

RESEARCH

Open Access



From genome to gene expression: the genomic landscape of a hybrid species of *Eucalyptus urophylla* × *Eucalyptus grandis* and its divergence from parental species hybrid

Guo Liu^{1,2}, Jianzhong Luo^{1,2*}, Wanhong Lu¹, Yan Lin¹, Lei Zhang³, Jingyi Pan¹, Jiangbo Zhai¹ and Anying Huang¹

Abstract

Background *Eucalyptus urophylla* × *Eucalyptus grandis* (*E. urograndis*) is a globally significant forest tree species renowned for its rapid growth, high yield, and exceptional wood production efficiency. A comparative analysis of its parental genomes, coupled with an in-depth investigation of the expression patterns of wood-related genes, will provide critical genomic resources to enhance research and utilization of this superior eucalypt hybrid species.

Results In this study, we present a draft genome assembly consisting of 592.09 Mb of data, with 99.91% anchored to 11 pseudochromosomes. The assembly achieved a contig N50 of up to 3.73 Mb and a scaffold N50 of up to 58.62 Mb. Gene annotation and evaluation revealed that the *E. urograndis* genome contains 32,151 genes, of which 93.50% were fully annotated using Benchmarking Universal Single-Copy Orthologs (BUSCOs). Based on evolutionary analysis, *E. grandis* and *E. urograndis* are estimated to have diverged approximately 2.90 million years ago (Mya). Additionally, 131 gene families were found to be significantly expanded, and 475 positively selected genes (PSGs) were identified in the *E. urograndis* genome. Furthermore, RNA sequencing (RNA-seq) technology was employed to analyze allele-specific expression patterns of key enzymes involved in cellulose, xylan, and lignin biosynthesis. Several allele-specific expression genes (ASEGs) were identified, potentially associated with heterosis in *E. urograndis*.

Conclusions The chromosomal-level genome assembly of *E. urograndis* presented in this study serves as a valuable genomic resource for eucalyptus molecular breeding, provides novel insights into its evolution, wood formation improvement, and adaptability, and enhances our understanding of genetic and molecular mechanisms underlying heterosis in *Eucalyptus* hybrids.

Keywords Genome, Chromosome-scale assembly, *Eucalyptus urophylla* × *Eucalyptus grandis*, Cellulose and lignin biosynthesis, Allele-specific expression

*Correspondence:

Jianzhong Luo
luojz@caf.ac.cn

¹Research Institute of Fast-growing Trees, Chinese Academy of Forestry, Zhanjiang 524022, China

²State Key Laboratory of Tree Genetics and Breeding Laboratory, Chinese Academy of Forestry, Beijing 100091, China

³Guangxi State-own Dongmen Forest Farm, Chongzuo 532199, China



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

Eucalyptus is one of the most important fast-growing timber species for the fiber, energy, and paper industries worldwide, and improvements in its genetic constitution largely determine the competitiveness of the corresponding industries [1]. Moreover, eucalyptus is a widely known effective reforestation tree species owing to its fast growth and high adaptability to various environments [2]. As the most important means of genetic improvement, hybrid breeding has been widely used to achieve great improvements in eucalyptus. Currently, *Eucalyptus* spp. and their hybrids are among the world's leading sources of woody biomass and are the main hardwoods used for pulpwood and timber [3]. In particular, almost all the main afforestation varieties in major eucalyptus plantations worldwide are hybrids [4, 5]. In general, hybrids are superior to parents in terms of growth rate, yield, disease resistance, and viability [6, 7]. Among the various *Eucalyptus* species, the hybrid species *Eucalyptus urophylla* × *Eucalyptus grandis* (*E. urograndis*) is the most popular and has a fast growth rate [8]. Additionally, it shows obvious heterosis, rapid growth, high yield, a straight shape, a neat forest phase, wide adaptability, 5.5 years of stock up to 213.80 m³/hm², and annual growth up to 40.95 m³/hm² [9]. The texture of *E. urograndis* timber is straight, the basic density of the finished material is 500 ± 20 kg/m³ [9], and the mechanical strength is about 80.82% [10, 11]. Furthermore, *E. urograndis* has a wide range of uses, such as for building materials, furniture, agricultural tools, fuelwood, communication poles, and pillars, thereby involving the construction, furniture, agriculture, communication, and manufacturing industries [12]. Considering the above, obtaining a high-quality genome of *E. urograndis* at the chromosomal level is of great commercial importance and essential for understanding the basis of its superior properties to extend these attributes to other hybrid species.

The formation of hybrids is highly important for the environmental adaptability of species. Hybridization is an evolutionary phenomenon that has fascinated biologists for centuries [13]. Before the advent of whole-genome sequencing, hybridization played a clear role in the evolutionary history of many extant taxa, particularly those of plants [14]. Studies have indicated that superparental expression genes may play a crucial role in the formation of heterosis, accompanied by significant quantitative and qualitative changes in gene expression between the hybrid and its parental lines [15]–[16]. Since all genetic material in the hybrid originates from its parents, no novel genes beyond those present in the parental genomes are introduced at the compositional level. Therefore, the emergence of heterosis is likely associated with the regulatory mechanisms governing gene expression [17]. A high-quality genome assembly would be a

valuable genomic resource for elucidating the genetic basis of the fast growth and adaptive evolution of *E. urograndis*. Furthermore, integrating transcriptomic technologies to compare and analyze the expression patterns of heterosis-related genes between hybrids and their parents represents an effective approach to uncovering the underlying mechanisms of heterosis formation.

These historic genetic concepts of heterosis (the dominance model, the overdominance model, and the epistasis concept) were developed at a time when not even DNA had been identified as the carrier of genetic information and, therefore, cannot be directly associated with molecular principles [18]. The difference in gene expression constitutes a critical source of phenotypic diversity, and the specificity of diploid hybrids arises from the co-expression of parental alleles at specific loci. The expression of alleles at specific loci may differ between hybrids and their parents, with five primary patterns identified: (1) alleles expressed in both parents but silenced in hybrids (parental co-silencing); (2) alleles expressed exclusively in one parent but not in the hybrid (parent-specific expression); (3) alleles expressed only in hybrids and absent in both parents (hybrid-specific expression); (4) alleles expressed in hybrids and one parent (single-parent expression, SPE); and (5) alleles expressed in both parents and hybrids [19]. The first four patterns correspond to differences in gene expression quality, specifically presence/absence variation (PAV), whereas the fifth pattern pertains to differences in gene expression quantity. Compared with the parental genotype, the gene expression effect of the F1 hybrid can be categorized into additive and non-additive expression patterns. Additive expression refers to the scenario where the hybrid exhibits an intermediate phenotype between the two parents, whereas non-additive expression encompasses four distinct patterns: overdominance, high-parent dominance, low-parent dominance, and underdominance. In recent years, increasing support for the notion that variation in the transcriptional regulation of hybrids compared with their parental inbred lines is related to increased performance of hybrids [20–22].

Although Shen et al. [23] have published a high-quality genome assembly of *E. urograndis* and investigated its evolutionary history and diversification of eucalyptus, the improved varieties of *E. urograndis* examined in this study were specifically derived from the identified paternal (G46) and maternal (U16) varieties. More importantly, this study integrated genomics and transcriptomics to investigate gene expression patterns in the cellulose, xylan and lignin biosynthesis pathways of *E. urograndis*, with the aim of elucidating the regulatory mechanisms underlying allele-specific gene expression in *E. urograndis* and its parental species. This research is expected to provide novel insights into the molecular

basis of heterosis in eucalyptus. In this study, we present a chromosomal-level genome assembly of *E. urograndis* (DH32-29), achieved through the integration of long-read single-molecule real-time (SMRT) PacBio long sequence reads and high-throughput chromosome conformation capture (Hi-C) data. We then used it to analyze hybridization dominance by genome annotation and gene family analyses. Furthermore, the evolutionary trajectories of *E. urograndis* and other 12 typical plant species, including *E. grandis* [24, 25], *E. melliodora* (GCA_004368105.3), *E. pauciflora* (GCA_007663325.1), *Solanum tuberosum* [26], *Zea mays* [27], *Oryza sativa* [28], *Phoenix dactylifera* [29], *Populus trichocarpa* [30], *Physcomitrella patens* [31], *Arabidopsis thaliana* [32], *Vitis vinifera* [33], and *Theobroma cacao* [34], were explored using comparative genomic analysis. Additionally, the genomes of *E. grandis* and *E. urograndis* were used to analyze collinearity; the *E. grandis* and *V. vinifera* genomes were combined to analyze the whole-genome duplication (WGD) of the *E. urograndis* genome. Based on the results of comparative genomics, functional annotation, expansion of gene families, and positive selection genes (PSGs) were determined. Furthermore, through the application of second-generation high-throughput transcriptome sequencing, the biosynthetic pathways of cellulose, xylan, and lignin have been extensively investigated, as these pathways are closely associated with the heterosis of wood properties in *E. urograndis*. Overall, the *E. urograndis* genome in this study provides valuable genomic resources for further research and utilization of this excellent hybrid eucalyptus species. Moreover, it could expand our understanding of the hybridization of woody plants, provide a powerful tool to accelerate cross-breeding, and enhance our understanding of comparative biology and biotechnology.

Results

Complete genome survey

A total of 93.82 Gb of clean stLFR (single-tube long-fragment reads) data were obtained after filtering using the BGISEQ-500 platform (BGI, Shenzhen, China). On the basis of the clean data, the genome size, heterozygosity, and repeat sequence percentage were estimated to be 566.72 Mb, 2.71%, and 59.00%, respectively, using 17-kmer analysis when the coverage was equal to 135. A repetitive peak and a heterozygous peak represented repeat sequences and heterozygous loci in the genome of *E. urograndis*, respectively, according to the GenomeScope profile (Supplementary Fig. S1A). Hence, *E. urograndis* presented high heterozygosity.

SOAPdenovo was used for the preliminary assembly of high-throughput sequencing data, and the results are shown in Supplementary Table S1. The genome size was 903.56 Mb; the contig N50 and scaffold N50 lengths

were 331 bp and 1.63 kb, respectively. The total number of contigs and scaffolds were 3,054,117 and 1,553,947, respectively. The GC content was 39.19%. Using the contig sequences (length ≥ 500) of the filtered data as the window for assembled genome sequences, the average GC content and average depth of nonduplicate fragments were recorded and visualized, which revealed that the genome was not contaminated by that of other species (Supplementary Fig. S1B). Additionally, to evaluate assembly quality and integrity as well as to detect potential sample contamination, the NT (Nucleotide Collection) comparison of the filtered data using BLAST demonstrated that the top six species were eucalyptus (Supplementary Table S2). Therefore, the *E. urograndis* samples used for genome sequencing were relatively pure, and subsequent analysis could be performed.

Chromosomal-level genome assembly

SMRT analysis with the PacBio platform generated 8,179,761 reads for 124.40 Gb, and the total length of the *E. urograndis* genome was 591.88 Mb, with a contig N50 of 3.73 Mb and a GC content of up to 39.44%. The minimum length of the contig was 1380 bp, and the largest contig size reached 12.70 Mb (Supplementary Table S1).

To assess the integrity of the genome assembly accurately, the GC distribution and depth/coverage were determined, and BUSCO analysis was performed. The distributions of the GC content and sequencing depth were relatively concentrated, with an average GC content of 39.44%, and the scatter plot distribution was similar to the Poisson distribution (Supplementary Fig. S1B). The alignment rate of all small fragment reads to the genome was about 96.36%, and the percentage of genome coverage was about 99.80%. BUSCO analysis revealed that the assembly captured 94.90% of the conserved single-copy orthologous genes, including 92.90% of the complete genes and 2.00% of the fragment genes (Supplementary Table S3). These results indicate good consistency and integrity, confirming the good quality of the chromosomal-level assembled genome of *E. urograndis*.

Using Hi-C sequencing, we obtained 652,905,028 raw read pairs, amounting to 130.58 Gb of Hi-C data. After the Hi-C reads were mapped against the assembly of the *E. urograndis* genome, 154,512,422 of valid pair reads, accounting for 70.88% of the mapping pair reads, were used for the Hi-C analysis. The contigs in the draft assembly were then anchored and oriented into a chromosomal-scale assembly using the Hi-C scaffolding approach. Juicer and 3D-DNA were used to construct chromosomal-level scaffolds (Supplementary Fig. S1C); the total length of the *E. urograndis* genome was 592.09 Mb, with a scaffold N50 of 58.62 Mb. Moreover, 11 chromosomes were generated, with lengths ranging

from 39.56 to 72.11 Mb, accounting for 99.91% of the assembly (Supplementary Table S4).

