Supplementary information for

The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole genome sequencing

Shaohua Xu[1,9], Ziwen He[1,9], Zhang Zhang[1,9], Zixiao Guo[1,9], Wuxia Guo[1], Haomin Lyu[1], Jianfang Li[1], Ming Yang[1], Zhenglin Du[2], Yelin Huang[1], Renchao Zhou[1], Cairong Zhong[3], David E. Boufford[4], Manuel Lerdau[5], Chung-I Wu[1,2,6], Norman C. Duke[7], **The International Mangrove Consortium**[8] & Suhua Shi[1]

[1] State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong, China

[2] Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[3] Hainan Dongzhai Harbor National Nature Reserve, Haikou, Hainan, China

[4] Harvard University Herbaria, Cambridge, Massachusetts, USA

[5] Departments of Environmental Sciences and of Biology, University of Virginia, Charlottesville, Virginia, USA

[6] Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

[7] Centre for Tropical Water and Aquatic Ecosystem Research, James Cook University, Townsville QLD, Australia

[8] Members of the consortium are listed at the end of the manuscript.

[9] These authors contributed equally to this work.

Correspondence should be addressed to S.S. (lssssh@mail.sysu.edu.cn).

Supplementary Note

1. Whole-genome sequencing, assembly and annotation

Materials preparation

The plants of the *Rhizophora* genus are typical mangroves in the Indo-West Pacific (IWP) and Atlantic-East Pacific (AEP) areas. *Rhizophora apiculata*, *R. stylosa*, and *R. mucronata* are distributed and dominant in the IWP area, while *R. mangle* is dominant in the AEP area. These species exhibit special characteristics that are beneficial for adapting to the intertidal zone, such as true vivipary, salt exclusion and special root systems. We randomly sampled one mature *R. apiculata* plant from Qinglan Harbor, Hainan, China (19°37'N, 110°48'E) for *de novo* genome sequencing and assembly. The *R. stylosa* sample was collected in Danzhou, Hainan, China (19°52'N, 109°32'E); a single *R. mucronata* individual was collected in Chaiya, Thailand (9°22'N, 99°16'E); and *R. mangle* was collected from a transplanted individual in Dongzhai Harbor, Hainan (19°57'N, 110°34'E) which was introduced from La Paz, Mexico (24°09'N, 110°20'W). Genomic DNA was extracted from leaves using the CTAB method [51]. Total RNA from leaves, roots, flowers and stems of *R. apiculata* was extracted using the modified CTAB method [52], for the gene annotation and genome completeness assessment.

Whole-genome Single Molecule Real Time (SMRT) sequencing

For the *de novo* whole-genome sequencing of *R. apiculata*, we obtained 16.23 Gb of Single Molecule Real Time (SMRT) long reads using the Pacific Biosciences RS II sequencing platform with C4 sequencing chemistry, P6 polymerase, and 25 SMRT cells (Supplementary Table 1; Supplementary Figs. 1-3). A 20-kb SMRT bell library was prepared from sheared genomic DNA using a 20-kb template library preparation workflow.

Whole-genome short-reads sequencing

For short-read sequencing of *R. apiculata*, ten libraries with different insert sizes (200 bp, 300 bp, 400 bp, 600 bp, 2 Kb, 5 Kb and 10 Kb; Supplementary Table 3) were prepared and 89.3 Gb of paired-end/mate-paired short reads were obtained using the HiSeq2000 platform. A total of 30 Gb of RNA sequences (library insert size of 300 bp) from *R. apiculata* tissues were also obtained. The raw reads were filtered in four steps:

(1) remove reads containing the Illumina TruSeq adaptor core sequence "GATCGGAAGA" with ≤ 1 mismatch in the 3' end;

(2) remove the duplicated reads from PCR amplification (if read 1 and read 2 of the two paired-end reads were identical in the first 30 bp);

(3) remove reads with length < 30 bp;

(4) remove single-end reads.

A single individual from each of the other three species, *R. stylosa*, *R. mucronata* and *R. mangle*, was also re-sequenced at lower depth, yielding in total 3.1 - 15.8 Gb of sequences with insert size of 300 bp (Supplementary Table 5; Supplementary Fig. 5).

De novo genome assembly

We assembled the *R. apiculata* genome *de novo* based on the PacBio long reads using four pieces of software: *falcon* (https://github.com/PacificBiosciences/FALCON/), *DBG2OLC* [53], *smartdenovo* (https://github.com/ruanjue/smartdenovo), and *wtdbg* (https://github.com/ruanjue/wtdbg). *smartdenovo* turned out to yield the best performance. *Quiver* [54] was used to further improve site-specific consensus accuracy for genome polishing. The Illumina reads were mapped to the polished assembly sequence using *BWA* [55]. After mapping, the SNPs as well as small insertion-deletions (indels) were called and corrected by *SAMTOOLS* [56] and in-house scripts. Finally, utilizing the 10 kb mate-pair sequencing data, we generated scaffolds and performed gap-filling using *SSPACE* 3.0 [57] with default parameter values except setting: -x 1 -m 50 -o 10 -z 200 -p 1. The statistics of the assembly of *R. apiculata* genome are shown in Supplementary Table 2.

Assessment of genome completeness

To evaluate the quality of our *de novo* assembly of the *R. apiculata* genome, we first aligned the assembled transcripts (151,828) to each scaffold using *BLAT* (v.34x12) [73] with default options. Our assembled scaffolds covered 96.95% of the assembled transcripts (Supplementary Table 7), indicating that most of the expressed genes were included in our genome assembly. The *CEGMA* (Core Eukaryotic Genes Mapping Approach; v2.4) [16] was also employed and 428 (93.45%) of the 458 core eukaryotic genes were present in our assembly (Supplementary Table 7). We further aligned the sequences of 79 randomly selected *R. apiculata* genes from our previous work [17] to the assembled genome using *BLAT* (v. 34x12), and succeeded in recovering 78 of them (Supplementary Table 7).

Gene structure prediction

To get high confidence *R. apiculata* gene models, we used three approaches to predict protein coding genes: homolog-based, *de novo*, and transcriptome-based predictions. Before the predictions, repeat sequences were masked throughout the genome using *RepeatMasker* (version 3.2.9) [58] and the RepBase library (version 16.08) [59]. We first aligned homologous proteins from six known whole genome sequences: *Oryza sativa, Mimulus guttatus, Sesamum indicum, Populus trichocarpa,* and *Eucalyptus grandis* (downloaded from Phytozome [http://www.phytozome.net] and Sinbase [http://ocri-genomics.org/Sinbase]) to the repeat-masked *R. apiculata* genome using exonerate (version 1.1.1) [60] and generated gene structures based on the homology alignments of proteins to the genome using Genewise (version 2.2.0) [61]. For *ab initio* gene prediction, we used the repeat-masked genome sequences as inputs with *Augustus* (version 3.2.2) [62] and *GeneMark-ET* (version 4.29) [63]. We used *Tophat* (version v2.1.1) [64] to map the RNA-seq sequences to the genome and *cufflinks* (version 2.2.1) [65] to map spliced transcripts to gene models. All gene models predicted from the above three approaches were integrated using *EVidenceModeler* (EVM) [66] into a weighted and non-redundant consensus of gene structures. In total 26,640 protein-coding genes were predicted with an average length of 2,838 bp, and average CDS length of 1,179 bp (Supplementary Table 9).

Gene function annotation

The predicted gene models were functionally annotated by their sequence similarity to genes and proteins in the NCBI nucleotide (Nt), non-redundant and UniProt/Swiss-Prot protein databases [67]. The gene models were also annotated by their protein domains using InterProScan [74]. All genes were classified by Gene Ontology (GO) (Supplementary Fig. 6), eukaryotic orthologous groups (KOG) (Supplementary Fig. 7) and Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways. The Gene ontology classification of each gene was obtained by aligning to the Pfam database using *HMMER2GO* (https://github.com/sestaton/HMMER2GO). The annotation of Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways was conducted by aligning to the KEGG database. A summary of the functional annotations is shown in Supplementary Table 10.

Transcription factor prediction

The program *iTAK* (v1.2; http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi) was used to identify and classify transcription factors (TFs). A total of 1,783 TFs belonging to 58 families were identified, among which MYB is the most abundant transcription factor family (Supplementary Table 12).

2. Divergence time and whole-genome duplication (WGD)

Estimation of species divergence time

To estimate divergence times among the Rhizophoreae species (a tribe which includes four exclusively mangrove genera *Bruguiera*, *Ceriops*, *Kandelia* and *Rhizophora*), we reconstructed phylogenetic trees using a data set including the *de novo* genome assembly of *Rh. apiculata*, three whole-genome re-sequencing data sets (*R. stylosa*, *R. mucronata* and *R. mangle*), five transcriptomes of their close relatives, and two published genomes of other species (*Ri. communis* and *P. trichocarpa* from Phytozome [http://www.phytozome.net]; Supplementary Table 6). The transcriptomes were first assembled by *Trinity* [75] and the coding region sequences (CDS) were aligned by *BLASTx* [76] against the assembled genome with an e-value threshold of 1×10^{-5}. The CDSs of *R. stylosa*, *R. mucronata* and *R. mangle* were obtained by mapping genome sequences to the genome of *R. apiculata*. Protein sequences were translated from CDS using a BioPerl script. Using *OrthoMCL* [70] we found 29,560 gene clusters, among which we defined those containing only one member from each species as single-copy ortholog groups. Putative orthologs were aligned by a combination of *PAL2NAL* [77] and *MUSCLE* [78]. After removing short (< 150 bp) orthologs or those with large dN (> 0.5) between *R. apiculata* and *P. trichocarpa*, we retained 590 single-copy orthologous groups for phylogenetic tree construction.

The phylogenetic tree was reconstructed by *PhyML* [20] with 1000 bootstrap replicates, having 100% support at all nodes (Supplementary Fig. 12a). We also randomly sampled 50 genes and reconstructed the tree again, repeating this process ten times. This exercise supported the same tree, suggesting that our whole-genome inference is not based on a sub-set of loci. For estimating divergence times, we applied a popular Bayesian method (see examples in: Yim *et al.*, 2014 [79]; Frantz *et al.*,

2013[80] and Ma *et al*., 2013 [3]) *MCMCTREE* from the package *PAML* (version 4.8) [22], with the key parameters set as "seq like (usedata = 1)," "HKY85+gamma (model = 4, alpha = 0.5)" and "independent rates (clock = 2)". The estimation was repeated a second time to ensure convergence. Two calibration nodes were included in the phylogeny to constrain the estimation of substitution rates in the *PAML* method. One was that the root node of Malpighiales, the common ancestor of Rhizophoraceae, Euphorbiaceae (*Ri. cimmunis*) and Salicaceae (*P. trichocarpa*), was placed at 105-120 Myr before present [24, 25]. The other (the red box 2 in Supplementary Fig. 12b) was that the common ancestor of *K. obovata*, *Ce. tagal* and *R. apiculata* was placed earlier than 38 Myr since the earliest convincing fossils of *Rhizophora* have been dated to the late Eocene [26, 27]. In addition, the known earliest convincing fossils of Rhizophoreae, which were found in the London Clay and dated to the early Eocene (47.8-56 Myr) [27, 28], were used to narrow the window of invasion.

Comparisons of divergence times estimated by different datasets / methods

To test the robustness of the estimation described in the previous paragraph, we repeated the analysis with data set of intact codons (1^{st} +2^{nd}+3^{rd} codon positions included) or only conserved codon positions (1^{st} +2^{nd} codon positions in each amino acid). Both the mode and 95% confidence interval of the estimated divergence times were robust to these methodological differences as presented in Supplementary Table 13 and Supplementary Fig. 13.

