the plant journal

SEB Society for Experimental Biology

The Plant Journal (2019) **97,** 779–794 doi: 10.1111/tpj.14159

RESOURCE

De novo genome assembly of the stress tolerant forest species *Casuarina equisetifolia* provides insight into secondary growth

Gongfu Ye[1,2,*,†], Hangxiao Zhang[3,†], Bihua Chen[1], Sen Nie[1], Hai Liu[4], Wei Gao[1], Huiyuan Wang[3], Yubang Gao[3] and Lianfeng Gu[3,*]

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

[2]*Fujian Casuarina Engineering Technology Research Center, Fuzhou, Fujian 350012 China,*

[3]*Basic Forestry and Proteomics Research Center, College of Forestry, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou 350002, China, and*

[4]*Fujian Forestry Investigations and Planning Institute, Fuzhou, Fujian 350003, China*

Received 2 August 2018; revised 7 November 2018; accepted 9 November 2018; published online 14 November 2018.

*For correspondence (e-mails yegongfu@126.com; lfgu@fafu.edu.cn).

†These authors contributed equally to this work.

SUMMARY

***Casuarina equisetifolia* (*C. equisetifolia*), a conifer-like angiosperm with resistance to typhoon and stress tolerance, is mainly cultivated in the coastal areas of Australasia. *C. equisetifolia*, making it a valuable model to study secondary growth associated genes and stress-tolerance traits. However, the genome sequence is unavailable and therefore wood-associated growth rate and stress resistance at the molecular level is largely unexplored. We therefore constructed a high-quality draft genome sequence of *C. equisetifolia* by a combination of Illumina second-generation sequencing reads and Pacific Biosciences single-molecule real-time (SMRT) long reads to advance the investigation of this species. Here, we report the genome assembly, which contains approximately 300 megabases (Mb) and scaffold size of N50 is 1.06 Mb. Additionally, gene annotation, assisted by a combination of prediction and RNA-seq data, generated 29 827 annotated protein-coding genes and 1983 non-coding genes, respectively. Furthermore, we found that the total number of repetitive sequences account for one-third of the genome assembly. Here we also construct the genome-wide map of DNA modification, such as two novel forms N^6-adenine (6mA) and N4-methylcytosine (4mC) at the level of single-nucleotide resolution using single-molecule real-time (SMRT) sequencing. Interestingly, we found that 17% of 6mA modification genes and 15% of 4mC modification genes also included alternative splicing events. Finally, we investigated cellulose, hemicellulose, and lignin-related genes, which were associated with secondary growth and contained different DNA modifications. The high-quality genome sequence and annotation of *C. equisetifolia* in this study provide a valuable resource to strengthen our understanding of the diverse traits of trees.**

Keywords: *Casuarina equisetifolia*, Pacific Biosciences single molecular real-time (SMRT) sequencing, Illumina sequencing, genome assembly, N^6-adenine, N^4-methylcytosine.

INTRODUCTION

The *Casuarina*ceae family has four genera (Sogo *et al.*, 2001), among which *Casuarina* is commercially cultivated along coastal areas as a windbreak shelterbelt (*Fagales, Casuarinaceae)* due to its outstanding performance against typhoons, desert, drought, and salinity (Tani and Sasakawa, 2003; Hu *et al.*, 2016). The subspecies *Casuarina equisetifolia* (*C. equisetifolia*) is an evergreen tree with remarkable resistance to typhoons, which typically occur in the tropical and subtropical coastlines of Southeast Asia, Australia, and the archipelagos in the Pacific Ocean Australia (Pinyopusarerk *et al.*, 2004; Wheeler *et al.*, 2011). *C. equisetifolia* has multiple uses, such as fuel wood, paper, and soil improvement (Zhong *et al.*, 2005, 2010). Its excellent wind resistance has enabled *C. equisetifolia*'s extensive introduction into other countries to stabilize moving sand. *C. equisetifolia* was introduced from Australia to China as early as 1897

and was mainly cultivated in tropical and subtropical zones (Yang *et al.*, 1995; Zhong *et al.*, 2010). The coastline *Casuarina* forest in Hainan is ~50 000 hm^2, and almost surrounds the island (Liu *et al.*, 2013). In total, the plantations of *C. equisetifolia* outstripped 300 000 hm^2 in southern China (Zhong and Bai, 1996; Zhong *et al.*, 2005). The cultivation of *C. equisetifolia* is mostly affected by temperature, and the Zhoushan islands of Zhejiang, where the average temperature is 17°C, is the farthest north region where it can be successfully cultivated in China (Chen *et al.*, 2005). Therefore low temperature is the major environmental factor limiting cultivation expansion, fast growth, and high yield in *C. equisetifolia* (Li *et al.*, 2017).

Previous studies on nitrogen fixation, tannins, chemicals, plantation ecosystems, and drought-and-salt resistance have been conducted due to the forestry importance of this species (Zhang *et al.*, 2009; Ogunwande *et al.*, 2011), although a whole genome of *Casuarina* at high resolution is still lacking. Therefore the research in *C. equisetifolia* is hampered by the lack of a draft genome. With the development of sequencing technologies, the combination of SMRT and second-generation sequencing is an ideal approach to assemble genomes (Au *et al.*, 2012; Koren *et al.*, 2012), since it can take full advantage high-throughput and long reads, respectively (Roberts *et al.*, 2013).

Here, we report the draft genome of *C. equisetifolia* with high-quality annotation analysis from a combination of PacBio SMRT sequencing and Illumina second-generation sequencing. The final genome assembly was about 300 Mb with an N50 scaffold size of 1.06 Mb. In total, there were 29 827 protein-coding genes, of which ~50% could be detected with fragments per kilobase of transcript per million (FPKM) higher than 3. Transposable elements (TE) consists of 33% of the genome. Comparing the genome of *C. equisetifolia* with other six tree species, with *A.thaliana* as an outgroup, we found out that *C. equisetifolia* and *J. regia* are two closely related species. In addition, we performed genome-wide analyses for DNA modification, such as N^6-methyldeoxyadenosine (6mA) and N4-methylcytosine (4mC). Finally, cellulose and lignin-related genes were investigated using the present genome assembly and transcriptome sequencing, which will be useful to investigate the fast growth of *C. equisetifolia*. Availability of *de novo* genome assembly of *C. equisetifolia* in this study will provide the foundation of genome-wide studies and contribute to investigations into potential genes involved in traits such as resistance to typhoons, drought, salt, and cold tolerance.

RESULTS

Morphological characteristics of *C. equisetifolia*

In China, the largest cultivated area of *C. equisetifolia* is the coastal region due to this species' wind resistance and wide adaptability. Therefore *C. equisetifolia* is well integrated into in the coastal region of the southeast coast, and is mainly planted in Zhejiang, Fujian, Guangdong, Guangxi, and Hainan provinces (Figure 1a). *C. equisetifolia* is an evergreen species and the branchlets or cylindrical cladodes of *C. equisetifolia* are very slender and internode ridges were surrounded with minute teeth-like reduced leaves (Figure 1b). Given that teeth-like leaves can reduce transpiration, the possibility that teeth-like leaves contributed to drought tolerant exists. Therefore there is this great adaptability, which makes an excellent species of windbreak and forest protection in the South China Sea.

Hybrid *de novo* genome assembly of *C. equis* ssp. *incana*

C. equis. ssp. *incana* is widely distributed and was therefore chosen to generate the reference genome sequence in this study (Zhong *et al.*, 2005, 2010). First, a genome survey with low depth Illumina short-read sequencing, about 76 gigabases (Gb) with 150 bp in length, was sequenced to estimate the genome size of *C. equis.* ssp. *incana.* The 17-mer frequency distribution analysis showed that the peak was at a depth of 211 (Figure 2a). The total number of k-mers was 63 453 274 679 in this subset. Therefore the genome size was estimated to be approximately 300 Mb using the following formula: the number of k-mers divided by the depth of peaks, which is consistent with the first genome size estimation reporting the genome size of *C. glauca* to be about 340 Mb (Schwencke *et al.*, 1998).

To construct high-quality genome assembly, we took full advantage of PacBio and second-generation sequencing since high-throughput data by Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA) can resolve the sequence depth and PacBio libraries are prepared to resolve the repetitive sequences in the genome and improve the assembly (Figure 2b). The first data set came from seven 2 × 150 bp Illumina HiSeq 2000 pair-end DNA libraries (Table S1) with different insert sizes (250 bp, 450 bp, 500 bp, 800 bp, 2 kb, 5 kb, and 10 kb) (Zhang *et al.*, 2014b), and single-molecule reads from PacBio sequencing libraries with a 20 kb insert size for *de novo* assembly of the *C. equis.* ssp. *incana* genome. In total, after deep whole genome short-read sequencing, these libraries produced about 168 Gb of data, after filtering out low-quality reads, representing 573-fold coverage of the *C. equis.* ssp. *incana* genome (Table S1). To improve genome assembly quality, the second data set was 4.2 Gb of initially filtered Pacific Bioscience SMRT long reads, representing a 14-fold coverage of the *C. equis.* ssp. *incana* genome.