Annotation of the *E. urograndis* genome

In this study, both homology searches in known repeat databases and *de novo* predictions were combined to identify both tandem repetitive sequences and interspersed repetitive sequences. A total of 281.02 Mb of repetitive sequences were generated, accounting for 47.48% of the entire genome assembly after redundancy removal (Supplementary Table S5). Long terminal repeats (LTRs) were the most abundant, amounting to 205.45 Mb, which is 34.68% of the genome. The second most common type of LTR was DNA elements (43.55 Mb), accounting for 7.36% of the genome, followed by long interspersed nuclear elements (LINEs) (17.73 Mb, 3.00% of the genome) and unknown repeats (11.79 Mb, 1.99% of the genome). Additionally, 0.32 Mb

(0.05%) of the *E. urograndis* genome contained short interspersed nuclear elements (SINEs). The spatial distribution of these tandem repeat sequences across the 11 chromosomes was variable, with higher densities located differently on different chromosomes (Fig. 1).

Recent studies have revealed that about 90.00% of genes in eukaryotic genomes are transcriptional genes, and only 0.01–0.02 of these transcriptional genes encode proteins, most of which are transcribed as ncRNAs [35, 36]. Although ncRNAs are not translated into proteins, they have important biological functions in plant immunity [37]. In the *E. urograndis* genome, annotation of ncRNA genes revealed 538 tRNAs, 144 snRNAs, 297 snoRNAs, 994 miRNAs, and 747 rRNAs with total lengths of 40,196, 20,226, 51,140, 93,638, and 162,195 bp, respectively (Supplementary Table S6). Among them, miRNAs are the most abundant (994 copies) and play vital roles in diverse biological processes, such as plant growth

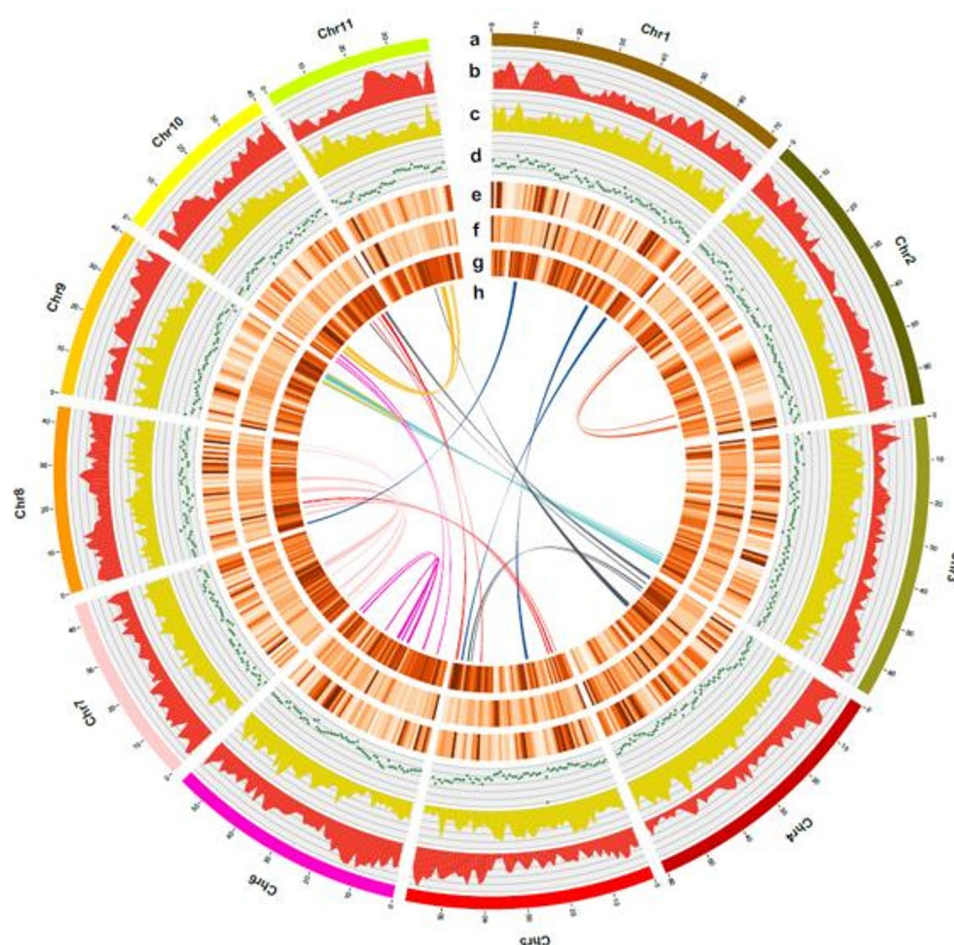


Fig. 1 Overview of the *Eucalyptus urograndis* genome. The genome features at 1 Mb intervals across the 11 chromosomes. Units on the circumference represent megabase values. **A** The outer layer of the colored blocks represents the chromosomes, and Chr1 ~ 11 represent the number of chromosomes. **B** Density of genes (0–1) **C** Density of repetitive sequences (0–1); **D** guanine-cytosine (GC) content (0–1); **E** density of DNA transposons (0–0.15) **F** density of tandem repeat sequences (0–0.1) **G** density of single nucleotide polymorphisms (SNPs) ($0-0.90 \times 10^{-2}$) and **H** self-collinearity analysis of the genome, and removed collinearity blocks with a span of less than 30 and those with a gene number of less than 10

and development and hormone and stress responses [36, 38–40]. However, rRNA has the longest total length among the four types of ncRNAs. In general, eukaryotic rRNAs are classified into four types (5 S, 5.8 S, 18 S, and 28 S) [41]. Among the four types of rRNAs, 5 S rRNA genes were the most abundant (365 copies, total length of 43,303 bp, amounting to 0.73×10^{-2} of the genome). However, 18 S rRNA genes were the longest (89,142 bp, amounting to 1.51×10^{-2} of the genome).

The combined results from homology-based, *de novo*, and RNA-seq methods for protein-coding gene prediction and functional annotation predicted a total of 32,151 protein-coding genes in the *E. urograndis* genome. The average lengths of the genes and CDSs were 4776.80 bp and 1322.38 bp, respectively. The average number of exons per gene, average exon length, and average intron length were 4.99, 265.25 bp, and 866.77 bp, respectively (Supplementary Fig. S2, Supplementary Table S7). As shown in Supplementary Table S7, the gene length, exon length, and intron length were greater in the *E. urograndis* genome than in the genomes of the other four species (*E. grandis*, *P. granatum*, *P. trichocarpa*, and *H. brasiliensis*). BUSCO analysis revealed that 93.50% of the single-copy orthologous genes retrieved from the assembly were fully annotated, including 88.80% of the complete single-copy genes and 4.70% of the complete duplicated BUSCOs (Supplementary Table S3). These results showed that the genome annotation accurately represented the completeness of the gene set of the *E. urograndis* genome.

The protein sequences predicted by gene structure were compared with those in five protein databases; a total of 31,019 genes (96.48% of all the predicted genes) were functionally annotated using the SwissProt, KEGG, TrEMBL, InterPro, and GO databases (E -values $\leq 10^{-5}$), and 18,099 genes were annotated in all five databases. KEGG pathway-based analysis assigned 25,010 genes to a total of 136 pathways, with the majority of these genes involved in metabolic pathways (ko01100, 4089 genes) and the biosynthesis of secondary metabolites (ko01110, 2633 genes) (Supplementary Fig. S3A). The phenylpropanoid biosynthesis pathway (ko00940), which is closely related to Lignin biosynthesis, was enriched with 338 genes. In particular, 690 genes were enriched in the plant-pathogen interaction pathway (ko04626), which was significantly related to disease resistance in *E. urograndis*. GO analyses were performed to classify the functions of the *E. urograndis* genes—19,906 genes were classified into 3478 GO terms and 45 subcategories (Supplementary Fig. S3B, Supplementary Table S8). According to the GO annotation, the main GO category of biological process (BP) comprised 1,920 GO terms. The terms metabolic process (GO:0008152, 8,473 genes), cellular process (GO:0009987, 6,385 genes), and single-organism process (GO:0044699, 4,915 genes) were enriched with a high

number of genes. The other main GO categories of cellular component (CC) included membrane (GO:0016020, 2,319 genes), cell part (GO:0044464, 2,085 genes), cell (GO:0005623, 2,085), and 422 other GO terms. Additionally, the molecular function (MF) category included binding (GO:0005488, 11,898 genes), catalytic activity (GO:0003824, 9,509 genes), transporter activity (GO:0005215, 1,032 genes), and 1,130 other GO terms.

Gene family analysis

The protein sequences of 13 species (*P. patens*, *P. trichocarpa*, *A. thaliana*, *T. cacao*, *E. grandis*, *E. urograndis*, *E. melliodora*, *E. pauciflora*, *V. vinifera*, *S. tuberosum*, *Z. mays*, *O. sativa*, and *P. dactylifera*) were used to identify the gene families using the OrthoMCL method. Consequently, 32,151 *E. urograndis* genes were clustered into 28,078 gene families, including 142 unique *E. urograndis* families, whereas 4,073 *E. urograndis*-specific genes were unclustered (Fig. 2A, Supplementary Table S9). In addition, the results revealed 1,454 gene families specific to four *Eucalyptus* species, 428 gene families specific to *E. grandis* and *E. urograndis*, and 577 gene families that were shared by the 13 species.

We also identified 13,098 gene families that were shared by four eucalyptus species (*E. urograndis*, *E. grandis*, *E. melliodora*, and *E. pauciflora*), and 236 gene families were unique to *E. urograndis*. Moreover, *E. urograndis* shared 14,738, 14,997, and 15,249 gene families with *E. grandis*, *E. pauciflora*, and *E. melliodora*, respectively (Fig. 2B).

Phylogenetic analysis and divergent time Estimation

To understand the phylogenetic position of *E. urograndis*, a phylogenetic tree was constructed on the basis of 577 single-gene families in the 13 sequenced plant genomes. The phylogenetic analysis revealed that *E. urograndis* and *E. grandis* were more closely related to *E. melliodora* than to *E. pauciflora* in the *Eucalyptus* subgenus, which is consistent with their phylogenetic classification on the basis of morphological characteristics. The Myrtales lineage represented by four *Eucalyptus* species formed a sister clade to Malvids with the basic rosoid lineage Vitales, whereas *P. trichocarpa* was grouped with Malvids (Fig. 2C). These results were consistent with the genomic analysis of *E. grandis* [25]. Moreover, on the basis of the gene family clustering results and phylogenetic relationships among species, the divergence times were estimated. Considering that the hybrid *E. urograndis* originated through artificial crossing, this estimation can, to some extent, reflect the divergence between its parental species. The results indicated that the parental species of *E. urograndis*, namely *E. urophylla* and *E. grandis*, diverged approximately 2.90 Mya. This finding

