucgenie.org/). (2012).

29 Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178-1186, doi:10.1093/nar/gkr944 (2012).

30 Sterck, L., Billiau, K., Abeel, T., Rouzé, P. & Van de Peer, Y. ORCAE: online resource for community annotation of eukaryotes. *Nature Meth.* **9**, 1041, doi:10.1038/nmeth.2242 (2012).

31 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

32 Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178-192, doi:10.1093/bib/bbs017 (2012).

33 Pervouchine, D. D., Knowles, D. G. & Guigo, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273-274, doi:10.1093/bioinformatics/bts678 (2013).

34 Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211-215, doi:10.1093/nar/gkn785 (2009).

35 Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783-795, doi:10.1016/j.jmb.2004.05.028 (2004).

36 Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581-1590 (2004).

37 Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175-182 (1998).

38 Rhee, S. Y. *et al.* The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31**, 224-228 (2003).

39 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337, doi:10.1093/bioinformatics/btp157 (2009).

40 Gardner, P. P. *et al.* Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* **39**, D141-145, doi:10.1093/nar/gkq1129 (2011).

41 Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439-441 (2003).

42 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664, doi:10.1101/gr.229202 (2002).

43 Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919-929 (2001).

44 Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15 Spec No 1**, R17-29, doi:10.1093/hmg/ddl046 (2006).

45 Song, D. *et al.* Computational prediction of novel non-coding RNAs in *Arabidopsis thaliana*. *BMC Bioinformatics* **10 Suppl 1**, S36, doi:10.1186/1471-2105-10-S1-S36 (2009).

46 Lu, Z. J. *et al.* Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* **21**, 276-285, doi:10.1101/gr.110189.110 (2011).

47 Carra, A. *et al.* Cloning and characterization of small non-coding RNAs from grape. *Plant J.* **59**, 750-763, doi:10.1111/j.1365-313X.2009.03906.x (2009).

48 Klevebring, D. *et al.* Genome-wide profiling of *Populus* small RNAs. *BMC Genomics* **10**, doi:10.1186/1471-2164-10-620 (2009).

49 Velasco, R. *et al.* The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nature Genet.* **42**, 833-+, doi:10.1038/Ng.654 (2010).

50 Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604, doi:10.1126/science.1128691 (2006).

51 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).

52 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

53 Prufer, K. *et al.* PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* **24**, 1530-1531, doi:10.1093/bioinformatics/btn223 (2008).

54 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-158, doi:10.1093/nar/gkm952 (2008).

55 Friedlander, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407-415, doi:10.1038/nbt1394 (2008).

56 Meyers, B. C. *et al.* Criteria for annotation of plant MicroRNAs. *Plant Cell* **20**, 3186-3190, doi:10.1105/tpc.108.064311 (2008).

57 Dai, X. & Zhao, P. X. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* **39**, W155-159, doi:10.1093/nar/gkr319 (2011).

58 Friedman, A. R. & Baker, B. J. The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.* **17**, 493-499, doi:10.1016/j.gde.2007.08.014 (2007).

59 Geddy, R. & Brown, G. G. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC Genomics* **8**, 130, doi:10.1186/1471-2164-8-130 (2007).

60 Amarasinghe, V., Dharmawardhana, P., Elser, J. & Jaiswal, P. Functional Annotation of Plant Genomes. in *Cereal Genomics II* (eds Pushpendra K. Gupta & Rajeev K. Varshney) Ch. 7, 155-176 (Springer Netherlands, 2013).

61 Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196-203, doi:10.1093/nar/gkp931 (2010).

62 Youens-Clark, K. *et al.* Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* **39**, D1085-1094, doi:10.1093/nar/gkq1148 (2011).

63 Jaiswal, P. Gramene database: a hub for comparative plant genomics. *Method. Mol. Biol.* **678**, 247-275, doi:10.1007/978-1-60761-682-5_18 (2011).