First, we generated the initial *de novo* genome assembly using DISCOVAR *de novo* (Love *et al.*, 2016) and Falcon assembler (Chin *et al.*, 2016) for Illumina and PacBio RSII reads, respectively (Figure 2b). After combination assembly using SSPACE (Boetzer *et al.*, 2010), the size for final merged genome version of *C. equisetifolia* was about

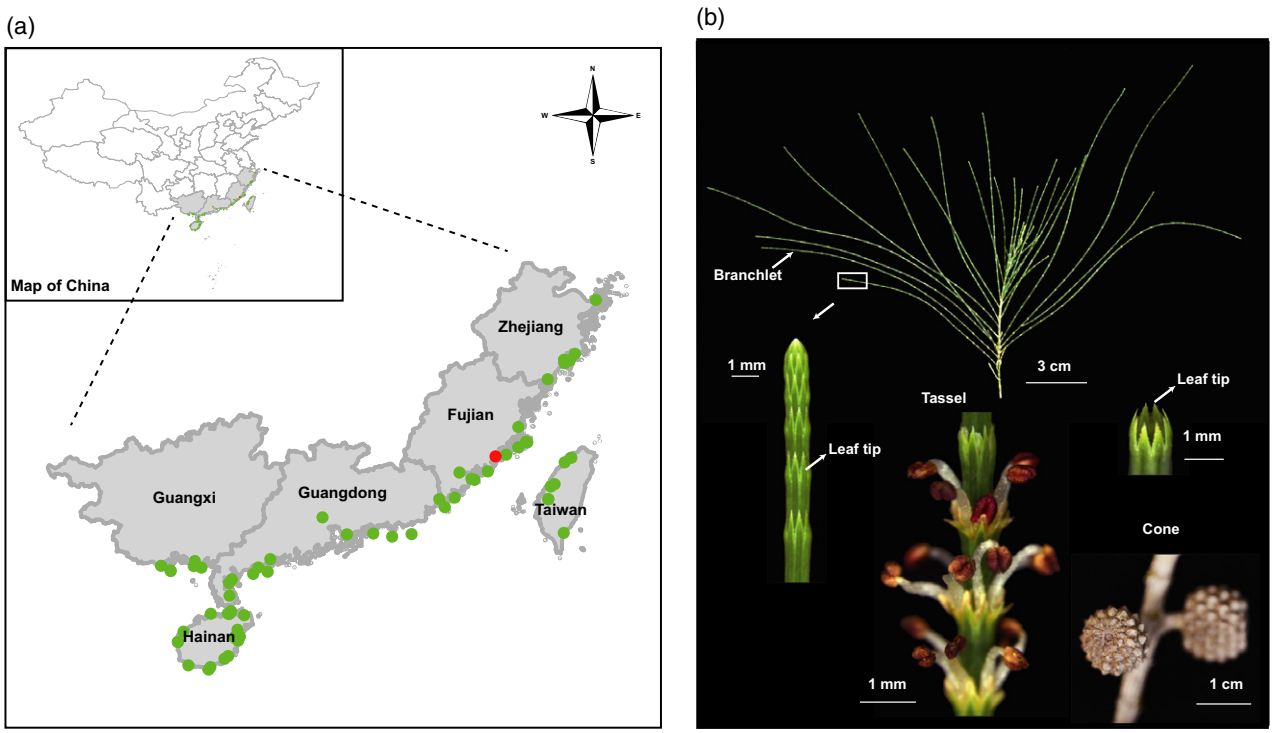

Figure 1. Morphological characteristics of *C. equisetifolia*. (a) The distribution of *C. equisetifolia* in the coastal region of south China. (b) Shoot silhouettes of tissues including branchlets, teeth-like leaves, and cones of *C. equisetifolia*.

301 Mb (Table S2). Contig N50 was about 464 kb, and 90% (in length) of the contigs fell into 718 contigs. The N50 for scaffold was 1.06 Mb and 90% (in length) of the assembly were contained in the 366 super-scaffolds, with the largest scaffold being 3.97 Mb (Table S2).

We then aligned the clean reads to the assembled sequence to obtain the base depth. With 10 kb as the window, the average depth and GC content of each window were calculated, and a GC depth scatter plot was made (Figure 2c). As can also be seen from the figure, since third-generation sequencing technology does not have an obvious GC bias, it can better cover some GC regions with higher or lower sequencing depth (i.e., the top and bottom of the ordinate), making the overall GC regional distribution more complete. GC analysis shows that the GC content of the sample is about 30–40% with no obvious bias (Figure 2c). The average depth of the scatter distribution is consistent with the depth of sequencing. We further compared the coverage of each base in the assembly sequence using the second-generation read. The overall single base depth was consistent with sequence depth (Figure 2d).

Assessment of the completeness of the draft genome assembly

In this study, we used Benchmarking Universal Single-Copy Orthologs (BUSCO) for the evaluation of assembly quality and completeness. About 95.1% of single-copy orthologs were included in our assembly, which confirms that the current version is relatively complete and accurate. To further evaluate the completeness of the assembled genome, we examined the proportion of a transcriptome data set from *C. equis.* ssp. *incana* and found that 94.1% of the reads could be mapped back to the genome of *C. equisetifolia* using TopHat (v2.0.11) (Trapnell *et al.*, 2009). Therefore, the assembled version of the genome represents the majority of the complete transcriptome, confirming its almost completeness.

Annotation of *C. equis.* ssp. *incana* genome assembly

Using a combination of *ab initio* prediction, homology-based search and coupled with transcriptome sequencing for identifying protein-coding genes to generate a global profile of transcriptome (Figure 2e), 29 827 protein-coding genes in the *C. equisetifolia* draft genome were acquired, with an average transcript length of 3256 bp and a mean of 4.68 exons per gene (Table S3). Functional annotation showed that 91.75% of protein-coding genes could be annotated by a publicly available protein database, e.g. InterPro, Swissprot, and TrEMBL, and appears to have known functions (Table S4). The percentage of protein-coding genes and the numbers of genes attributed to the Gene Ontology (GO) terms are shown in Figure 3(a). These genes included diverse functions, such as catalytic activity, binding, metabolic process, and response to stimulus and

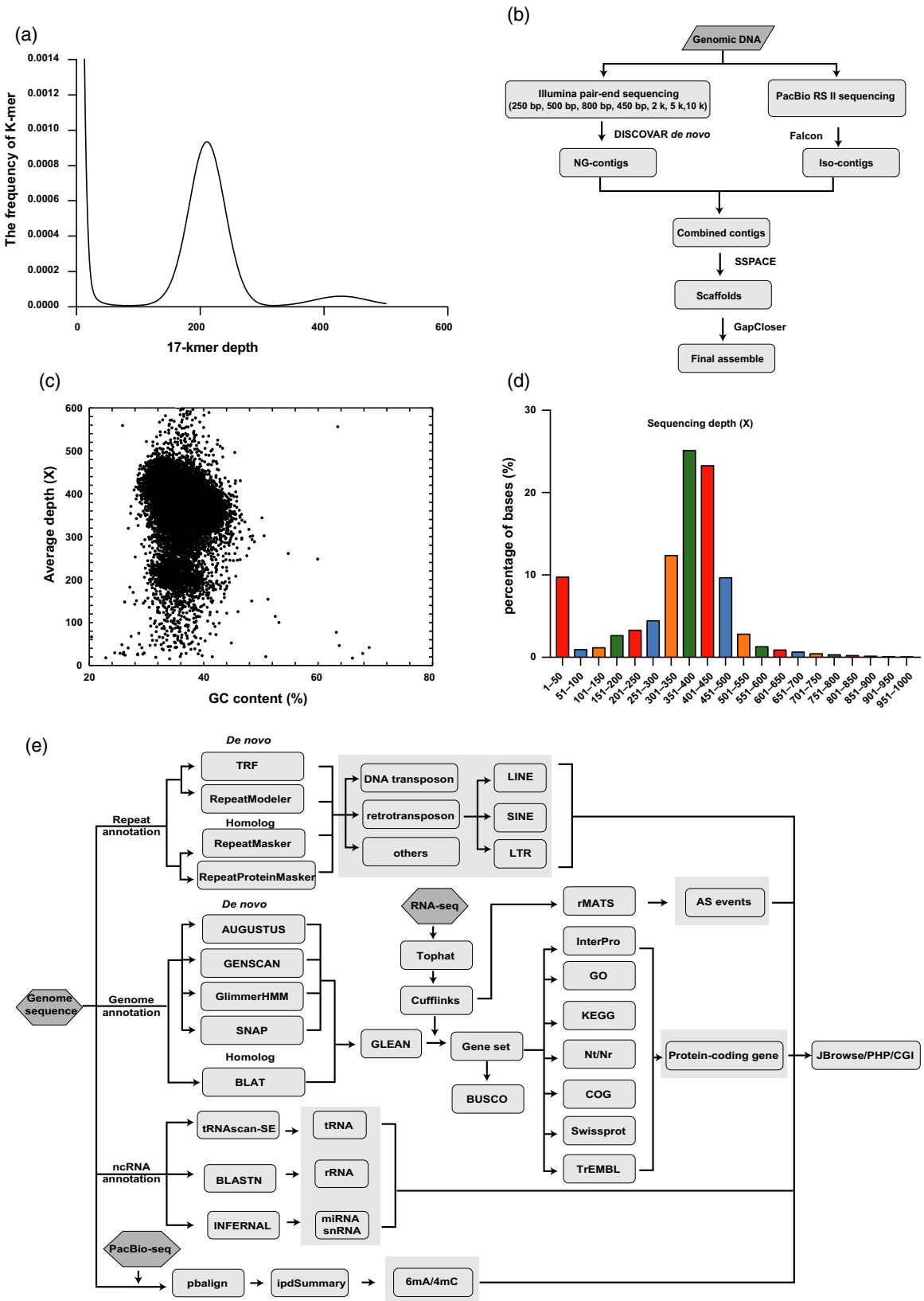

Figure 2. Genome survey and the pipeline of assembly and annotation. (a) Distribution frequency of the 17-kmer graph for genome size estimation. Density plot of the frequency of unique 17-kmer for each kmer depth (*x* axis) is plotted. (b) Flow chart of genome assembly using Illumina paired-end sequencing and the PacBio RSII platform. (c) Distribution of GC depth. (d) Histogram of single base depth chart. (e) Flow chart of the gene annotation pipeline.

others. With regard to non-coding RNA, 168 microRNAs (miRNAs), 595 tRNA, 365 rRNAs, and 855 small nuclear RNAs (snRNAs) were identified (Figure 3b; Table S5).

To produce annotation of transposable elements (TEs), we identified retrotransposons and DNA transposons in the *C. equisetifolia* genome. Long terminal repeats (LTRs), short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) are the most common TEs in the *C. equisetifolia* genome, and we annotated about 33% sequences as TEs (Table S6). Consistent with the percentage in other plant species (Liu *et al.*, 2018), the most abundant repeat elements were LTRs with the highest ratio of 23% of the total assembly, of which 35.6% and 42.8% of LTR is *Gypsy* and *Copia*, respectively. The ratios were ~2.32%, (DNA transposons), ~2.16% (LINEs), ~0.04% (SINEs), and ~0.0006 (other types) of the total assembly, respectively. Finally, 6.79% of the total assembly represented uncharacterized repeats, which were not assigned to any type of TE or repetitive sequence (Figure 3c).

