the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) doi: 10.1111/tpj.16201

RESOURCE

Chromosome-scale *de novo* genome assembly and annotation of three representative *Casuarina* species: *C. equisetifolia, C. glauca,* and *C. cunninghamiana*

Yong Zhang[1], Yongcheng Wei[1], Jingxiang Meng[1], Yujiao Wang[1], Sen Nie[2], Zeyu Zhang[3], Huiyuan Wang[3], Yongkang Yang[3], Yubang Gao[3], Ji Wu[3], Tuhe Li[3], Xuqing Liu[3], Hangxiao Zhang[3] and Lianfeng Gu[3,*]

[1]*Research Institute of Tropical Forestry, Chinese Academy of Forestry, Guangzhou 510520, China,*

[2]*Fujian Academy of Forestry Sciences, Fuzhou, Fujian 350012, China, and*

[3]*College of Forestry, Basic Forestry and Proteomics Research Center, Fujian Agriculture and Forestry University, Fuzhou 350002, China*

Received 26 February 2022; revised 15 February 2023; accepted 13 March 2023.
*For correspondence (e-mail lfgu@fafu.edu.cn).

SUMMARY

Australian pine (*Casuarina* spp.) is extensively planted in tropical and subtropical regions for wood production, shelterbelts, environmental protection, and ecological restoration due to their superior biological characteristics, such as rapid growth, wind and salt tolerance, and nitrogen fixation. To analyze the genomic diversity of *Casuarina*, we sequenced the genomes and constructed *de novo* genome assemblies of the three most widely planted *Casuarina* species: *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. We generated chromosome-scale genome sequences using both Pacific Biosciences (PacBio) Sequel sequencing and chromosome conformation capture technology (Hi-C). The total genome sizes for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* are 268 942 579 bp, 296 631 783 bp, and 293 483 606 bp, respectively, of which 25.91, 27.15, and 27.74% were annotated as repetitive sequences. We annotated 23 162, 24 673, and 24 674 protein-coding genes in *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*, respectively. We then collected branchlets from male and female individuals for whole-genome bisulfite sequencing (BS-seq) to explore the epigenetic regulation of sex determination in these three species. Transcriptome sequencing (RNA-seq) revealed differential expression of phytohormone-related genes between male and female plants. In summary, we generated three chromosome-level genome assemblies and comprehensive DNA methylation and transcriptome datasets from both male and female material for three *Casuarina* species, providing a basis for the comprehensive investigation of genomic diversity and functional gene discovery of *Casuarina* in the future.

Keywords: *Casuarina*, whole-genome assembly, chromosome conformation capture technology, PacBio single-molecular long-read technology (SMRT) sequencing, whole-genome bisulfite sequencing.

INTRODUCTION

The 96 members of the Casuarinaceae family are grouped into four genera: *Allocasuarina* L. A. S. Johnson, *Casuarina* L., *Ceuthostoma* L. A. S. Johnson, and *Gymnostoma* L. A. S. Johnson (Sogo et al., 2001). *Casuarina* is comprised of 17 species (karyotypes: $2n = 18$) (Jarolímová, 1994; Yasodha et al., 2004), most of which originate in southeast Asia, tropical and subtropical coastal Australia, and the Pacific islands (Ho et al., 2002; Wheeler et al., 2011). Many *Casuarina* species are cultivated as pioneer trees along tropical and subtropical coastal areas for use as windbreaks and sand-shifting controls due to their high tolerance for extreme environmental conditions, such as typhoons, high salinity, low-nutrient sandy soils, and drought (Scotti-Campos et al., 2016; Van der Moezel et al., 1989). The leaves of *Casuarina* trees are reduced to lanceolate scales, with their needle-like branchlets being the actual photosynthetic organs, thereby reducing evapotranspiration and contributing to the adaptation of these trees to high temperatures (Zhong et al., 2013) and strong

winds or typhoons from the sea (Zhong et al., 2010). Large *Casuarina* trees, especially *C. equisetifolia* L., *C. glauca* Sieber ex Spreng., and *C. cunninghamiana* Miq., are widely cultivated in poor habitats due to their superior biological characteristics, such as rapid growth, wind and salt tolerance, and nitrogen fixation (Goel & Behl, 2005; He et al., 2005; Pinyopusarerk et al., 2004). Extensive studies have been performed on the nitrogen-fixing root nodules of *C. glauca* (Abdel-Lateif et al., 2013) and *C. cunninghamiana* (Davenport, 1960), which are derived from a symbiotic relationship with the nitrogen-fixing actinomycete *Frankia*. Other research topics include the dissection of salt stress tolerance in *C. equisetifolia* (Ngom et al., 2016; Tani & Sasakawa, 2003). However, it is difficult to explore the molecular basis of these physiological responses without a reference genome. We previously reported the first scaffold-level genome assembly of *C. equisetifolia*, achieving a scaffold N50 size of 1.06 Mb, largely based on short Illumina reads (Ye et al., 2019). With the development of long-read technologies, both the contiguity and precision of genome assemblies can now be improved. Moreover, genome sequences or assemblies have not been reported for *C. glauca* or *C. cunninghamiana*.

Like other members of the Casuarinaceae (Broadhurst, 2012), *C. equisetifolia* is predominantly a dioecious species, but monoecy does occur. The proportion of monoecious individuals in a given stand varies considerably from less than 10% (Luechanimitchit & Luangviriyasaeng, 1996) to as high as 80%, observed on the island of Guam (Schlub et al., 2010). The male and female inflorescences of monoecious *C. equisetifolia* individuals flower synchronously, and a selfing rate of up to 42% was detected, underscoring their high self-compatibility (Zhang et al., 2016). *C. glauca* and *C. cunninghamiana* are also primarily dioecious (Wilson & Johnson, 1989); however, their reproductive biology has not been extensively described.

Previous studies reported multiple methods of identifying *Casuarina* species (Ghosh et al., 2011). The extent of genetic diversity in *C. equisetifolia*, *C. cunninghamiana*, and *C. glauca* has been investigated using random amplified polymorphic DNA markers (Ho et al., 2002; Ndoye et al., 2011), inter-simple sequence repeat (ISSR) markers (Rasmi et al., 2011), ISSR-PCR (Yasodha et al., 2004), sequence-characterized amplified region (SCAR) markers (Ghosh et al., 2011), and expressed sequence tag (EST)-SSRs (Li et al., 2018; Xu et al., 2018; Yu et al., 2020; Zhang et al., 2020). The transcripts expressed in the nodules of *C. glauca* have been identified by comparing sets of ESTs prepared from roots with or without nodules (Hocher et al., 2006). Several studies have also focused on the mechanisms of salt tolerance (Fan et al., 2018; Tani & Sasakawa, 2006; Wang, Zhang, Fan, et al., 2021; Wang, Zhang, Qiu, et al., 2021), cold tolerance (Li et al., 2017), and wood formation (Vikashini et al., 2018; Ye et al., 2019) in *C. equisetifolia*. Genes related to actinorhizal nodule development in *C. glauca* have also been reported (Clavijo et al., 2015; Diédhiou et al., 2014; Ghodhbane-Gtari et al., 2019; Hocher et al., 2011; Laplaze, Duhoux, et al., 2000; Laplaze, Ribeiro, et al., 2000; Obertello et al., 2003; Péret et al., 2007; Zhong et al., 2013), in work that has benefited greatly from the availability of a genetic transformation system for this species (Clavijo et al., 2015; Le et al., 1996; Péret et al., 2007; Santi et al., 2003; Svistoonoff et al., 2010; Zhong et al., 2013). Knockdown of genes of interest has been achieved by RNA interference to investigate nodulation mechanisms in *C. glauca*; for example, knockdown of a chalcone synthase gene decreased flavonoid levels and resulted in severely impaired nodulation (Abdel-Lateif et al., 2013).

A chromosome-scale genome assembly of *Casuarina* is still lacking, which hampers functional genomics-empowered discovery of mechanism behind stress tolerance in *C. equisetifolia* and the nitrate signaling pathway of *C. glauca* and *C. cunninghamiana*. Our goal was therefore to produce chromosome-scale genomes to accelerate molecular studies of *Casuarina*. Here, we report chromosome-scale genome assemblies of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* obtained using a combination of Pacific Biosciences (PacBio; Menlo Park, CA, USA) single-molecule real-time (SMRT) long-read sequencing, high-throughput chromosome conformation capture sequencing (Hi-C), and Illumina (San Diego, CA, USA) short-read sequencing. We also performed genome-wide bisulfite sequencing (BS-seq) analyses to identify the differences in DNA methylation profiles between male and female plants from the three species, whose epigenetics were previously unexplored. Finally, we investigated the relationship between gene expression and DNA methylation using the above BS-seq data and by producing transcriptome sequencing (RNA-seq) datasets, offering preliminary clues about epigenetic regulation of gene expression in *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. In conclusion, the availability of *de novo* chromosome-scale genome assemblies for the three species developed in this work will complement the previously developed genetic transformation system for *C. glauca* and facilitate the elucidation of the molecular mechanisms involved in stress tolerance, wood formation, nitrate signaling, and sex determination in *Casuarina*.

RESULTS

Morphology and genome survey of the three species

C. equisetifolia, *C. glauca*, and *C. cunninghamiana* are important species used in ecological restoration. The branchlets of the three species present different morphologies (Figure 1a), with *C. equisetifolia* producing the