64 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

65 Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467, doi:10.1038/nature06148 (2007).

66 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).

67 Kersting, A. R., Bornberg-Bauer, E., Moore, A. D. & Grath, S. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol. Evol.* **4**, 316-329, doi:10.1093/gbe/evs004 (2012).

68 Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, doi:10.1038/ng.740 (2011).

69 Robbertse, B., Yoder, R. J., Boyd, A., Reeves, J. & Spatafora, J. W. Hal: an automated pipeline for phylogenetic analyses of genomic data. *PLoS Curr.* **3**, RRN1213, doi:10.1371/currents.RRN1213 (2011).

70 Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041-1052, doi:10.1006/jmbi.2000.5197 (2001).

71 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

72 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

73 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564-577, doi:10.1080/10635150701472164 (2007).

74 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

75 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

76 Wang, H. C. *et al.* Rosid radiation and the rapid rise of Angiosperm-dominated forests. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3853-3858, doi:10.1073/pnas.0813376106 (2009).

77 Zhu, X. Y. *et al.* Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids. *BMC Evol. Biol.* **7**, 217, doi:10.1186/1471-2148-7-217 (2007).

78 Wurdack, K. J. & Davis, C. C. Malpighiales phylogenetics: gaining ground on one of the most recalcitrant clades in the Angiosperm tree of life. *Annu. Rev. Ecol. Evol. Syst.* **96**, 1551-1570, doi:10.3732/ajb.0800207 (2009).

79 Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100, doi:10.1038/nature09916 (2011).

80 Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3, doi:10.1186/gb-2012-13-1-r3 (2012).

81 D'Hont, A. *et al.* The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-217, doi:10.1038/nature11241 (2012).

82 Duarte, J. M. *et al.* Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, Vitis and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61, doi:10.1186/1471-2148-10-61 (2010).

83 Lee, E. K. *et al.* A functional phylogenomic view of the seed plants. *PLoS Genet.* **7**, e1002411, doi:10.1371/journal.pgen.1002411 (2011).

84 Zhang, N., Zeng, L., Shan, H. & Ma, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.* **195**, 923-937, doi:10.1111/j.1469-8137.2012.04212.x (2012).

85 THE_ANGIOSPERM_PHYLOGENY_GROUP. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* **161**, 105-121, doi:10.1111/j.1095-8339.2009.00996.x (2009).

86 Moore, M. J. *et al.* Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *Int. J. Plant Sci.* **172**, 541-558 (2011).

87 Soltis, D. E. *et al.* Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98**, 704-730, doi:10.3732/ajb.1000404 (2011).

88 Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4623-4628, doi:10.1073/pnas.0907801107 (2010).

89 Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725-732, doi:10.1038/nrg2600 (2009).

90 Murat, F. *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545-1557, doi:10.1101/gr.109744.110 (2010).

91 Abrouk, M. *et al.* Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479-487, doi:10.1016/j.tplants.2010.06.001 (2010).

92 Salse, J. *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.* **15**, 122-152, doi:10.1016/j.pbi.2012.01.001 (2012).

93 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

94 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

95 Proost, S. *et al.* i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11, doi:10.1093/nar/gkr955 (2012).

96 Fostier, J. *et al.* A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27**, 749-756, doi:10.1093/bioinformatics/btr008 (2011).

97 Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177-190, doi:10.1093/molbev/mss214 (2013).

98 Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5737-5742, doi:10.1073/pnas.0900906106 (2009).

99 Van Bel, M. *et al.* Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* **158**, 590-600, doi:10.1104/pp.111.189514 (2012).

100 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).

101 Bremer, B. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* **161**, 105-121 (2009).

102 Gandolfo, M. A., Nixon, K. C. & Crepet, W. L. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). *Am. J. Bot.* **85**, 964-974, doi:10.2307/2446363 (1998).

103 Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18724-18728, doi:10.1073/pnas.0909766107 (2010).

104 Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5897-5902, doi:10.1073/pnas.1001225107 (2010).