We also employed the software *r8s* [81] to date the divergence time. A phylogeny with branch length and constrains on nodes is necessary for the calculation in *r8s*. We built such a phylogenetic tree using two different nucleotide substitution models: HKY85+G and GTR+I+G, which produced similar branch lengths. In *r8s*, there are three methods: LF (Langley-Fitch), PL (Penalized likelihood) and NPRS (nonparametric rate smoothing). Each method is implemented using three algorithms: TN, Powell and Quewt. We conducted our calculations using the TN algorithm for LF and PL methods and the Powell algorithm for the NPRS method, as recommended by the *r8s* user manual. In addition, the dataset including 1^{st} +2^{nd}+3^{rd} codon positions of each amino acid and the same time calibrations described above were used in this time dating. The results obtained from r8s were similar to those from *MCMCTREE* (Supplementary Table 14; Supplementary Fig. 13).

Whole-genome duplication identification and dating

Self-alignment was performed on protein sequences of *R. apiculata* using *BLASTp* (with an e-value cutoff of 1.0×10^{-5}, identity $\geq 40\%$) and duplicated blocks were identified using *MCScanX* [19]. Only collinear blocks with at least five paired genes were accepted as duplicated blocks in this study. Under this criterion, 377 duplicate blocks were identified in the genome of *R. apiculata*, covering 74.01% of the genome, accounting for 40.71% of the total genes. This provides convincing evidence for a whole genome duplication (WGD) event. The collinear blocks were visualized using *Circos* (v0.65) [71] and shown in Figure 1.

To gauge the age of the WGD event, we examined the distribution of synonymous nucleotide substitution rates (dS) between paralogous genes. Consistent with the prediction of a single WGD event, the dS distribution shows a single peak (Figure 2b). By comparing the dS distribution among

the paralogs within the *R. apiculata* genome with that of orthologous genes between *R. apiculata* and its relatives, we can roughly estimate the likely date of the WGD event (Figure 2b). To estimate the date more precisely, we plot branch length (nucleotide substitution rate) distributions of paralogous gene pairs within the *R. apiculata* genome, whose peak (L_peak) is a good prediction of the WGD date (Supplementary Fig. 14). By comparing the branch length between the WGD event (L_peak) and the closest node (L_node), the time t1 between the WGD and the node can be calculated by (L_peak - L_node)/μ, where L_node is the branch length between the node and the present, and μ is the average mutation rate on this branch (Supplementary Fig. 12b). Both L_node and μ were estimated from previous analyses of divergence time. Given the divergence time of the closest node (t2), we can date the WGD event by adding t1 to t2 (when the WGD is more ancient than the closest node) or subtracting t1 from t2 (when the WGD is more recent than the closest node). The dS peak between pairs of *R. apiculata* paralogous genes is at lower divergence than that of orthologs between *R. apiculata* and *P. trichocarpa*, but higher than that of other dS pairs, indicating that the WGD event occurred previous to the split of the common ancestor of Rhizophoraceae. The branch length distribution peak between *R. apiculata* paralogs is 0.17, which puts the WGD event at about 69 Mya (Supplementary Fig. 12b).

Gene retention after WGD

As the peak of the dS distribution between pairs of paralogous genes within the *R. apiculata* genome is 0.35 (Figure 2b), the paralogs produced in the Rhizophoraceae specific WGD are expected have dS values near the peak. To pick out the paralogous genes produced in this WGD rather than more ancient events, we filtered all the blocks except those with median dS in the 0.25-0.7 range for further analysis. The 121 remaining collinear blocks comprise 2,878 pairs of paralogous genes (paired genes) and 14,274 single genes (genes without paralogs). With single genes as control, we tested GO enrichment of these 2,878 paired genes. We found 93 GO terms in the type "Biological Process" significantly enriched for genes that remained duplicated ($q<0.05$). When focusing on terms at level 2, we found that eight GO terms were enriched, including related to regulation processes such as "regulation of biosynthetic process" and "signaling", as well as related to stress response ("response to stimulus").

3. Gene family analysis

The *OrthoMCL* method [70] was used to infer orthologous and paralogous gene groups in the genomes of *R. apiculata* as well as three inland species downloaded from Phytozome (http://www.phytozome.net), namely *A. thaliana* [82], *Ri. communis* [83] and *P. trichocarpa* [84]. For genes with alternative splicing, the longest transcripts were selected. Proteins of these four species were then combined to perform an all-*vs*-all comparison using *BLASTp* with an e-value cutoff of 1×10^{-10}. The results were fed into a stand-alone *OrthoMCL* program with the default MCL inflation parameter of 2.0.

A total of 26,640 protein-coding genes in *R. apiculata* are classified into 17,806 families, with 10,054 families shared by all the four species and 432 families private to *R. apiculata* consisting of only one member (Supplementary Fig. 7). After gene family clustering, *CAFE* [46] was used to analyze

gene family expansion and contraction along the phylogeny of the four species. The phylogenetic tree topology and branch lengths were used to infer the significance of change in gene family size for each branch. We found that 112 or 118 gene families experienced expansion or contraction with P values smaller than 0.05 (Supplementary Fig. 20). The GO enrichment analysis of the expanded gene families is shown in Supplementary Fig. 21.

4. Gene tandem duplication in *R. apiculata*

We defined tandem duplicated genes as those that are homologous with *BLAST* e-values at least $1e^{-20}$ and also are separated by no more than five other loci. We identified 2,963 such genes (11% of total genes) belonging to 792 tandem duplicated regions in the *R. apiculata* genome. Gene Ontology (GO) enrichment analysis of the tandem duplicated genes indicates that 79 GO terms are significantly enriched after applying Fisher's exact test and multiple correction (Supplementary Table 20).

Some duplicated genes are particularly notable for their probable role in vivipary. Two genes in the GA biosynthesis pathway, ent-kaurene synthase (*KS*) and GA3β-hydroxylase (*GA3ox*), were found to have experienced copy number expansion in the genome of *R. apiculata* compared to its three inland relatives. *KS* had significantly expanded to nine copies in *R. apiculata* (Supplement Figure 16), whereas there are at most four copies in the related inland species. *GA3ox* had experienced a unique tandem duplication event. The two expanded genes could result in an elevated biosynthesis of gibberellin (GA). GA and abscisic acid (ABA) are two main regulators of seed dormancy and germination, with ABA inhibiting germination while GA promoting it. Consistent with the effect of ABA on germination, previous study has reported a decreased ABA concentration [85] in Rhizophoreae mangroves' seeds and hypocotyls. The gene copy number expansion in GA biosynthesis pathway may increase the GA concentration in seeds and hypocotyls, and further promote embryonic development. Another piece of evidence comes from *SAE2* (SUMO-activating enzyme 2), which shows signs of positive selection (see below). Mutants affecting the single genes encoding SAE2 are embryonic lethal in *Arabidopsis* [39]. SAE2 conjugates with SAE1 as a heterodimer to activate SUMO (small ubiquitin-like modifier), which then attaches to a target protein functioning as a post-translational modification. Studies have shown a role for SUMO in the modulation of the ABA signal transduction pathway [86]. The altered regulation of ABA signaling by SAE2 and expanded gene copy number in GA biosynthesis pathway may contribute to viviparous embryo development in Rhizophoreae mangroves (Fig. 4d).

5. Amino Acid usage

We collected available whole genome protein sequences of 47 inland dicotyledons according to CoGepedia (https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes) and compared their amino acid (AA) compositions with *R. apiculata* and other two mangrove species from other clades (*Avicennia marina* and *Sonneratia alba*).

6. Inference of protein evolution rates from dN/dS ratios

To detect genes that may have undergone positive selection, we calculated dN/dS ratios among proteins from our Rhizophoraeae genomes. We used the sequences of the nine species shown in Figure 2c to identify orthologs using *OrthoMCL* by retaining orthologous groups with a single copy in each species as well as those multi-copy orthologs that had only a single copy in the outgroup (*A. thaliana*). The latter orthologs were obtained as follows: we first aligned (using *BLAST*) single copy genes of the outgroup against all copies from other species and selected the best hits as representative orthologs. The prepared orthologs were then aligned using *MUSCLE* and codon sequences were obtained using *PAL2NAL* [77]. We initially obtained 4,414 aligned orthologs and discarded those with fewer than 50 codons after removing sites with ambiguous data. The final data set contained 4,079 high-confidence orthologs.

The alignments were used as input for *CODEML* in the *PAML* 4.8 package to detect positively-selected genes (PSGs). Setting the branch of *R. apiculata* as foreground, we used the branch-site model to calculate the likelihood of the null model (no site in the foreground is positively selected) and alternative model (existing sites in the foreground are positively selected), and then computed the likelihood ratio. To remove false discoveries, the Benjamini-Hochberg correction [87] for multiple testing was performed (FDR < 0.05). As described in the main text, we also performed this analysis on the internal branch ancestral to the Rhizophoraeae clade using 255 seed genes (Supplementary Table 17).

We also used *CODEML* to calculate dN/dS ratio (ω) for each Rhizophoraceae branch. All 4,079 orthologs were concatenated into a single super gene. The "free-ratio" model [88, 89] was used. To identify rapidly evolving genes in *R. apiculata*, we used sequences from five species: *A. thaliana*, *Ri. communis*, *P. trichocarpa*, *Ca. brachiata* and *R. apiculata*. We ran *CODEML* with a null model hypothesizing that the *Ca. brachiata* and *R. apiculata* branches have the same dN/dS ratio. The alternative hypothesis allowed these two branches to have different dN/dS. The likelihood ratio test (LRT) was used to select rapidly evolving genes in *R. apiculata* ($p<0.05$). We found 319 genes with significantly higher dN/dS ratio in *R. apiculata* than *Ca. brachiata*. We further clustered all orthologous genes by GO terms and found that almost all GO terms have higher median dN/dS ratios in *R. apiculata* than *C. brachiata*. Seven of these GO terms show a two-fold or larger increase, the most conspicuous ones being "cell redox homeostasis" and "cellular homeostasis" (Supplementary Table 16).

7. Analysis of transcription profiles

To explore gene expression patterns in response to salt stress, we sequenced a series of *R. apiculata* transcriptomes under different NaCl treatments. Seedlings were collected from Hainan Island and first cultivated with 1/2 Hoagland's nutrient solution [90] on clean sand for at least seven days. The seedlings were then divided into three groups and treated with different NaCl concentrations (0 mM, 250 mM and 500 mM NaCl in 1/2 Hoagland's nutrient solution) for seven days. Treatment length is based on our previous experiments. Liang et al. (2012) [9] have shown that with high salinity treatment in *Ceriops tagal* no obvious gene differential expression arises until about eight days. Total RNA from healthy young leaves and roots from plants under each treatment (two individuals each)

were extracted using the Plant RNA Kit (OMEGA) and sequenced using the HiSeq 2000 platform. Two independent biological replicates were examined.

Short reads from each transcriptome were aligned to the corresponding reference genome by *TopHat* (version v2.1.1) [64] and assembled by *Cufflinks* (version 2.2.1) [65]. Two programs, *Cuffmerge* and *Cuffdiff* from the *Cufflinks* package were used to compare the differential expression profiles between different tissues and conditions. The Benjamini-Hochberg correction [87] for multiple testing was used (q-value < 0.05) in the analysis. The genes significantly differentiated in expression (q-value < 0.05) with a fold change greater than two were selected for further study.