Multiple isoforms from a single locus can be generated by alternative splicing (AS), which is a post-transcription mechanism to increase the fitness of plants (Liu *et al.*, 2017; Zhang *et al.*, 2017). At present, genome-wide AS identification has been investigated in several plant species, including *Arabidopsis thaliana* (Filichkin *et al.*, 2010), *Zea mays* (Wang *et al.*, 2016), *Phyllostachys edulis* (Wang *et al.*, 2017), *Salvia miltiorrhiza* (Xu *et al.*, 2015b), and *Sorghum bicolor* (Abdel-Ghany *et al.*, 2016). In this study, the transcriptome atlas of *C. equis.* ssp. *incana* was used to construct gene annotation. In addition, we also used this transcriptome atlas to identify AS in *C. equis.* ssp. *incana*. In total, the analysis of the transcriptome data identified 2524 AS genes, which included 3386 AS events (Table S7). GO enrichment analysis for these AS genes showed that the enrichment GO for biological process included regulation of transcription (5.81E-05), regulation of RNA metabolic process (6.82E-05), chromatin modification (8.56E-03) *et al.* (Figure 3d). Comparing with *C. cunninghamiana*, we also identified 566 differential splicing events. For example, CCG002044 was a B3 domain-containing gene, which affects transcription and included intron retention events in the fourth intron. Based on RNA-seq, we found that the isoforms with the retention intron in *C. equis.* ssp. *incana* was more abundant than that in *C. cunninghamiana* (Figure 3e). More work needs to be performed to determine if these differential splicing events contribute to phenotype variation.

In total, we identified 119 genes that were associated with a stress response (Table S8). Among these 119 stress response genes, the average number of exons for these genes was 4.8. Rapid AS responses have been reported to link to diverse stress responses (Calixto *et al.*, 2018; Sanyal *et al.*, 2018). In total, there were 12 stress response genes (10%) including AS events, that might contribute to the stress response.

Ortholog clustering and syntenic blocks analysis

Single-copy and multiple-copy orthologs were identified from eight sequenced genomes, including *J. regia*, *P. mume*, *P. bretschneideri*, and *Z. jujuba*, by cluster analysis of gene families using OrthoFinder (Figure 4a). The number of single-copy orthologs in *C. equisetifolia* is similar to that in *P. mume* and higher than other analyzed species. The number of multiple-copy orthologs in *C. equisetifolia* is smaller than other analyzed species (Figure 4a). Ortholog analysis revealed that *C. equisetifolia*, *P. bretschneideri*, *P. mume*, *Z. jujuba*, and *J. regia* share a core set of 11 331 gene families, and 225 gene families are *C. equisetifolia* specific (Figure 4b). Gene family expansion and contraction analysis identified 954 gene families that are expanded, while 4101 gene families have been lost in *C. equisetifolia* (Figure 4c). Phylogenetic reconstruction showed that *C. equisetifolia* was closer to *J. regia* and separated from the common ancestor from 31.6 million years ago (Figure 4d).

Within species syntenic block analysis in *C. equisetifolia* showed that there were 57 syntenic blocks and the average collinear genes for each block is nine. In total, there are 510 collinear genes in all blocks (Figure 4e). Intergenomic analyses between *C. equisetifolia* and *J. regia* genomes indicated that the two genomes possess 1234 conserved collinear blocks (with an average of 12 genes per collinear block) covering 14 589 collinear genes (Figure 4f).

Genome-wide landscape of 6mA and 4mC

In this study, we generated PacBio RSII reads using unamplified gDNA to investigate the genome-wide DNA modification landscape, using landmarks such as 6mA (Greer *et al.*, 2015; Zhang *et al.*, 2015; Liang *et al.*, 2018) and 4mC (Iyer *et al.*, 2011). In total, we found 6956 and 9517 sites for 6mA and 4mC, respectively (Table S9). The consensus sequences including the upstream and downstream 4-bp of 6mA and 4mC sites are shown in Figure 5(a and b), respectively. Among these, there were about 25% and 21% sites within the gene regions for 6mA and 4mC, respectively. Multiple sites of DNA modification in one gene were also revealed (Figure 5c). For example, CCG000741, annotated as a transcriptional repressor complex, was found to include four 6mA and one m4C site, and two of which existed in the exon region (Figure 5d), implying potential regulatory mechanisms about this gene. The most two significantly enriched motifs for 6mA were AMBGA and ARGYA (Figure 5e). Additionally, the most two significant motifs of 4mC were CTCGTTK and TGCMGR (Figure 5f). The enrichment GO for 6mA-methylated genes was an assembly of spliceosomal tri-snRNP, which included CCG0 17037 (pre-mRNA-processing factor 6-like), CCG017960 (U4 U6 small nuclear ribonucleo Prp31), CCG014580 (pre-mRNA-processing factor 6-like), and CCG028114 (pre-mRNA-processing-splicing factor 8-like). The enrichment

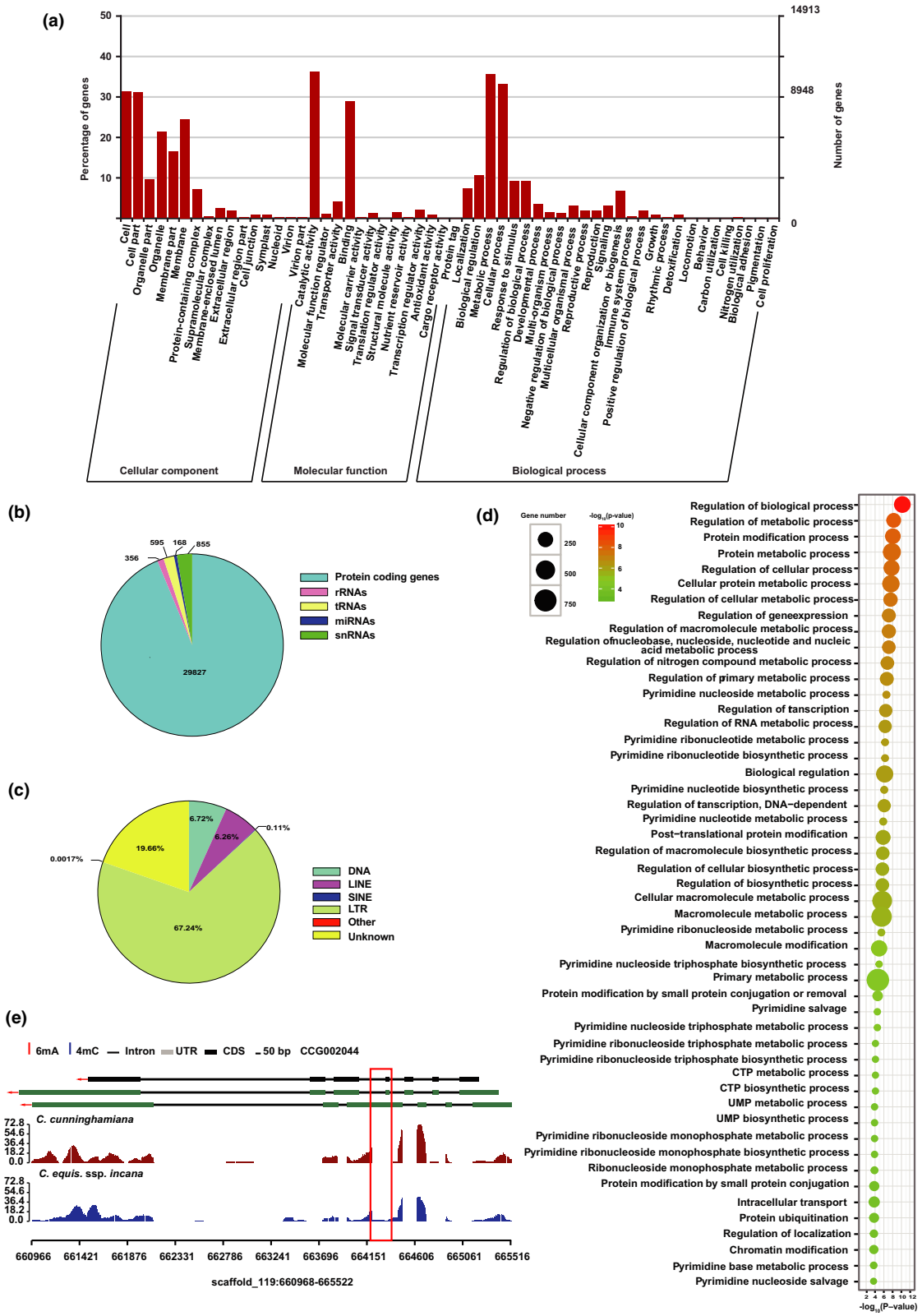

Figure 3. Gene annotation of *C. equis.* ssp. *incana*. (a) Distribution of Gene Ontology (GO) terms including the percentage of the protein-coding genes and the numbers of genes attributed to the GO terms. (b) Pie chart shows the percentage of protein-coding genes and non-coding RNA, including miRNAs, tRNA, rRNAs, and snRNAs. (c) The ratios of different types of TE. (d) GO enrichment analysis for all AS genes. (e) Wiggle plot shows the differential splicing event of CCG002044.