105 Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytol.* **192**, 266-301, doi:10.1111/j.1469-8137.2011.03794.x (2011).

106 Heled, J. & Drummond, A. J. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61**, 138-149, doi:DOI 10.1093/sysbio/syr087 (2012).

107 Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273-279, doi:10.1093/nar/gkh458 (2004).

108 Dubchak, I., Poliakov, A., Kislyuk, A. & Brudno, M. Multiple whole-genome alignments without a reference organism. *Genome Res.* **19**, 682-689, doi:10.1101/gr.081778.108 (2009).

109 Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721-731, doi:10.1101/gr.926603 (2003).

110 Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107, doi:10.1101/gr.809403 (2003).

111 Hurst, L. D., Williams, E. J. & Pal, C. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* **18**, 604-606 (2002).

112 Davila Lopez, M., Martinez Guerra, J. J. & Samuelsson, T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One* **5**, e10654, doi:10.1371/journal.pone.0010654 (2010).

113 Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res.* **11**, 2009-2019, doi:10.1101/gr.212701 (2001).

114 Rodgers-Melnick, E. *et al.* Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Res.* **22**, 95-105, doi:10.1101/gr.125146.111 (2012).

115 Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155 (2000).

116 Tibbits, J. F. G., McManus, L. J., Spokevicius, A. V. & Bossinger, G. A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. *Plant Mol. Biol. Rep.* **24**, 81-91, doi:10.1007/bf02914048 (2006).

117 Grattapaglia, D. & Bradshaw, H. D. Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Can. J. For. Res.* **24**, 1074-1078 (1994).

118 Praca, M. M., Carvalho, C. R., Ribeiro, C. & Novaes, D. B. Nuclear DNA content of three *Eucalyptus* species estimated by flow and image cytometry. *Aus. J. Bot.* **57**, 524-531, doi:10.1071/bt09114 (2009).

119 Bennetzen, J. L., Ma, J. X. & Devos, K. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.-London* **95**, 127-132, doi:10.1093/aob/mci008 (2005).

120 Agren, J. A. & Wright, S. I. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Res.* **19**, 777-786, doi:10.1007/s10577-011-9229-0 (2011).

121 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).

122 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

123 Lloyd, J. & Meinke, D. A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. *Plant Physiol.* **158**, 1115-1129, doi:10.1104/pp.111.192393 (2012).

124 Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32-42, doi:10.1038/nrg2899 (2011).

125 Goh, T. *et al.* VPS9a, the common activator for two distinct types of Rab5 GTPases, is essential for the development of *Arabidopsis thaliana*. *Plant Cell* **19**, 3504-3515, doi:10.1105/tpc.107.053876 (2007).

126 Grima-Pettenati, J., Soler, M., Camargo, E. & Wang, H. Transcriptional regulation of the lignin biosynthetic pathway revisited: New players and insights. in *Adv. Bot. Res.* Vol. 61 (ed In Lise Jouanin and Catherine Lapierre) 173-218 (Academic Press, 2012).

127 Zhao, Q. & Dixon, R. A. Transcriptional networks for lignin biosynthesis: more complex than we thought? *Trends Plant Sci.* **16**, 227-233, doi:10.1016/j.tplants.2010.12.005 (2011).

128 Demura, T. & Fukuda, H. Transcriptional regulation in wood formation. *Trends Plant Sci.* **12**, 64-70 (2007).

129 Zhong, R., Lee, C. & Ye, Z.-H. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends Plant Sci.* **15**, 625-632 (2010).

130 Hussey, S. G., Mizrachi, E., Creux, N. M. & Myburg, A. A. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Front. Plant. Sci.* **4**, 325, doi:10.3389/fpls.2013.00325 (2013).

131 Nilsson, J. *et al.* Dissecting the molecular basis of the regulation of wood formation by auxin in hybrid aspen. *Plant Cell* **20**, 843-855 (2008).