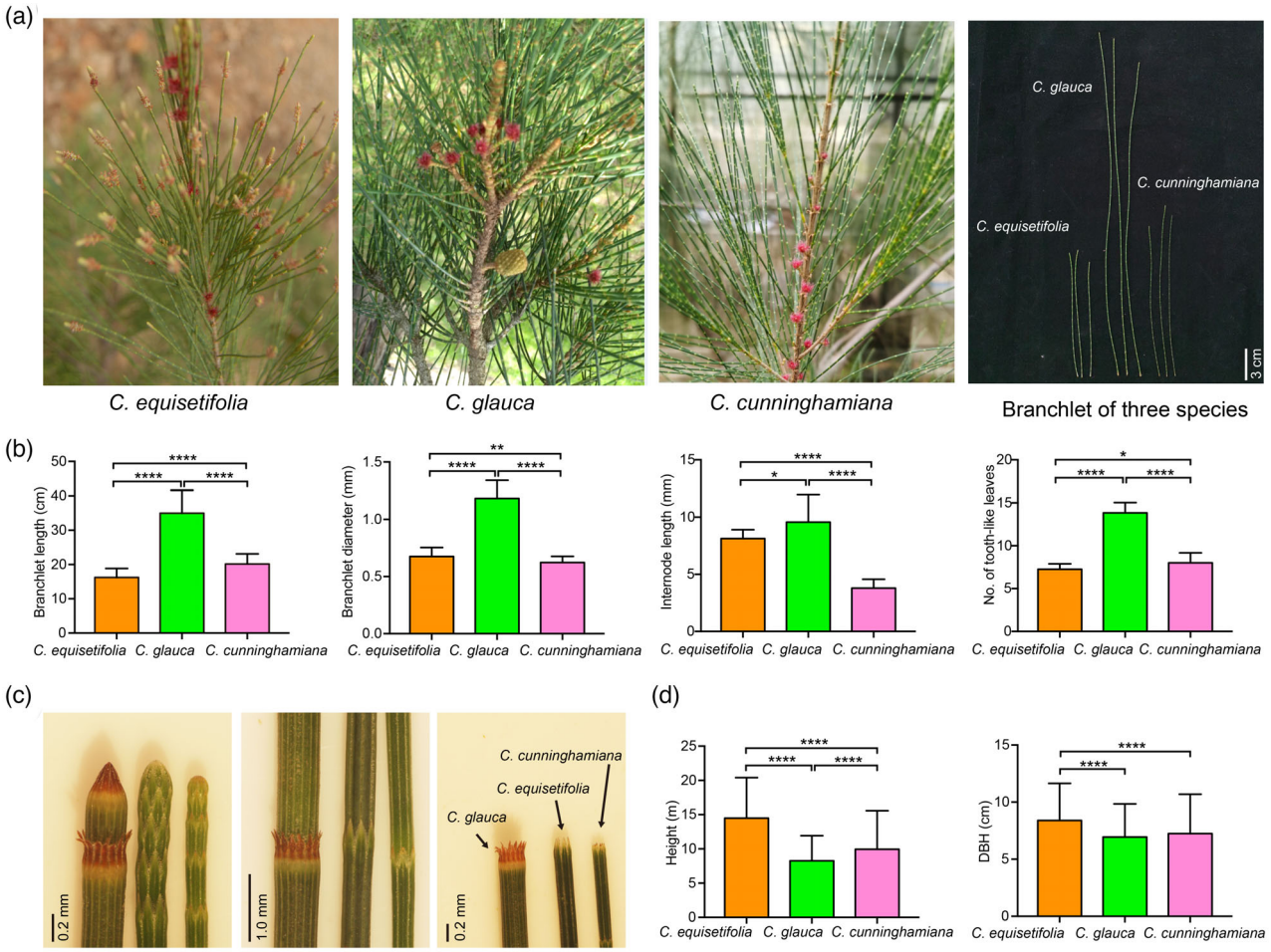

Figure 1. Morphological characteristics of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. (a) Branchlets of each species. (b) Branchlet length, diameter, internode length, and number of tooth-like leaves present in 5-year-old plants of each species. Asterisks represent significant differences (Student's *t*-test). (c) Branchlets observed under a microscope. (d) Height and diameter at breast height (DBH) of 6-year-old plants.

shortest branchlets, *C. glauca* forming the thickest and longest branchlets, and *C. cunninghamiana* presenting the thinnest branchlets with the shortest internodes (Figure 1b). The branchlets of each species are surrounded by tiny tooth-like leaves, whose diminutive nature may help reduce water loss; however, the tooth-like leaf tips and leaf sheaths differ between the species (Figure 1b,c), suggesting an underlying genetic diversity. *C. glauca* has the most tooth-like leaves of the three species (Figure 1c).

C. equisetifolia is taller and has a greater diameter at breast height (DBH) than *C. glauca* or *C. cunninghamiana* (Figure 1d). *C. equisetifolia* trees can be male, female, or monoecious, while *C. glauca* and *C. cunninghamiana* trees are almost exclusively diecious. The extent of phenotypic variation observed above within *Casuarina* suggests underlying genetic differences among the three species. We selected only female individuals from diecious species for genome sequencing. Epigenetic regulatory mechanisms play an important role in sex determination of *Asparagus officinalis* (Li et al., 2021). Therefore, we separately collected male and female samples to characterize their genome-wide methylation profiles so as to focus on the epigenetic effects rather than on genetic determinants of sex determination (Figure 2).

We performed a genome survey using 87 983 462, 83 278 102, and 82 681 769 paired-end reads of 150 bp in length to estimate the genome sizes of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*, respectively. In this study, we selected a *k*-mer value of 19 to cover most of the genomic and repetitive sequences in the three species (Figure S1). Based on the frequency distribution of 19-mers, we estimated the genome sizes of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* to be approximately 315 Mb, 313 Mb, and 328 Mb, respectively. We also estimated the heterozygosity rates for *C. equisetifolia* (0.66%), *C. glauca* (2.47%), and *C. cunninghamiana* (1.75%). Vigorous trees tend to be highly heterozygous and display a fitness superiority over homozygous individuals

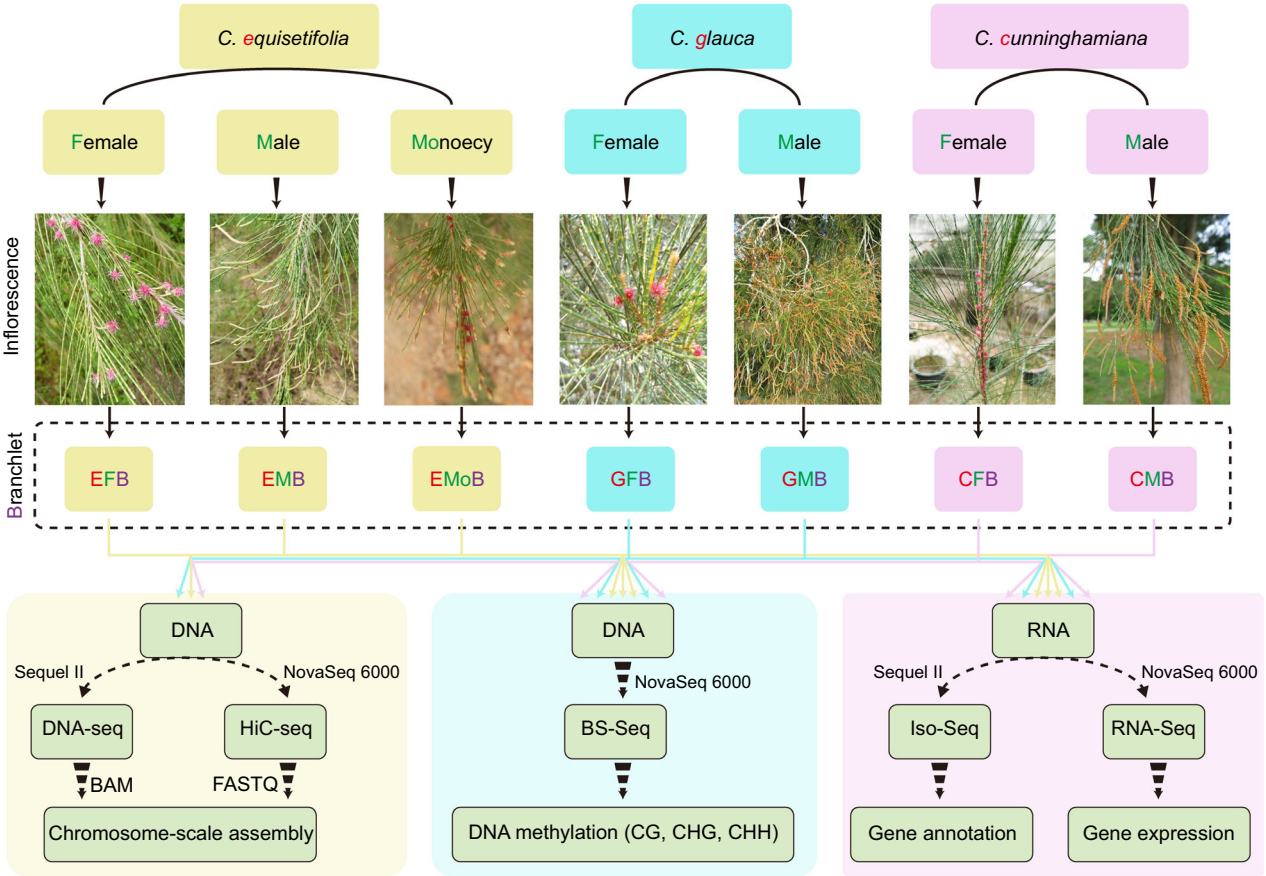

Figure 2. Flowchart of genome sequencing, BS-seq, Iso-seq, and RNA-seq for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. Samples collected from female branchlets from *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* were designated EFB, GFB, and CFB, respectively, while those collected from male branchlets of the three species were designated EMB, GMB, and CMB, respectively. Samples collected from the monoecious branchlets of *C. equisetifolia* were designated EMoB.

(Mitton, 2001). Consistent with a heterozygous advantage, *C. glauca* trees, with the highest heterozygosity, have significantly longer and thicker branchlets than trees of the other two species (Figure 1b).

De novo genome assembly for the three species using PacBio single-molecule sequencing

We generated *de novo* genome assemblies for each species using PacBio SMRT long-read sequencing technology. In total, PacBio SMRT sequencing provided approximately 408×, 331×, and 332× coverage of the *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* genomes, respectively. We constructed an initial string graph assembly to generate contig sequences using the genome assembler FALCON 1.2.4 (Chin et al., 2016), which is designed for PacBio SMRT long-read data (Figure S2). PacBio long reads can be used to overcome high levels of repetitiveness and heterozygosity. We detected high heterozygosity in our preliminary survey of *C. glauca* (2.47%) and *C. cunninghamiana* (1.75%). Thus, we employed Purge Haplotigs (Roach et al., 2018) to remove redundant sequences and generate de-duplicated assemblies for the three species (Figure S2). The contig N50 values were 5 391 619 bp for *C. equisetifolia*, 1 118 783 bp for *C. glauca*, and 2 232 746 bp for *C. cunninghamiana*. The largest contigs for each species were 13 802 205 bp (*C. equisetifolia*), 6 073 744 bp (*C. glauca*), and 12 282 031 bp (*C. cunninghamiana*), and the final combined contig lengths were 268 859 079 bp (*C. equisetifolia*), 296 399 783 bp (*C. glauca*), and 293 352 106 bp (*C. cunninghamiana*). We calculated the GC content and average depth per 10-kb sliding window based on an alignment of paired-end clean Illumina short reads to the assembled contigs above using BWA 0.7.15 (Li & Durbin, 2009). The GC contents of the newly assembled genomes were identical (36.7%) in three species (Figure S3).