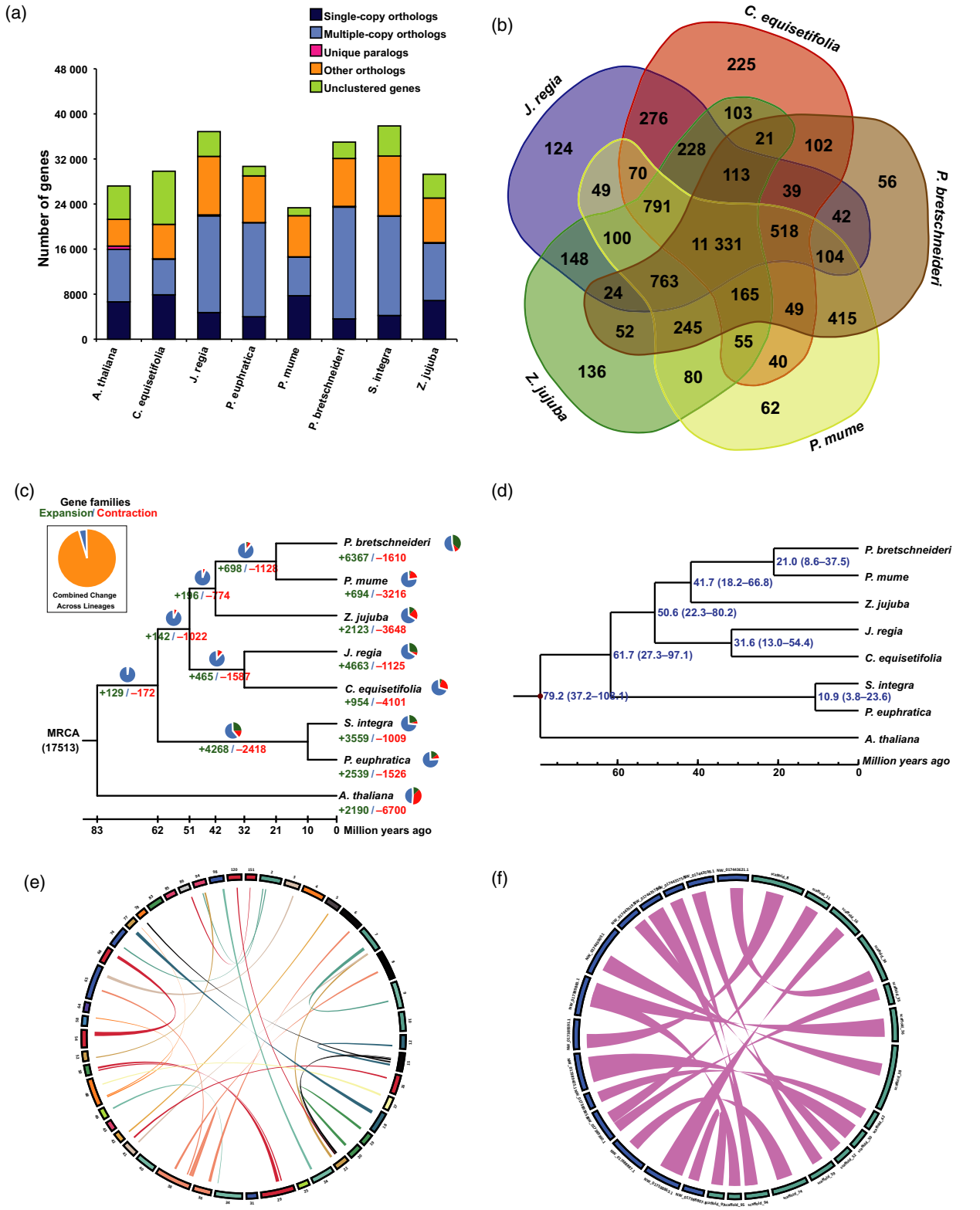

Figure 4. Analysis of phylogenetic relationships, gene family and syntenic blocks. (a) Number of paralogous genes in *A. thaliana*, *C. equisetifolia*, *J. regia*, *P. euphratica*, *P. mume*, *P. bretschneideri*, *S. integra*, *and Z. jujuba*. (b) Venn diagram comparison of gene families in the four analyzed plants: *C. equisetifolia*, *J. regia*, *P. mume*, *P. bretschneideri*, and *Z. jujuba*. (c) Estimation of phylogenetic relationships and the number of gene families presenting expansion and contraction. (d) Phylogenetic tree of seven tree species with *A. thaliana* presented as an outgroup. The number indicates the evolution divergence time. (e) Circos plot showing paralogous gene pairs in *C. equisetifolia*. (f) Circos plot showing synteny blocks between *C. equisetifolia* and *J. regia*.

GO for 4mC-methylated genes was energy coupled proton transport, oxidative phosphorylation, and ATP biosynthetic processes (Figure 5g). It is worth noting that genes with 6mA (17%) and 4mC (15%) showed an overlap with genes including AS events (Figure 5h). The expression levels for genes including 6mA and 4mC modifications were higher than the total expression level (Figure 5i). A previous study showed that an epigenetic mechanism can affect biotic and abiotic stresses (Annacondia *et al.*, 2018). We revealed three genes (CCG009774, CCG009840 and CCG027382) and one stress response gene (CCG002621) including modification of 6mA and 4mC, respectively. It will be interesting to determine if stresses can induce the epigenetic change in these stress response genes.

Morphological characteristics of different subspecies and chlorophyll synthesizing genes

At present, *C. equisetifolia* subsp. *incana* (*C. equis* ssp. *incana*) with three other different subspecies: *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca*, were four major variations with wide distribution. Internode lengths of these subspecies ranged from 4.1 mm to 5.3 mm. The cladode diameters were 0.7, 0.6, 0.55, and 0.9 mm for *C. equis* ssp. *incana*, *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca*, respectively (Figure 6a). In *C. equis.* subsp. *incana*, the numbers of whorled leaves were different among natural variations, which usually were in whorls of 6–9 per node, and the number of whorled leaves of *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca* were 6–8, 6–10, and 10–17, respectively (Figure 6a). The strong reduced leaf will reduce photosynthetic area, which will modify the light harvest efficiency together with chlorophyll content (Kuusk *et al.*, 2017). Leaf sheaths were fused to each other and the shoot axis was completely surrounded by the extended leaf sheath, which is the major site for photosynthesis (Figure 6a). The mean concentration of chlorophyll content of branchlet varied from 1.2 to 3.0 mg g^{-1} among subspecies, *C. cunninghamiana* was highest and *C. glauca* was lowest (Figure 6b). Scanning electron microscopy showed that internode ridges in cylindrical cladodes were separated by furrows (Figure 6c). Furthermore, the leaf tips of *C. glauca* covered the internode ridges rather than the furrows, as in three other subspecies (Figure 6c). It will be interesting to investigate the mechanism underlying the differences. Leaf trichomes were widely distributed in *C. equis.* ssp. *incana* and *C. equis.* ssp. *equisetifolia*, while they decreased greatly in *C. cunninghamiana* and *C. glauca* (Figure 6c). In total, we identified 44 chlorophyll-related genes, including genes associated with chlorophyll catabolic process, chlorophyll biosynthetic process, and chlorophyllase activity (Table S10). Three genes (CCG009847, CCG010379, and CCG000523) included AS events, and two genes (CC G028520, CCG005040) had 6mA modifications.

Anatomical characterization and identification of genes related to the cell wall

Casuarina equisetifolia presents fast early growth rates (about 2–3 m per annum in height) in cultivation (BRIEF, 2006), and cell wall is the major determinant of plant growth and development (Farrokhi *et al.*, 2006). The fifth internode of culms was obtained from branchlet, and lignified cell walls were stained (Figure 7a). Branchlet sections showed that cladodes of *C. equisetifolia* surrounded by mesophyll, and pith, sclerenchyma, phloem fibers, xylem, and partly parenchyma below the mesophyll were lignified (Figure 7a). Two mesophyll cells, surrounded by sclerenchyma, formed one ridge, which was separated by furrows hiding the stomata. Outside the cambial ring, there were phloem and phloem fibers, while inside there were xylem and pith, which mad up the stele of *C. equisetifolia*. Among *C. equis.* ssp. *incana* and three other subspecies: *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca*, the number of mesophyll cells was different and corresponded with the number of whorled leaves (Figure 7a). Lignin content increases gradually from *C. equis.* ssp. *incana* to *C. glauca* (Figure 7b), which might result in different growth rates. The fibre of *C. equisetifolia* is non-septate under a scanning electron microscope, the mean fiber length of sample trees varied from 676 to 850 μm, and the width varied from 17 to 20 μm (Figure 7c). Both fiber length and width gradually decreased from the base to the top (Figure 7c). The ratio of fiber length to width is reported to be strongly associated with the anti-typhoon performance (Xu *et al.*, 2015a). *C. equis.* ssp. *incana*, *C. equis.* ssp. *equisetifolia*, and *C. glauca* have relatively higher ratios of fiber length to width, with the average ratios of 40, 42, and 40, respectively (Figure 7c). *C. cunninghamiana* is relatively lower with a value of 35, which may relate to their various wind resistance.

The assembled genome and gene annotation made it possible to identify genes involved in cell wall biosynthesis (such as cellulose synthase) and secondary cell wall (such as lignin). Using the related genes from *Populus* as reference sequences, we identified 23 lignin-, 8 cellulose-, and 12 hemicellulose-related genes from *C. equis.* ssp. *incana* genome assembly, respectively (Figure 7d). Among the four subspecies, *C. equis.* ssp. *incana* and *C. cunninghamiana* have distinct features. *C. equis.* ssp. *incana* has a relatively higher growth rate, more excellent wind resistance (Van der Moezel *et al.*, 1989), and its height can reach 6–12 m (Ndoye *et al.*, 2011). *C. cunninghamiana* is highest at 20–35 m (BRIEF, 2006) and enormously resistant to wind, which makes it the most important timber in the coastal areas (Zhong *et al.*, 2010). So we subsequently calculated the expression levels of genes involved in cell wall and secondary cell wall biosynthesis in *C. equis.* ssp. *incana* and *C. cunninghamiana* and revealed lignin-, cellulose-, and hemicellulose-related genes presented diverse