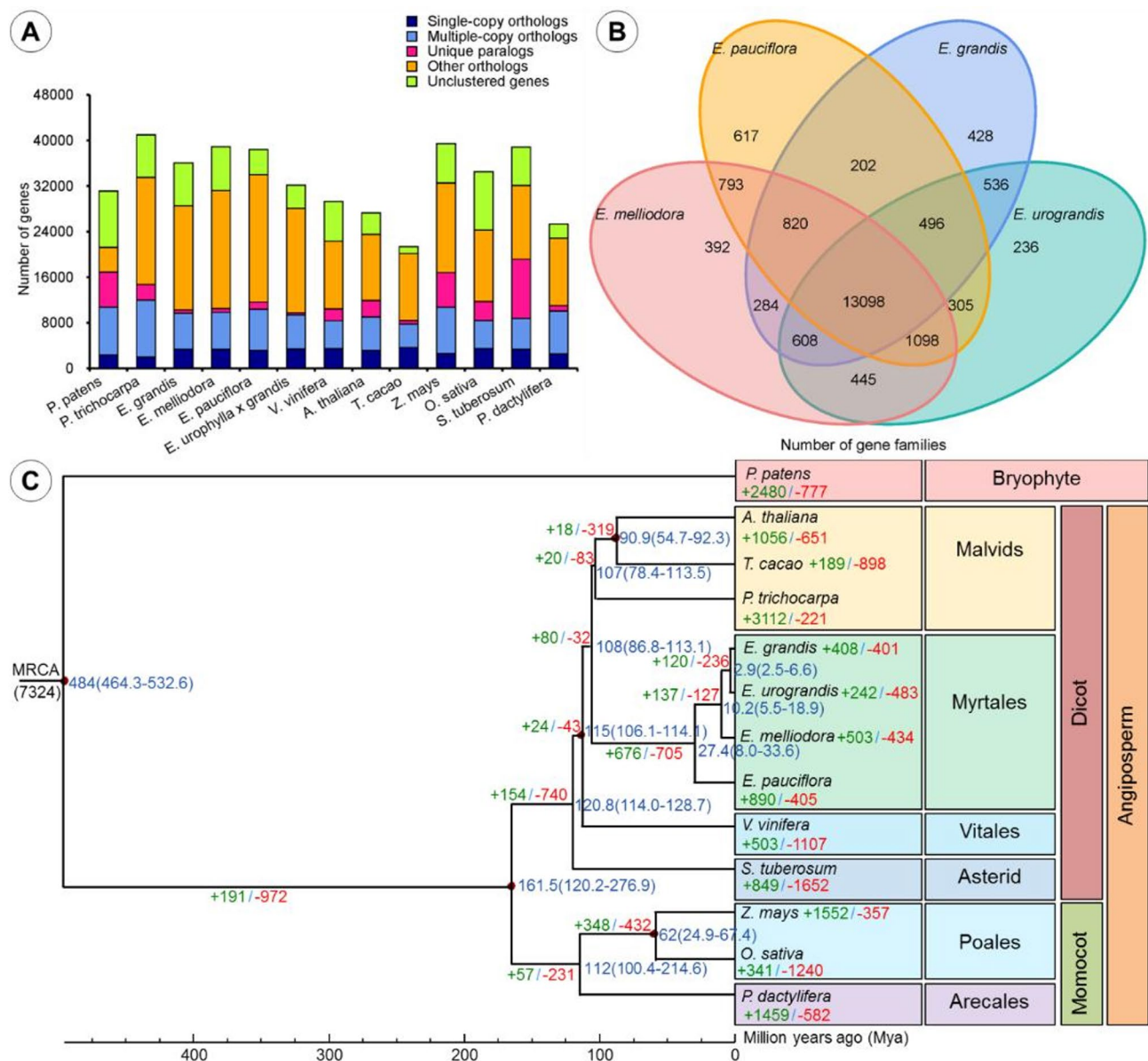


Fig. 2 Gene families and evolution of the *E. urograndis* genome. **A** Comparison of homologous gene numbers. The x-axis indicates the species, and the y-axis indicates the number of genes. The dark blue color represents single-copy orthologs; the light blue color represents multicopy orthologs; the pink color represents unique paralogs; the light yellow color represents other orthologs; and the green color represents unclustered genes. **B** Venn diagram of shared and unique homologous gene families among the four eucalyptus species (*E. urograndis*, *E. grandis*, *E. melliodora*, and *E. pauciflora*). **C** Phylogenetic tree of 13 plant species. The blue numbers denote the divergence time of each node (Mya, million years ago); the green and red numbers on the branch represent the total number of expanded and contracted gene families, respectively

implied that *E. urophylla* and *E. grandis* began to diverge as early as 2.90 Mya.

Expansion and contraction of gene families

Gene families expand and contract in plants because plants experienced selection pressure during evolution [42]. Hence, these processes play major roles in the phenotypic diversification of plants [43, 44]. In this study, the expansion and contraction of the gene families in *E. urograndis* were analyzed by comparing the gene families

in the other 12 representative species, and the results are shown in Fig. 2C. Phylogenetic analysis of 6,899 gene families revealed 242 expanded gene families encompassing 344 genes and 483 contracted gene families encompassing 618 genes in *E. urograndis*. Among the 242 expanded gene families, 131 gene families significantly expanded ($p < 0.05$). The *E. urograndis* genome presented the least gene family expansion and the most contractions among the genomes of the other three eucalyptus species.

Furthermore, we used the 131 gene families (828 genes) annotated to the KEGG and GO databases that were significantly expanded. KEGG pathway analysis grouped 131 gene families into 79 pathways (Supplementary Table S10). In particular, 20 genes were enriched in the phenylpropanoid biosynthesis pathway (ko00940), of which 16 genes encoded cinnamyl alcohol dehydrogenase (CAD; EC:1.1.1.195). CAD catalyzes the final step in phenylpropanoid synthesis, leading to the production of lignin monomers, and is closely related to plant growth and development and resistance to pathogen invasion [45]. The remaining four genes encoded cytochrome P450 (C3'H; EC:1.14.1496), which plays a key role in plant development and defense. Additionally, 18 significantly expanded genes encoding sucrose synthase (SUSY; EC:2.4.1.13) were enriched in the starch and sucrose metabolism pathway (ko00500). The Susy enzyme plays a central role in source-sink coordination and carbon flow in trees. Gessler et al. [46] and Dominguez et al. [47] demonstrated that SUSY activity influences C allocation to developing woody tissues and maintains the C balance in the whole tree. Moreover, the plant-pathogen interaction pathway (ko04626) was enriched with seven significantly expanded genes, including three genes encoding PTI1-like tyrosine-protein kinase 3 (PTK3; EC:2.7.11.1) and four genes encoding calmodulin (CALM), which possibly explains the wide suitability and high disease resistance of *E. urograndis*.

Positive Selection of Genes in the *E. urograndis* genome

The screening and analysis of PSGs in the genome can aid in understanding the specific evolutionary adaptability of a species [48]. A total of 475 PSGs were identified in the *E. urograndis* genome. KEGG pathway annotation revealed high enrichment of PSGs related to metabolic pathways (64 genes, ko01100), biosynthesis of secondary metabolites (35 genes, ko01110), and biosynthesis of amino acids (10 genes, ko01230) (Supplementary Tables S11). The 20 significantly enriched pathways with *p* values of annotated and enriched genes are shown in Supplementary Fig. S4. Four PSGs were enriched in the phenylpropanoid biosynthesis pathway (ko00940), and weijuan_GLEAN_10017689, encoding 4-coumarate-CoA ligase (4CL; EC:6.2.1.12), is key to the phenylpropanoid pathway and participates in monolignol biosynthesis through p-coumaroyl-CoA production. Furthermore, weijuan_GLEAN_10016490, encoding caffeic acid 3-O-methyltransferase (COMT; EC:2.1.1.68), catalyzes the multi-step methylation reactions of hydroxylated monomeric lignin precursors and is believed to occupy a pivotal position in the lignin biosynthetic pathway. The gene of weijuan_GLEAN_10029076 encodes cinnamoyl CoA reductase (CCR; EC:1.2.1.44), which is the first enzyme in the monolignol-specific

branch of the lignin biosynthetic pathway, where it converts feruloyl-CoA to coniferaldehyde [42]. In addition, 6 PSGs were enriched in the plant-pathogen interaction pathway (ko04626), including 3 PSGs encoding a calcium-binding protein (CML), and weijuan_GLEAN_10010339, weijuan_GLEAN_10032007, and weijuan_GLEAN_10022765 encoding WRKY41, LRR receptor-like serine/threonine-protein kinase, and nitric oxide synthase (NOS; EC:1.14.13.39), respectively. This pathway could be attributed to the strong environmental adaptability and disease resistance of *E. urograndis*.

Collinearity and whole genome duplication (WGD) analysis

To investigate the natural evolutionary course of *E. urograndis*, we first performed a collinearity analysis between the genomes of *E. urograndis* and *E. grandis*, and the results are shown in Fig. 3A. The collinearity analysis between *E. urograndis* and *E. grandis* revealed 39,042 paralogous gene pairs (59.60%) in 636 syntenic regions. Hence, the *E. urograndis* genome showed high synteny with the *E. grandis* genome.

Chromosomes 1, 2, 5, 9, and 11 of the *E. urograndis* genome presented high collinearity with chromosomes 2, 3, 4, 8, and 9 of the *E. grandis* genome, respectively (Fig. 3B). However, the other six chromosomes of *E. urograndis* underwent a fusion event with the *E. grandis* chromosomes. For example, chromosome 8 of *E. urograndis* shares many paralogous gene pairs with chromosomes 11 and 1 of *E. grandis* (Fig. 3B). Moreover, chromosomal recombination events during the evolution of tree species and the formation of hybrids were confirmed. Additionally, more translocation events were detected in chromosomes 3, 4, 6, 7, 8, and 10 of *E. urograndis* than in the *E. grandis* chromosomes. Thus, *E. urograndis* may have experienced chromosomal fusion, inversion, or other rearrangement events.

To estimate the WGD events that occurred during the natural evolutionary course of *E. urograndis*, the *E. grandis* and *V. vinifera* genomes were analyzed to explore the WGD history of the *Eucalyptus* genus (Fig. 3C). By determining the distribution of the synonymous substitution rate (*K*s) using syntenic ortholog pairs within each genome, species differentiation was found to occur between the genomes of *E. urograndis* and *E. grandis* at about 0.03, those of *E. grandis* and *V. vinifera* at about 1.21, and those of *E. urograndis* and *V. vinifera* at about 1.23. The distributions of *K*s values of paralogous pairs in the genomes of *E. urograndis*, *E. grandis*, and *V. vinifera* presented two clear peaks in all three genomes (recent WGD and ancient WGD). Moreover, the *K*s peak values of the three genomes for the ancient WGD events were very close, indicating that they experienced gamma events (γ), the hexaploidization event shared by core eudicots [47, 48]. Using the γ time (117 ± 1 Mya)

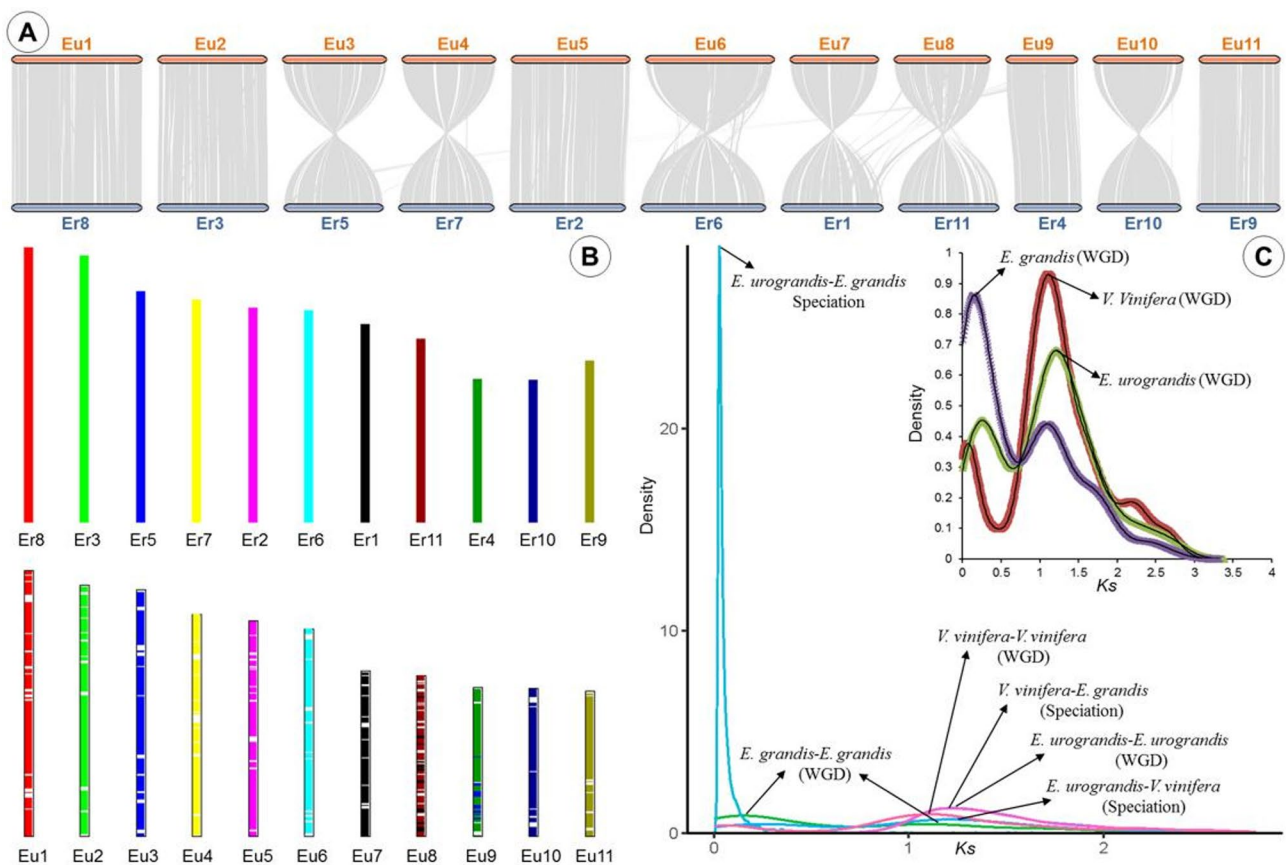


Fig. 3 Comparative genomic analyses. **A** Collinearity patterns between the genomic regions of *E. urograndis* (Eu1~Eu11) and *E. grandis* (Er1~Er11). **B** Bar graph illustration of collinearity blocks on chromosomes between genomes. The same color indicates the corresponding chromosomes between the two genomes. **C** Distribution of synonymous substitution levels (K_s) of syntenic orthologous (speciation) and paralogous genes (WGD) between *E. urograndis* and *E. grandis* or *V. vinifera*. Notes: Eu1 to Eu11 represent the chromosome numbers 1 to 11 of *E. urograndis*, and Er1 to Er11 represent the chromosome numbers 1 to 11 of *E. grandis*

[49] and peak K_s values (1.11) of *V. vinifera*, the synonymous substitutions per site per year were estimated as 4.75×10^{-9} for *Eucalyptus*; thus, the time of speciation was estimated at about 2.68 Mya, which is close to the divergence time (2.90 Mya) between *E. urograndis* and *E. grandis*. Furthermore, the WGD of *E. urograndis* was estimated to have occurred about 127.79 Mya. Additionally, the ancient WGD of *E. grandis* was estimated to have occurred at 114.87 Mya in this study, which is close to the value (105.90–113.90 Mya) reported by Myburg et al. [25]. However, the densities of the three genomes for recent WGD events were relatively low, indicating a small-scale duplication event.