There are 391-875 genes differently expressed in *R. apiculata* roots or leaves (Supplementary Figs. 16-17). We mapped these genes to Gene Ontology, focusing on categories with no fewer than five genes. We tested for category enrichment using Fisher's exact test requiring FDR less than 0.05 (Supplementary Fig. 18). In leaves, as salt concentration increase, DEGs are enriched in GO terms related to carbohydrate metabolism, such as carbohydrate metabolic process (GO:0005975), polysaccharide metabolic process (GO:0005976), cellular glucan metabolic process (GO:0006073), glucan biosynthetic process (GO:0009250). Carbohydrate metabolic processes play important roles in high salinity tolerance by providing energy for development and regulating signal transduction (Gupta and Kaur, 2005). In roots, DEGs are enriched for "response to oxidative stress" (GO:0006979), defense response (GO:0006952) and response to biotic stimulus (GO:0009607). The results suggest root tissue play a major role in response to the environment stress and protect plants from damage.

8. Heterozygosity and demographic history

Estimation of heterozygosity

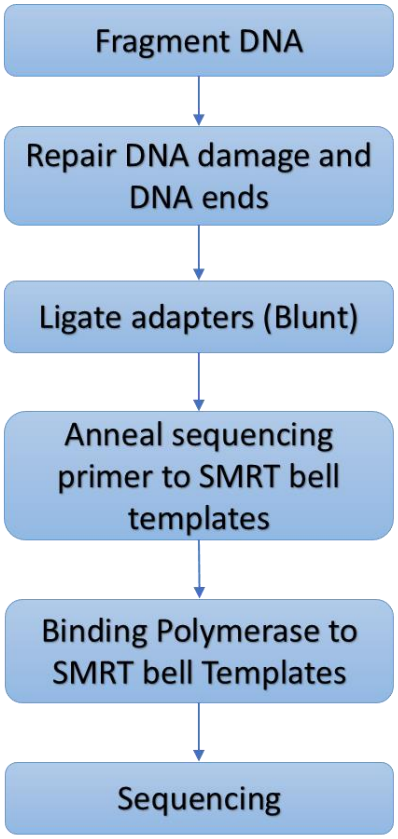
The heterozygosity levels of the four *Rhizophora* species (*R. apiculata, R. stylosa. R. mucronata* and *R. mangle*) were calculated by calling heterozygous sites. Reads from high-quality short insert libraries were mapped to the *de novo* assembled genome using the aligner *bowtie2* [72] using default parameters. After removal of potential PCR duplicated, single-end mapped and improperly paired reads the alignments were analyzed for SNP calling. All sites that met sequencing depth criteria (20-200× for *R. apiculata*, 15-80× for *R. mangle* and *R. stylosa*, 10-25× for *R. mucronata*) were analyzed. To exclude sequencing errors, only the sites with minor reads mapping depth larger than 0.15 were used. More than 99.9% of heterozygous sites can be retained according to the binomial function, assuming that the two alleles are equally sequenced. Heterozygosity was estimated as the number of identified heterozygous sites divided by the total number of sites meeting our read depth criteria. As a result, the heterozygosities are 5.51×10^{-4}, 4.95×10^{-4}, 3.11×10^{-4} and 3.82×10^{-4} per bp for *R. apiculata*, *R. mangle*, *R. stylosa* and *R. mucronata*, respectively.

Effective population size analysis

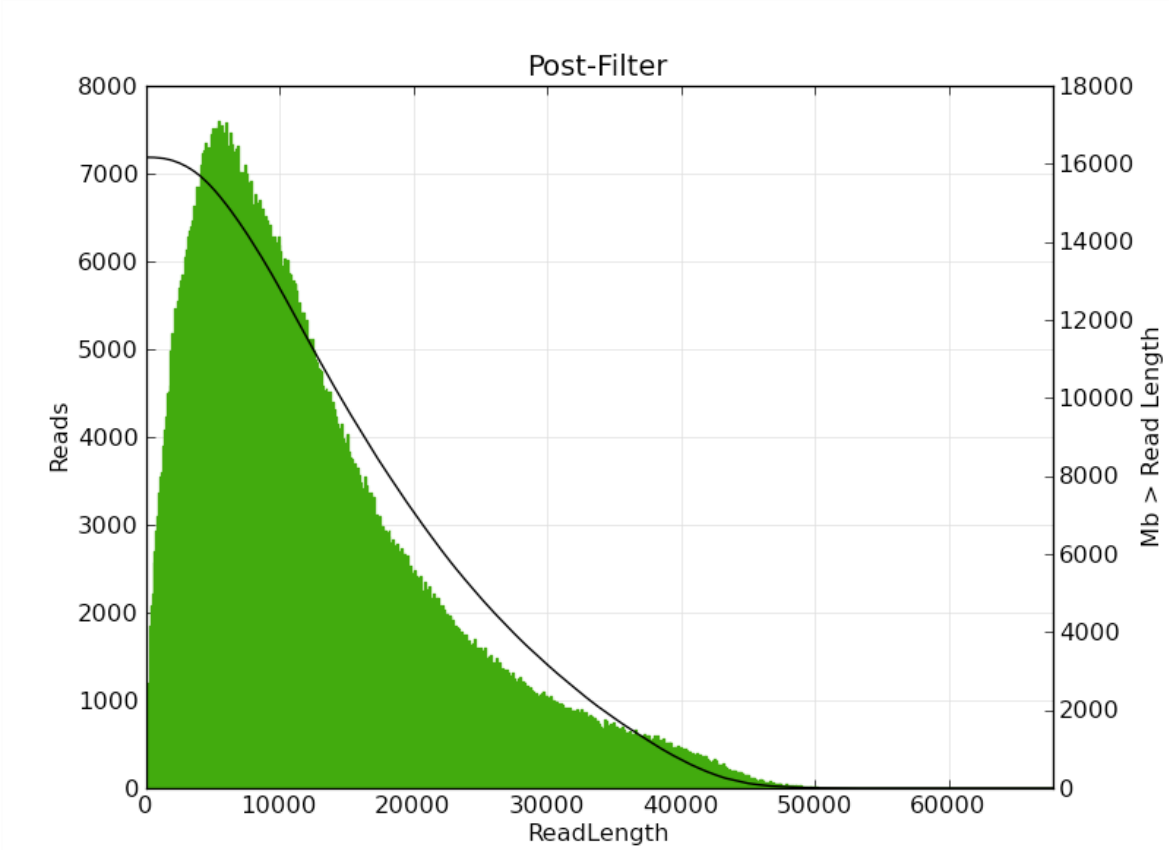
The Pairwise Sequentially Markovian Coalescent (*PSMC*) model [49] provides a tool to divide the genome into segments to estimate the distribution of the time to the most recent common ancestor (TMRCA) across the genome, by drawing on information from the varying local density of

heterozygous sites. Its usefulness was validated by simulations and it has been used in many recent studies (Prado-Martinez *et al.*, 2013 [91]; Zhao *et al.*, 2013 [92]; Hung *et al.*, 2014 [93]; Groenen *et al.*, 2012 [94]; Kelley *et al.*, 2014 [95]; Albert *et al.*, 2013 [96]). We set the parameters of *PSMC* as "-N25 -t500 -r5 -p "4+25*2+4+6"." The generation time was 20 years, and the mutation rate for each species was set as 1.6×10^{-8} per site per generation as estimated from analyses of divergence time (see in the section 2: Divergence time and whole-genome duplication). The result shows a decrease in effective population size over the last 10^5-10^6 years. We quantified uncertainty in our estimates by analyzing 100 bootstrap replicates for each individual shown in Supplementary Fig. 22.

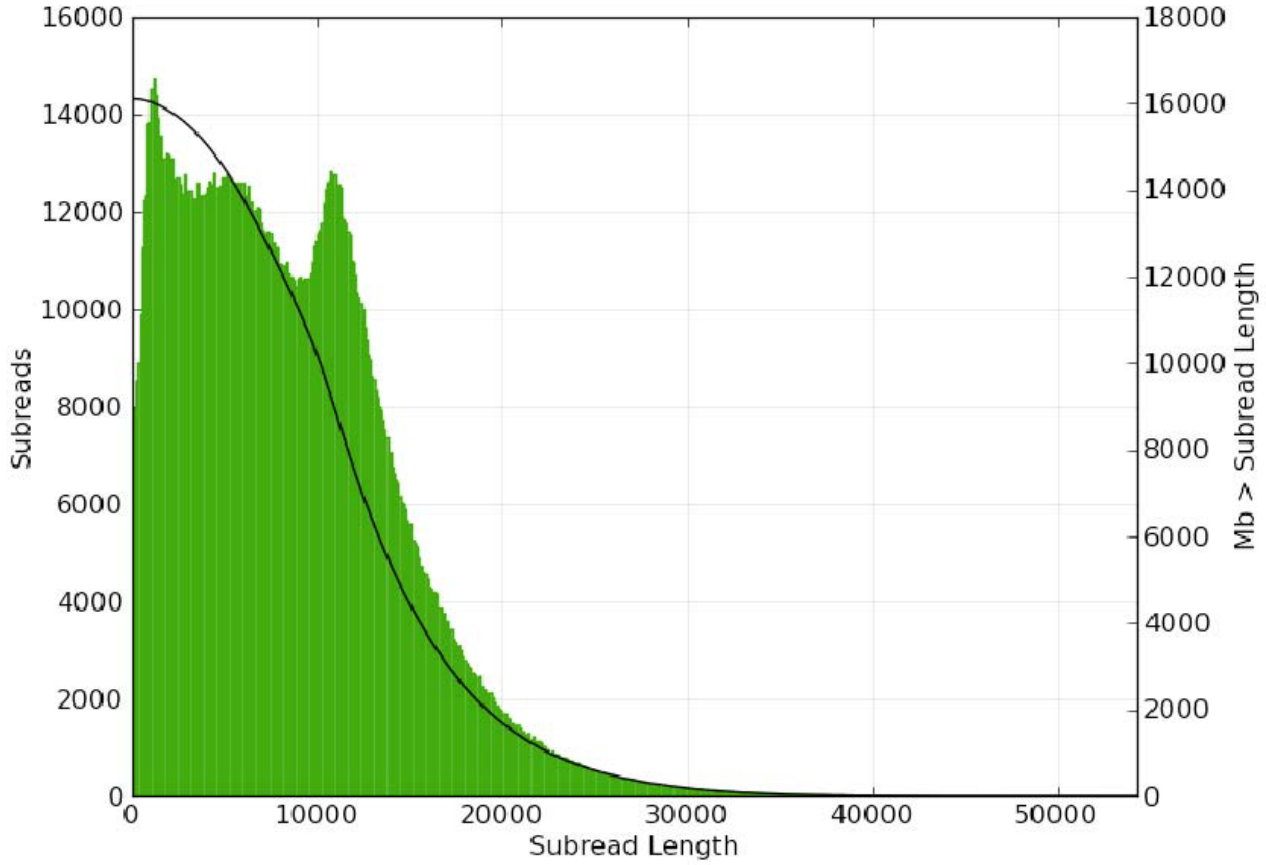
Supplementary Figures



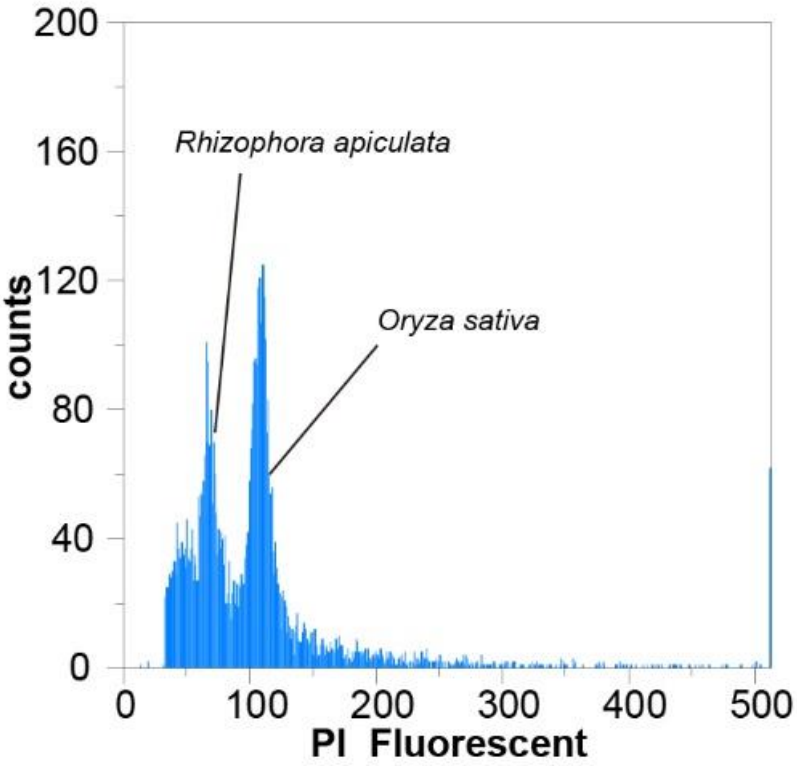
Supplementary Figure 1 | The workflow of PacBio SMRT library preparation.