Characterization of the Hi-C-assisted genome assemblies

The Hi-C method was originally proposed to investigate 3D genome organization by measuring the frequency of physical contacts between any pair of genomic loci within the

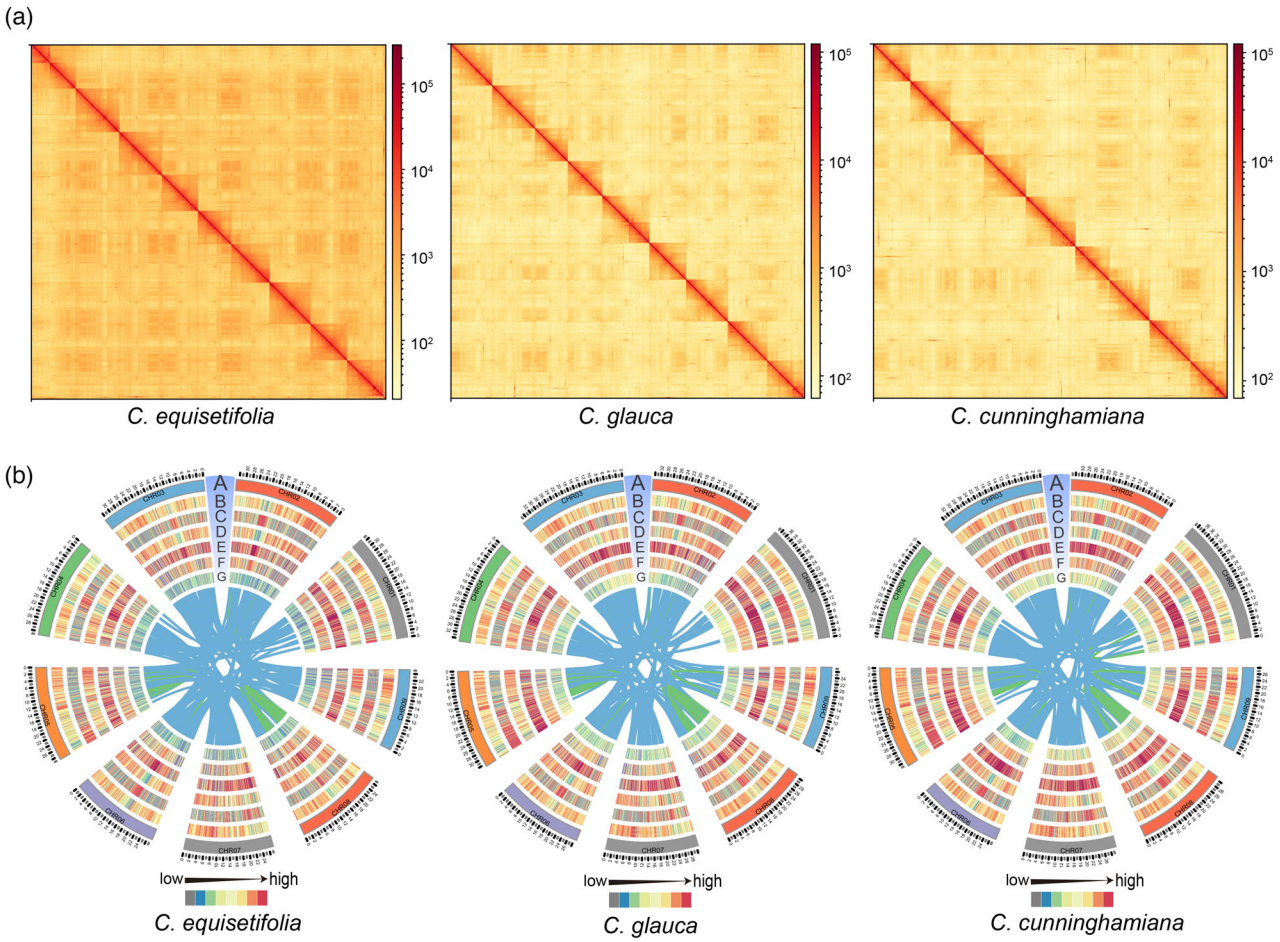

Figure 3. Features of the anchored assemblies among chromosomes 1–9 for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. (a) Heatmap showing the two-dimensional chromosome interaction matrix from Hi-C. (b) Circos plots summarizing genome features: from outermost to innermost, the circles show the number and size of the chromosomes (A), GC content (B), transposable element (TE) density (C), protein-coding gene density (D), CG methylation (E), CHG methylation (F), and CHH methylation (G). Paralogous gene pairs are linked in the center of the Circos plot. Green and blue indicate intrachromosomal and interchromosomal paralogous gene pairs, respectively.

nucleus (Lieberman-Aiden et al., 2009). More recently, Hi-C technology has been increasingly applied to the assembly of chromosome-scale scaffolds in helping to orient large contigs along pseudochromosomes (Zhang et al., 2021). Here, we produced a Hi-C interaction map to resolve the position of the contigs generated by *de novo* genome assembly. In total, our Hi-C libraries generated 5 337 502 (*C. equisetifolia*), 4 951 212 (*C. glauca*), and 3 438 248 (*C. cunninghamiana*) Illumina short reads. The Juicer pipeline extracted 2 301 842 (*C. equisetifolia*), 1 273 081 (*C. glauca*), and 874 678 (*C. cunninghamiana*) valid pairs (representing di-tags), of which 1 822 798 (*C. equisetifolia*), 1 035 503 (*C. glauca*), and 757 771 (*C. cunninghamiana*) had unique di-tags. For each species, we then loaded unique di-tags and a FASTA file from FALCON representing the draft assembly into the 3D *de novo* assembly (3D-DNA) pipeline to assign each draft assembly to a chromosome, resulting in the anchoring of contigs along nine pseudochromosomes (Figure 3). The final sizes of these assembled pseudomolecules were 266 390 111 bp (*C. equisetifolia*), 295 462 384 bp (*C. glauca*), and 292 499 405 bp (*C. cunninghamiana*). The leftover unanchored scaffolds only accounted for 2 552 468 bp (*C. equisetifolia*), 1 169 399 bp (*C. glauca*), and 984 201 bp (*C. cunninghamiana*) of sequence. This resource provides a valuable reference for the functional investigation of their genomes and for the genetic improvement of *Casuarina* species.

To the best of our knowledge, structural variants (SVs) among the three species have not been explored. To explore the diversity of *Casuarina*, we performed SV prediction and analysis at the whole-chromosome level in the three species. From pairwise comparisons between species, we identified 7 572 081 (*C. equisetifolia* versus *C. glauca*), 8 264 972 (*C. equisetifolia* versus *C. cunninghamiana*), and 5 573 389 (*C. glauca* versus *C. cunninghamiana*) single-nucleotide variations (SNVs) (Figure S4). We also

identified 401 227, 423 448, and 306 514 deletions and 397 978, 429 429, and 316 537 insertions in *C. equisetifolia* versus *C. glauca*, *C. equisetifolia* versus *C. cunninghamiana*, and *C. glauca* versus *C. cunninghamiana*, respectively (Figure S4). The SVs among the three species were widely distributed across the nine pseudochromosomes (Figures S5, S6, and S7). SVs tended to cluster in specific genomic regions rather than being evenly distributed along the genome. For example, many SVs were concentrated on chromosome 1 from 29 605 000 bp to 29 610 000 bp in the pairwise comparison between *C. glauca* and *C. cunninghamiana* (Figure S7). It will be interesting to investigate the association between genetic and phenotypic differences (Figure 1) in the future. These SVs among the three genomes will serve as an important resource for describing the genomic variation in *Casuarina*.

Evaluation of genome assembly revealed high completeness values and improved contiguity

To assess the quality of our chromosome-level genome assemblies, we calculated the Benchmarking Universal Single-Copy Orthologs (BUSCO) score using the Eudicotyledons dataset to search for conserved single-copy orthologs (SCOs). We identified 98.0, 97.9, and 97.9% of SCOs as complete single-copy genes in the *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* genome assemblies, respectively. These high values confirmed the high quality of all three chromosome-level genome assemblies. We also produced RNA-seq datasets from male and females branchlets of each species (Figure 2). To validate the Hi-C genome assemblies, we mapped all clean Illumina short reads to the assembled genomes, yielding overall alignment rates of 93.5 (*C. equisetifolia*), 93.0 (*C. glauca*), and 95.5% (*C. cunninghamiana*). These high rates of alignment further confirmed that the three genome assemblies cover most expressed genes in these species.

We previously reported a scaffold-level genome assembly for *C. equisetifolia* and released a draft genome (Ye et al., 2019). We wished to compare this older version of the genome to our improved assembly generated in this study using mimimap2 (2.15-r905) (Li, 2018). We determined that 272 323 437 bp of sequence from the older version were aligned to 247 633 663 bp of current genome versions, representing 92% of the current pseudochromosome-scale assembly. Notably, 603 previously unanchored scaffolds appear to have been incorporated into the nine pseudochromosomes (from chr1 to chr9) in the new assembly, comprising 90 (chr 1), 74 (chr 2), 77 (chr 3), 61 (chr 4), 77 (chr 5), 48 (chr 6), 58 (chr 7), 62 (chr 8), and 56 (chr 9) unaligned scaffolds (Figure S8). We also evaluated the quality of the *de novo* assembly using the long terminal repeat (LTR) Assembly Index (LAI), which evaluates assembly contiguity using LTR retrotransposons (Ou et al., 2018). We obtained LAI values of 21.52 (*C. equisetifolia*), 20.53 (*C. glauca*), and 20.55 (*C. cunninghamiana*). LAI scores above 20 indicate a high quality of the assembly (Ou et al., 2018). Among the three genomes, *C. equisetifolia* had the highest LAI, indicative of high-quality assembly of all intergenic and repetitive sequences. A previous scaffold-level genome assembly for *C. equisetifolia* based on Illumina short reads (Ye et al., 2019) achieved an LAI of 10.87, reflecting the limited ability of Illumina short reads to resolve intergenic and repetitive sequences. Thus, the chromosome-scale assemblies presented here substantially improved contiguity, which will facilitate analyses from functional and evolutionary perspectives in the future.

Annotation of repetitive sequences and protein-coding genes

Genome size is positively correlated with the abundance of repeat sequences and transposable elements (TEs) (Michael, 2014). Long-read technologies solve the major challenge of assembling the typically highly repetitive TE-rich regions in plant genomes. We annotated high-quality TEs throughout the three assembled genomes. In total, we classified 25.91 (*C. equisetifolia*), 27.15 (*C. glauca*), and 27.74% (*C. cunninghamiana*) of each genome as repetitive sequences. The fraction of TEs detected here was thus close to the average of 30% seen in other plant genomes (Lee & Kim, 2014). LTR TEs, including *Copia* and *Gypsy*, were the most common retroelements. We also detected a relatively small fraction of uncharacterized repeats, accounting for 5.25, 6.19, and 6.70% of the total assemblies of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*, respectively.

Using several tools (Figure S9), we annotated 23 161 (*C. equisetifolia*), 24 672 (*C. glauca*), and 24 672 (*C. cunninghamiana*) protein-coding genes. The average transcript lengths were 1816 nucleotides (nt) (*C. equisetifolia*), 1834 nt (*C. glauca*), and 1830 nt (*C. cunninghamiana*). Transcripts consisted of an average of six exons in all three species. The average size of predicted proteins was 452 amino acids (aa) (*C. equisetifolia*), 445 aa (*C. glauca*), and 455 aa (*C. cunninghamiana*).