Allele-specific expression associated with Cellulose, xylan, and Lignin Biosynthesis in *E. urograndis* and its parental species

Shao et al. [50] reported that allele-specific expression (ASE), or an imbalance between the expression levels of two parental alleles in a hybrid, has been suggested as a mechanism of heterosis. Cellulose and lignin are the

major components in forest wood and non-food biomass [51], and lignin is the second most abundant polymer in nature. Cellulose and lignin are the key components of plant cell walls and provide remarkable strength to wood. Understanding the biosynthetic mechanism of cellulose and lignin in *E. urograndis*, the most popular hybrid species, is important. In this study, the analysis of wood chemical components in *E. urograndis* and its parental species revealed that the levels of three chemical components in *E. urograndis* were more closely aligned with those of the maternal parent (*E. urophylla*) and were significantly different from those of the paternal parent (*E. grandis*) (Supplementary Fig. S5). These findings are consistent with the conclusions drawn by Shen et al. [23], who reported greater maternal heritability of chemical components in hybrid wood than paternal heritability during genetic analysis of the F1 generation of *E. urograndis*. Based on a comparative genomic analysis of *E. urograndis* and 12 other species, we identified putative functional homologs of genes encoding six enzymatic steps in cellulose biosynthesis from sucrose (Fig. 4 and

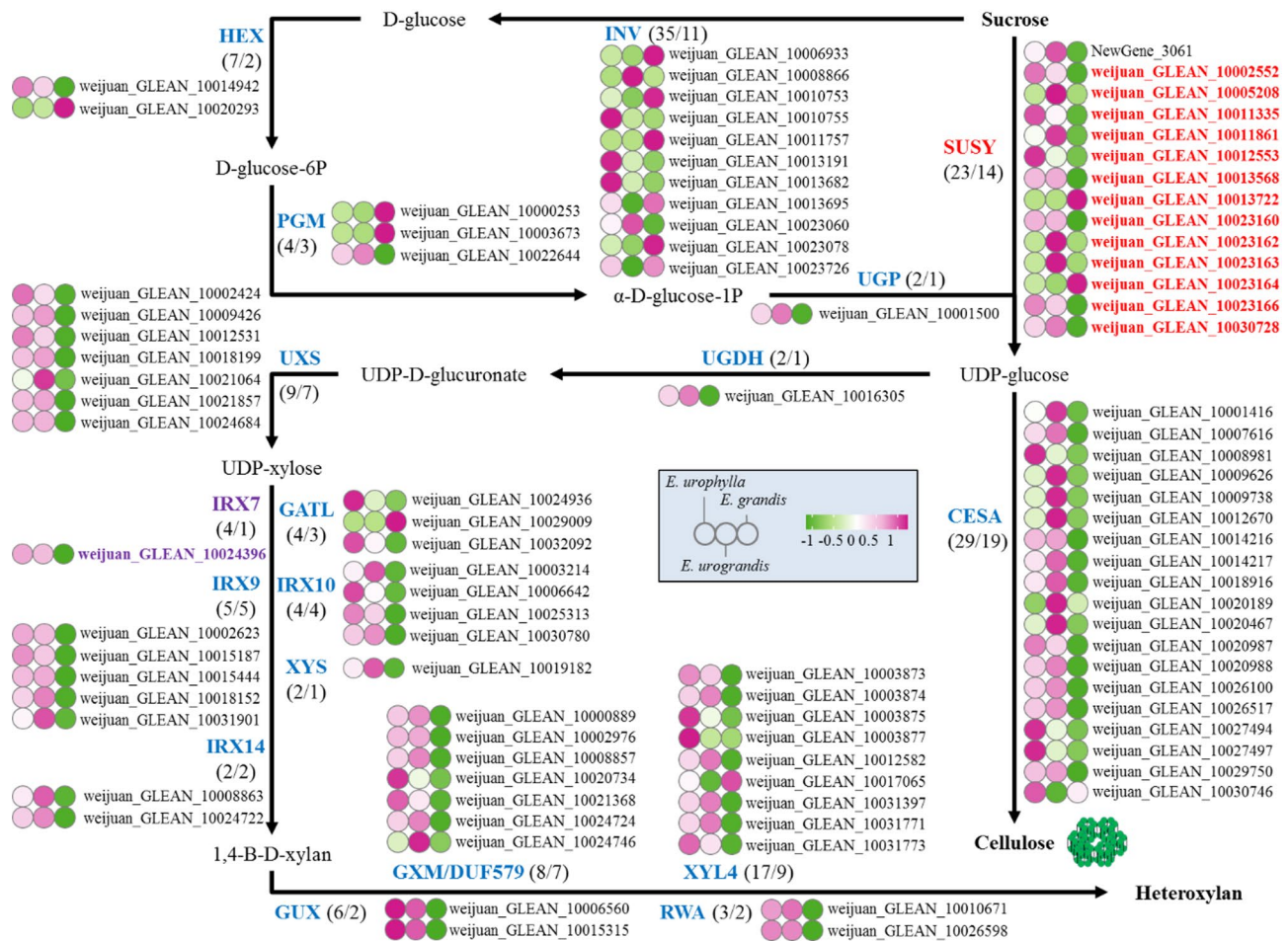


Fig. 4 Genes involved in the cellulose and xylan biosynthesis in the immature xylem tissue of the *E. urograndis* genome. Relative expression profile (green-deeppink scale) of secondary cell-wall-related genes implicated in cellulose and xylan biosynthesis. The relative expression levels for all genes represent the average values derived from three biological replicates. The numbers in parentheses represent the number of homologous genes in the gene family/the number of DEGs. The bold red or purple fonts symbolizes the DEGs of significantly expanded genes or PSGs. Detailed protein names, annotation and RNA-seq expression data are provided in Supplementary Table S12. Notes: CESA: cellulose synthase A; DUF: domain of unknown function; GATL: galacturonosyltransferase-like; GUX: xylan alpha-glucuronosyltransferase; GXM: glucuronoxylan 4-O-methyltransferase; HEX: hexokinase; INV: beta-fructofuranosidase; IRX: irregular xylem; PGM: phosphoglucomutase; RWA: reduced wall acetylation; SUSY: sucrose synthase; UGDH: UDPglucose 6-dehydrogenase; UGP: UTP-glucose-1-phosphate uridylyltransferase; UXS: UDP-glucuronate decarboxylase; XYL: xylan 1,4-beta-xylosidase; XYS: 1,4-beta-D-xylan synthase

Supplementary Table S12). The *CesA* gene family comprises 29 homologous genes, including 9 *CesA* genes (*CesA1*, *CesA3* to 10). Among them, 10 and 4 genes homologous to *CesA1* and *CesA5*, respectively, were present. The *CesA1* and *CesA5* genes are associated with the formation of primary walls in *Arabidopsis* [52]. Notably, the first key enzyme in the direct production pathway of UDP-glucose was *Susy*, and 23 homologous genes were identified in the *Susy* gene family in this study, of which 18 genes were significantly expanded compared with the genes in the other 12 typical plant species. In this study, we analyzed DEGs involved in the cellulose biosynthesis pathway of *E. urograndis* and its parental species. The results revealed that among the 14 DEGs of the *SUSY* gene family, seven DEGs exhibited overdominance,

whereas four DEGs demonstrated high-parent dominance. The differential expression patterns of these genes in *E. urograndis* and its parents may be associated with the increased cellulose content observed in the three tree species. The indirect pathway for the production of UDP-glucose comprises four key enzymes—INV (beta-fructofuranosidase; EC:3.2.1.26), HEX (hexokinase; EC:2.7.1.1), PGM (phosphoglucomutase; EC:5.4.2.2), and UGP (UTP-glucose-1-phosphate uridylyltransferase; EC:2.7.7.9), for which 35, seven, four, and five homologous genes, respectively, are found in the *E. urograndis* genome. On the basis of the transcriptomic analysis of immature xylem in *E. urograndis* and its parental species, a total of 50 DEGs in the cellulose biosynthesis pathway were identified. Among these, 25 DEGs exhibited overdominance,

whereas six and eight DEGs displayed high-parent dominance and low-parent dominance, respectively. Additionally, 11 DEGs were underdominant. Notably, among the 19 DEGs of the *CesA* gene family, 14 DEGs were overdominance, and seven DEGs were silenced in *E. grandis*, suggesting that these seven DEGs conform to the SPE model. Moreover, seven *SUS* genes, two *INV* genes, one *PGM* gene, and one *UGP* gene displayed overdominance. Specifically, the expression levels of three *SUS* genes (weijuan_GLEAN_10005208, weijuan_GLEAN_10023162, and weijuan_GLEAN_10023163) and one *INV* gene (weijuan_GLEAN_10008866) in *E. urograndis* were more than 10-fold higher compared to those in the parental species. These findings may be associated with the higher cellulose content observed in hybrid *E. urograndis* than in its parental species.

Hemicellulose xylan is a plant cell wall polysaccharide that is widely considered the second most abundant plant biopolymer on earth after cellulose [53]. Heteroxylan synthesis occurs in the Golgi apparatus through the coordinated action of several enzyme classes (as shown in Fig. 4). In this study, one gene of *IRX7* (irregular xylem 7, weijuan_GLEAN_10024396) was identified as a PSG. Furthermore, analogous to the cellulose biosynthesis pathway, 44 DEGs involved in xylan biosynthesis exhibited significant differential expression between *E. urograndis* and its parents. Among these DEGs, 24 demonstrated overdominance, 14 showed high-parent dominance, and five and one exhibited low-parent dominance and underdominance, respectively. Notably, only one DEG from the *GATL* (galacturonosyl transferase-like) gene family (weijuan_GLEAN_10029009) and only one DEG (weijuan_GLEAN_10017065) from the *XYL4* (xylan 1,4-beta-xylosidase 4) gene family (weijuan_GLEAN_10017065) presented relatively high expression levels in *E. grandis*. Furthermore, three DEGs were silenced in *E. grandis*, suggesting that these three DEGs conform to the SPE model. This finding aligns with the observed trend in hemicellulose content changes among *E. urograndis* and its parents.

Lignin is synthesized from phenylalanine, which is a complex process involving 12 enzymes and a series of steps—deamination, hydroxylation, methylation, and reduction (Fig. 5 and Supplementary Table S12). Lignin monomers (H lignin, G lignin, and S lignin) are produced in the cytoplasm and transported across the cell membrane to be polymerized in the cell wall [54]. CCR and CAD are the last two enzymes in the monolignol synthesis pathway; one gene of the *CCR* gene family (30 homologous genes) was identified as a PSG, and 14 genes of the *CAD* gene family (36 homologous genes) and four homologous genes in the *C3H* gene family were identified as significantly expanded genes. In addition, two genes belonging to the *4CL* (20 homologous genes,

weijuan_GLEAN_10020758) and *COMT* (28 homologous genes, weijuan_GLEAN_10022151) gene families were identified as PSGs. From the analysis of DEGs in the lignin biosynthesis pathway of immature xylem in *E. urograndis* and its parental species, 17 DEGs exhibited underdominance, whereas 16 DEGs showed low-parent dominance. Additionally, nine DEGs demonstrated overdominance, and seven DEGs exhibited high-parent dominance. Notably, the number of upregulated genes in *E. grandis* was the greatest, potentially contributing to its significantly higher lignin content than those of *E. urophylla* and *E. urograndis*. Specifically, one DEG (NewGene_9791) of *CAD* and one DEG of *CCR* (weijuan_GLEAN_10023993) were silenced in *E. urograndis*, suggesting that these two DEGs conform to the parental co-silencing model (Fig. 5). These findings may be associated with the lower lignin content in *E. urograndis* and the higher lignin content observed in *E. grandis*.