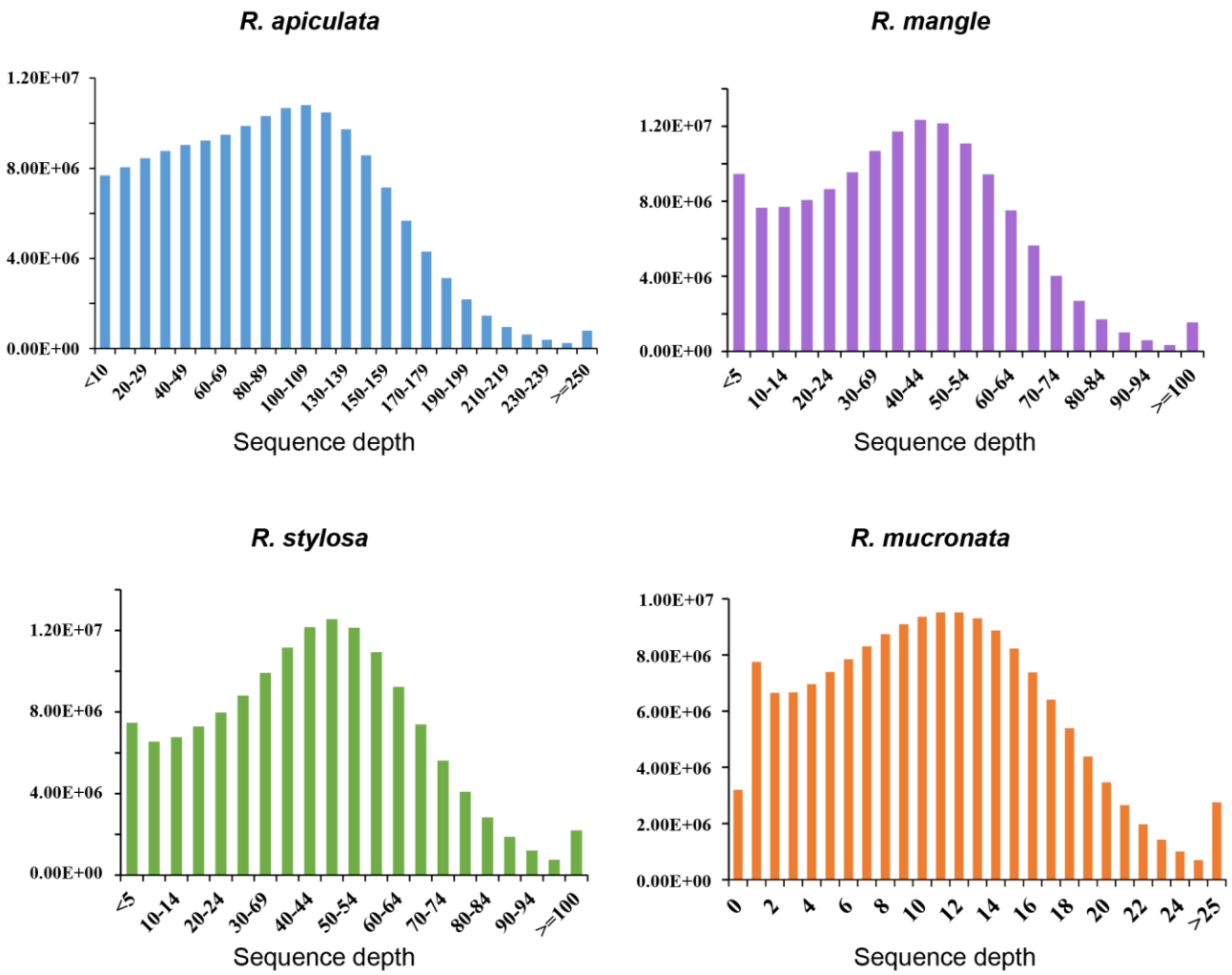
Supplementary Figure 2 | The length distribution of PacBio SMRT P6-C4 reads. Reads with quality value less than 0.75 were filtered. The x axis is read length. The green histogram shows the number of reads in bins of length intervals, while the black line shows number of reads with length larger than x bp. The mean length of the filtered reads is 12,866 bp, and the N50 is 18,074 bp.



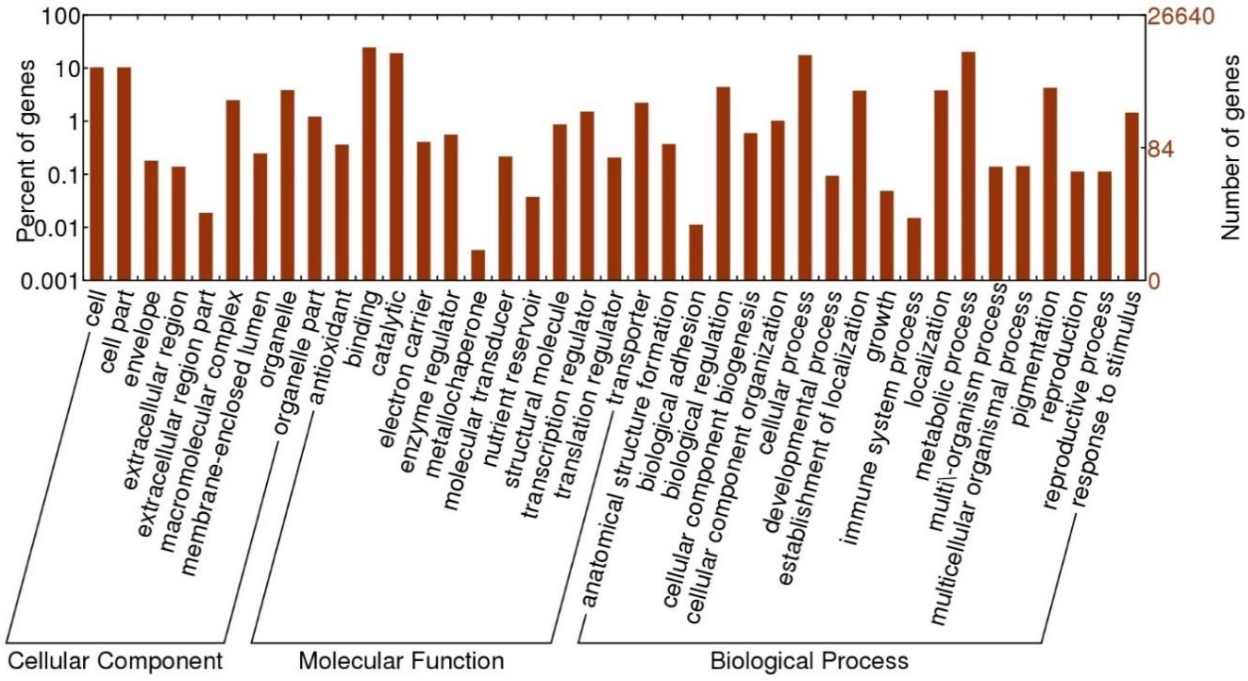
Supplementary Figure 3 | Histogram of PacBio SMRT P6-C4 subread lengths. Subread is generated when the adapters located within a read are removed. The x axis is subread length. The green histogram shows the number of reads in bins of length intervals, while the black line shows number of subreads with length larger than x bp.



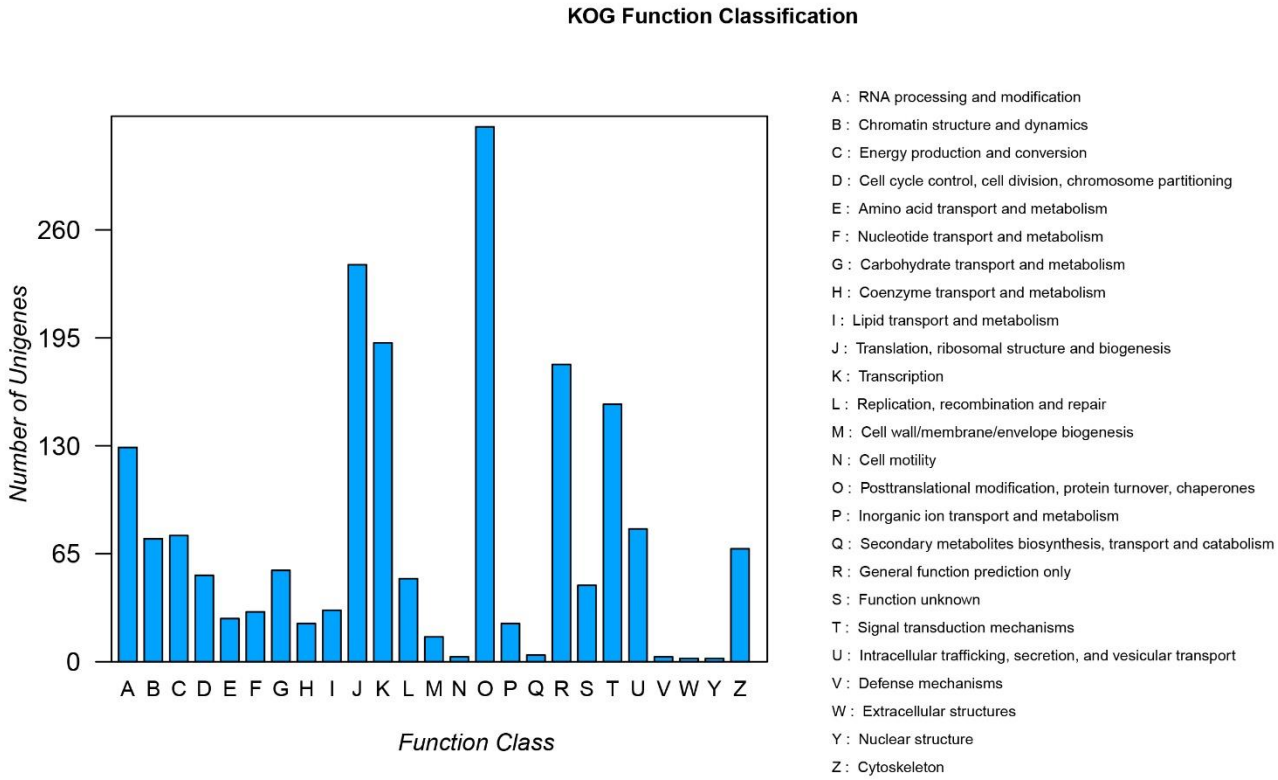
Supplementary Figure 4 | *Rhizophora apiculata* genome size estimation. Genome size was estimated by counting the number of nuclei in suspension using flow cytometry. *Oryza sativa* subsp. *japonica* cv. Nipponbare (1*C* = 442 Mb) was used as the internal standard. The genome size of *R. apiculata* was estimated basing on the relative value of PI Fluorescent of the two peaks.