132 Sundberg, B., Uggla, C. & Tuominen, H. in *Cell and Molecular Biology of Wood Formation* (eds R Savidge, J Barnett, & R. Napier) 169-188 (BIOS Scientific Publishers, Oxford, UK, 2000).

133 Marjamaa, K., Kukkola, E. M. & Fagerstedt, K. V. The role of xylem class III peroxidases in lignification. *J. Exp. Bot.* **60**, 367-376, doi:10.1093/jxb/ern278 (2009).

134 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-238, doi:10.1093/nar/gkn663 (2009).

135 Dvortsov, I. A. *et al.* Carbohydrate-binding properties of a separately folding protein module from beta-1,3-glucanase Lic16A of *Clostridium thermocellum*. *Microbiology* **155**, 2442-2449, doi:10.1099/mic.0.026930-0 (2009).

136 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-141, doi:10.1093/nar/gkh121 [doi] (2004).

137 Martin, D. *et al.* Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biology* **10**, 226 (2010).

138 Keszei, A., Brubaker, C. L. & Foley, W. J. A molecular perspective on terpene variation in Australian Myrtaceae. *Aus. J. Bot.* **56**, 197-213, doi:10.1071/bt07146 (2008).

139 Moore, B. D., Foley, W. J., Wallis, I. R., Cowling, A. & Handasyde, K. A. Eucalyptus foliar chemistry explains selective feeding by koalas. *Biol. Lett.* **1**, 64-67, doi:10.1098/rsbl.2004.0255 (2005).

140 He, C. R., Murray, F. & Lyons, T. Monoterpene and isoprene emissions from 15 *Eucalyptus* species in Australia. *Atmos. Environ.* **34**, 645-655, doi:10.1016/s1352-2310(99)00219-8 (2000).

141 Padovan, A., Keszei, A., Wallis, I. R. & Foley, W. J. Mosaic eucalypt trees suggest genetic control at a point that influences several metabolic pathways. *J. Chem. Ecol.* **38**, 914-923, doi:10.1007/s10886-012-0149-z (2012).

142 Külheim, C. *et al.* The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytol.* **191**, 1041-1053, doi:10.1111/j.1469-8137.2011.03769.x (2011).

143 Padovan, A., Keszei, A., Kollner, T. G., Degenhardt, J. & Foley, W. J. The molecular basis of host plant selection in *Melaleuca quinquenervia* by a successful biological control agent. *Phytochemistry* **71**, 1237-1244, doi:10.1016/j.phytochem.2010.05.013 (2010).

144 Keeling, C. I. & Bohlmann, J. Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytol.* **170**, 657-675, doi:10.1111/j.1469-8137.2006.01716.x (2006).

145 Herde, M. *et al.* Identification and regulation of TPS04/GES, an *Arabidopsis* geranyllinalool synthase catalyzing the first step in the formation of the insect-induced volatile C16-Homoterpene TMTT. *Plant Cell* **20**, 1152-1168 (2008).

146 Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212-229, doi:10.1111/j.1365-313X.2011.04520.x (2011).

147 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281-D288, doi:10.1093/nar/gkm960 (2008).

148 Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5857-5864, doi:10.1073/pnas.95.11.5857 (1998).

149 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1-19, doi:10.1186/1471-2105-5-113 (2004).

150 Saitou, N. & Nei, M. The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).

151 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599, doi:10.1093/molbev/msm092 (2007).

152 Diaz-Riquelme, J., Lijavetzky, D., Martinez-Zapater, J. M. & Carmona, M. J. Genome-wide analysis of MIKCC-Type MADS Box genes in grapevine. *Plant Physiol.* **149**, 354-369, doi:10.1104/pp.108.131052 (2009).

153 De Bodt, S., Theissen, G. & Van de Peer, Y. Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* **23**, 1293-1303, doi:10.1093/molbev/msk016 (2006).