Discussion

Genomic data help characterize the history of hybridization and the genetic basis of speciation [55]. The importance of eucalypt species and their interspecific hybrids has been demonstrated in forestry programs owing to their wood quality and ability to adapt to diverse environmental conditions [56]. In this study, a chromosomal-scale genome assembly of *E. urograndis* was prepared using the BGI-SEQ500 platform, PacBio-SMRT technology, and Hi-C-assisted assembly technology. A high-quality reference genome for *E. urograndis* was obtained with a scaffold N50 of up to 58.62 Mb, and 99.91% of the sequences were anchored to 11 pseudochromosomes. Genome surveys revealed that the *E. urograndis* genome presented high heterozygosity, implying high genetic variability. According to the results of the genome assembly quality assessment, the *E. urograndis* genome presented better quality, with the mapping rate of all small fragment reads to the genome being 96.36% and the percentage of genome coverage reaching 99.80%. Conserved single-copy orthologous genes were assembled using the BUSCO method. The assembled genome can serve as a high-quality reference genome for *E. urograndis* to support studies of molecular breeding, genetics, and evolution in hybrids of *E. urophylla* and *E. grandis*, as well as the *Eucalyptus* genus.

Traditionally, the genome size of hybrid offspring is determined by the genome size of both the male and female parents [57]. The assembly size of *E. urograndis* was 592.09 Mb in this study, which is close to the mid-parent value (588.03 Mb) of *E. urophylla* (559.53 Mb, provided by the Guangdong Academy of Forestry) and *E. grandis* (616.53 Mb, published on April 2, 2021). This confirmed that *E. grandis* and *E. urophylla* had good hybridization affinity during the formation of offspring,

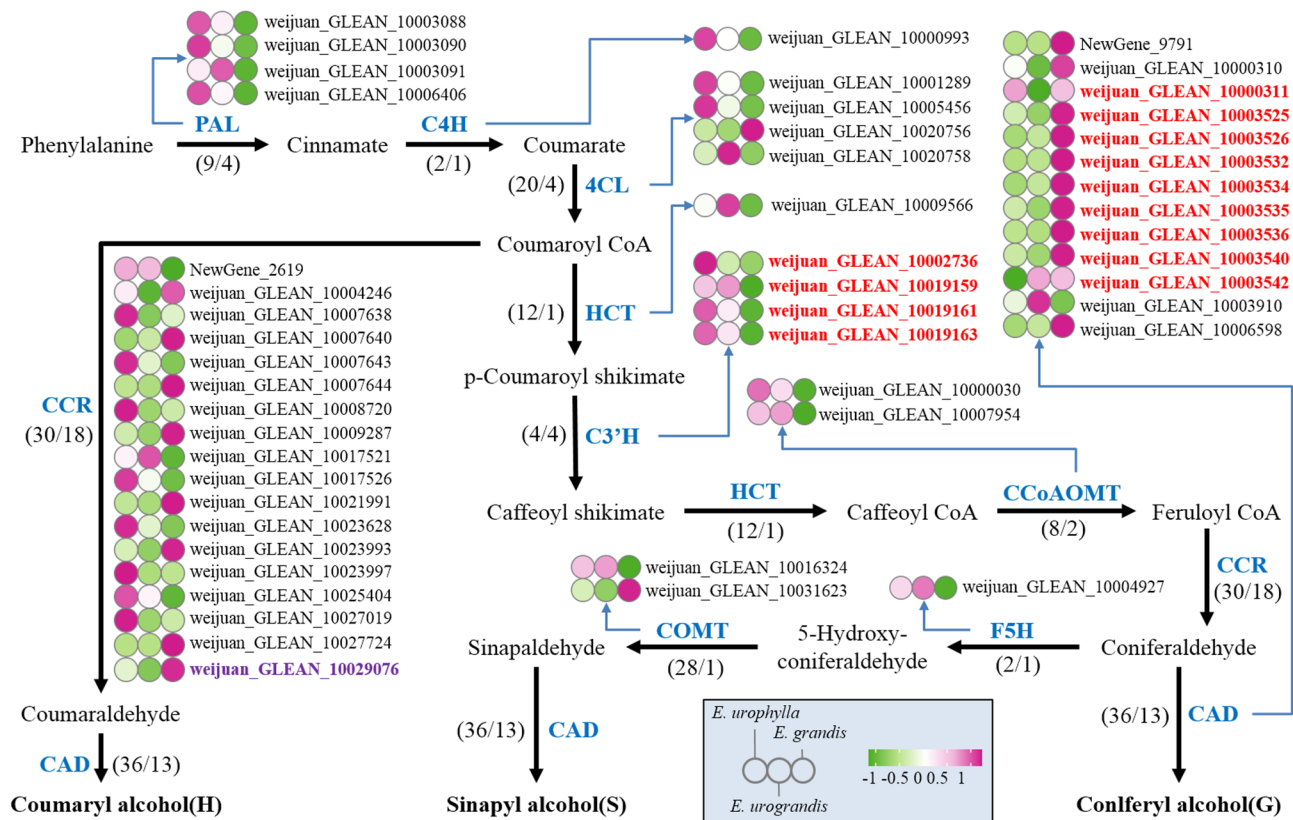


Fig. 5 Genes involved in the lignin biosynthesis in the immature xylem tissue of the *E. urograndis* genome. Relative expression profile (green-deeppink scale) of secondary cell-wall-related genes implicated in cellulose and xylan biosynthesis. The relative expression levels for all genes represent the average values derived from three biological replicates. The numbers in parentheses represent the number of homologous genes in the gene family/the number of DEGs. The bold red or purple fonts symbolizes the DEGs of significantly expanded genes or PSGs. Detailed protein names, annotation and RNA-seq expression data are provided in Supplementary Table S12. Notes: 4CL: 4-coumarate-CoA ligase; C3'H: 5-O-(4-coumaroyl)-D-quinic acid 3'-monooxygenase; C4H: trans-cinnamate 4-monooxygenase; CAD: cinnamyl-alcohol dehydrogenase; CCoAOMT: caffeoyl-CoA O-methyltransferase; CCR: cinnamoyl-CoA reductase; COMT: caffeic acid 3-O-methyltransferase; F5H: ferulate-5-hydroxylase; HCT: shikimate O-hydroxycinnamoyltransferase; PAL: phenylalanine ammonia-lyase

which is primarily governed by stable chromosomal inheritance, additive gene expression effects, and a balanced distribution of combinatorial ability. This result aligns with findings in other species, such as *Scirpus maritimus*, where the genome size of hybrids is generally intermediate compared to that of their parental species [58]. Moreover, there is a notable discrepancy in genome size between the two studies (545.93 Mb in [23] versus 592.09 Mb in this study). K-mer-based genome size estimations from Lötter et al. [59] reported a haplotype assembly size of 443.20 Mb for *E. urophylla*, 482.30 Mb for *E. grandis*, and 477.80 Mb for the F1 hybrid. According to these predictions, the genome size of the F1 hybrid exceeds the average of its parental species (462.75 Mb), a result consistent with the findings of this study, which also observed that the genome size of *E. urograndis* (592.09 Mb) is larger than the average of *E. grandis* and *E. urophylla* (588.03 Mb). Compared with the draft assembly of *E. grandis*, the genome assembly of *E. urograndis* presented better continuity and greater coverage. The

E. urograndis genome presented a contig N50 size of 3.73 Mb and a scaffold N50 size of 58.62 Mb, which was greater than that of *E. grandis* (contig N50 size, 0.61 Mb; scaffold N50 size, 58.49 Mb). The percentage of embryophyte BUSCO genes in *E. urograndis* was 94.90%, which was slightly greater than that in *E. grandis* (92.30%). The number of protein-coding genes in *E. urograndis* was 32,151, which is lower than that in *E. grandis* v2.0 (36,349). This might be due to the smaller genome size of *E. urograndis* than that of *E. grandis*.

Transposable elements, essential elements in plant genomes, can move around the genome by either “cut-paste” (DNA transposons) or “copy-paste” mechanisms (RNA transposons) [60]. The percentage of transposable elements in the *E. urograndis* genome reached 45.01%, which was slightly greater than that in *E. grandis* (44.50%). A substantial portion of eukaryotic genomes consists of highly repeated non-coding DNA sequences [61]. The repetition rate in the *E. urograndis* genome was 47.48%, which is close to that reported for the genomes

of flowering cherry (*Prunus yedoensis*, 47.20%) [62], *E. melliodora* (47.97%), and *E. pauciflora* (46.66%). This may be attributed to the analogous regulatory roles of repetitive sequences in the stress responses of these species.

Interspecific hybridization is one of the main mechanisms of plant speciation. The merging of two genomes from different subspecies, species, or even genera is frequently accompanied by WGD [13]. In this study, we analyzed collinearity and WGD events between *E. urograndis* and *E. grandis*. The results of collinearity analysis revealed a high degree of collinearity between the *E. urograndis* and *E. grandis* genomes, suggesting that the *E. urograndis* genome inherited many genes from its male parent (*E. grandis*). In this study, *E. urograndis* was generated via artificial crossing. The estimation of the divergence time of *E. urograndis* can, to some extent, provide insights into the divergence between its parental species, namely *E. urophylla* and *E. grandis*. The species differentiation of *E. urophylla* and *E. grandis* was estimated to have occurred at the end of the Pliocene epoch (~2.90 Mya). During the Pliocene period (5.30–2.59 Mya), the biological world was similar to the modern world; the plant kingdom comprised the same species as those found in the modern world, and the first human-like animals appeared at the end of this period. Additionally, the number of synonymous substitutions per site per year was estimated to be 4.75×10^{-9} for *Eucalyptus*, which is slightly greater than that for Laurales (4.21×10^{-9}) [43] and less than that for Ranunculales (6.98×10^{-9}) [63]. The ancient γ WGD event might have been related mainly to environmental adaptation in the Cretaceous [42], and the recent small-scale duplication event would have been related mainly to artificial breeding and environmental changes.

E. urograndis is widely used in the forestry industry because of its great potential for use as timber, short rotation, high basic density, high Lignin and cellulose contents, and high mechanical strength. In general, the wood of angiosperm trees contains 42–55% cellulose and 20–25% lignin [64]; however, *E. urograndis* wood contains 68.41% and 29.94% cellulose and lignin contents, respectively [11]. Therefore, identification of the cellulose and lignin biosynthetic pathways in *E. urograndis* from the perspective of genome evolution is necessary. In this study, the expansion and contraction of genes and PSGs in the *E. urograndis* genome, as well as the cellulose and Lignin biosynthetic pathways, were identified. Six enzymatic steps involving 126 genes involved in cellulose biosynthesis from sucrose were identified. Genomic analysis revealed that *E. urograndis* has 29 *CesA* genes; however, the *E. grandis* genome has 16 *CesA* genes [25]. Additionally, the 29 *CesA* genes in the *E. urograndis* genome comprised 9 variants (*CesA1*, *CesA3*–*CesA10*), which were comparable to the 10 *CesA* genes in the *A. thaliana*

genome [32], but *CesA2* was absent. Studies have shown that *CesA2*, 6, 5, and 9 are functionally redundant and involved in the synthesis of primary wall cellulose. Burn et al. [65] reported that the function of *CesA2* was less obvious in *Arabidopsis*, which might be related to the lack of *CesA2* in the *E. urograndis* genome. Lignin biosynthesis from phenylalanine involves 12 enzymes and 317 genes in *E. urograndis*. Among them, four gene families were contracted, and two gene families were under positive selection. The evolutionary relationships of these candidate genes might be related to the high cellulose and lignin contents of *E. urograndis*.

The higher the cellulose content is, the higher the biomass energy yield [66]. However, lignin is an important factor affecting pulp yield and quality when wood is used in the papermaking industry [67]. The modification of lignin-related genes and/or lignin composition in wood will help improve our current knowledge of lignification and facilitate chemical pulping and bleaching processes, thereby lowering their energy demand and environmental impact. Analysis of the lignin biosynthetic pathways in this study revealed that one DEG, NewGene_9791, associated with *CAD*, and one DEG, weijuan_GLEAN_10023993, associated with *CCR*, exhibited silenced expression in *E. urograndis*. These findings provide valuable genetic information for the future directed improvement of pulp raw material forests. Furthermore, this research will provide a theoretical basis for changing the chemical composition and structure of *E. urograndis* wood through genetic engineering in the future to produce new varieties that are more conducive to energy conservation and would help reduce pollution.