Collinearity of the three genome assemblies

We performed a global genomic comparison among the three high-quality genomes. We observed extensive global collinearity among the three species (Figure 4a). Indeed, we identified 14 101 orthologous genes among the three species (Figure S10). To gain a better understanding of secondary cell wall thickening in these species, we analyzed the families of cellulose, hemicellulose, and lignin biosynthesis-related genes in three species based on homology with black cottonwood (*Populus trichocarpa*) (Table S1). Phylogenetic analysis indicated an absence of some orthologous groups of genes related to hemicellulose (Figure S11, green branch) and lignin

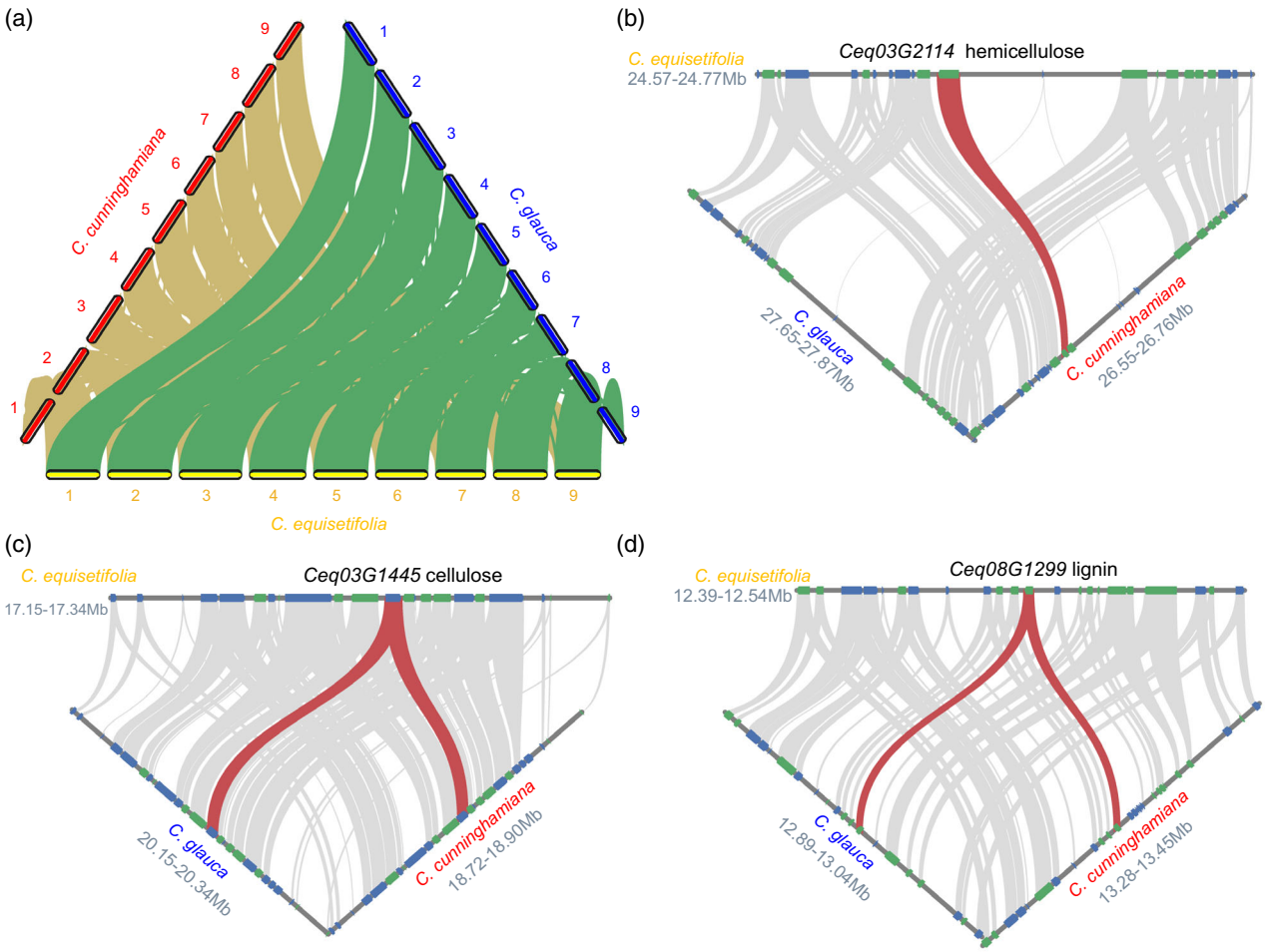

Figure 4. Global collinearity and microsynteny visualization for the three species. (a) Visualization of pairwise synteny among the nine chromosomes of the three species. Lines represent homologous blocks syntenic between *C. equisetifolia* and *C. glauca* (green) or *C. cunninghamiana* (yellow). (c, d) Microsynteny visualization of selected hemicellulose- (b), cellulose- (c), and lignin-related (d) genes. Blue/green blocks and gray lines correspond to genes and syntenic gene pairs, respectively. Red lines in (b)–(d) represent syntenic gene pairs for hemicellulose (Ceq03G2114), cellulose (Ceq03G1445), and lignin (Ceq08G1299), respectively.

biosynthesis (Figure S12, green branch). Furthermore, a microsynteny visualization revealed that a hemicellulose-related gene associated with secondary growth (Figure 4b) was not syntenic between *C. equisetifolia* and *C. glauca*, potentially explaining the differences in the height and DBH of these two species (Figure 2d). Furthermore, the depth of coverage in this region was 54× (*C. cunninghamiana*), 69× (*C. equisetifolia*), and 71× (*C. glauca*). Thus, the loss of collinearity of Ceq03G2114 in *C. glauca* is not caused by problems with genome assembly, as *C. glauca* had more reads than the other two species in this region. However, phylogenetic analysis revealed that most other orthologous groups among the three species show one-to-one correspondence (Figures S11 and S12, blue branch). For example, we observed microsynteny for both a cellulose-related gene (Figure 4c) and a lignin-related gene (Figure 4d).

Methylation profiling by whole-genome BS-seq revealed differentially methylated regions in sex determination

The proportion of monoecious individuals in *C. equisetifolia* varies from less than 10% to as much as 80% (Schlub et al., 2010), suggesting the potential influence of epigenetics in its regulation. To explore this possibility, we collected branchlets from male and female individuals of each species, as well as from monoecious *C. equisetifolia* individuals, for whole-genome BS-seq and RNA-seq analyses (Figure 2). Cytosine methylation at the 5′ position in the CG, CHG, and CHH contexts plays an important role in epigenetic regulation. Here, we examined the DNA methylation profile of chromosome 1 in each of the three species using BS-seq to obtain a preliminary estimate of their entire methylation profiles. We determined that the methylation level in the CG and CHG contexts is higher than in the CHH context (Figure S13).

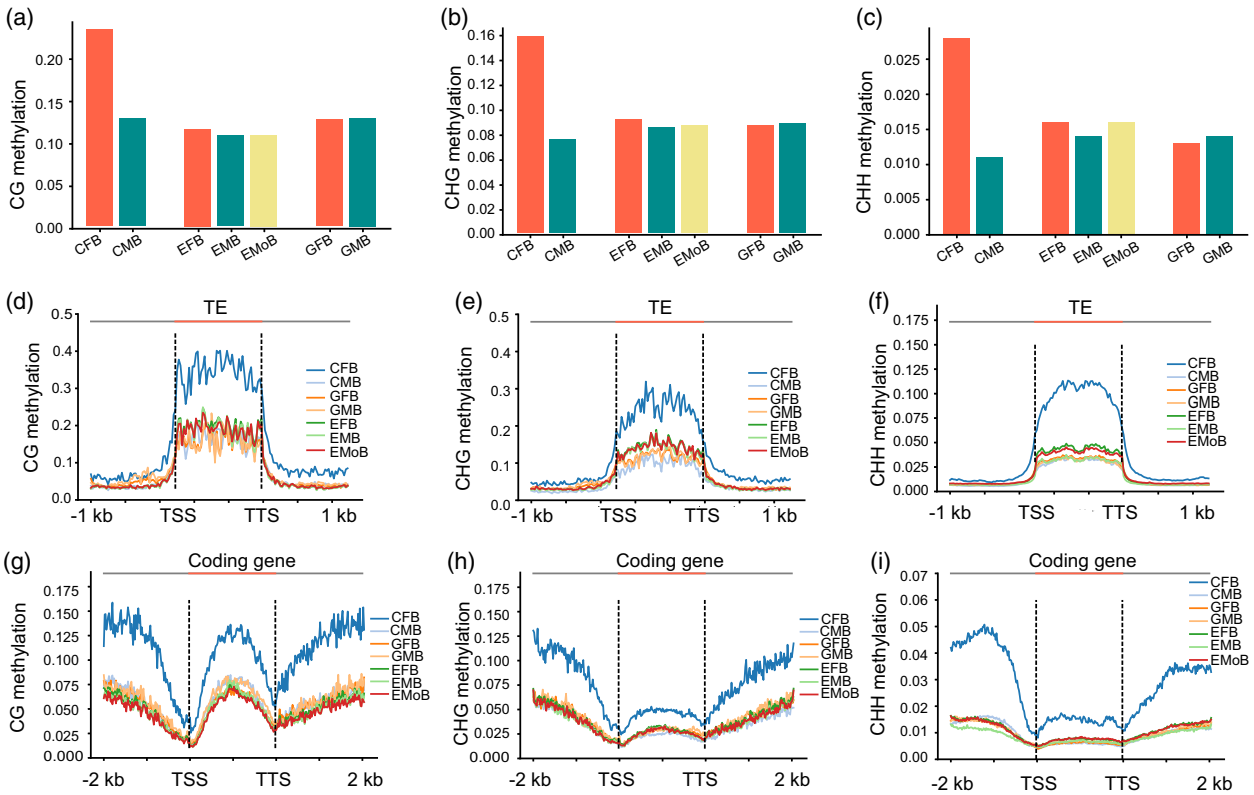

Figure 5. Whole-genome methylation profiles for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. (a–c) Histograms showing the genome-wide methylation of (a) CG, (b) CHG, and (c) CHH. (d–f) Genome-wide methylation profiles of transposable elements (TEs) in (d) CG, (e) CHG, and (f) CHH contexts. (g–i) Genome-wide methylation profiles of protein-coding genes in (g) CG, (h) CHG, and (i) CHH contexts.

The examination of methylation levels at the chromosome scale also enabled the elucidation of potential differences between the male and female samples in the three species. Accordingly, we subtracted the methylation level measured in the female sample from that of the male sample and plotted the methylation differences. In *C. cunninghamiana* (CFB), we detected more regions with higher CG, CHG, and CHH methylation levels in the female samples compared to the male samples (Figure S14). Of the three methylation contexts, the overall difference between the male and female samples in all three species was largest for CG methylation, while the difference in CHG methylation was relatively small.

In all three species, the methylation level in the female samples of *C. cunninghamiana* (CFB) was the highest in all three contexts, which was significantly different from the pattern in male samples (Figure 5a–c). Overall, the methylation level in the CG context (about 12% in *C. equisetifolia* and *C. glauca*) was the highest, followed by CHG (about 8% in the three species) and CHH (about 1.5% in the three species).

We also determined the overall methylation level of TEs (Figure 5d–f). TEs presented higher DNA methylation levels than protein-coding genes (Figure S15). The metaplots of methylation levels of TEs were relatively similar between samples, with the exception of CFB samples. The methylation levels in the CG (Figure 5d), CHG (Figure 5e), and CHH contexts (Figure 5f) along the TE bodies were about 20, 15, and 3.5%, respectively. We observed a drastic reduction in methylation both upstream and downstream of the TE bodies.