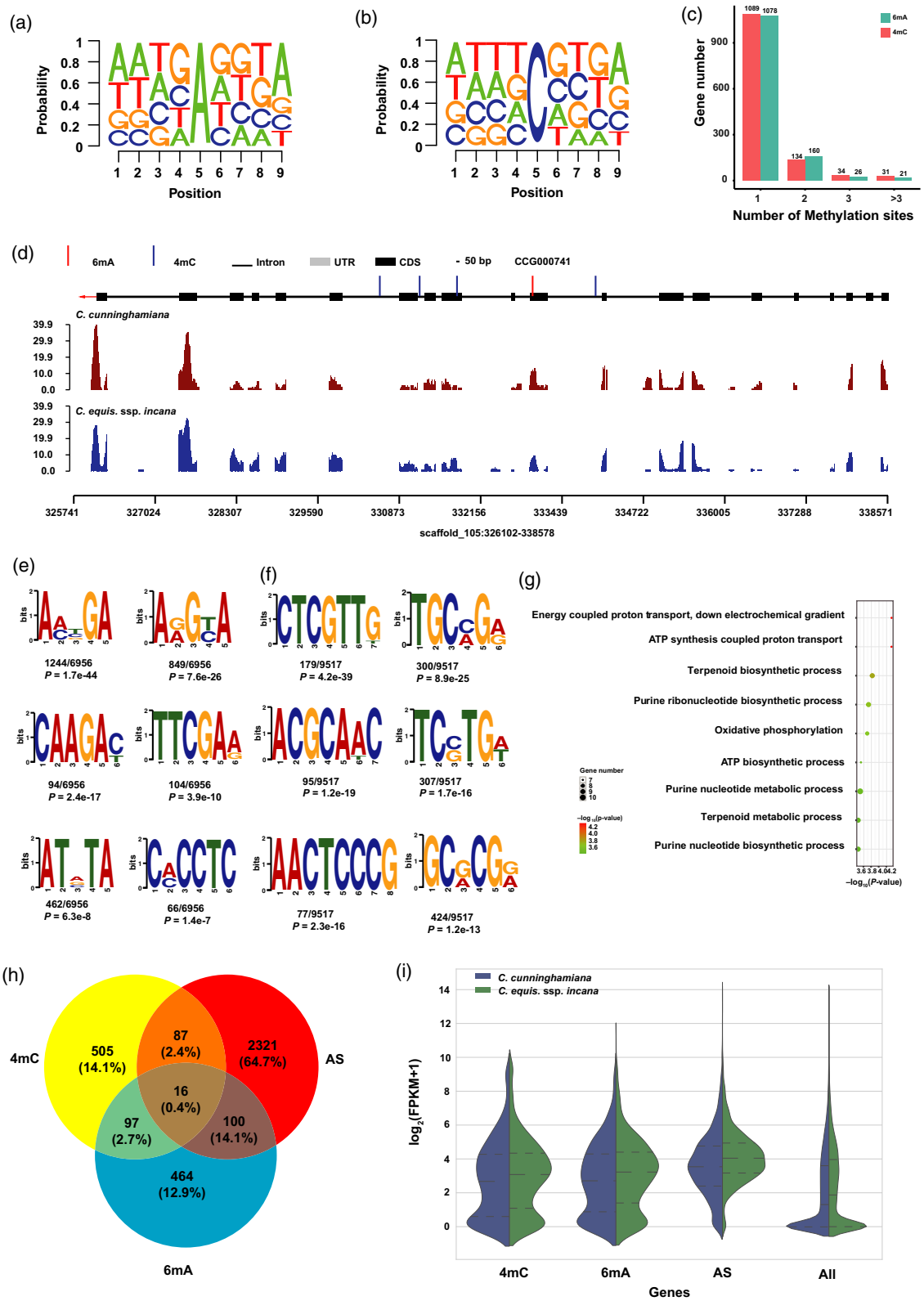

Figure 5. Genome-wide identification of DNA modifications (a) Sequence logo represents the motif containing all 6mA sites. (b) Sequence logo represents the motif containing all 4mC sites. (c) Counts of genes possessing different quantities of 6mA and 4mC sites. (d) Study of the 6mA and 4mC sites in CCG000741. (e) Consensus motif for 6mA sites. (f) Consensus motif for 4mC sites. (G) GO enrichment analysis for 4mC modification genes. (h) Overlapping genes with 6mA, 4mC, and AS. (i) Expression level for genes with 6mA/4mC modification and AS. 'All' indicates the expression levels for the total genes.

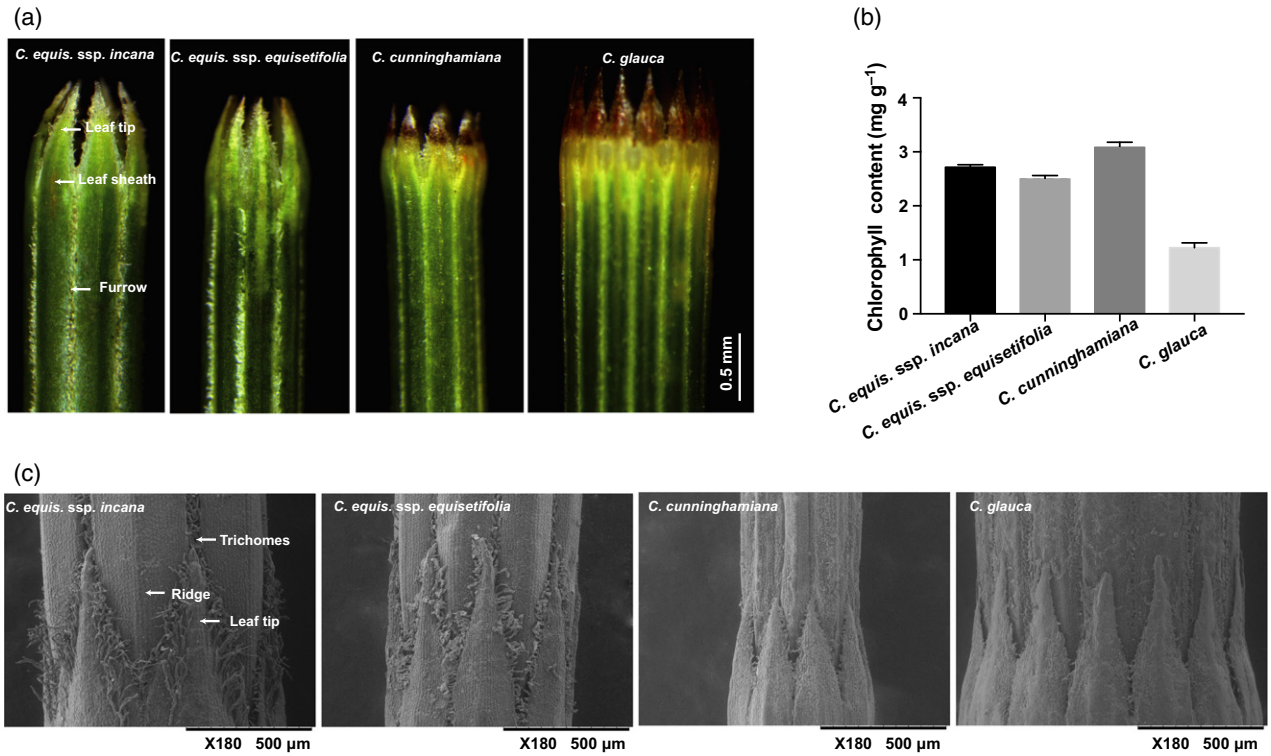

Figure 6. Morphological characteristics of four subspecies of *C. equisetifolia*. (a) The internode under microscope shows the teeth-like leaf tips, leaf sheath, and furrow in the different subspecies. (b) Concentration of chlorophyll in *C. equis.* ssp. *incana*, *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca*. (C) Scanning electron microscope observations of degenerate leaves and branchlets in *C. equis.* ssp. *incana*, *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca*. The whorled leaves are inserted in the node, the ridge is running along the internode, and trichomes are present in grooves.

expression profiles, which are shown in Figure 6(d). Most lignin- and cellulose-related genes showed much higher expressional level in *C. equis.* ssp. *incana*, which was consistent with the higher growth rate. In this study, we found out one caffeic acid *O*-methyltransferase (CCG020411.1) involved in the lignin biosynthetic process, which included AS events (intron retention) (Figure 7e). Besides, we also revealed that lignin-related genes included DNA modification, such as 6mA and 4mC. A previous study showed that over-expression of *KNAT1* represses lignin deposition (Sunaryo and Fischer, 2009). We found 6mA modification in the intron of CCG017382 (*KNAT1*) and exons of CCG 009869 (beta-1,4-xylosyltransferase IRX10L) (Figure 7f). The expression levels of *KNAT1* and CCG009869 in *C. cunninghamiana* were higher than those in *C. equis.* ssp. *incana* (Figure 7f). MYB46-like transcription factors mediate xylem-specific transcription by MYB46-like transcription factors in *Populus* (Winzell *et al.*, 2010). We found 4mC modifications in the exons of CCG008448 (*MYB46*) and CCG015969 (cellulose synthase) (Figure 7g). The expression levels of *MYB46* and cellulose synthase in *C. cunninghamiana* were lower than those in *C. equis.* ssp. *incana* (Figure 7g). It will be of interest to investigate the regulation of 6mA and 4mC in the biosynthesis of the cell wall and secondary cell wall.

DISCUSSION

Since the first forest tree genome, *Populus trichocarpa*, was sequenced in 2006 (Tuskan *et al.*, 2006); several other forest tree species have been sequenced including the fast-growing timber *Phyllostachys heterocycla* (Peng *et al.*, 2013), *Eucalyptus grandis* (Myburg *et al.*, 2014), *Picea abies* (Nystedt *et al.*, 2013), the living fossil *Ginkgo biloba* (Guan *et al.*, 2016), *Juglans regia* (Martínez-García *et al.*, 2016), *Populus euphratica* (Ma *et al.*, 2013), *Populus pruinosa* (Yang *et al.*, 2017), and the shrub willow *S. suchowensis* (Dai *et al.*, 2014). These forest tree genomic studies improved our understanding of the tree development, adaptation, and even the molecular basis of the distinct species. However, genome resources for forest trees are still limited due to the complexity of tree genomes, and only few forest tree species have the complete genome sequenced to date. For example, assembly of tree genomes such as *Picea abies* (Nystedt *et al.*, 2013) and loblolly pine (Neale *et al.*, 2014) is a challenging task using short reads due to high repetitiveness.

In this study, we revealed that the total number of TEs in *C. equis.* ssp. *incana* only account for 33% genome sequences. At the same time, we coupled the long reads sequencing technologies with highly accurate short