Compared with other eucalyptus hybrids, *E. urograndis* has greater adaptability to the environment. This might be due to the evolution of intricate mechanisms to recognize and defend itself against potential pathogens. According to the evolutionary analysis of the plant-pathogen interaction pathway in *E. urograndis*, 6 out of 43 enzymes were identified as significantly expanded genes or PSGs; these six enzymes belong to PTI, which is the primary response against pathogen invasion in plants. Rojas et al. [68] revealed that primary metabolism is involved in regulating plant defense against pathogens. In this study, the analysis of the expanded gene families and PSGs revealed that carbohydrate, amino acid, and lipid metabolism pathways were more enriched in *E. urograndis* than in the other 13 plants. These results may help elucidate the reasons for the greater adaptability of *E. urograndis* and provide a useful direction for improving genetics and breeding using molecular biological techniques.

As a key raw pulp material species globally, eucalyptus is characterized by low production costs, high pulp-yielding yields, and excellent papermaking properties. The

primary chemical components of plant fiber raw materials include cellulose, hemicellulose, and lignin. In this study, gene families related to the biosynthesis of cellulose, xylan, and lignin were systematically analyzed through an integration of gene annotation analysis and mRNA-seq expression profiling. Cellulose constitutes the primary component of plant raw materials and represents the most valuable fraction in pulping and papermaking processes [51]. The cellulose content in fibrous raw materials serves as the key indicator for evaluating the suitability of materials for pulping and papermaking. Lignin, a major constituent of the plant cell wall, functions to reinforce the cell wall and bind fibers together. However, when used as a raw material for pulping, a lower lignin content in wood generally correlates with superior pulping and papermaking performance. Hemicellulose refers to the collective term for non-cellulosic high glycans present in plant cell walls. In general, retaining hemicellulose during pulping is desirable to increase the pulp yield. The results of this study demonstrated that the hybrid (*E. urograndis*) presented a higher cellulose content, higher hemicellulose content and lower lignin content than its parental species. Additionally, the analysis of gene expression profiles revealed that the number of DEGs associated with the cellulose and hemicellulose biosynthesis pathways predominantly displayed patterns of overdominance and high-parent dominance. In contrast, within the lignin biosynthesis pathway, the majority of DEGs were characterized by underdominance and low-parent dominance. These findings are largely consistent with the hypothesis of “direction-shifting ASEs” (shift expression direction across tissues/conditions) proposed by Shao et al. [50], which suggest that under specific spatiotemporal conditions, hybrids can selectively express advantageous alleles at particular genetic loci, thereby conferring heterosis. Similarly, ASE also shows a potential association with fruit-quality-related heterosis in two pear hybrid varieties [69]. Moreover, Zhan et al. [70] found that 758 ASE genes (ASEGs) were significantly associated with heterosis quantitative trait loci (QTLs). High-frequency variations in the promoter regions of heterosis-related ASEGs were also detected between the parental lines of two maize hybrids. These results demonstrate the role of ASE in heterosis and provide a robust pangenome reference for high-resolution allele discovery and association mapping.

Conclusions

In this study, we constructed a high-quality chromosomal-level genome assembly of *E. urograndis* (DH32-29) using the BGI-SEQ500 platform, PacBio-SMRT technology, and Hi-C-assisted assembly technology. The resulting draft genome assembly spans 592.09 Mb, with 99.91% of the data anchored onto 11 pseudochromosomes. This

assembly achieved a contig N50 of 3.73 Mb and a scaffold N50 of 58.62 Mb. Gene annotation and evaluation indicated that the *E. urograndis* genome contains 32,151 genes, of which 93.5% were fully annotated by BUSCOs. Additionally, 47.48% of the genome consists of repeat sequences, and the functions of 96.48% of the genes could be predicted. Based on evolutionary analysis, *E. grandis* and *E. urograndis* are estimated to have diverged approximately 2.90 Mya. Furthermore, 131 gene families were significantly expanded, and 475 PSGs were identified in the *E. urograndis* genome. Among these, 48 significantly expanded genes and 15 PSGs potentially associated with the fast growth and high disease resistance of *E. urograndis* were screened. Moreover, comparative genomic analysis revealed highly conserved synteny between the genomes of *E. urograndis* and *E. grandis*, while the *E. urograndis* genome exhibited chromosomal fusion, inversion, and other rearrangement events. Additionally, RNA-seq technology was employed to analyze allele-specific expression patterns of key enzymes involved in cellulose, xylan, and lignin biosynthesis. Several genes displaying allele-specific expression were identified, which may contribute to heterosis in *E. urograndis*. Collectively, this study provides a foundational resource and offers significant insights into the genetic and molecular mechanisms underlying *Eucalyptus* heterosis, thereby accelerating selective breeding in *Eucalyptus* species.

Materials and methods

Plant material

The improved varieties of *E. urograndis* (NO. Gui S-SC-EUG-001–2009), specifically the well-known dominant *Eucalyptus* hybrid clonal line DH32-29, provided by the Guangxi State-owned Dongmen Forest Farm, were utilized for whole-genome sequencing in this study. Young leaves were collected from about 3-year-old eucalyptus plants in autumn for genome sequencing. Young plantlets were used for Hi-C library construction and sequencing. Three fresh tissue samples, namely, the leaves, xylem, and cambium, from Gui S-SC-EUG-001-2009 were collected for RNA-seq, which could help with genome annotation. In addition, for the RNA-seq analysis, the immature xylem at the diameter at breast height (DBH) level of each tree was collected from *E. urograndis* and its identified paternal (G46) and maternal (U16) plants. Three biological replicates were obtained for each sample.

Sequencing, assembly, and assessment

Genomic DNA was isolated from fresh leaves using a QIAamp DNA purification kit (Qiagen, Germany) according to the manufacturer's instructions. The integrity and quality of the extracted DNA were evaluated using 1% gel electrophoresis. The DNA concentration was assessed using a Pultton DNA/Protein

Analyzer (Plextech, USA). DNA samples with a total amount ≥ 20.0 μg , $1.80 < \text{OD}_{260/280} < 2.00$, and concentration > 12.50 $\text{ng}/\mu\text{L}$ were used to construct the sequencing libraries.

For long-read sequencing, we constructed a SMRTbell Library with a fragment size of 20 kb using the SMRTbell template preparation kit 1.0 (PacBio, USA). The PacBio Sequel II-continuous long reads (CLR) sequencing library was sequenced using a PacBio Sequel system (Pacific Biosciences, Menlo Park, CA, USA) with version 3.0 chemistry, and data from one SMRT cell were generated.

To obtain a chromosome-scale genome assembly, we constructed a Hi-C Library for sequencing. The genomic DNA from the leaf samples was fixed with 1% formaldehyde, and the fixation was terminated with 0.2 M glycine. A Hi-C library was prepared following the Hi-C library protocol [71], followed by sequencing using a BGISEQ-500 sequencing platform (BGI, Shenzhen, China).

Tissue collection and RNA sequencing

To facilitate the prediction of protein-coding genes, total RNA was extracted from three different tissues of *E. urograndis*, namely, leaves, xylem, and cambium, using the TRIzol reagent (Invitrogen, CA, USA). The same clonal line (DH32-29) was utilized for both genome and transcriptome sequencing. RNA integrity and quantity were evaluated using an Agilent 2100 Bioanalyzer (Agilent, USA). RNA samples with a total amount ≥ 10.00 μg , $\text{OD}_{260/230} > 2.00$, and concentration > 20.00 $\text{ng}/\mu\text{L}$ were used to construct the sequencing libraries. The three RNA-Seq libraries, which were prepared using the NEB-Next Ultra RNA Library Prep Kit (Illumina, USA) following the manufacturer's protocol, were sequenced on a BGISEQ-500 sequencing platform (BGI, Shenzhen, China), which produced 97.65 Gb of raw data. The quality control (QC) of the raw reads was performed using a QC pipeline for RNA-Seq data—RNA-QC-Chain—and 93.82 Gb of clean RNA-Seq data were obtained for further analyses.

Genome assembly and quality assessment

On the basis of the quality-filtered reads, the genome size, heterozygosity, and repeat sequence information of the *E. urograndis* genome were estimated using k-mer analysis. The k-mer count frequencies were computed using Jellyfish (v2.2.10) [72], with $k = 17$ and a maximum k-mer count of 10,000. The k-mer distribution was measured and plotted using GenomeScope [73]. The genome size was calculated using the formula $G = N17\text{-mer}/D17\text{-mer}$, where N17-mer is the total number of 17-mers and D17-mer denotes the peak frequency of 17-mers. The short reads were assembled *de novo* into contigs and scaffolds using SOAPdenovo2 [74]. Gaps in the initial

assembly were filled using TGS-Gapcloser (v1.12) [75] with the parameters “avg_ins=364, max_ins=500, and min_ins=260”. The draft assembly was then anchored and oriented into a chromosomal-scale assembly using the Hi-C scaffolding approach. First, the raw Hi-C reads were filtered using HiC-Pro (v2.8.0) [76]. Then, 3D-DNA (v170123) [77] with the parameters “-m haploid -s 0 -c 24” was used to anchor the primary contigs and scaffolds into chromosomes. The inter/intrachromosomal contact maps were built and visualized using Juicebox [78]. To further improve the integrity and accuracy of the genome assembly, we employed TGS-GapCloser, which uses low-depth ($\geq 10\times$) single-molecule sequencing long reads without any error correction to close the gaps in the draft assembly. The long sequences were split into three groups, namely, total reads (with options -min_idy 0.2, -min_match 200 -r_round 1), reads with a length ≥ 20 kb (with options -min_idy 0 -min_match 0 -r_round 3), and reads with a length of 2–20 kb (with options -min_idy 0 -min_match 0 -r_round 3), and each group was used to fill the corresponding aligned gaps.

The completeness of the genome assembly was assessed using Benchmarking sets of Universal Single-Copy Orthologs (BUSCO v5.4.3) and GC content analyses. The single-copy orthologs of Embryophyta_odb9 (BUSCO, v2.0) [79] were searched against the assembled genome using the BUSCO tool. The GC content and average sequencing depth across the genome were also measured with 10 kb non-overlapping sliding windows, and windows harboring more than 50% N were filtered. No external contamination was detected in the genome.

Repetitive sequence annotation

The repetitive sequences in the *E. urograndis* genome were annotated using both homology searches in known repeat databases and *de novo* predictions. Known repeats were identified using RepeatMasker (v3.3.0) [80] and the RepBase TE library (v14.06) [81]. The RepeatProteinMask (v3.2.2) implemented in RepeatMasker was used to detect the TE-relevant proteins. Novel repeats were predicted using RepeatModeler (<http://www.repeatmasker.org>) on the basis of the *de novo* repeat library constructed with LTR_Finder [82] and RepeatScout (1.0.6) [83]. In addition, the tandem repeat finder (TRF, v4.09) [84] was used to identify the tandem repeats in the genome with the parameters “Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPeriod=2000”.

Gene prediction and annotation

Based on the repeat-masked genome, we employed *de novo*, homology-based, and transcriptome-assisted predictions to detect protein-coding genes. *De novo* gene prediction was performed using Augustus (v2.7) [85] and Genscan [86] with the default settings. For

homology-based prediction, protein sequences of *E. grandis*, *Citrus sinensis*, *Populus trichocarpa*, *Prunus persica*, *Punica granatum*, *Solanum lycopersicum*, *Theobroma cacao*, and *Hevea brasiliensis* were downloaded from the NCBI database and aligned to the *E. urograndis* genome using tBLASTn (E value $\leq 1e-5$). The homologous genome sequences were then aligned against the matching proteins using GeneWise (v2.4.0) [87] for accurate spliced alignments. Transcriptomic data were generated from three RNA-Seq Libraries constructed from three different tissue samples, namely, leaves, xylem, and cambium. A total of 36.45 Gb of clean data were aligned to the assembled genome sequences using HISAT2-StringTie (v2.0.10) [88], and the putative transcript structures were detected using StringTie (v2.1.1) [89]. The candidate protein-coding regions within the transcript sequences were then predicted using TransDecoder (v5.5.0) (<https://github.com/TransDecoder/TransDecoder/>). Finally, genes predicted using the above methods were merged into a consensus gene set using Glean [90]. The completeness of the gene set was evaluated using BUSCO v5.4.3 software, and the gene set of embryogenic plants was selected as the reference. For gene function annotation, we aligned all the genes of *E. urograndis* to the TrEMBL, SwissProt, Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and InterProScan databases. Additionally, for non-coding RNA prediction, 500 tRNAs were identified using tRNAscan-SE 2.0 [91]. Moreover, 438 snRNAs and 994 miRNAs were annotated by the infernal tool [92] in the Rfam database [93]. Additionally, 747 rRNAs were identified via homology searches against closely related species using the BLASTN tool.