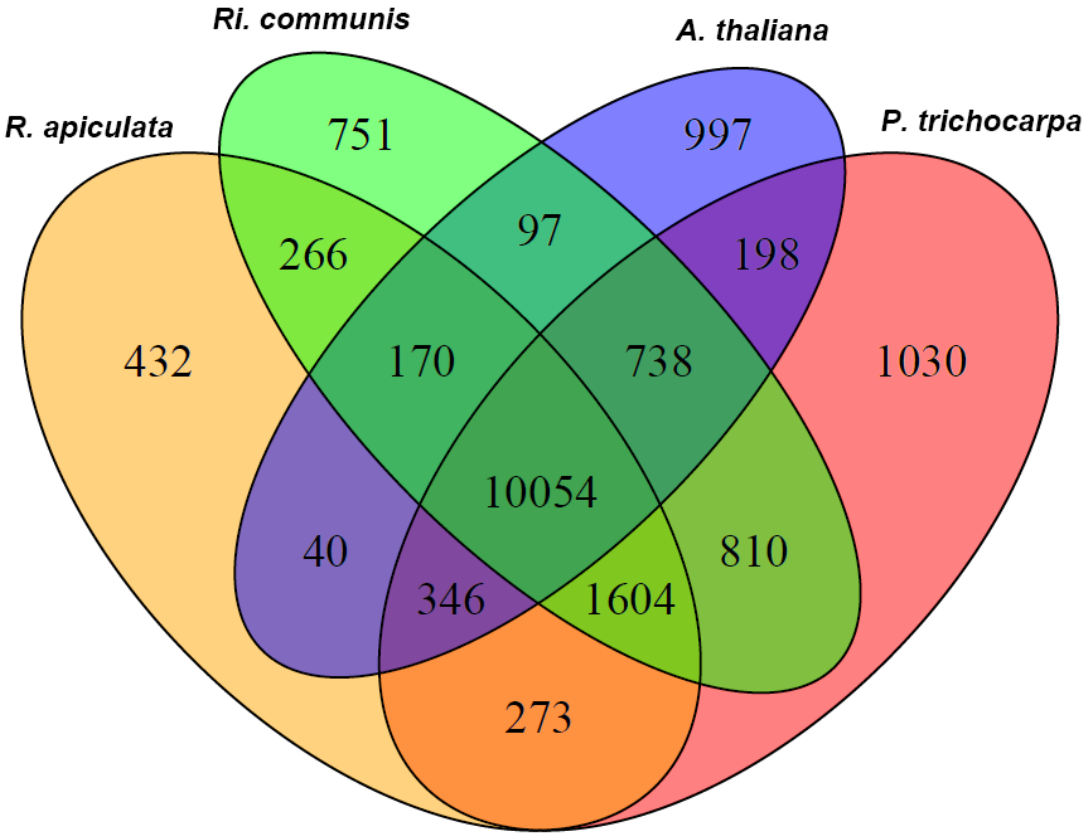
Supplementary Figure 5 | Sequencing depth distributions of four *Rhizophora* genomes. In each species, short reads from small insert-size libraries (200, 300, 400 and 600 bp in *R. apiculata*, 300 bp in other three species) were mapped to the *R. apiculata* reference genome using *bowtie2* [72].



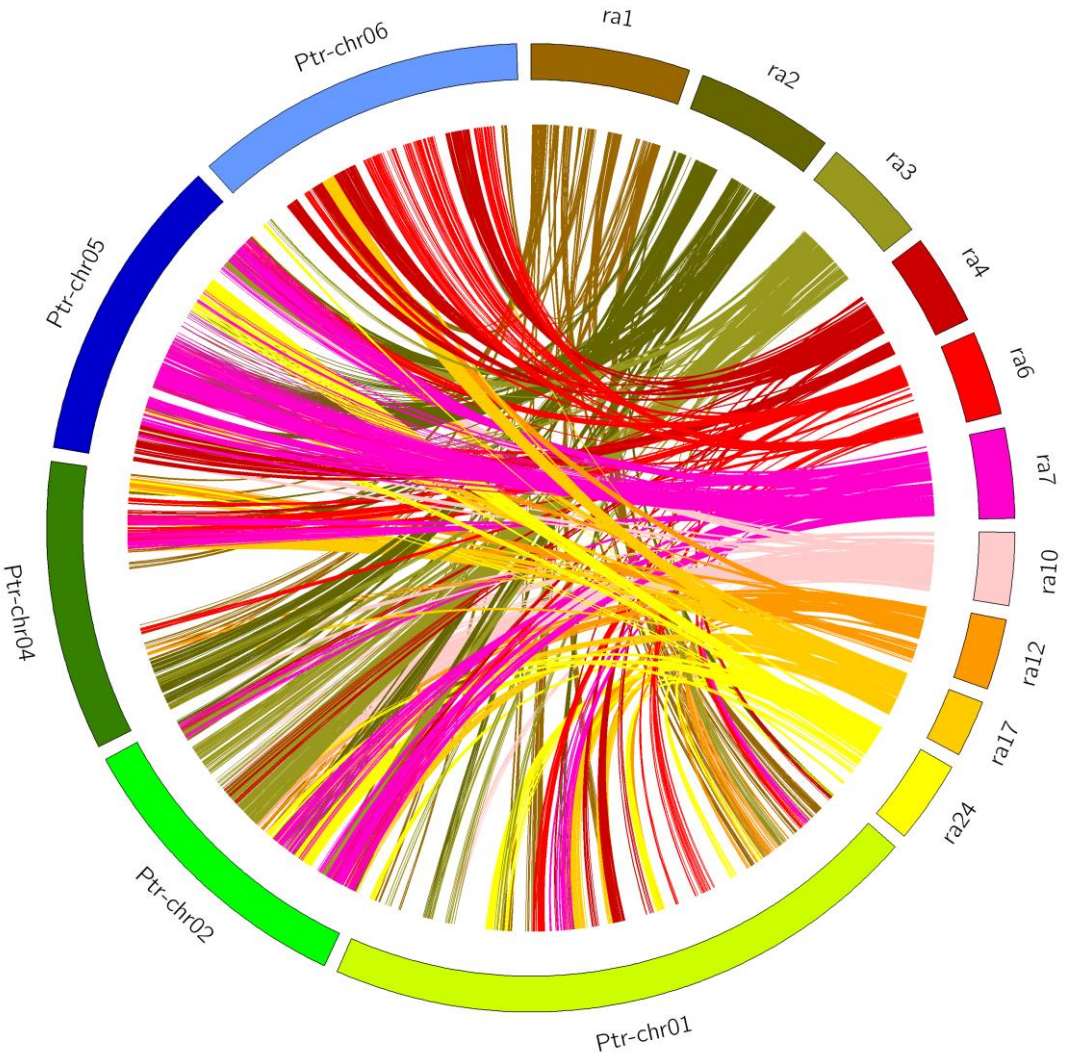
Supplementary Figure 6 | Gene Ontoloty (GO) annotation of protein-coding genes in the *R. apiculata* genome. The plot was generated using WEGO [97] (http://wego.genomics.org.cn/cgi-bin/wego/index.pl).



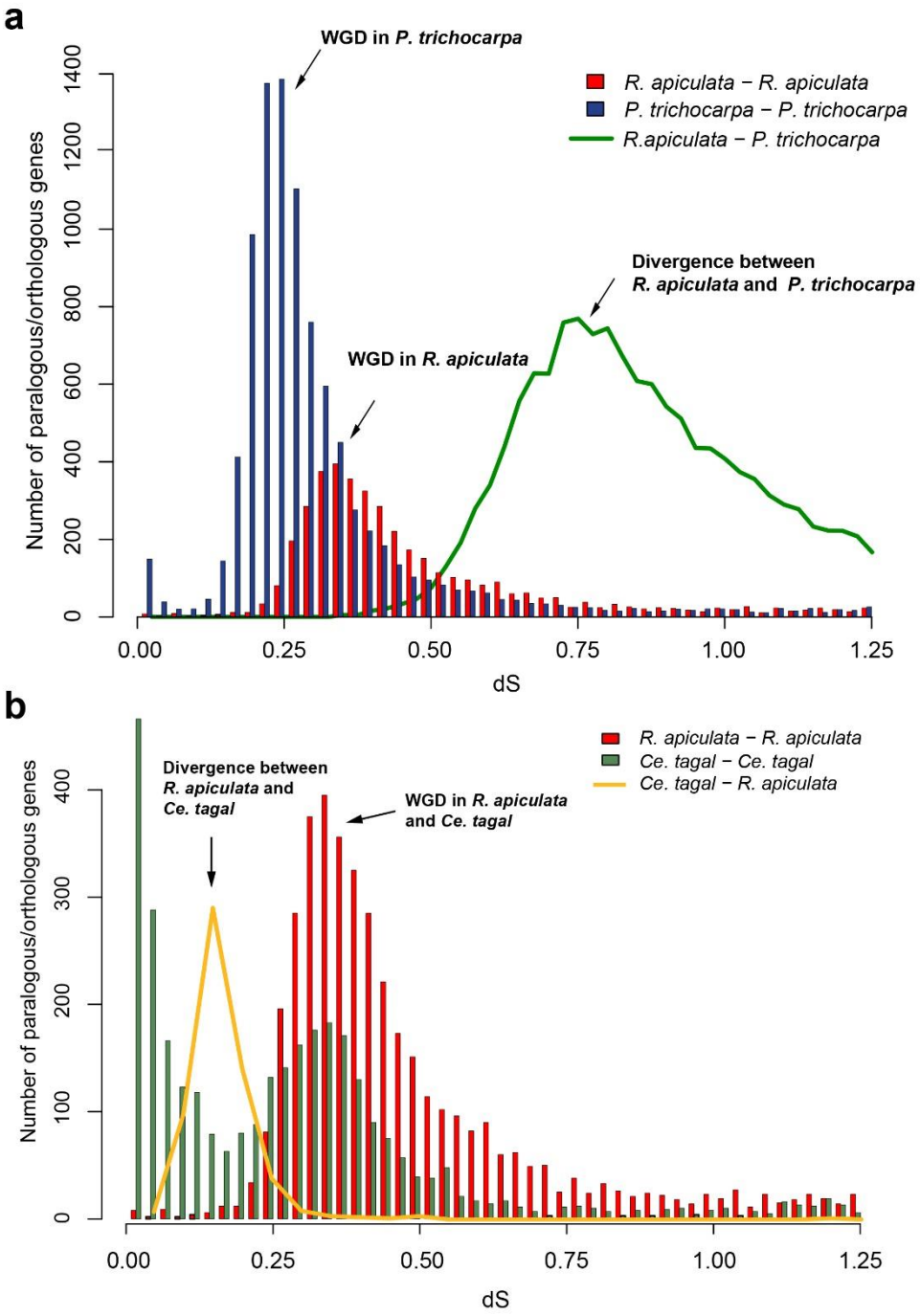
Supplementary Figure 7 | KOG (euKaryotic Orthologous Groups) classifications of protein-coding genes in the *R. apiculata* genome. The protein-coding genes were compared against the KOG database to assign functional classifications.