We also generated metaplots for methylation levels along the body of protein-coding genes and 2-kb upstream and downstream regions. The methylation levels of both the gene body and the flanking regions were higher in CFB than in the other samples (Figure 5g–i). The center of the gene body was more highly methylated in the CG context (Figure 5g) than at the transcription start sites (TSSs) and transcription termination sites (TTSs) in all samples; however, we detected no clear trend for methylation in the CHG (Figure 5h) and CHH (Figure 5i) contexts.

We defined differentially methylated regions (DMRs) across the genome to explore differences between male and female samples in the three species. For the pairwise comparisons of the male (CMB) and female (CFB) samples collected from *C. cunninghamiana* samples, most DMRs

were more highly methylated in the female tissues than in the male tissues (defined as a methylation gain) (Figure S16a), which was consistent with the methylation distribution along the TE and gene bodies (Figure 5). *C. glauca* also harbored more DMR gain events in the male (GMB) than in female samples (GFB) in the CG and CHG contexts. For *C. equisetifolia*, we observed comparable numbers of CG and CHG DMRs among the male (EMB), female (EFB), and monoecious (EMoB) samples (Figure S16a); however, the number of methylation gain and loss events in the CHH context was disproportionate, with more gain events between EFB and EMB and more loss events between EFB and EMoB. We classified DMRs into intergenic and genic types, based on their location on the chromosome. We identified loci that overlapped with DMR events. The percentage of genes with gain or loss DMRs showed a similar trend to that of DMR events (Figure S16b). In total, there were 210 common genes in all three species (CFB versus CMB, EFB versus EMB, and GFB versus GMB) that showed gain events and another 252 common genes in three species with loss events.

Finally, we focused on the distribution of DMRs around the gene coding regions and 1-kb regions upstream of the TSS and downstream of the TTS. For CG methylation, we observed that the DMRs are more frequent in the gene body than in their flanking regions (Figure S17). The distribution of CHG DMRs was similar across the gene bodies and flanking regions, while the promoter region was strongly enriched in CHH DMRs. For example, we noticed that the expansin-like B1 gene (*EXLB1*) has a lower methylation level in its gene body and promoter region in the female samples than in the male samples in both *C. equisetifolia* and *C. glauca* (Figure S18); however, the opposite was true for *C. cunninghamiana*.

Transcriptomic regulation in the three species

To quantify gene expression, we performed strand-specific RNA-seq of the male and female samples from the three species (Figure 2) and estimated expression as fragments per kilobase of transcript per million mapped reads (FPKM) values. A Spearman correlation analysis revealed a clear distinction between the three species based on their transcript levels using one-to-one-to-one orthologs across the three species (Figure S19a). The three replicate samples for each species clustered together and showed a high correlation, based on a principal component analysis (Figure S19b).

We identified differentially expressed genes (DEGs) using a *P*-value of <0.01 and a fold change value of >1.5 or <1/1.5 as the cutoffs by comparing expression profiles from female samples to those of male samples for each species (Figure S20). We identified 2765 downregulated genes and 3304 upregulated genes in female samples (CFB) from *C. cunninghamiana* compared to matching male samples (CMB) and 2199 downregulated genes and 4031 upregulated genes in female samples (GFB) from *C. glauca* compared to matching male samples (GMB). By contrast, we obtained comparable numbers of upregulated and downregulated genes in the pairwise comparisons (EFB versus EMB and EMoB versus EMB) of *C. equisetifolia* samples, whereas the EMoB samples had more upregulated genes than the female samples (EFB), with 3251 downregulated genes and 4301 upregulated genes in the EMoB samples compared to EFB (Figure S20).

We used *Z*-scores to normalize gene expression and *k*-means clustering analysis to classify these DEGs into eight clusters, as shown in Figures S21 and S22. The first group in the heatmap represents DEGs with a high expression in GMB samples (Figure S21) with a clear enrichment of Gene Ontology (GO) terms such as 'ethylene-activated signaling pathway', 'response to auxin', and 'isoprenoid biosynthetic process' (Figure S22). The second group was comprised of genes highly expressed in CFB samples, with an enrichment of GO terms associated with 'amino acid transmembrane transport' and 'positive regulation of transcription elongation from RNA'. The third group represented genes highly expressed in GFB samples, with an enrichment of the GO terms 'mRNA transcription', 'RNA-directed DNA methylation', and 'cell differentiation'. The fourth group was formed by highly expressed genes in EMoB samples, with enrichment in GO terms associated with 'gibberellin catabolic process', 'auxin biosynthetic process', 'cell differentiation', and 'regulation of transcription'. The fifth group was comprised of genes highly expressed in the female samples CFB and GFB, with enrichment of the GO terms 'cell division', 'cellulose catabolic process', 'mitotic spindle organization', and 'mitotic cell cycle'. The sixth group of genes was highly expressed in EMB samples, with enrichment of the GO terms 'ATP synthesis coupled proton transport', 'cell wall biogenesis', and 'response to auxin'. Group seven contained genes highly expressed in CMB samples, with enrichment of GO terms associated with the 'cytokinin biosynthetic process', 'RNA processing', and 'ethylene-activated signaling pathway'. The last group was comprised of genes highly expressed in EFB samples, with enrichment in the GO terms 'sexual reproduction', 'cellulose biosynthetic process', and 'ethylene-activated signaling pathway'. For example, Ccu03G1202 and Ccu08G1699 were annotated as having transcription factor activity and possibly being involved in the ethylene response. Both genes were highly expressed in CMB samples and expressed at low levels in CFB samples (Figure S23). Another example was Ccu06G1569, annotated as an auxin-responsive gene and highly expressed in female (CFB) compared to male trees (CMB) for *C. cunninghamiana*, which we validated by RT-qPCR (Figure S23). It will be interesting to investigate sex determination related to

phytohormone-related genes using genetic transformation to bridge the gap between gene and phenotype. We also extracted the promoter sequences from all genes from the eight groups above for HOMER (v4.11.1) motif analysis. The promoters of genes from several groups harbored a CGTA motif (Figure S24). The CGTA motif is a potential jasmonic acid (JA)-responsive element (Qayyum et al., 2022). JA was reported to regulate flower development (Yuan & Zhang, 2015). The association between the CGTA motif and flower development can now be investigated to identify potential transcription factors that bind to this motif.

Association between transcription and DNA methylation in the three species

To explore the relationship between gene expression and DNA methylation, we integrated our RNA-seq and BS-seq data (Figure S25). We noted that non-expressed genes have highly methylated promoters and gene bodies in all three species. Methylation in the CG context in particular showed a clear repressive effect in all three species. We had detected a higher methylation level in the female samples from *C. cunninghamiana* (CFB) compared to the male samples (CMB) (Figure 6). CHH methylation accumulated around the promoter regions in *C. cunninghamiana* genes, which was associated with a gradual increase in gene expression.

We classified the genes overlapping with hypermethylated regions detected in CFB samples into four expression types (no, low, middle, and high). Interestingly, we observed an enrichment of GO terms related to 'sexual reproduction', 'recognition of pollen', and 'pollen tube growth' in three groups of expressed genes (Figure S26), suggesting that these genes might be involved in sex determination in *C. cunninghamiana*. For example, Ccu05G0051 and Ccu09G1599 are annotated as having cysteine synthase activity by BLAST2GO, a function that is involved in pollen tube growth by double fertilization forming a zygote and endosperm. Hypermethylation may regulate gene expression. Indeed, we observed lower expression of Ccu05G0051 and Ccu09G1599 in CFB samples, which we validated by RT-qPCR (Figure S23). In summary, we provide preliminary clues about the association between DNA methylation and sex determination.

DISCUSSION

Multiple species of *Casuarina* were introduced into China for cultivation as pioneer trees in the construction of coastal shelterbelts for environmental protection and ecological restoration. With the aim of exploring the genomic diversity of *Casuarina*, we used both PacBio Sequel sequencing and Hi-C technology to assemble chromosome-scale genomes for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. The genome of each of the three species was assembled into nine pseudochromosomes, providing an important resource for researchers to explore the genetic diversity of this genus. We investigated SVs between the three species, providing useful information for pangenome analysis of *Casuarina* in the future. The effect of structural variation on key stress tolerance or nitrogen-fixing root nodulation traits can be assessed by genome-wide association studies when structural variations across multiple genetic materials of *Casuarina* are available.

Furthermore, we annotated all repetitive sequences and protein-coding genes in each species. TEs presented higher DNA methylation levels than coding genes (Figure 3), which was consistent with DNA methylation predominantly occurring at TEs as epigenetic surveillance systems (Michael, 2014). We also obtained preliminary clues to the epigenetic regulation of sex determination in the three species using BS-seq and RNA-seq. Most DEGs identified in this study between female and male samples showed enrichment of GO terms associated with phytohormone metabolism (Figure S22), suggesting that phytohormone-related genes might be critical for the regulation of sex determination. However, genetic transformation should be applied to investigate sex determination in relation to phytohormone-related genes.

Collectively, the chromosome-level genome assemblies of three representative species coupled with comprehensive DNA methylation and transcriptome datasets will pave the way for the *Casuarina* research community to explore the genetic diversity of these species and effectively elucidate the molecular mechanisms involved in stress tolerance, wood formation, nitrogen-fixing root nodulation, and sex determination.

EXPERIMENTAL PROCEDURES

Sample collection and genome survey

Young needle-like branchlets were harvested from the progeny of test trials for the three species (*C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*) in Dianbai, Maoming, Guangdong Province, China (111°01′27.99″E; 21°28′6.69″N), for use in PacBio genome sequencing. All branchlet samples were immediately frozen in liquid nitrogen, transported to the laboratory, and subjected to DNA and RNA extraction.