Comparative and evolutionary genomic analysis

To construct the phylogenetic tree, we identified orthologous gene families by comparing the protein and cDNA sequences of *E. urograndis* and 12 sequenced plant genomes—*E. grandis*, *S. tuberosum*, *Z. mays*, *O. sativa*, *P. dactylifera*, *Populus trichocarpa*, *P. patens*, *A. thaliana*, *E. melliodora*, *E. pauciflora*, *V. vinifera*, and *Theobroma cacao*. First, BLASTP software was employed to compare the protein sequences of the 13 species, with the E -value threshold set as $1e-5$ and the inflation parameter (I) set at 2.0. OrthoMCL software [94] was used to cluster the gene families of all the species. The gene set data were downloaded from Ensemble or NCBI. The genes obtained from databases with frameshifts or fewer than 50 amino acids were removed, and for protein-coding genes with alternative splicing isoforms, only the longest protein sequence prediction was selected for their representatives. Additionally, the genome assembly of *E. urograndis* was compared with the published genomes of other species, including *E. grandis*, *E. melliodora*, and

E. pauciflora. Second, the protein sequences of single-copy gene families were compared using MUSCLE [95], and the protein sequences were transcribed into coding sequences (CDSs) on the basis of the alignment sequence; the phase 1 locus was extracted and concatenated into a supergene. MrBayes 3.2.4 (<https://sourceforge.net/projects/mrbayes/files/mrbayes/3.2.4/>) was used to define the orthologous and paralogous relationships among all the organisms. Using the single-copy orthologous genes, a phylogenetic tree was generated on the basis of the Bayes model using PhyML (v3.0) [96] with 500 bootstrap replications. The MCMCTree program implemented in the PAML package [97] was used to predict the divergence times. The divergence times were estimated using the approximate method with fossil calibrations from the TimeTree database (<http://www.timetree.org/>). Third, we identified the orthologous groups among these 13 species using all-to-all BLAST (E value $\leq 1e-5$, identity $\geq 80\%$) and identified the expanded and contracted gene families using CAFÉ 5 [98]. GO and KEGG enrichment analyses were performed to identify the functional implications of the expanded and contracted genes (Fisher's exact test, adjusted p value < 0.05). CodeML in PAML was used for positive selection analysis, and the "branch-site" model was selected, with the foreground branches being *E. urograndis* and *E. grandis* and the background branches being *E. melliodora*, *E. pauciflora*, and *P. trichocarpa*. Through the likelihood ratio test, a p -value less than 0.05 indicates that the gene has undergone positive selection.

Collinearity and Whole-Genome duplication analysis

To reveal the collinearity relationships between *E. urograndis* and *E. grandis*, we aligned the chromosomes of *E. urograndis* with those of *E. grandis* using the LASTZ tool (v1.04.22) [99] with default options. Chromosomal collinearity was constructed using mapped regions with lengths > 2 kb and visualized using Circos (v0.69) [100]. Additionally, syntenic blocks within the *E. urograndis* and *E. grandis* genomes were identified using MCScanX [101] with default parameters based on Diamond v0.9.29.130 software [102], which was used to compare the gene sequences of the two species to determine similar gene pairs with an E value cutoff of $1e-5$ and a C score ≥ 0.5 , where the value of the C score was filtered using JCVI v0.9.13 software [103]. The K_s (synonymous substitutions per site) values between collinear genes were estimated using the CodeML approach as implemented in the PAML package. The K_s and 4DTv (fourfold synonymous third-codon transversion rate) methods were used to identify WGD events. WGD v1.1.1 software [104] and a custom script (<https://github.com/JinfengChen/Scripts>) were used to identify WGD events in *E. urograndis*, *E. grandis*, *E. pauciflora*, *C. citriodora*, and *V. vinifera*.

RNA sequencing and differentially expressed gene analysis

Total RNA was isolated from the nine immature xylem samples, including three biological replicates from each from DH32-29, *E. urophylla* and *E. grandis*, using an RNAPrep Pure Plant Kit (TIANGEN Biotech, Beijing, China) with DNase I treatment to remove genomic DNA. Then, the qualified RNA was processed for Library construction. To ensure the quality of the Library, a Qubit 2.0 and Agilent 2100 instruments were used to examine the concentration of cDNA and the insert size. The qualified library was sequenced by the high-throughput sequencing platform in PE150 mode.

FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) was applied to measure the expression level of a gene by StringTie using maximum flow algorithm [105]. Differential expression analysis was performed using DESeq2 (v1.31.16) [106], and DEGs were identified on the basis of the criteria of a $|\log_2(\text{fold change, FC})| \geq 1$ and a false discovery rate (FDR) < 0.01 . Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were subsequently conducted for the identified DEGs.

Abbreviations

4CL	4-coumarate-CoA ligase
ASE	Allele-specific expression
ASEGs	Genes showing allele-specific expression
BP	Biological process
BUSCOs	Benchmarking universal single-copy orthologs
C3'H	5-O-(4-coumaroyl)-D-quinic acid 3'-monooxygenase
C4H	Trans-cinnamate 4-monooxygenase
CAD	Cinnamyl-alcohol dehydrogenase
CALM	Calmodulin
CC	Cellular component
CCoAOMT	Caffeoyl-CoA O-methyltransferase
CCR	Cinnamoyl-CoA reductase
CesA	Cellulose synthase A
COMT	Caffeic acid 3-O-methyltransferase
DEGs	Differentially expressed genes
DUF	Domain of unknown function
<i>E. urograndis</i>	<i>Eucalyptus urophylla</i> × <i>Eucalyptus grandis</i>
F5H	Ferulate-5-hydroxylase
GATL	Galacturonosyltransferase-like
GUX	Xylan alpha-glucuronosyltransferase
GXM	Glucuronoxylan 4-O-methyltransferase
HCT	Shikimate O-hydroxycinnamoyltransferase
HEX	Hexokinase
Hi-C	High-throughput chromosome conformation capture
INV	Beta-fructofuranosidase
IRX	Irregular xylem
LINES	Long interspersed nuclear elements
LTRs	Long terminal repeats
MF	Molecular function
Mya	Million years ago
PAL	Phenylalanine ammonia-lyase
PAV	Presence/absence variation
PGM	Phosphoglucomutase
PSGs	Positive selection genes
PTK3	PTI1-like tyrosine-protein kinase 3
RWA	Reduced wall acetylation
SINES	Short interspersed nuclear elements
SMRT	Long-read single-molecule real-time
SNPs	Single nucleotide polymorphisms
SPE	Single-parent expression

SUSY	Sucrose synthase
UGDH	UDPglucose 6-dehydrogenase
UGP	UTP-glucose-1-phosphate uridylyltransferase
UXS	UDP-glucuronate decarboxylase
WGD	Whole-genome duplication
XYL	Xylan 1,4-beta-xylosidase
XYS	1,4-beta-D-xylan synthase

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-025-07371-3>.

Additional file 1. Fig S1-S5. Fig. S1. Assembly results of the *E. urograndis* genome. (A) Genomescope analysis of heterozygosity; (B) Scatter diagram of GC content and sequencing depth. (C) Hi-C interaction heatmap of *E. urograndis*. The darker the red color, the stronger the interaction intensity. The interaction intensity within the same chromosome is more pronounced than between staining, and the chromosome boundary is more obvious, resulting in an ideal mounting effect. Fig. S2. Annotation quality comparison of protein-coding genes. The messenger RNA (mRNA) length, CDS length, exon length, and intron length among 5 species: *E. urograndis*, *E. grandis*, *P. granatum*, *P. trichocarpa*, *H. brasiliensis*. Fig. S3. The KEGG enrichment (A) and GO annotation (B) results of the *E. urograndis* genome. Fig. S4. The 20 significantly enriched pathways with p-values of annotated and enriched genes in the *E. urograndis* genome. Fig. S5. The results of three major main wood chemical compositions components of in the immature xylem of *E. urophylla*, *E. urograndis* and *E. grandis*. The same letters in the figure indicate no significant difference, and the different letters indicate significant differences. The data was subjected to the method of multiple comparison of LSD. Error bars indicate standard error derived from three replicates.

Additional file 2. Tables S1-S12. Table S1. Assembly statistics of *E. urograndis* genome. Table S2. Results of NT comparison. Table S3. Results of BUSCO assessment of genome assembly. Table S4. The lengths of 11 chromosomes in the *E. urograndis* genome. Table S5. Statistics results of repetitive sequences. Table S6. Statistical results of non-coding RNA. Table S7. Statistical results of gene structure prediction. Table S8. GO enrichment analysis of all the *E. urograndis* genes. Table S9. Clustering results of gene families in the *E. urograndis* genome. Table S10. KEGG pathway analysis of the 131 expansion and contraction of gene families. Table S11. Positive selection genes in the *E. urograndis* genome. Table S12. Key enzymes and paralogous genes identified in cellulose, xylan and lignin biosynthesis pathways in the *E. urograndis* genome.

Acknowledgements

We appreciate the comments from Prof. Jihua Ding (Huazhong Agricultural University, College of Horticulture and Forestry Sciences) and the information about the *E. urophylla* genome provided by Prof. Weihua Zhang (Guangdong Academy of Forestry). We gratefully acknowledge the valuable comments and four anonymous reviewers that helped to improve our manuscript.

Authors' contributions

G.L. and J. L. designed and managed the project. W.L., L. Z., Y.L., J.P., J.B., and A.H. participated in material collecting and processing. W.L. and Y.L. performed bioinformatics analyses. G.L. and J.Z. wrote the manuscript. J.P. and A.H. contributed to validation works. J.L., G.L., and W.L. revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was financially supported by the Fundamental Research Funds of CAF (CAFYBB2023MB034) and the National Key R&D Program of China (Grant No. 2022YFD2200203 and 2023YFD2201001).

Data availability

The genomic data and transcriptomic data generated in this study were deposited in the Genome Sequence Archive (GSA) of the China National Center for Bioinformation (CNCB) under accession number of CRA017352 and CRA024878, respectively. All data generated or analyzed during this study are included in this manuscript and supplementary file.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 June 2025 / Accepted: 8 September 2025