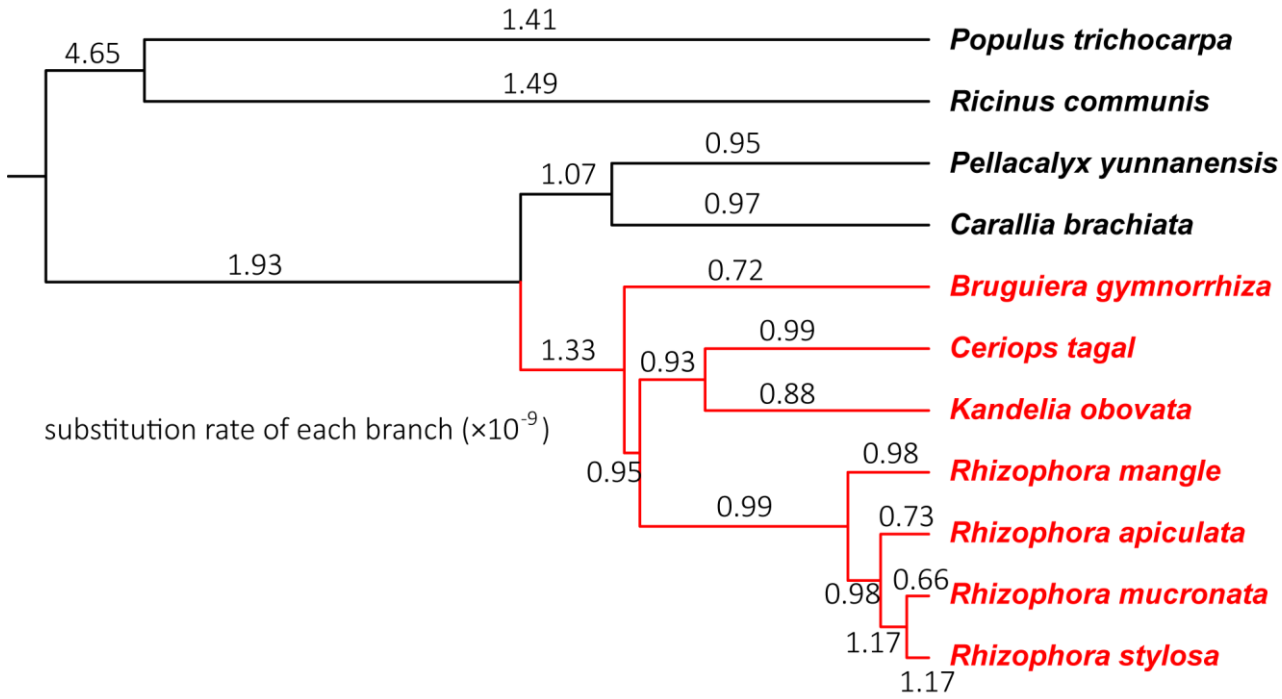
Supplementary Figure 8 | Shared and unique gene families in *R. apiculata*, *P. trichocarpa*, *Ri. communis* and *A. thaliana*. Gene family clustering was performed using *OrthoMCL* [70].



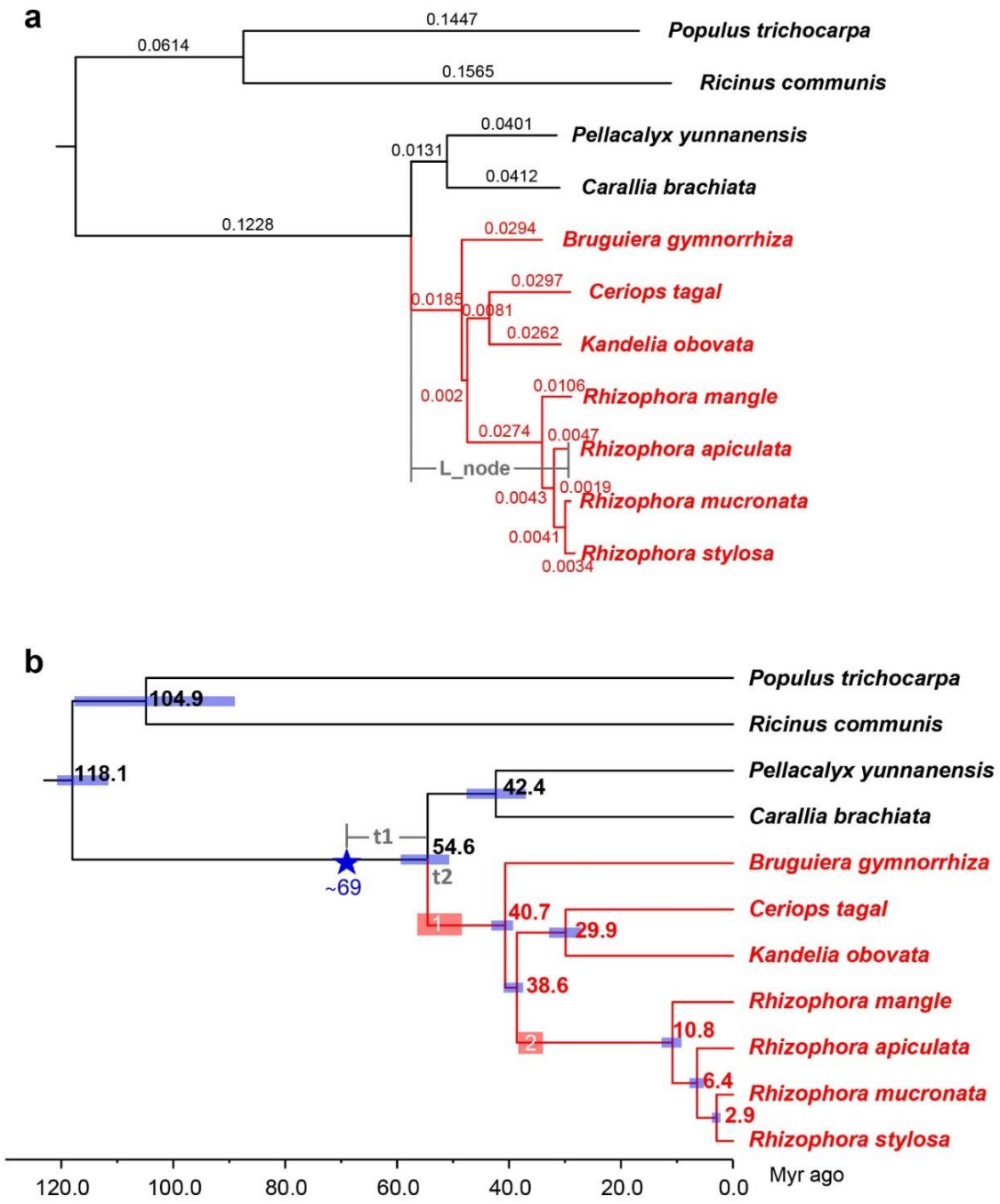
Supplementary Figure 9 | Inter-species syntenic blocks between *R. apiculata* and *P. trichocarpa*. The five chromosomes of *P. trichocarpa* and ten scaffolds of *R. apiculata* which contain the largest syntenic blocks are shown. Each line connects a pair of homologous genes and a cluster of such lines indicates a collinear block.



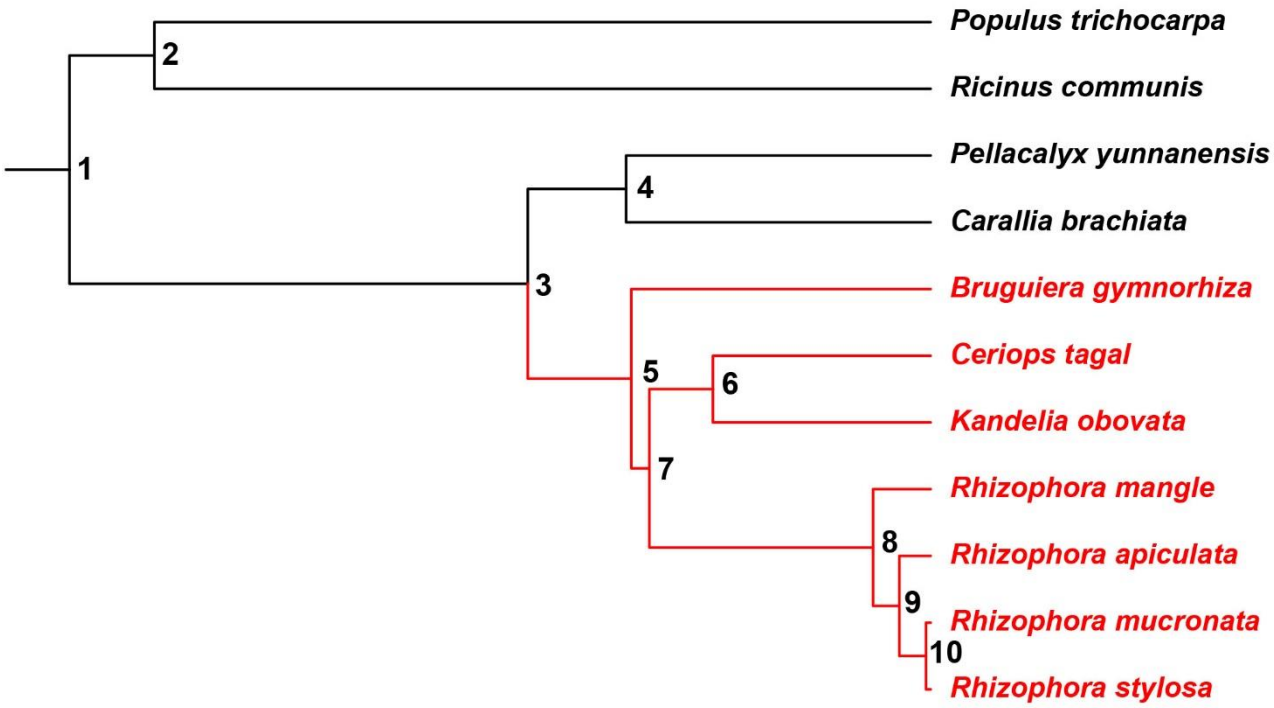
Supplementary Figure 10 | The distributions of inter-species (solid lines) and intra-species (bars) synonymous divergence (dS). Peaks of intra-species dS distribution indicate WGD events, and peaks of inter-species dS distribution indicate speciation events. **a**, We estimated that *R. apiculata* diverged from *P. trichocarpa* before the WGD event in Rhizophoreae clade. Consistent with the estimation, the dS peak (dS=0.35) within *R. apiculata* genome was absent in *P. trichocarpa*. Instead, a peak (dS=0.25) in the dS distribution of *P. trichocarpa* genome indicates another WGD event specific in *P. trichocarpa*. **b**, The WGD event which occurred before the divergence of *R. apiculata* and *Ce. tagal* was shared by the two species.