Genomic DNA (gDNA) from the branchlets was extracted using a DNA extraction kit (TIANGEN, Beijing, China). DNA concentration was determined using a Qubit dsDNA HS assay kit. DNA purity and integrity were confirmed using agarose gel electrophoresis on a Gel Doc XR+ (Bio-Rad Laboratories, Hercules, CA, USA) before being subjected to high-throughput sequencing. Illumina libraries suitable for generating paired-end 150-bp reads were constructed and sequenced using a NovaSeq 6000 platform (Illumina, San Diego, CA, USA). Low-quality reads were removed if they had more than three unidentified (N) bases and/or more than 20% of bases were of low quality (Phred Quality Score < 5). Reads with more than eight bases matching adapter sequences were also discarded. Genome heterozygosity, repeat content, and

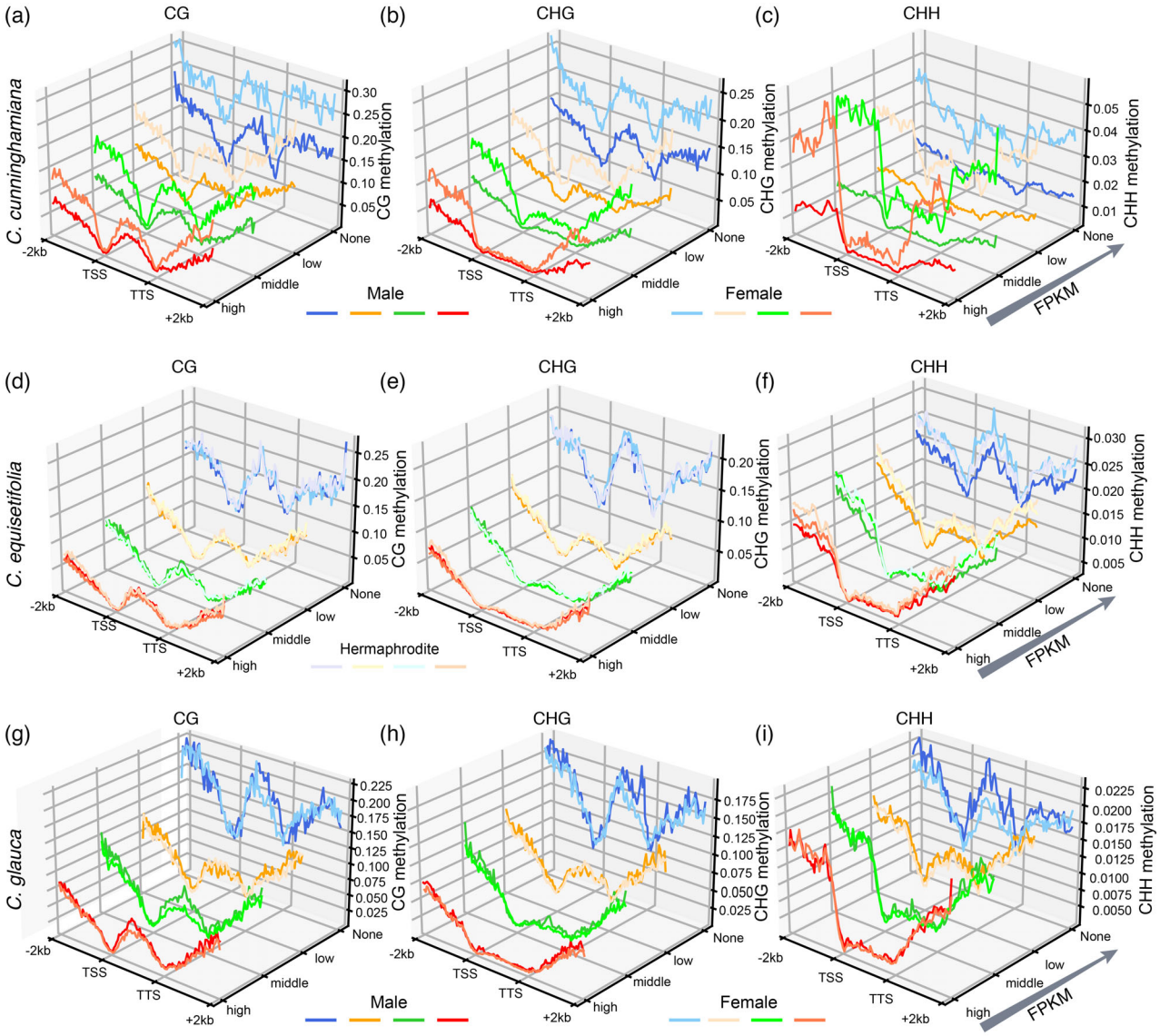

Figure 6. DNA methylation of gene bodies (from TSS to TTS) and flanking regions in the three species. The genes were divided into four categories according to their expression levels (from low to high).

size of sequencing reads were estimated using a *k*-mer-based statistical approach in GenomeScope v1 (Vurture et al., 2017).

Genome sequencing using a PacBio sequel II platform

Whole-genome sequencing libraries were constructed with a targeted insert size of 15 kb. To check the quality of the gDNA, 1 μl gDNA from each sample was used for a NanoDrop measurement (Thermo Fisher Scientific, Waltham, MA, USA). An A_{260}/A_{280} ratio of 1.8–2.0 with an A_{260}/A_{230} ratio greater than 2.0 was required before proceeding with sample processing. Furthermore, 1 μl gDNA was used to measure the DNA concentration by Qubit. The main gDNA band for each sample should be larger than 40 kb, with no obvious degradation or RNA present. Each sample was purified using 0.8× AMPure PB beads, and then g-TUBEs (Covaris, Woburn, MA, USA) were used to shear the gDNA. Agarose gel electrophoresis was used to confirm that the size of the fragmented bands was about 15 kb.

Finally, a Template Prep Kit (Pacific Biosciences, Menlo Park, CA, USA) was used to construct SMRTbell libraries. First, the template was digested with *Exo*VII to remove single-stranded DNA contamination, after which the template was repaired with DNA Damage Repair Mix to remove impurities and repair nicks and gaps in the DNA chain. End Repair Mix was then used to fill in and repair the 5′ or 3′ of DNA double-stranded ends and simultaneously add phosphoric acid. The above products were purified and recovered with AMPure PB beads and their concentrations were determined using Qubit. Next, SMRT adaptors with hairpin structures were added. The products were digested with *Exo*III and *Exo*VII to remove excess adaptors and library molecules with incomplete structures, after which the products were again purified, recovered with AMPure PB beads, and quantified using Qubit. A Bluepippin apparatus was then used to select fragments of the expected size from the library. The products were purified and recovered with AMPure PB beads. The quality of the libraries was determined using an Agilent 2100 Bioanalyzer system

(Agilent Technologies, Santa Clara, CA, USA). Finally, all libraries were sequenced on a PacBio Sequel II platform in Continuous Long reads (CLR) sequencing mode.

De novo genome assembly

The FALCON *de novo* assembler v1.2.4 (Chin et al., 2016) was used to call high-quality consensus sequences based on Sequel subreads using the following parameters: length_cutoff = 5000, length_cutoff_pr = 10 000 (Figure S2). Then, pbmm2/gcpp from the SMRT Link software (https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/) was used to correct potential errors in the contig sequences. Furthermore, the contigs were corrected with Illumina short reads using the Burrows–Wheeler Alignment tool 0.7.15 (Li & Durbin, 2009) and Pilon 1.22 (Walker et al., 2014) with default parameters. Pairs of contigs (regional duplication) with high degrees of heterozygosity were identified and reassigned using purge_haplotigs 1.0.2 (Roach et al., 2018) to improve the genome assembly.

Construction and analysis of Hi-C libraries

Following the standard protocol described previously (Belton et al., 2012) with some modifications, Hi-C libraries were constructed using the original genome sequencing samples as starting material. After being ground in liquid nitrogen, the samples were crosslinked in 4% (w/v) formaldehyde at room temperature under a vacuum for 30 min. The crosslinking reaction was quenched with the addition of 2.5 M glycine for 5 min, after which the samples were placed on ice for 15 min. The samples were centrifuged at 1239 × *g* at 4°C for 10 min; the resulting pellet was washed with 500 μl phosphate-buffered saline and then centrifuged for 5 min at 1239 × *g*. Each pellet was resuspended in 20 μl of lysis buffer (1 M Tris–HCl, pH 8, 1 M NaCl, 10% [v/v] CA-630, and 13 units protease inhibitor), and the supernatant was then centrifuged at 4956 × *g* at room temperature for 10 min. The pellet was washed twice in 100 μl ice-cold 1× NEB buffer and then centrifuged for 5 min at 4956 × *g*. Nuclei were resuspended in 100 μl NEB buffer, solubilized with diluted SDS, and then incubated at 65°C for 10 min. After quenching of the SDS with Triton X-100, the samples were digested overnight with a 4-cutter restriction enzyme, *Dpn*II (400 units), on a rocking platform at 37°C. Finally, the Hi-C libraries from *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana* were sequenced on an Illumina NovaSeq 6000 platform.

To analyze the Hi-C data, paired-end FASTQ files from the three species were mapped to their respective *de novo* assembled contigs using Juicer 1.6.2 (Durand et al., 2016) with default parameters to identify valid pairs. 3D-DNA v1 with default parameters (Dudchenko et al., 2017) was then applied to generate chromosome-length assemblies. Collinearity between the three reference genomes was plotted using the JCVI utility libraries v1.0.9 (Tang et al., 2008). The LAI was calculated using LTR_retriever v2.9.0 with default parameters (Ou et al., 2018).

RNA extraction, construction of libraries for RNA-seq, and bioinformatics analysis

Total RNA was extracted from collected samples from the three species using a Dynabeads mRNA DIRECT Kit (Thermo Fisher Scientific). In total, 40 ng mRNA was used for the construction of each RNA-seq library. Strand-specific transcriptome libraries were then generated using the dUTP method, as follows. First, mRNA was sheared at high temperature, after which first-strand cDNA was synthesized by reverse transcription. During the synthesis of second-strand cDNAs, dTTP was replaced by dUTP in the PCR. Next, after end repair and A-tailing, the Illumina Truseq adaptor was added. The USER enzyme was then added to digest dUTP in the second cDNA strand. Finally, the digested products were purified, and Illumina P5 and P7 primers were used to amplify RNA-seq libraries by PCR before sequencing on an Illumina NovaSeq 6000 platform to generate 150-bp paired-end reads.

Clean RNA-seq reads were first mapped to the respective chromosome-scale genome using HISAT2 version 2.0.3-beta (Kim et al., 2019) with default parameters. The read summary for each gene was calculated using featureCounts v2.0.3 (Liao et al., 2014), and the number of reads was normalized to the FPKM value (Mortazavi et al., 2008; Trapnell et al., 2012). The *P*-value and false discovery rate (FDR) were calculated using edgeR v3.12.1 (Robinson et al., 2010). A fold change of >1.5 or <1/1.5 and an FDR of <0.01 were used as the cutoffs for the identification of DEGs between male and female branchlets. An enrichment analysis of GO terms was performed using clusterProfiler 3.18.0 (Yu et al., 2012) using a P_{adj} cutoff of 0.05.

Construction of libraries for Iso-seq and bioinformatics analysis

PacBio Iso-seq was performed on total RNA extracted from the three species. First, the quality of total RNA was checked using an Agilent RNA 6000 Nano Kit, ensuring a concentration of ≥300 ng μl^{-1} and an RNA integrity number (RIN) of ≥8.5. The RNA bands were required to be intact, with no degraded RNA or DNA contamination. Next, double-stranded cDNA synthesis was performed using a SMARTer PCR cDNA Synthesis Kit (Takara Bio, Kusatsu, Japan). PCR amplification was performed on the cDNA to obtain a sufficient amount for each sample. Finally, a Template Prep Kit (PacBio) was used for Iso-seq library construction. All cDNAs were enzymatically repaired, flattened, purified, and then annealed to a SMRT adaptor. The final products were purified and subjected to exonuclease processing to remove impurities, such as small fragments and adapter dimers. The final Iso-seq libraries were obtained after two rounds of purification with AMPure PB beads. After Qubit quantification and Agilent 2100 quality control, the Iso-seq libraries were sequenced on a PacBio Sequel II platform.

Highly accurate consensus sequences were generated using the ccs module of SMRT Link with default parameters. The isoseq3 refine module from IsoSeq v3 (https://github.com/PacificBiosciences/IsoSeq) was then used to remove the poly(A) tail and primers. Next, bam2fasta v1.3.0 from the BAM2fastx tool (https://github.com/PacificBiosciences/bam2fastx) was used to convert the PacBio Iso-seq reads to FASTA files, which were corrected using LoRDEC (v0.7) with the following command: lordec-correct -k 19 -s 3 (Salmela & Rivals, 2014). The corrected long reads were aligned to the assembled genome using minimap2-2.17 (Li, 2018) with the '-ax splice -uf' option.