Published online: 27 October 2025

References

- Vilasboa J, Da Costa CT, Fett-Neto AG. Rooting of eucalypt cuttings as a problem-solving oriented model in plant biology. *Prog Biophys Mol Biol*. 2019;146:85–97.
- Du K, Xia Y, Zhan D, Xu T, Lu T, Yang J, et al. Genome-wide identification of the *Eucalyptus urophylla* GATA gene family and its diverse roles in chlorophyll biosynthesis. *Int J Mol Sci*. 2022;23(9):5251.
- Paiva JA, Prat E, Vautrin S, Santos MD, San-Clemente H, Brommonschenkel S, et al. Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. *BMC Genomics*. 2011;12:137.
- Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ Jr. High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biol*. 2011;11:65.
- Retief ECL, Stanger TK. Genetic parameters of pure and hybrid populations of *Eucalyptus grandis* and *E. urophylla* and implications for hybrid breeding strategy. *South Forests*. 2009;71(2):133–40.
- Hochholdinger F, Hoecker N. Towards the molecular basis of heterosis. *Trends Plant Sci*. 2007;12(9):427–32.
- Wu X, Liu Y, Zhang Y, Gu R. Advances in research on the mechanism of heterosis in plants. *Front Plant Sci*. 2021;12:745726.
- Shinya T, Iwata E, Nakahama K, Fukuda Y, Hayashi K, Nanto K, et al. Transcriptional profiles of hybrid *Eucalyptus* genotypes with contrasting lignin content reveal that monolignol biosynthesis-related Gnes regulate wood composition. *Front Plant Sci*. 2016;7:443.
- Shi D. Analysis of characteristics of several *Eucalyptus Urophlla* × *E. grandis* species and their direction of production. *Forestry Sci Technol*. 2019;44(6):41–3.
- Gonçalves FG, Oliveira JTS, Lucia RMD, Sartório RC. Estudo de Algumas propriedades mecânicas Da Madeira de Um Híbrido clonal de *Eucalyptus urophylla* × *Eucalyptus grandis*. *Rev Árvore*. 2009;33(3):501–9.
- Rocha MEL, Ristau ACP, Cruz MSFV, Oliveira Neto CFd M, Malavasi MM. Growth dynamics of container seedlings of *Eucalyptus grandis* × *Eucalyptus urophylla* and *Hymenaea courbaril* L. *Rev Ceres*. 2022;69(4):425–35.
- Wang Y, Chen F, Ma Y, Zhang T, Sun P, Lan M, et al. An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (*Camellia sinensis*). *Hortic Res*. 2021;8(1):176.
- Glombik M, Bačovský V, Hobza R, Kopecký D. Competition of parental genomes in plant hybrids. *Front Plant Sci*. 2020;11:200.
- Taylor SA, Larson EL. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat Ecol Evol*. 2019;3(2):170–7.
- Zhang J, Zhang W, Ding C, Zhao J, Su X, Yuan Z, et al. Non-additive gene expression in carbon and nitrogen metabolism drives growth heterosis in *Populus deltoides*. *Plant Cell Environ*. 2025;48(5):3529–43.
- Pitz M, Baldauf JA, Piepho HP, Hochholdinger F. Nonadditive gene expression contributing to heterosis in partially heterozygous maize hybrids is predominantly regulated from heterozygous regions. *New Phytol*. 2025;247(2):669–83.
- Wan J, Wang Q, Zhao J, Zhang X, Guo Z, Hu D, et al. Gene expression variation explains maize seed germination heterosis. *BMC Plant Biol*. 2022;22(1):301.
- Gu Z, Gong J, Zhu Z, Li Z, Feng Q, Wang C, et al. Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. *Nat Genet*. 2023;55(10):1745–56.
- Xu CL, Sun XM, Zhang SG. Mechanism on differential gene expression and heterosis formation. *Hereditas (Beijing)*. 2013;35(6):714–26.
- Hochholdinger F, Yu P. Molecular concepts to explain heterosis in crops. *Trends Plant Sci*. 2025;30(1):95–104.
- Xie J, Wang W, Yang T, Zhang Q, Zhang Z, Zhu X, et al. Large-scale genomic and transcriptomic profiles of rice hybrids reveal a core mechanism underlying heterosis. *Genome Biol*. 2022;23(1):264.
- Botet R, Keurentjes JJB. The role of transcriptional regulation in hybrid vigor. *Front Plant Sci*. 2020;11:410.
- Shen C, Li L, Ouyang L, Su M, Guo K. *E. urophylla* × *E. grandis* high-quality genome and comparative genomics provide insights on evolution and diversification of *Eucalyptus*. *BMC Genomics*. 2023;24(1):223.
- Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, et al. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytol*. 2015;206(4):1283–96.
- Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. *Nature*. 2014;510:356–62.
- Kyriakidou M, Achakkagari SR, Gálvez López JH, Zhu X, Tang CY, Tai HH, et al. Structural genome analysis in cultivated potato taxa. *Theor Appl Genet*. 2020;133(3):951–66.
- Nie S, Wang B, Ding H, Lin H, Zhang L, Li Q, et al. Genome assembly of the Chinese maize elite inbred line RP125 and its EMS mutant collection provide new resources for maize genetics research and crop improvement. *Plant J*. 2021;108(1):40–54.
- Wang L, Zhao L, Zhang X, Zhang Q, Jia Y, Wang G, et al. Large-scale identification and functional analysis of *NLR* genes in blast resistance in the tetep rice genome sequence. *Proc Natl Acad Sci U S A*. 2019;116(37):18479–87.
- Fang Y, Wu H, Zhang T, Yang M, Yin Y, Pan L, et al. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. *PLoS ONE*. 2012;7(5):e37164.
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*. 2006;313(5793):1596–604.
- Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, Haas FB, et al. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J*. 2018;93(3):515–33.
- Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol Plant*. 2022;15(8):1247–50.
- Magris G, Jurman I, Fornasiero A, Paparelli E, Schwöpe R, Marroni F, et al. The genomes of 204 *Vitis vinifera* accessions reveal the origin of European wine grapes. *Nat Commun*. 2021;12(1):7240.
- Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D 3, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*. 2013;14(6):r53.
- Rai MI, Alam M, Lightfoot DA, Gurha P, Afzal AJ. Classification and experimental identification of plant long non-coding RNAs. *Genomics*. 2019;111(5):997–1005.
- Yu Y, Zhang Y, Chen X, Chen Y. Plant noncoding rnas: hidden players in development and stress responses. *Annu Rev Cell Dev Biol*. 2019;35:407–31.
- Nadarajah KK, Abdul Rahman NSN. The role of non-coding RNA in rice immunity. *Agronomy*. 2021;12(1):39.
- D'Ario M, Griffiths-Jones S, Kim M. Small rnas: big impact on plant development. *Trends Plant Sci*. 2017;22(12):1056–68.
- Martinez G, Köhler C. Role of small RNAs in epigenetic reprogramming during plant sexual reproduction. *Curr Opin Plant Biol*. 2017;36:22–8.
- Tang J, Chu C. Micronas in crop improvement: fine-tuners for complex traits. *Nat Plants*. 2017;3:17077.
- Lafontaine DL. Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nat Struct Mol Biol*. 2015;22(1):11–9.
- Zhang L, Wu S, Chang X, Wang X, Zhao Y, Xia Y, et al. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ*. 2020;43(12):2847–56.
- Shang J, Tian J, Cheng H, Yan Q, Li L, Jamal A, et al. The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol*. 2020;21(1):200.
- Zhang L, Li X, Ma B, Gao Q, Du H, Han Y, et al. The Tartary buckwheat genome provides insights into Rutin biosynthesis and abiotic stress tolerance. *Mol Plant*. 2017;10(9):1224–37.
- Barakat A, Bagniewska-Zadworna A, Choi A, Plakkat U, DiLoreto DS, Yellanki P, et al. The cinnamyl alcohol dehydrogenase gene family in *Populus*: phylogeny, organization, and expression. *BMC Plant Biol*. 2009;9:26.
- Gessler A. Sucrose synthase - an enzyme with a central role in the source-sink coordination and carbon flow in trees. *New Phytol*. 2021;229(1):8–10.

47. Dominguez PG, Donev E, Derba-Maceluch M, Bünder A, Hedenström M, Tomášková I, et al. Sucrose synthase determines carbon allocation in developing wood and alters carbon flow at the whole tree level in aspen. *New Phytol.* 2021;229(1):186–98.
48. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. *BMC Biol.* 2017;15(1):98.
49. Wang Z, Li L, Ouyang L. Efficient genetic transformation method for *Eucalyptus* genome editing. *PLoS ONE.* 2021;16(5):e0252011.
50. Shao L, Xing F, Xu C, Zhang Q, Che J, Wang X, et al. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc Natl Acad Sci U S A.* 2019;116(12):5653–8.
51. Liu G, Xie Y, Shang X, Wu Z. Expression patterns and gene analysis of the cellulose synthase gene superfamily in *Eucalyptus grandis*. *Forests.* 2021;12(9):1254.
52. Carroll A, Mansoori N, Li S, Lei L, Vernhettes S, Visser RG, et al. Complexes with mixed primary and secondary cellulose synthases are functional in *Arabidopsis* plants. *Plant Physiol.* 2012;160(2):726–37.
53. Curry TM, Peña MJ, Urbanowicz BR. An update on xylan structure, biosynthesis, and potential commercial applications. *The Cell Surface.* 2023;9:100101.
54. Miao YC, Liu CJ. ATP-binding cassette-like transporters are involved in the transport of lignin precursors across plasma and vacuolar membranes. *Proc Natl Acad Sci U S A.* 2010;107(52):22728–33.
55. Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. *Mol Ecol.* 2016;25(11):2337–60.
56. Souza DMSC, Avelar MLM, Fernandes SB, Silva EO, Duarte VP, Molinari LV, et al. Spectral quality and temporary immersion bioreactor for *in vitro* multiplication of *Eucalyptus grandis* × *Eucalyptus urophylla*. *3 Biotech.* 2020;10(10):457.
57. Boutte J, Maillet L, Chaussepied T, Letort S, Aury JM, Belser C, et al. Genome size variation and comparative genomics reveal intraspecific diversity in *Brassica Rapa*. *Front Plant Sci.* 2020;11:577536.
58. Deng HK, Luo L, Wang RQ, Gao SY, Zhang WJ. Genome size determination of *Scirpus Mariqueter* and its related species. *Guihaia.* 2023;10:1838–48.
59. Lötter A, Duong TA, Candotti J, Mizrahi E, Wegrzyn JL, Myburg AA. Haplogenome assembly reveals structural variation in *Eucalyptus* interspecific hybrids. *Gigascience.* 2022;12:giad064.
60. Ramakrishnan M, Satish L, Sharma A, Kurungara Vinod K, Emamverdian A, Zhou M, et al. Transposable elements in plants: recent advancements, tools and prospects. *Plant Mol Biol Rep.* 2022;40:628–45.
61. Cooper G, Adams K. *The cell: A molecular approach.* 2nd ed. Sunderland (MA): Sinauer Associates; 2000.
62. Baek S, Choi K, Kim GB, Yu HJ, Cho A, Jang H, et al. Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biol.* 2018;19(1):127.
63. Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and morphinan production. *Science.* 2018;362(6412):343–7.
64. Suzuki S, Li L, Sun YH, Chiang VL. The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* 2006;142(3):1233–45.
65. Burn JE, Hocart CH, Birch RJ, Cork AC, Williamson RE. Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. *Plant Physiol.* 2002;129(2):797–807.
66. Madhu P, Nithiyesh Kumar C, Anojkumar L, Matheswaran M. Selection of biomass materials for bio-oil yield: a hybrid multi-criteria decision making approach. *Clean Techn Environ Policy.* 2018;20:1377–84.
67. Santos RB, Hart PW, Jameel H, Chang H. Wood based lignin reactions important to the biorefinery and pulp and paper industries. *Biores.* 2013;8(1):1456–77.
68. Rojas CM, Senthil-Kumar M, Tzin V, Mysore KS. Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. *Front Plant Sci.* 2014;5:17.
69. Li Q, Qiao X, Li L, Gu C, Yin H, Qi K, et al. Haplotype-resolved T2T genome assemblies and pangenome graph of Pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant Commun.* 2024;5(10):101000.
70. Zhan W, Cui L, Yang S, Zhang K, Zhang Y, Yang J. Natural variations of heterosis-related allele-specific expression genes in promoter regions lead to allele-specific expression in maize. *BMC Genomics.* 2024;25(1):476.
71. Lafontaine DL, Yang L, Dekker J, Gibcus JH. Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. *Curr Protoc.* 2021;1(7):e198.
72. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27(6):764–70.
73. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33(14):2202–4.
74. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de Novo assembler. *Gigascience.* 2012;1(1):18.
75. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-gapcloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience.* 2020;9(9):giaa094.
76. Servant N, Varoquaux N, Lajoie BR, Viana E, Chen CJ, Vert JP, et al. HiC-pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 2015;16:259.
77. Zhang Y, Xiong Y, Xiao Y. 3dDNA: a computational method of building DNA 3d structures. *Molecules.* 2022;27(18):5936.
78. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 2016;3(1):99–101.
79. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647–54.
80. Saha S, Bridges S, Magbanua ZV, Peterson DG. Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Trop Plant Biol.* 2008;1:85–96.
81. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA.* 2015;6:11.
82. Xu Z, Wang H. Ltr_finder: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007;35(Web Server issue):W265–8.
83. Price AL, Jones NC, Pevzner PA. De Novo identification of repeat families in large genomes. *Bioinformatics.* 2005;21(Suppl 1):i351–8.
84. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27(2):573–80.
85. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics.* 2011;27(6):757–63.
86. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997;268(1):78–94.
87. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res.* 2004;14(5):988–95.
88. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37(8):907–15.
89. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. *Nat Protoc.* 2016;11(9):1650–67.
90. Chan KCK, Xu X, Wang X, Gu J, Loy CC. Glean: generative latent bank for image super-resolution and beyond. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(3):3154–68.
91. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 2021;49(16):9077–96.
92. Nawrocki EP, Eddy SR. Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.* 2013;10(7):1170–9.
93. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-coding RNA analysis using the Rfam database. *Curr Protoc Bioinformatics.* 2018;62(1):e51.
94. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13(9):2178–89.
95. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
96. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59(3):307–21.
97. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24(8):1586–91.
98. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics.* 2021;36(22–23):5516–8.
99. Harris RS. Improved pairwise alignment of genomic DNA [Ph.D. Thesis]. Ann Arbor, MI: The Pennsylvania State University; 2007.

100. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19(9):1639–45.
101. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 2012;40(7):e49.
102. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59–60.
103. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. *Science.* 2008;320(5875):486–8.
104. Zwaenepoel A, Van de Peer Y. Wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics.* 2019;35(12):2153–5.
105. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with stringtie. *PLoS Comput Biol.* 2022;18(6):e1009730.
106. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.