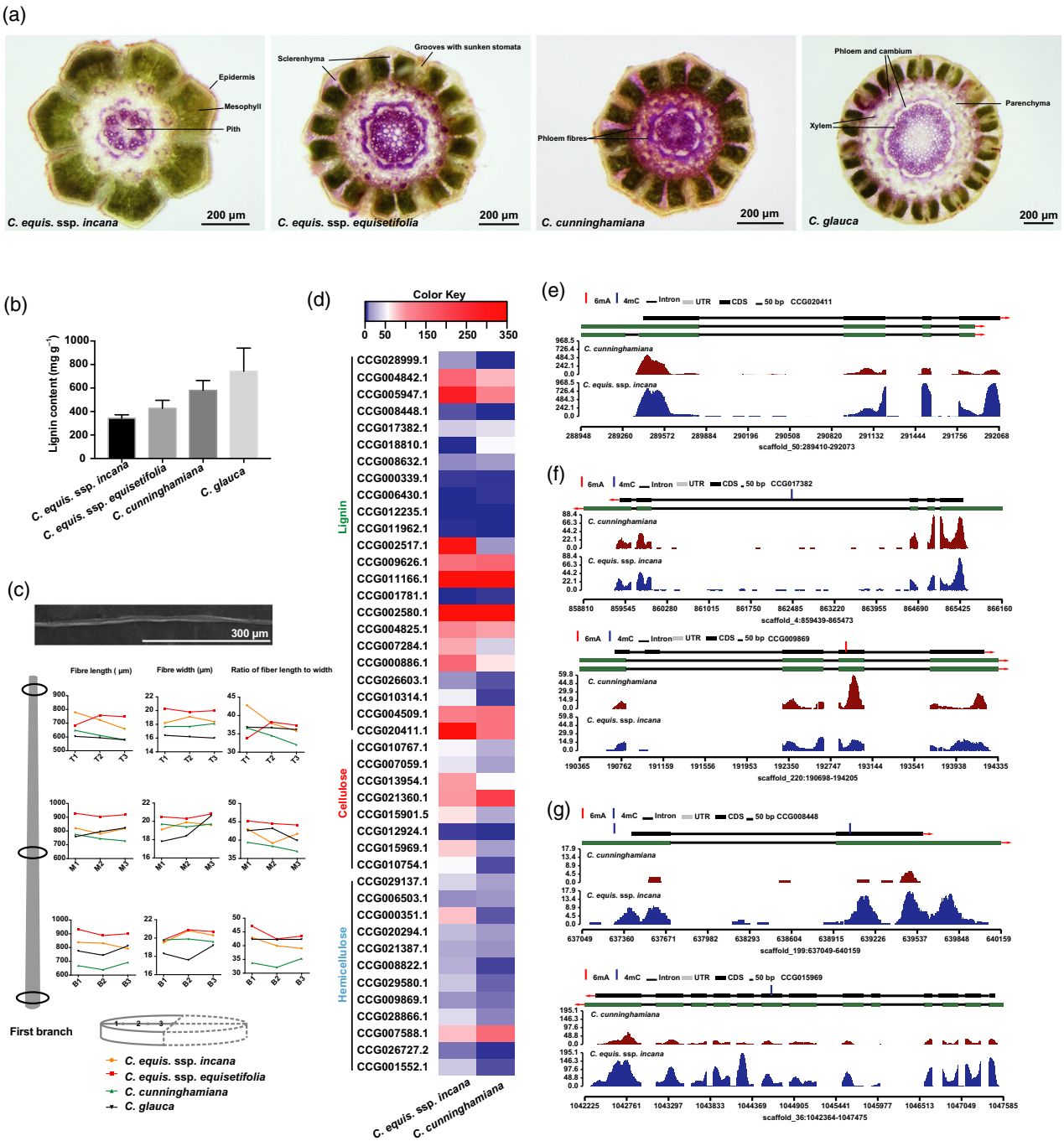

Figure 7. Anatomy of *C. equisetifolia* and investigation of fast-growth related genes. (a) Free-hand transverse sections of the fifth internode and lignified tissues were stained with phloroglucinol. Lignified tissues mainly include the pith, sclerenchyma, phloem fibres, xylem, and partly parenchyma below the mesophyll. (b) Lignin content of branchlets of four subspecies. (c) The fiber of *C. equisetifolia* was observed under a scanning electron microscope, the length and width of fibers was analyzed by three sections of the branch, and the corresponding ratio of fiber length to width was calculated. B stands for base, M for medial and T for top of branch, respectively. Numbers 1 to 3 indicate disks from pith to bark, separately. (d) Heatmap of fast-growth related genes mainly including lignin-, cellulose-, and hemicellulose-related genes. (e–g) Wiggle plot of fast-growth related genes including AS sites or 6mA/4mC modifications.

sequencing, which enabled a good assembly of *C. equis.* ssp. *incana.* To take full advantage of both the short reads from Illumina and long reads from PacBio technology, several tree genomes were assembled using a combination of Illumina and PacBio reads to produce a high-quality genome assembly. For example, a chromosome-scale of the apple reference genome has been reported following hybrid assembly, consisting of 2150 contigs with an N50 of

620 kb (Daccord *et al.*, 2017). The desert poplar showed high tolerance to a saline environment (Ma *et al.*, 2013). The assembly of *Populus alba* consisted of 464M of the genome including 37 901 protein-coding genes using 320× Illumina data and 30× PacBio RS raw data, which was highly collinear with that of *P. trichocarpa* (Ma *et al.*, 2018). The contig and scaffold N50 size of this assembly reach to 26.5 Kb and 459 Kb, respectively (Ma *et al.*, 2018). In this study the N50 contig for Illumina assembly with the reads from 250-bp, 500-bp, and 800-bp insertion libraries was 92 kb, and PacBio read only assembly was 41 kb, respectively. The combined assembly resulted in contig N50 of 464 kb, which showed greater improvement than Illumina or PacBio only assembly. In a previous study, ~×96 and ~×70 coverage from PacBio long reads was obtained to result in contig N50 of 5.81 Mb and 5.36 Mb for zebra finch and hummingbird, respectively (Korlach *et al.*, 2017). However, we only sequenced ~×14 coverage using PacBio-based long reads for the hybrid assembly. Therefore, higher coverage by PacBio sequencing still has potential improvements in contig N50 for the present genome. The hybrid assembly using both PacBio data and Illumina reads generated 301 Mb with an N50 scaffold of 1.06 Mb in *C. equis.* ssp. *incana*. In the future, chromosome conformation capture sequencing (Hi-C) will be a valuable method to further improve the assembly contiguity (Lieberman-Aiden *et al.*, 2009; Dudchenko *et al.*, 2017). Despite the need for chromosome-scale assembly, the present version covers approximately 94% of the RNA-seq reads, which could be mapped to the genome of *C. equis.* ssp. *incana*. Therefore, most of the gene regions are included in the present version, and the genome assembly is mostly complete, therefore providing a solid foundation and great value to the *C. equisetifolia* research community for future genomics studies.

Though *C. equis.* ssp. *incana* was selected as a reference sequence, three other subspecies also have wide distribution. Scanning electron microscopy revealed that the morphology of *C. equis.* ssp. *incana*, *C. equis.* ssp. *equisetifolia*, *C. cunninghamiana*, and *C. glauca* differs greatly (Figure 6c). Especially, leaf trichomes were widely distributed in *C. equis.* ssp. *incana* and *C. equis.* ssp. *equisetifolia*, while leaf trichomes decreased greatly in *C. cunninghamiana* and *C. glauca* (Figure 6c). The deep longitudinal furrows usually had invisible stomata, therefore the trichomes covered in deep longitudinal furrows may affect transpiration in subspecies, which included different leaf trichomes. As the draft reference genome of *C. equis.* ssp. *incana* has been completed, this genome resource will be useful for re-sequencing several subspecies to find potential single-nucleotide polymorphism (SNP), which may contribute to unravelling the mechanism underlying the phenotype and the stress-tolerance diversity of *C. equisetifolia* (Figure 6c).

Casuarina equisetifolia possesses an outstanding resistance to wind and tolerance to salt, making it an ideal model species to study abiotic and biotic stresses. For example, *C. equisetifolia* can withstand up to 500 mM NaCl (Tani and Sasakawa, 2003). Therefore, it will be very interesting to investigate the mechanism of salt tolerance with the aim of translating it to other species. However, in this study, we only sequenced total RNA of branchlet in the field under normal conditions. In the future, it will be interesting to investigate the RNA profile upon stress treatment. In addition, PacBio Iso-seq and nanopores RNA direct sequencing was proven to provide higher value information about full-length isoforms resulting from AS and alternative polyadenylation sites (APA) (Wang *et al.*, 2017; Garalde *et al.*, 2018). In particular, nanopores RNA direct sequencing could detect modified bases, such as RNA N^6-methyladenosine (m^6A) and 5-methylcytosine (5-mC) (Garalde *et al.*, 2018). Therefore, more powerful technologies such as Iso-seq or nanopores RNA direct sequencing should be used in the future to provide a comprehensive characterization of the total RNA profile under stress treatment in *C. equisetifolia*.

DNA modifications, such as 6mA, were discovered in *Chlamydomonas* (Fu *et al.*, 2015), *Escherichia coli* (Fang *et al.*, 2012), *Arabidopsis* (Liang *et al.*, 2018), *C. elegans* (Greer *et al.*, 2015) and *Drosophila* (Zhang *et al.*, 2015). Previous studies showed that PacBio SMRT sequencing works well to detect DNA modifications (Fang *et al.*, 2012; Liang *et al.*, 2018; Zhu *et al.*, 2018). So the assembly genome and PacBio long reads in this study provided unprecedented insights into the DNA modifications in *C. equis.* ssp. *incana*. Here, we presented a genome-wide landscape of DNA modification and identified 6956 and 9517 sites for 6mA and 4mC, respectively. In Arabidopsis, there are 29 811 adenines including 6mA in total (Liang *et al.*, 2018). Given the relatively low coverage of PacBio data of *C. equis.* ssp. *incana*, the real number of DNA modifications may be underestimated. Interestingly, we found that genes involved in the regulation of lignin and cellulose deposition contained 6mA and 4mC modifications (Figure 7f). At the same time expression for these genes showed differences between *C. equis.* ssp. *incana* and *C. cunninghamiana*. In the future, it will be interesting to investigate if the change in expression levels is related to the change in 6mA and 4mC between *C. equis.* ssp. *incana* and *C. cunninghamiana*.

EXPERIMENTAL PROCEDURES

Genomic DNA extraction

Genomic DNA was obtained from branchlets of *C. equis.* subsp. *incana*, which were collected from CHIHU state-owned protection forest farm of Fujian province (24°35′N, 118°55′E). All samples were immediately frozen in liquid nitrogen for DNA extraction using the cetyltrimethylammonium bromide (CTAB) method. The branchlets of *C. equis.* ssp. *incana* and *C. cunninghamiana* were

sampled and snap frozen in liquid nitrogen for RNA extraction. RNAprep Pure Plant Kit (Tiangen, Cat. #DP441, China) was used to isolate total RNA. The integrity of RNA samples was evaluated using Agilent 2100 Bioanalyzer and samples with RNA Integrity Number (RIN) values higher than 8 were used for downstream RNA-seq library construction.

Library construction and sequencing for genome assembly

Seven paired-end Illumina whole-genome sequencing (WGS) libraries with different insert sizes (250 bp, 450 bp, 500 bp, 800 bp, 2 kb, 5 kb, and 10 kb) were constructed according to the published standard protocol (Zhang *et al.*, 2014a) and sequenced with Illumina HiSeq 2000 sequencers in 1gene company. PacBio sequencing libraries with an insert size of 20 kb were constructed according to the manufacturer's recommendations. Sequencing was performed using a PacBio RSII Sequencer with P6/C4 chemistry system.