Repeat and structural annotation

Extensive *de-novo* TE Annotator (EDTA) v1.8.3 (Ou et al., 2019) was employed to annotate repetitive sequences, using the following parameters: -step all -t 30 -sensitive 1 --anno 1. To assist with structural annotation, a transcriptome-based annotation was used. The BAM alignment results from both RNA-seq and Iso-seq were then used as an input for Stringtie (v1.3.3) for genome-based transcript assembly using default parameters (Pertea et al., 2015). The non-redundant reference gene structures were generated and integrated with MAKER software 2.31.11 (Cantarel et al., 2008) (Figure S2). The GO terms associated with each gene were determined using BLAST2GO 1.3.11 (Conesa et al., 2005).

Construction of libraries for BS-seq and bioinformatics analysis

gDNA samples were purified to remove impurities such as proteins and salts. First, gDNA was fragmented into 150–450-bp dsDNA, as determined by their electrophoretic profile. The fragmented DNA samples were then end-filled and subjected to 3′-end d(A) tailing. The adapters were then ligated, after which the products were separated using 2% (w/v) agarose gel electrophoresis and treated with an EZ DNA Methylation-Gold kit (Zymo Research, Orange, CA, USA) to convert cytosines (C) to thymines (T). Finally, all products were PCR-amplified and purified for sequencing as 150-bp paired-end reads.

The bismark_genome_preparation module from the bismark software v0.22.1 (Krueger & Andrews, 2011) was used to index the genomes of the three species, and bismark and the bismark_methylation_extractor command were used to generate CX report files, including the 5mC methylation information, using default options. DMRs were identified using the R package DMRcaller 3.10 (Catoni et al., 2018) with default parameters. DMRs were defined as 100-bp regions with differences in CG, CHG, and CHH methylation levels greater than or equal to 0.4, 0.2, and 0.1, respectively.

RT-qPCR validation

Total RNA (1 μg) from CFB was reverse transcribed into cDNA using a PrimeScript™ RT reagent Kit with gDNA Eraser (TaKaRa, RR047A) for validation of gene expression. All cDNAs were diluted 10×, and 1–2 μl of 10× diluted cDNA was used as template in a 20-μl qPCR system. *Actin* was used as a reference gene. qPCR was performed using a Hieff qPCR SYBR Green Master Mix (YEASEN, 11202ES08) on an Agilent M × 3005P Real-Time PCR System, following the manufacturers' instructions. All qPCR primers are listed in Table S2.

AUTHOR CONTRIBUTIONS

YZ and LG conceived and designed the study. YZ, YW, JM, YW, SN, BW, and JW performed the experiments. ZZ, HW, YY, YG, XL, HZ, and LG analyzed the high-throughput sequencing data and performed genome assembly and annotation. YZ and LG wrote the manuscript. All authors have read and approved the final version of this manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant no. 31770716), the Fujian Forest Seedling Technology Project of *Casuarina*, and the Forestry Peak Discipline Construction Project of Fujian Agriculture and Forestry University (72202200205).

CONFLICTS OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The raw data used in this study have been deposited in the NCBI Sequence Read Archive under BioProject PRJNA796484. The assembled genomes are available from NCBI under accession numbers GCA_028551395.1 (*C. glauca*), GCA_028551485.1 (*C. cunninghamiana*), and GCA_028551475.1 (*C. equisetifolia*). The data in this study are also available at forestry.fafu.edu.cn/pub/Casuarina.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Frequency distribution map of 19-mer analysis.

Figure S2. Flowchart describing the chromosome-scale assembly procedure followed for *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*.

Figure S3. Distribution of GC contents in *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*. The highest density in the scatterplot is highlighted with a red dashed line.

Figure S4. Histograms showing SNVs, deletions, and insertions of pairwise cross-species comparisons.

Figure S5. Distribution of SVs in nine chromosomes from a *C. equisetifolia* versus *C. cunninghamiana* comparison. The red line represents regions including more than 1000 SVs per 10-kb bin.

Figure S6. Distribution of SVs in nine chromosomes from a *C. equisetifolia* versus *C. glauca* comparison. The red line represents regions including more than 1000 SVs per 10-kb bin.

Figure S7. Distribution of SVs in nine chromosomes from a *C. glauca* versus *C. cunninghamiana* comparison. The red line represents regions including more than 1000 SVs per 10-kb bin.

Figure S8. Clustering of scaffolds into nine chromosomes. Each chromosome is numbered (shown in red).

Figure S9. Flowchart illustrating the procedure for the annotation of protein-coding genes of *C. equisetifolia*, *C. glauca*, and *C. cunninghamiana*.

Figure S10. Venn diagram showing the extent of overlap for orthologous pairs among three species.

Figure S11. Phylogenetic analysis of orthologous groups for cellulose-related and hemicellulose-related genes in the three species.

Figure S12. Phylogenetic analysis of orthologous groups for lignin-related genes in the three species.

Figure S13. Methylation profiles along chromosome 1.

(a) CG, (b) CHG, and (c) CHH contexts. The *x*-axis and *y*-axis represent the linear genome coordinate and methylation level in the three species, respectively.

Figure S14. Comparison of methylation levels in the male, female, and monoecious samples of the three species. Female, male, and monoecious samples are designated F, M, and Mo, respectively. The *x*-axis and *y*-axis represent the linear genome coordinate and the difference in methylation level between two samples, respectively. The *y*-axis was generated by subtracting the methylation level of the female sample from that of the male sample or subtracting that of the male or female sample from that of the monoecious sample. Top, middle, and bottom panels correspond to CG (a), CHG (b), and CHH (c), respectively.

Figure S15. Correlation between DNA methylation and three genome elements (protein-coding genes, all, and TEs). All, entire genome; TE, transposable element.

Figure S16. Histogram representing the number of DMRs in the three species. A higher methylation level in the female sample relative to the male sample was defined as a methylation gain, whereas a lower methylation level in the female sample relative to the male sample was defined as a methylation loss. (a) The *y*-axis represents the number of DMR events. (b) The *y*-axis represents the number of genes including DMR events.

Figure S17. Distribution of DMRs based on pairwise comparisons between the male and female samples. The 1-kb upstream and downstream sequences were selected for the transcription start site (TSS), middle, and transcription termination site (TTS) regions, respectively. The red and blue lines represent higher and lower methylation in the female sample, respectively.

Figure S18. CG/CHG/CHH distribution along the expansin-like B1 gene and flanking regions in the three species.

Figure S19. Clustering of RNA-seq data from the three species. (a) Heatmap of the RNA-seq data from all samples. (b) Principal component analysis of the RNA-seq data. A $\log_2$(FPKM+1) transformation was applied to the FPKM values, and a PCA dimensionality reduction was performed.

Figure S20. Volcano plots of all differentially expressed genes in the three species.

Figure S21. Heatmap clustering analysis of all differentially expressed genes between the male and female samples of the three species.

Figure S22. Enriched GO terms in the eight categories of clustered differentially expressed genes.

Figure S23. RNA-seq and RT-qPCR validation of five selected genes.

Figure S24. Motif analysis of promoter regions from all genes differentially expressed between the male and female samples among the three species.

Figure S25. DNA methylation over the gene body (from TSS to TTS) and flanking upstream/downstream 2-kb regions in the male and female samples. The genes were divided into four categories according to their expression (from low to high).

Figure S26. GO enrichment for the four expression types (no, low, middle, and high) corresponding to hypermethylated regions in the CFB sample.

Table S1. The cellulose-related, hemicellulose-related, and lignin-related genes in the three species.

Table S2. Primer sequences for RT-qPCR validation.

REFERENCES

Abdel-Lateif, K., Vaissayre, V., Gherbi, H., Verries, C., Meudec, E., Perrine-Walker, F. et al. (2013) Silencing of the chalcone synthase gene in C asuarina glauca highlights the important role of flavonoids during nodulation. *New Phytologist*, **199**, 1012–1021.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. & Dekker, J. (2012) Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.

Broadhurst, L.M. (2012) Genetic diversity and population genetic structure in fragmented Allocasuarina verticillata (Allocasuarinaceae)–implications for restoration. *Australian Journal of Botany*, **59**, 770–780.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B. et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**, 188–196.

Catoni, M., Tsang, J.M., Greco, A.P. & Zabet, N.R. (2018) DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Research*, **46**, e114.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A. et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**, 1050–1054.

Clavijo, F., Diedhiou, I., Vaissayre, V., Brottier, L., Acolatse, J., Moukouanga, D. et al. (2015) The casuarina NIN gene is transcriptionally activated throughout Frankia root infection as well as in response to bacterial diffusible signals. *New Phytologist*, **208**, 887–903.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Davenport, H.E. (1960) Haemoglobin in the root nodules of Casuarina cunninghamiana. *Nature*, **186**(4725), 653–654.

Diédhiou, I., Tromas, A., Cissoko, M., Gray, K., Parizot, B., Crabos, A. et al. (2014) Identification of potential transcriptional regulators of actinorhizal symbioses in Casuarina glauca and Alnus glutinosa. *BMC Plant Biology*, **14**, 1–13.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C. et al. (2017) De novo assembly of the Aedes aegypti genome using hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S. et al. (2016) Juicer provides a one-click system for analyzing loop-resolution hi-C experiments. *Cell Systems*, **3**, 95–98.

Fan, C., Qiu, Z., Zeng, B., Li, X. & Xu, S. (2018) Physiological adaptation and gene expression analysis of Casuarina equisetifolia under salt stress. *Biologia Plantarum*, **62**, 489–500.

Ghodhbane-Gtari, F., Nouioui, I., Hezbri, K., Lundstedt, E., D'angelo, T., McNutt, Z. et al. (2019) The plant-growth-promoting actinobacteria of the genus nocardia induces root nodule formation in Casuarina glauca. *Antonie Van Leeuwenhoek*, **112**, 75–90.

Ghosh, M., Chezhian, P., Sumathi, R. & Yasodha, R. (2011) Development of SCAR marker in Casuarina equisetifolia for species authentication. *Trees*, **25**, 465–472.

Goel, V. & Behl, H. (2005) Growth and productivity assessment of Casuarina glauca Sieb. Ex. Spreng on sodic soil sites. *Bioresource Technology*, **96**, 1399–1404.

He, X., Critchley, C., Ng, H. & Bledsoe, C. (2005) Nodulated N2-fixing Casuarina cunninghamiana is the sink for net N transfer from non-N2-fixing Eucalyptus maculata via an ectomycorrhizal fungus Pisolithus sp. using 15NH4+ or 15NO3− supplied as ammonium nitrate. *New Phytologist*, **167**, 897–912.

Ho, K., Yang, J. & Hsiao, J. (2002) An assessment of genetic diversity and documentation of hybridization of casuarina grown in Taiwan using RAPD markers. *International Journal of Plant Sciences*, **163**, 831–836.