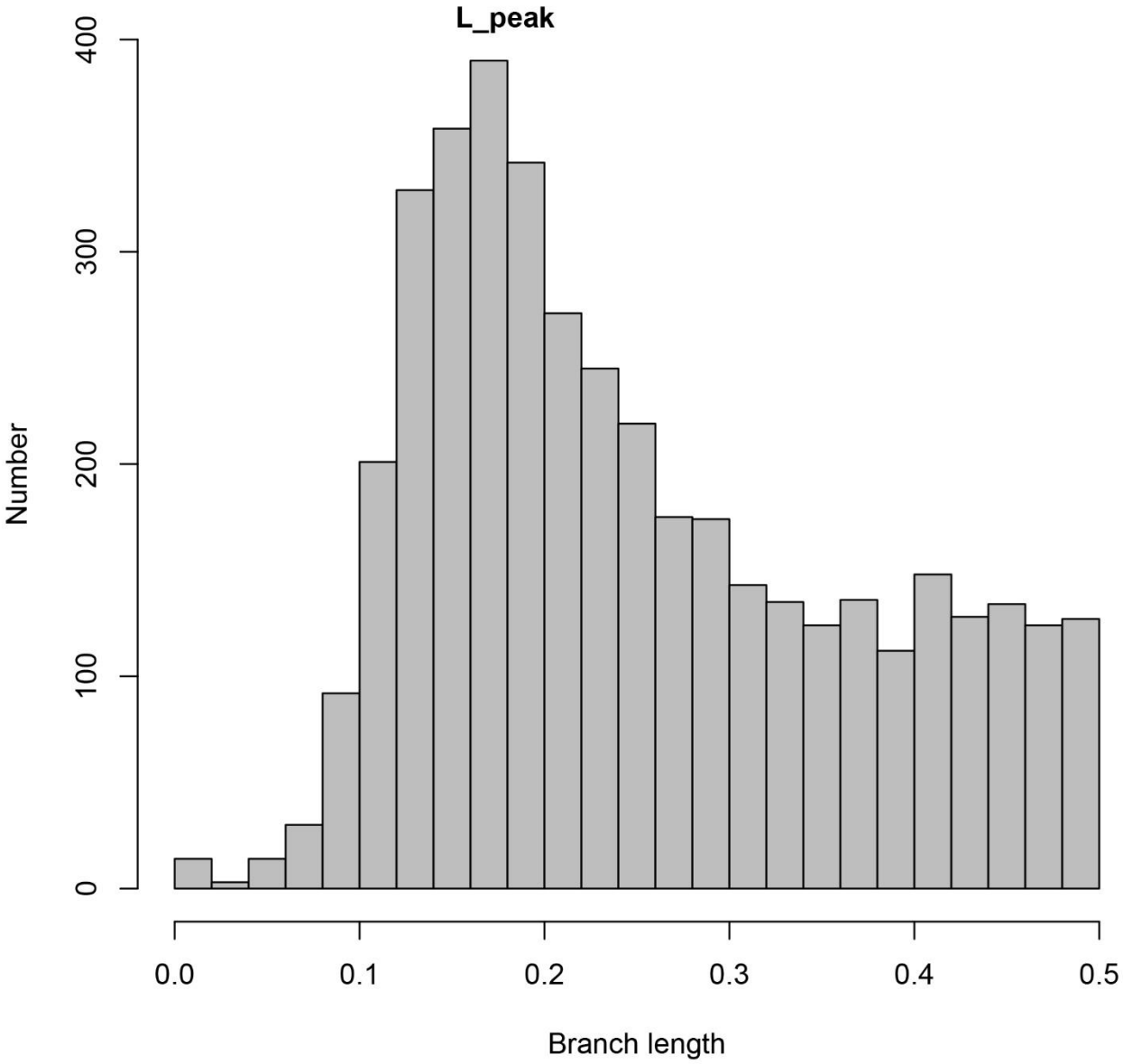
Supplementary Figure 11 | Substitution rate (per site per year) estimates for all lineages. In each branch, substitution rate was estimated via dividing branch length by the time span of the branch. Taxa in red are mangrove species.



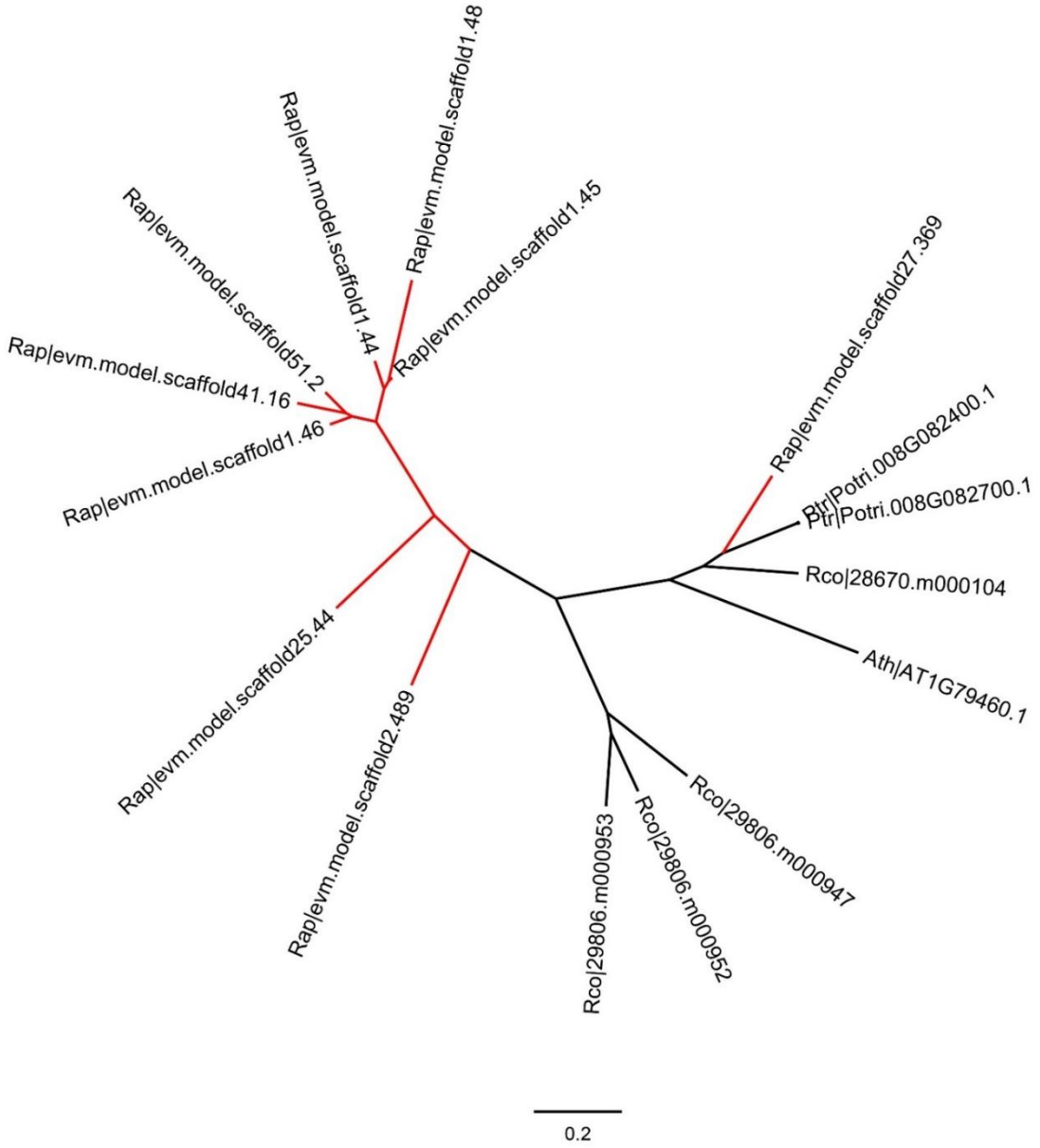
Supplementary Figure 12 | Dating the species divergence and whole-genome duplication event. **a**, Phylogenetic tree: the numbers above each branch represent branch lengths with the HKY85+gamma model and 1000 bootstraps. All nodes are 100% supported. Red branches and taxa annotations indicate mangrove species. **b**, The divergence times of the Rhizophoreae group. The blue bars show 95% credible intervals. Red branches and names represent mangrove species. The blue star shows the whole genome duplication event with the time below. Red rectangles with numbers represent the earliest known fossil records of mangrove lineages: 1) Hypocotyl fossils of *Bruguiera* are known from the London Clay and are identified as *Palaeobruguiera* in early Eocene (47.8-56 Myr ago) [27, 28]; 2) the oldest records of Rhizophora were the upper Eocene (33.9-38.0 Myr ago) [26, 27].



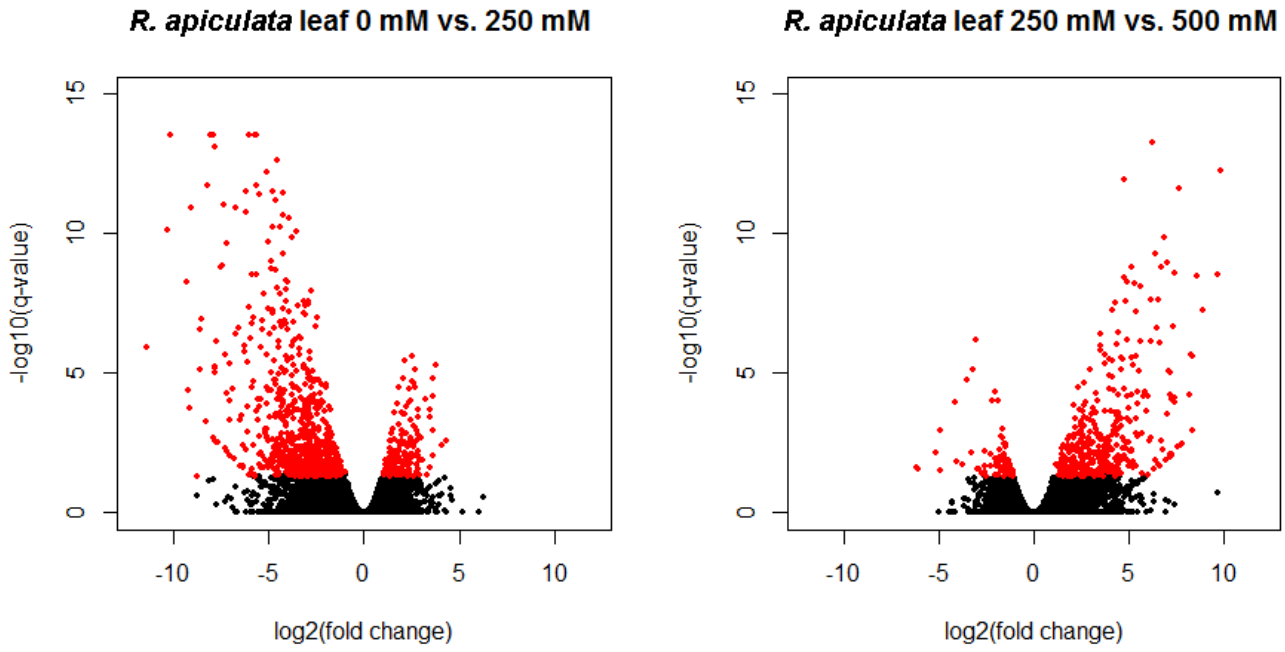
Supplementary Figure 13 | A phylogenetic tree showing node numbers. The nodes are numbered in order to present the divergence times of each node clearly in the Supplementary Tables 13-14.