De novo genome assembly

The 17-kmer depth analysis was analyzed using clean reads from the Illumina platforms with Jellyfish (2.1.4) using the options '-m 17 -s 4294967296 -t 21 -c 8 –C' (Marcais and Kingsford, 2011). To obtain a more accurate assembly, sequencing data from each platform were firstly assembled separately, and then the two assemblies were subsequently combined to generate a complete genome (Figure 2e). Illumina data were assembled with DISCOVAR *de novo* (v52488) with option: MAX_MEM_GB=800 NUM_THREADS=5 (Love *et al.*, 2016). Falcon (v1.8.2) (Chin *et al.*, 2016) was used to the PacBio reads after adjust with the options as: length_cutoff = 6000; length_cutoff_pr = 3500; pa_HPCdaligner_option = -v -B128 -t16 -e.70 -l1000 -s1000; pa_DBsplit_option = -x500 -s200; falcon_sense_option = –output_multi –min_idt 0.70 –min_cov 4 –max_n_read 200 –n_core; overlap_filtering_setting = –max_diff 100 –max_cov 100 –min_cov 2 –bestn 10 –n_core 8, and then merged these results with the software named SSPACE (v3.0) with the parameter '-T 8' (Boetzer *et al.*, 2010), which was used to build a scaffold and GapCloser was for closed gap to obtain the final assembly with the thread number 16.

Gene and repetitive sequence annotation

We combined homolog homology, *de novo* prediction and RNA-seq data to predict protein-coding genes. Then, using the GLEAN software (Elsik *et al.*, 2007), gene sets predicted by various methods were integrated into a non-redundant and more complete gene annotation (Figure 2e). Gene function annotation included several protein databases (SwissProt, TrEMBL, KEGG, Nt, Nr, InterPro, and BLAST2GO) to annotate the proteins in the gene set and obtain the functional information.

During annotation of non-coding RNAs, the tRNAscan-SE software (Lowe and Eddy, 1997) was used to search for tRNA sequences in the genome based on the structural characteristics. As rRNAs are highly conserved, rRNA sequences were identified using BLASTN by searching for reference sequences with related species. The miRNA and snRNA sequence information of the genome were annotated using INFERNAL software (Nawrocki and Eddy, 2013).

For repetitive sequence annotation, TRF (Tandem Repeats Finder, v4.04) (Benson, 1999) with the options: Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, MaxPerid = 2000, and RepeatModeler (v1.04) (Chen, 2004) were adopted to construct a *de novo* repetitive sequence features library on the basis of *C. equis.* ssp. *incana* genome sequences. Then RepeatMasker (v3.2.9) (Tempel, 2012) and RepeatProteinMask (v3.2.2) (Chen, 2004) were used to classify the types of repetitive sequences.

Evolution analysis

Gene families were clustered with OrthoFinder (v1.1.5) (Emms and Kelly, 2015) and single-copy gene families were used to analyze the evolutionary relationship with PhyML (v3.0) based on the maximum likelihood method with 100 replication. The divergence time was estimated by MCMCTREE (v4.4) (He *et al.*, 2013). Analysis of the expansion and contraction of the gene family was carried out by CAFE (Computational Analysis of gene Family Evolution, v1.6) (De Bie *et al.*, 2006).

RNA-seq bioinformatics analysis

RNA-seq reads were mapped to the assembled genome using TopHat (v2.0.11) with options : –read-mismatches 5 –read-gap-length 5 –read-edit-dist 5 -p 10 -r 50 -a 8 (Trapnell *et al.*, 2009) with anchor length more than 8-nt and further assembled using Cufflinks v2.1.1 (Trapnell *et al.*, 2012) with the following option: -F 0.05 -A 0.01 -I 100000 –min-intron-length 30. Then above transcript assemble were combined using the Cuffmerge utility with the following option: –min-isoform-fraction 0.01. Then the above assembled transcripts and the RNA-seq aligned reads were loaded into rMATS.3.2.2 (Shen *et al.*, 2014) using the following parameters: -t paired -len 125 -a 8 -c 0.0001 -analysis U. The output included several types of AS events including intron retention, exon skipping, alternative acceptor, and alternative donor, and the differential AS events were defined with $P < 0.05$.

Fiber length measurements

The four studied samples (from *C. equis.* ssp. *incana*, *C. equisetifolia*, *C. cunninghamiana* and *C. glauca*, respectively) for fiber length measurements were collected from trees planting in CHIHU state-owned protection forest farm of Fujian province (24°35′N, 118°55′E) on January 19, 2018. The first branches of the above four subspecies were collected at about 2 m from ground level. The diameter at the base ranged from 30 to 40 cm wide, and based on the annual rings, they were 7, 8, 8, and 10 years old, respectively. After each branch was harvested, three disks (200 mm thick) were collected from the top, medium, and base of the branch, so that 12 disks in total were obtained for each sample. Due to the radial variation of wood properties, each disk was separated into three equal sections from pith to bark with a razor blade. For fiber length determination, samples were macerated in a solution of 30% hydrogen peroxide and 100% glacial acetic acid for 48 h at 60°C until totally bleached. Fibers were separated from each other at 30 000 rpm for 3 min (GBJ-A fiber standard dissociation device, Yongxing test instrument Co., Ltd, China), and finally the lengths and widths of 50 000 fibers per sample were measured using a fiber analyzer (MORFI COMPACT, Techpap, France).

Determination of chlorophyll content

We used our previously described method to determine chlorophyll content (Zhang *et al.*, 2018). Approximately 0.1 g of fresh branchlets were cut into about 4-cm long segments and then placed into a solution containing ethanol, acetone and distilled water. Samples were protected from light until totally bleached, absorbance was then measured at 645 and 663 nm using a Thermo Scientific Multiskan GO microplate spectrophotometer (ThermoFisher Scientific, San Jose, CA, USA; product code: 51119300). For each subspecies, 10 replicates were included to determine chlorophyll content.

Anatomical characterization

Degenerate blades and fibers were observed under a scanning electron microscope (Model TM3030Plus Tabletop microscope, Hitachi, Japan). Free-hand cut sections of the fifth internode from random branchlets of distinct subspecies were prepared for lignin visualization. Sections were stained with phloroglucinol-HCl (1% (weight/volume) phloroglucinol in 6 M HCl) for 5 min and then observed under a Leica stereomicroscope (Leica M205FA, Wetzlar, Germany).

Lignin content measurements

We used our previously described method to measure lignin content (Zhang *et al.*, 2018). In brief, about 0.1 g of branchlets were ground into powder in liquid nitrogen, and then hydrolyzed with 95% ethanol. Lignin was dissolved with bromine acetyl acetic and absorbance was measured at 280 nm using a Thermo Scientific Multiskan GO microplate spectrophotometer (ThermoFisher Scientific, product code: 51119300). A standard curve was created from serial dilutions of alkali lignin (Sigma-Aldrich, Poole, Dorset, UK; 370959). Each sample included five replications, and the sample lignin content was determined by comparison with the standard curve.

Detecting DNA modifications

Firstly, reads from PacBio sequencing libraries were mapped to the assembled genome sequences by pbalign (https://github.com/PacificBiosciences/pbalign) using the following settings: –minMatch 12 –bestn 10 –minPctSimilarity 70.0 –refineConcordantAlignments. The above BAM alignments were loaded into ipdSummary (https://github.com/PacificBiosciences/kineticsTools) and DNA modification sites with $P < 0.01$ were regarded as real modification sites.

GO enrichment analysis

Blast2GO was used to annotate GO vocabulary (Conesa *et al.*, 2005) with the default option. Overrepresented GO terms were defined with the BiNGO plugin [39] for Cytoscape [54] with a $P < 0.05$. For *P*-value correction, we selected the false discovery rate (FDR) correction method.

Searching for homologous genes

The InParanoid algorithm (Ostlund *et al.*, 2010) was used to search for the homologs in the *Populus* database (http://popgenie.org) to acquire all homologous genes to *C. equisetifolia* using the default options.

Availability of supporting data

Raw data are available from NCBI Sequence Read Archive (SRA) with accession numbers SRP145409 under BioProject PRJNA450482. The assembled genome has been deposited at NCBI under accession number RDRV00000000. The assembly genome and annotation can be browsed at http://forestry.fafu.edu.cn/db/Casuarinaceae/. The browser combines the genome assembly with the gene annotations, and has tracks for RNA-seq.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (2016YFD0600106 and 2018YFD0600101), the National Natural Science Foundation of China Grant (Grant No. 31570674), the Natural Science Foundation of Fujian Province (Grant No. 2018J01608) and the International Science and Technology Cooperation and Exchange Fund from Fujian Agriculture and Forestry University (KXGH1701). This work was supported by The Key Laboratory of Timber Forest Breeding and Cultivation for Mountainous Areas in Southern China, Fujian Forest Seedling Technology Project. We thank Chao Ma for project coordination.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

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

Table S1. Summary of library construction for *C. equisetifolia*.

Table S2. Statistic of genome assembly.

Table S3. Gene annotation summary.

Table S4. Summary of functional annotation.

Table S5. Identification of non-coding genes.

Table S6. Annotation of repeat sequences.

Table S7. List of genes with alternative splicing events.

Table S8. List of 119 stress genes.

Table S9. List of genome-wide analysis of DNA modification.

Table S10. List of chlorophyll-related genes.

REFERENCES

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A. and Reddy, A.S. (2016) A survey of the *Sorghum* transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706.

Annacondia, M.L., Magerøy, M.H. and Martínez, G.J.P.P. (2018) Stress response regulation by epigenetic mechanisms: changing of the guards. *Physiol. Plant.* **162**, 239–250.

Au, K.F., Underwood, J.G., Lee, L. and Wong, W.H. (2012) Improving PacBio long read accuracy by short read alignment. *PLoS ONE*, **7**, e46679.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2010) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.

BRIEF, I. (2006) *Casuarina* equisetifolia (beach she-oak) C. cunninghamiana (river she-oak). *Traditional Trees of Pacific Islands: Their Culture, Environment, and Use*, 227.

Calixto, C.P., Guo, W., James, A.B., Tzioutziou, N.A., Entizne, J.C., Panter, P.E., Knight, H., Nimmo, H., Zhang, R. and Brown, J.W.J.T.P.C. (2018) Rapid and dynamic alternative splicing impacts the Arabidopsis cold response transcriptome. *Plant Cell* **30**, 1424–1444.

Chen, N. (2004) Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, **5**, 4.10.11–14.10.14.