Hocher, V., Alloisio, N., Auguy, F., Fournier, P., Doumas, P., Pujic, P. et al. (2011) Transcriptomics of actinorhizal symbioses reveals homologs of the whole common symbiotic signaling cascade. *Plant Physiology*, **156**, 700–711.

Hocher, V., Auguy, F., Argout, X., Laplaze, L., Franche, C. & Bogusz, D. (2006) Expressed sequence-tag analysis in Casuarina glauca actinorhizal nodule and root. *New Phytologist*, **169**, 681–688.

Jarolímová, V. (1994) Chromosome counts of some Cuban angiosperms. *Folia Geobotanica*, **29**, 101–106.

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915.

Krueger, F. & Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.

Laplaze, L., Duhoux, E., Franche, C., Frutz, T., Svistoonoff, S., Bisseling, T. et al. (2000) Casuarina glauca prenodule cells display the same differentiation as the corresponding nodule cells. *Molecular Plant-Microbe Interactions*, **13**, 107–112.

Laplaze, L., Ribeiro, A., Franche, C., Duhoux, E., Auguy, F., Bogusz, D. et al. (2000) Characterization of a Casuarina glauca nodule-specific subtilisin-like protease gene, a homolog of Alnus glutinosa ag12. *Molecular Plant-Microbe Interactions*, **13**, 113–117.

Le, Q., Bogusz, D., Gherbi, H., Lappartient, A., Duhoux, E. & Franche, C. (1996) Agrobacterium tumefaciens gene transfer to Casuarina glauca, a tropical nitrogen-fixing tree. *Plant Science*, **118**, 57–69.

Lee, S.-I. & Kim, N.-S. (2014) Transposable elements and genome size variations in plants. *Genomics and Informatics*, **12**, 87–97.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H.-B., Li, N., Yang, S.-Z., Peng, H.-Z., Wang, L.-L., Wang, Y. et al. (2017) Transcriptomic analysis of Casuarina equisetifolia L. in responses to cold stress. *Tree Genetics & Genomes*, **13**, 1–15.

Li, N., Zheng, Y.-Q., Ding, H.-M., Li, H.-P., Peng, H.-Z., Jiang, B. et al. (2018) Development and validation of SSR markers based on transcriptome sequencing of Casuarina equisetifolia. *Trees*, **32**, 41–49.

Li, S.-F., Lv, C.-C., Lan, L.-N., Jiang, K.-L., Zhang, Y.-L., Li, N. et al. (2021) DNA methylation is involved in sexual differentiation and sex chromosome evolution in the dioecious plant garden asparagus. *Horticulture Research*, **8**, 198.

Liao, Y., Smyth, G.K. & Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A. et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Luechanimitchit, P. & Luangviriyasaeng, V. (1996) Study of sex ratio and relationship between growth and sex in Casuarina equisetifolia in Thailand. Recent Casuarina research and development. In: Pinyopusarerk, K., Turnbull, J.W. & Midgley, S.J. (Eds.) *CSIRO forestry and forest products*, Melbourne, Australia, pp. 30–32.

Michael, T.P. (2014) Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*, **13**, 308–317.

Mitton, J.B. (2002) Heterozygous advantage. In: *Encyclopedia of life sciences*. Chichester: John Wiley & Sons. Available from: https://doi.org/10.1038/npg.els.0001760; http://www.els.net/

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, **5**, 621–628.

Ndoye, A., Sadio, O. & Diouf, D. (2011) Genetic variation of Casuarina equisetifolia subsp equisetifolia and C. equisetifolia subsp incana populations on the northern coast of Senegal. *Genetics and Molecular Research*, **10**, 36–46.

Ngom, M., Gray, K., Diagne, N., Oshone, R., Fardoux, J., Gherbi, H. et al. (2016) Symbiotic performance of diverse Frankia strains on salt-stressed Casuarina glauca and Casuarina equisetifolia plants. *Frontiers in Plant Science*, **7**, 1331.

Obertello, M., SY, M.O., Laplaze, L., Santi, C., Svistoonoff, S., Auguy, F. et al. (2003) Actinorhizal nitrogen fixing nodules: infection process, molecular biology and genomics. *African Journal of Biotechnology*, **2**, 528–538.

Ou, S., Chen, J. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, **46**, e126.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R., Hellinga, A.J. et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, **20**, 1–18.

Péret, B., Swarup, R., Jansen, L., Devos, G., Auguy, F., Collin, M. et al. (2007) Auxin influx activity is associated with Frankia infection during actinorhizal nodule formation in Casuarina glauca. *Plant Physiology*, **144**, 1852–1862.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

Pinyopusarerk, K., Kalinganire, A., Williams, E. & Aken, K.M. (2004) Evaluation of international provenance trials of Casuarina equisetifolia. ACIAR Technical Reports. pp. 1–106.

Qayyum, Z., Noureen, F., Khan, M., Khan, M., Haider, G., Munir, F. et al. (2022) Identification and expression analysis of stilbene synthase genes in Arachis hypogaea in response to methyl Jasmonate and salicylic acid induction. *Plants*, **11**, 1776.

Rasmi, R., Arjunan, D., Arunachalam, S. & Paulraj, S. (2011) Analysis of genetic relationship in superior individuals of Casuarina equisetifolia L. using ISSR markers. *International Journal of Integrative Biology*, **11**, 73.

Roach, M.J., Schmidt, S.A. & Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 1–10.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Salmela, L. & Rivals, E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**, 3506–3514.

Santi, C., Svistoonoff, S., Constans, L., Auguy, F., Duhoux, E., Bogusz, D. et al. (2003) Choosing a reporter for gene expression studies in transgenic actinorhizal plants of the Casuarinaceae family. *Plant and Soil*, **254**, 229–237.

Schlub, R., Mersha, Z., Aime, C., Badilles, A., Cannon, P., Marx, B. et al. (2010) Guam ironwood (Casuarina equisetifolia) tree decline conference and follow-up. In: Zhong, C., Pinyopusarerk, K., Kalinganire, A. & Franche, C. (Eds.) *Proceedings of the 4th International Casuarina Workshop*. China: Haikou, pp. 239–246.

Scotti-Campos, P., Duro, N., da Costa, M., Pais, I.P., Rodrigues, A.P., Batista-Santos, P. et al. (2016) Antioxidative ability and membrane integrity in salt-induced responses of Casuarina glauca Sieber ex Spreng. in symbiosis with N2-fixing Frankia Thr or supplemented with mineral nitrogen. *Journal of Plant Physiology*, **196**, 60–69.

Sogo, A., Setoguchi, H., Noguchi, J., Jaffré, T. & Tobe, H. (2001) Molecular phylogeny of Casuarinaceae based on rbcL and matK gene sequences. *Journal of Plant Research*, **114**, 459–464.

Svistoonoff, S., Gherbi, H., Nambiar-Veetil, M., Zhong, C., Michalak, Z., Laplaze, L. et al. (2010) Contribution of transgenic Casuarinaceae to our knowledge of the actinorhizal symbioses. *Symbiosis*, **50**, 3–11.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. & Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tani, C. & Sasakawa, H. (2003) Salt tolerance of Casuarina equisetifolia and Frankia Ceq1 strain isolated from the root nodules of C. equisetifolia. *Soil Science and Plant Nutrition*, **49**, 215–222.

Tani, C. & Sasakawa, H. (2006) Proline accumulates in Casuarina equisetifolia seedlings under salt stress. *Soil Science and Plant Nutrition*, **52**, 21–25.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R. et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nature Protocols*, **7**, 562–578.

Van der Moezel, P., Walton, C., Pearce-Pinto, G. & Bell, D. (1989) Screening for salinity and waterlogging tolerance in five casuarina species. *Landscape and Urban Planning*, **17**, 331–337.

Vikashini, B., Shanthi, A. & Dasgupta, M.G. (2018) Identification and expression profiling of genes governing lignin biosynthesis in Casuarina equisetifolia L. *Genes*, **676**, 37–46.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. et al. (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Wang, Y., Zhang, J., Qiu, Z., Zeng, B., Zhang, Y., Wang, X. et al. (2021) Transcriptome and structure analysis in root of Casuarina equisetifolia under NaCl treatment. *PeerJ*, **9**, e12133.

Wang, Y., Zhang, Y., Fan, C., Wei, Y., Meng, J., Li, Z. et al. (2021) Genome-wide analysis of MYB transcription factors and their responses to salt stress in Casuarina equisetifolia. *BMC Plant Biology*, **21**, 1–17.

Wheeler, G.S., Taylor, G.S., Gaskin, J. & Purcell, M.F. (2011) Ecology and management of Sheoak (casuarina spp.), an invader of coastal Florida, USA. *Journal of Coastal Research*, **27**, 485–492.

Wilson, K. & Johnson, L. (1989) Casuarinaceae. *Flora of Australia*, **3**, 100–174.

Xu, X., Zhou, C., Zhang, Y., Zhang, W., Gan, X., Zhang, H. et al. (2018) A novel set of 223 EST-SSR markers in casuarina L. ex Adans.: polymorphisms, cross-species transferability, and utility for commercial clone genotyping. *Tree Genetics & Genomes*, **14**, 1–8.

Yasodha, R., Kathirvel, M., Sumathi, R., Gurumurthi, K., Archak, S. & Nagaraju, J. (2004) Genetic analyses of casuarinas using ISSR and FISSR markers. *Genetica*, **122**, 161–172.

Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H., Gao, W. et al. (2019) De novo genome assembly of the stress tolerant forest species Casuarina equisetifolia provides insight into secondary growth. *The Plant Journal*, **97**, 779–794.

Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, **16**, 284–287.

Yu, W., Zhang, Y., Xu, X., Zhong, C., Wei, Y., Meng, J. et al. (2020) Molecular markers reveal low genetic diversity in Casuarina equisetifolia clonal plantations in South China. *New Forests*, **51**, 689–703.

Yuan, Z. & Zhang, D. (2015) Roles of jasmonate signalling in plant inflorescence and flower development. *Current Opinion in Plant Biology*, **27**, 44–51.

Zhang, X., Chen, S., Shi, L., Gong, D., Zhang, S., Zhao, Q. et al. (2021) Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. *Nature Genetics*, **53**, 1250–1259.

Zhang, Y., Hu, P., Zhong, C., Wei, Y., Meng, J., Li, Z. et al. (2020) Analyses of genetic diversity, differentiation and geographic origin of natural provenances and land races of Casuarina equisetifolia based on EST-SSR markers. *Forests*, **11**, 432.

Zhang, Y., Zhong, C., Han, Q., Jiang, Q., Chen, Y., Chen, Z. et al. (2016) Reproductive biology and breeding system in Casuarina equisetifolia (Casuarinaceae)–implication for genetic improvement. *Australian Journal of Botany*, **64**, 120–128.

Zhong, C., Mansour, S., Nambiar-Veetil, M., Bogusz, D. & Franche, C. (2013) Casuarina glauca: a model tree for basic research in actinorhizal symbiosis. *Journal of Biosciences*, **38**, 815–823.

Zhong, C., Zhang, Y., Chen, Y., Jiang, Q., Chen, Z., Liang, J. et al. (2010) Casuarina research and applications in China. *Symbiosis*, **50**, 107–114.