154 Leseberg, C. H., Li, A. L., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **378**, 84-94, doi:10.1016/j.gene.2006.05.022 (2006).

155 Lehti-Shiu, M. D., Zou, C., Hanada, K. & Shiu, S. H. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* **150**, 12-26, doi:10.1104/pp.108.134353 (2009).

156 Clark, S. E. Cell signalling at the shoot meristem. *Nat. Rev. Mol. Cell Biol.* **2**, 276-284, doi:10.1038/35067079 (2001).

157 Chen, M. K., Wilson, R. L., Palme, K., Ditengou, F. A. & Shpak, E. D. ERECTA family genes regulate auxin transport in the shoot apical meristem and forming leaf primordia. *Plant Physiol.* **162**, 1978-1991, doi:10.1104/pp.113.218198 (2013).

158 Tanaka, H. *et al.* Novel receptor-like kinase ALE2 controls shoot development by specifying epidermis in *Arabidopsis*. *Development* **134**, 1643-1652, doi:10.1242/dev.003533 (2007).

159 Colcombet, J., Boisson-Dernier, A., Ros-Palau, R., Vera, C. E. & Schroeder, J. I. *Arabidopsis* SOMATIC EMBRYOGENESIS RECEPTOR KINASES1 and 2 are essential for tapetum development and microspore maturation. *Plant Cell* **17**, 3350-3361, doi:10.1105/tpc.105.036731 (2005).

160 Tanaka, H. *et al.* Abiotic stress-inducible receptor-like kinases negatively control ABA signaling in *Arabidopsis*. *Plant J.* **70**, 599-613, doi:10.1111/j.1365-313X.2012.04901.x (2012).

161 Li, J. *et al.* BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell* **110**, 213-222 (2002).

162 Gómez-Gómez, L. & Boller, T. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Mol Cell.* **5**, 1003-1011 (2000).

163 Song, W. Y. *et al.* A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* **270**, 1804-1806 (1995).

164 Kachroo, A., Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. Allele-specific receptor-ligand interactions in *Brassica* self-incompatibility. *Science* **293**, 1824-1826 (2001).

165 Escobar-Restrepo, J. M. *et al.* The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. *Science* **317**, 656-660 (2007).

166 Lehti-Shiu, M. D., Zou, C., Hanada, K. & Shiu, S. H. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* **150**, 12-26 (2009).

167 Shiu, S. H. *et al.* Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16**, 1220-1234 (2004).

168 Naithani, S., Chookajorn, T., Ripoll, D. R. & Nasrallah, J. B. Structural modules for receptor dimerization in the S-locus receptor kinase extracellular domain. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12211-12216, doi:10.1073/pnas.0705186104 (2007).

169 Stein, J. C., Dixit, R., Nasrallah, M. E. & Nasrallah, J. B. SRK, the stigma-specific S locus receptor kinase of *Brassica*, is targeted to the plasma membrane in transgenic tobacco. *Plant Cell* **8**, 429-445 (1996).

170 Nasrallah, J. B. Recognition and rejection of self in plant self-incompatibility: comparisons to animal histocompatibility. *Trends Immunol.* **26**, 412-418 (2005).

171 Sasaki, G. *et al.* Multiple receptor-like kinase cDNAs from liverwort *Marchantia polymorpha* and two charophycean green algae, *Closterium ehrenbergii* and *Nitella axillaris*: Extensive gene duplications and gene shufflings in the early evolution of streptophytes. *Gene* **401**, 135-144, doi:10.1016/j.gene.2007.07.009 (2007).

172 Lehti-Shiu, M. D. & Shiu, S. H. Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 2619-2639, doi:10.1098/rstb.2012.0003 (2012).

173 Novaes, E. *et al.* High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**, 312 (2008).

174 Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F. & Myburg, A. A. De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* **11**, 681, doi:10.1186/1471-2164-11-681 (2010).

175 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).