Chen, Y., Wang, G. and Zhou, J. (2005) Advances in the study of stress resistance of *Casuarina equisetifolia*. *Chinese Bulletin of Botany*, **22**, 746–752.

Chin, C.S., Peluso, P., Sedlazeck, F.J. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, **13**, 1050–1054.

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

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D. and Velasco, R.J.N.G. (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099.

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G.A., Milne, R., Chen, Y., Wan, Z. and Wang, Z. (2014) The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res.* **24**, 1274.

De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

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

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S. and Weinstock, G.M. (2007) Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157.

Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M.C. and Jabado, O.J. (2012) Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232.

Farrokhi, N., Burton, R.A., Brownfield, L., Hrmova, M., Wilson, S.M., Bacic, A. and Fincher, G.B. (2006) Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. *Plant Biotechnol. J.* **4**, 145–167.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.-K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.* **20**, 45–58.

Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X. and Doré, L.C. (2015) N 6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell*, **161**, 879–892.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P. and Warland, A. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods*, **15**, 201–206.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.-H., Aravind, L., He, C. and Shi, Y. (2015) DNA methylation on N 6-adenine in C. elegans. *Cell*, **161**, 868–878.

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W. and Liang, X. (2016) Draft genome of the living fossil Ginkgo biloba. *Gigascience*, **5**, 49.

He, N., Zhang, C., Qi, X. *et al.* (2013) Draft genome sequence of the mulberry tree Morus notabilis. *Nat. Commun.* **4**, 2445.

Hu, P., Zhong, C., Zhang, Y., Jiang, Q., Chen, Y., Chen, Z., Pinyopusarerk, K. and Bush, D. (2016) Geographic variation in seedling morphology of *Casuarina* equisetifolia subsp. equisetifolia (*Casuarina*ceae). *Aust. J. Bot.* **64**, 160–170.

Iyer, L.M., Abhiman, S. and Aravind, L. (2011) Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **101**, 25–104.

Koren, S., Schatz, M.C., Walenz, B.P. *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700.

Korlach, J., Gedman, G., Kingan, S.B., Chin, C.-S., Howard, J.T., Audet, J.-N., Cantin, L. and Jarvis, E.D.J.G. (2017) De Novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience*, **6**, 1–16.

Kuusk, V., Niinemets, Ü. and Valladares, F. (2017) A major trade-off between structural and photosynthetic investments operative across plant and needle ages in three Mediterranean pines. *Tree Physiol.* **38**, 543–557.

Li, H.-B., Li, N., Yang, S.-Z., Peng, H.-Z., Wang, L.-L., Wang, Y., Zhang, X.-M. and Gao, Z.-H. (2017) Transcriptomic analysis of *Casuarina* equisetifolia L. in responses to cold stress. *Tree Genet. Genomes*, **13**, 7.

Liang, Z., Shen, L., Cui, X., Bao, S., Geng, Y., Yu, G., Liang, F., Xie, S., Lu, T. and Gu, X. (2018) DNA N6-Adenine methylation in Arabidopsis thaliana. *Dev. Cell*, **45**, 406–416.e403.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J. and Dorschner, M.O. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Liu, C., Ran, Y., Tao, Y., Ning, X., Bai, L., Ye, H., Li, X. and Li, L. (2013) The present situation investigation of coastline *Casuarina* forest in Hainan Island. *Forest Resources Management*, **4**, 102–118.

Liu, X., Mei, W., Soltis, P.S., Soltis, D.E. and Barbazuk, W.B. (2017) Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Mol. Ecol. Resour.* **17**, 1243–1256.

Liu, Q., Chang, S., Hartman, G.L. and Domier, L.L. (2018) Assembly and annotation of a draft genome sequence for *Glycine latifolia*, a perennial wild relative of soybean. *Plant J.* **95**, 71–85.

Love, R.R., Weisenfeld, N.I., Jaffe, D.B., Besansky, N.J. and Neafsey, D.E. (2016) Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genom.* **17**, 187.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955.

Ma, T., Wang, J., Zhou, G., Yue, Z., Hu, Q., Chen, Y., Liu, B., Qiu, Q., Wang, Z. and Zhang, J.J.N.C. (2013) Genomic insights into salt adaptation in a desert poplar. *Nat. Commun.* **4**, 2797.

Ma, J., Wan, D., Duan, B., Bai, X., Bai, Q., Chen, N. and Ma, T. (2018) Genome sequence and genetic transformation of a widely distributed and cultivated poplar. *Plant Biotechnol.* https://doi.org/10.1111/pbi.12989.

Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Martínez-García, P.J., Crepeau, M.W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K.A., Paul, R., Butterfield, T.S., Britton, M.T. and Reagan, R.L. (2016) The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **87**, 507–532.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A. *et al.* (2014) The genome of Eucalyptus grandis. *Nature*, **510**, 356–362.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Ndoye, A.L., Sadio, O. and Diouf, D. (2011) Genetic variation of *Casuarina* equisetifolia subsp equisetifolia and C. equisetifolia subsp incana populations on the northern coast of Senegal. *Genet. Mol. Res.* **10**, 36–46.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E. and Liechty, J.D. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59.

Nystedt, B., Street, N.R., Wetterbom, A. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.

Ogunwande, I.A., Flamini, G., Adefuye, A.E., Lawal, N.O., Moradeyo, S. and Avoseh, N.O. (2011) Chemical compositions of *Casuarina* equisetifolia L., Eucalyptus toreliana L. and Ficus elastica Roxb. ex Hornem cultivated in Nigeria. *S. Afr. J. Bot.* **77**, 645–649.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O. and Sonnhammer, E.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–D203.

Peng, Z., Lu, Y., Li, L. *et al.* (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). *Nat. Genet.* **45**, 456–461. 461.e451-461.e452.

Pinyopusarerk, K., Kalinganire, A., Williams, E. and Aken, K.M. (2004) *Evaluation of International Provenance Trials of Casuarina equisetifolia.* ACIAR Technical Report No. 58. Australian Centre for International Agricultural Research, Canberra. Citeseer.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing. *Genome Biol.* **14**, 405.

Sanyal, R.P., Misra, H.S. and Saini, A.J.J.O.E.B. (2018) Heat-stress priming and alternative splicing-linked memory. *J. Exp. Bot.* **69**, 2431–2434.

Schwencke, J., Bureau, J.-M., Crosnier, M.-T. and Brown, S. (1998) Cytometric determination of genome size and base composition of tree species of three genera of *Casuarina*ceae. *Plant Cell Rep.* **18**, 346–349.

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA*, **111**, E5593–E5601.

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

Sunaryo, W. and Fischer, U. (2009) In silico expression analysis of the Arabidopsis KNAT1 gene and its homologs in poplar. In *Review of forests, wood products and wood biotechnology of Iran and Germany* (Kharazipour, A.R., Schöpper, C., Müller, C. and Euring, M., eds). Göttingen, Germany: Universitätsdrucke Göttingen.

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

Tempel, S. (2012) Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and

transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Tuskan, G.A., Difazio, S., Jansson, S. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.

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

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C. and Ware, D. (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708.

Wang, T., Wang, H., Cai, D., Gao, Y., Zhang, H., Wang, Y., Lin, C., Ma, L. and Gu, L. (2017) Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (Phyllostachys edulis). *Plant J.* **91**, 684–699.

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

Winzell, A., Aspeborg, H., Wang, Y. and Ezcurra, I. (2010) Conserved CA-rich motifs in gene promoters of Pt× tMYB021-responsive secondary cell wall carbohydrate-active enzymes in Populus. *Biochem. Biophys. Res. Comm.* **394**, 848–853.

Xu, X.Y., Xiao, L., Wang, M.H. and Zhang, H.X. (2015a) A comprehensive evaluation system for anti-typhoon performance of trees in coastal areas. *J. Zhejiang Univ. Sci. B*, **32**, 516–522.

Xu, Z., Peters, R.J., Weirather, J., Luo, H., Liao, B., Zhang, X., Zhu, Y., Ji, A., Zhang, B. and Hu, S. (2015b) Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. *Plant J.* **82**, 951–961.

Yang, J., Chang, T., Chen, T. and Chen, Z. (1995) Provenance trial of *Casuarina* equisetifolia in Taiwan. 1. Seed weight and seedling growth. *Bulletin of Taiwan Forestry Research Institute*, **10**, 195–207.

Yang, W., Wang, K., Zhang, J., Ma, J., Liu, J. and Ma, T. (2017) The draft genome sequence of a desert tree Populus pruinosa. *Gigascience*, **6**, 1–7.

Zhang, L.H., Ye, G.F., Lin, Y.M., Zhou, H.C. and Zeng, Q. (2009) Seasonal changes in tannin and nitrogen contents of *Casuarina* equisetifolia branchlets. *J. Zhejiang Univ. Sci. B*, **10**, 103–111.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J. and Meredith, R.W.J.S. (2014a) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.

Zhang, G.J., Li, C., Li, Q.Y. *et al.*; Avian Genome Consortium. (2014b) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P. and Liu, J. (2015) N6-methyladenine DNA modification in *Drosophila*. *Cell*, **161**, 893–906.

Zhang, H., Lin, C. and Gu, L. (2017) Light regulation of alternative Pre-mRNA splicing in plants. *Photochem. Photobiol.* **93**, 159–165.

Zhang, H., Wang, H., Zhu, Q. *et al.* (2018) Transcriptome characterization of moso bamboo (Phyllostachys edulis) seedlings in response to exogenous gibberellin applications. *BMC Plant Biol.* **18**, 125.

Zhong, C. and Bai, J. (1996) Introduction trials of casuarinas in southern China. In *Recent Casuarina Research and Development. Proceedings of the Third International Casuarina Workshop. Danang, Vietnam* (Pinyopusarerk, K.T., Turnbull, J.W. and Midgley, S.J., eds). Canberra, Australia: CSIRO, pp. 191–195.

Zhong, C., Bai, J. and Zhang, Y. (2005) Introduction and conservation of *Casuarina* trees in China. *Forest Res.* **18**, 345–350.

Zhong, C.L., Zhang, Y., Chen, Y., Jiang, Q.B., Chen, Z., Liang, J.F., Pinyopusarerk, K., Franche, C. and Bogusz, D. (2010) *Casuarina* research and applications in China. *Symbiosis*, **50**, 107–114.

Zhu, S., Beaulaurier, J., Deikus, G., Wu, T.P., Strahl, M., Hao, Z., Luo, G., Gregory, J.A., Chess, A. and He, C. (2018) Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res.* **28**, 1067–1078.