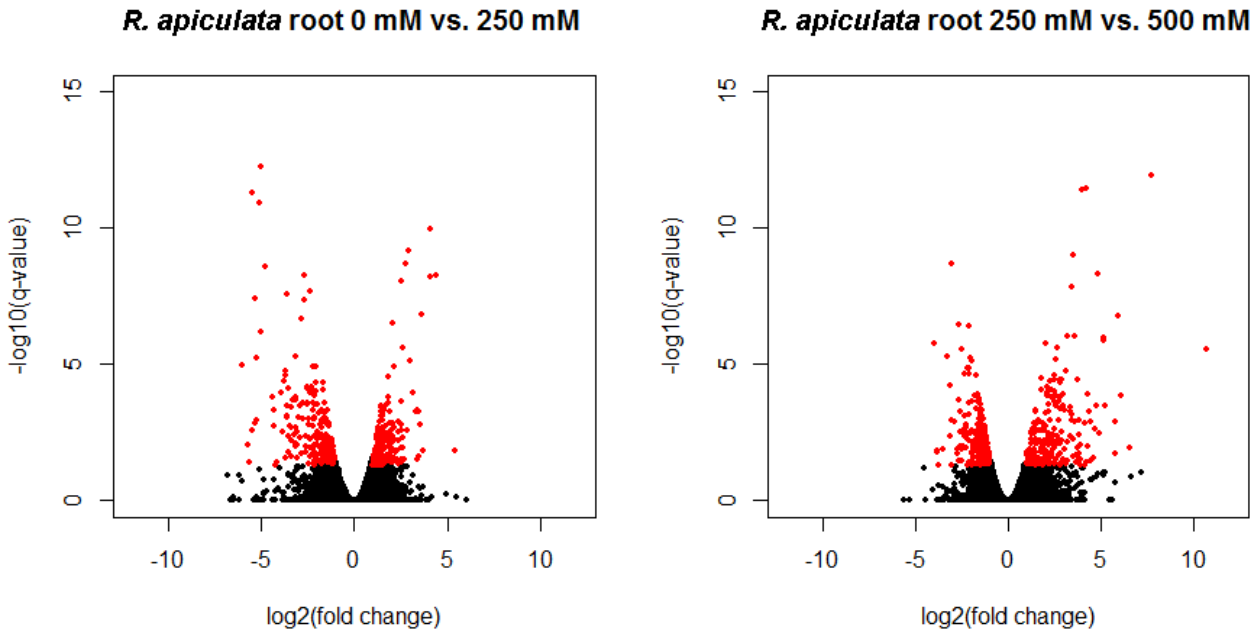
Supplementary Figure 14 | Branch length distribution of the orthologous gene pairs in *R. apiculata*. L_peak is 0.17 which was used to date the WGD event (see Supplementary Note).



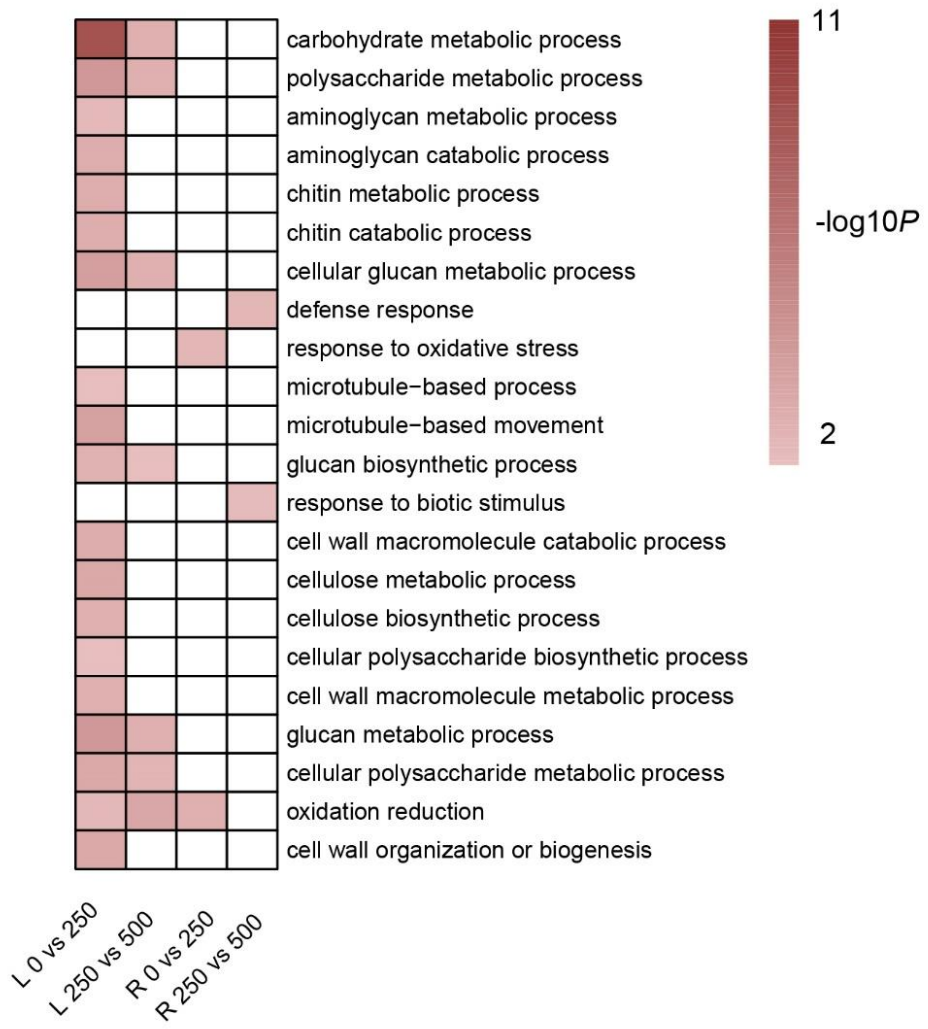
Supplementary Figure 15 | Copy number expansion of ent-kaurene synthase (KS) gene in *R. apiculata*. The annotation of each terminal branch consists of a three-letter abbreviation indicates the species and the gene ID. Namely, Rap represents *R. apiculata*, Ath represents *A. thaliana*, Rco represents *Ri. communis* and Ptr represents *P. trichocarpa*. The nine copies of KS genes in *R. apiculata* were clustered into two clades. One clade (clade A) includes one copy and its orthologs t in *A. thaliana* and other species, the other clade (clade B) includes eight paralogous copies of *R. apiculata*. The two clades diverged before the divergence of *R. apiculata* and *A. thaliana*. We excluded the possibility that the clade B are different genes by searching against the *A. thaliana*'s proteome. The copy number expansion is due to the tandem duplication in clade B., which occurred after the divergence between *R. apiculata* and *Ca. brachiata*.



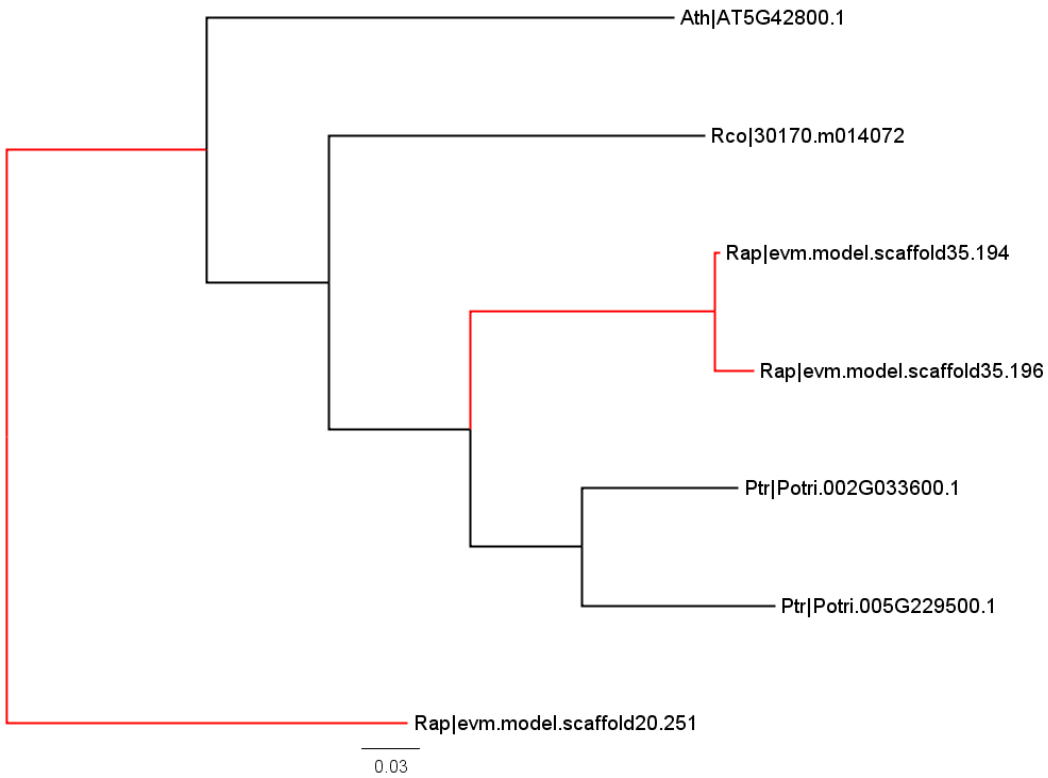
Supplementary Figure 16 | Volcano plots of gene expression profiles in leaves of *R. apiculata*. Red dots represent significantly differentially expressed genes (with q-value < 0.05 and fold change > 2). The left panel shows gene expression level changes when concentration of NaCl changed from 0 mM to 250 mM, and the right panel shows gene expression level changes when concentration of NaCl changed from 250 mM to 500 mM.



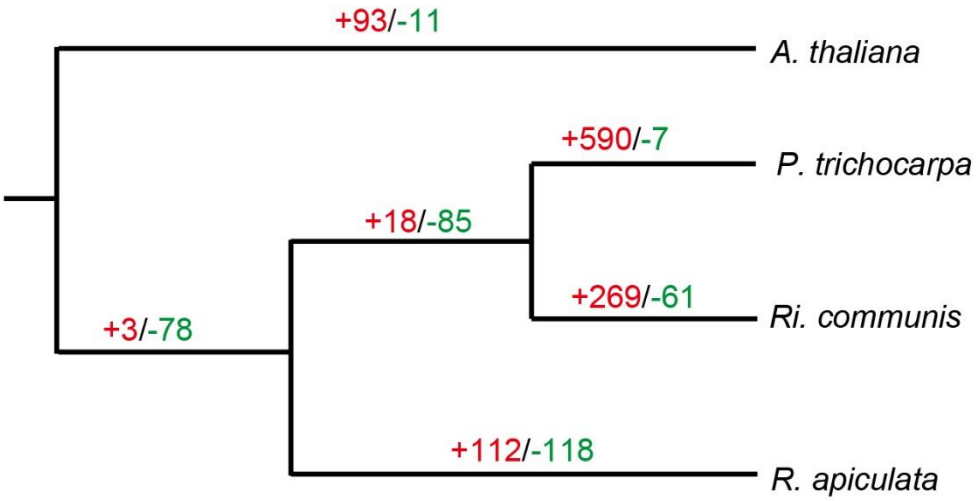
Supplementary Figure 17 | Volcano plot of gene expression profiles in roots of *R. apiculata*. Red dots represent significantly differentially expressed genes (with q-value < 0.05 and fold change > 2). The left panel shows gene expression level changes when concentration of NaCl changed from 0 mM to 250 mM, and the right panel shows gene expression level changes when concentration of NaCl changed from 250 mM to 500 mM.



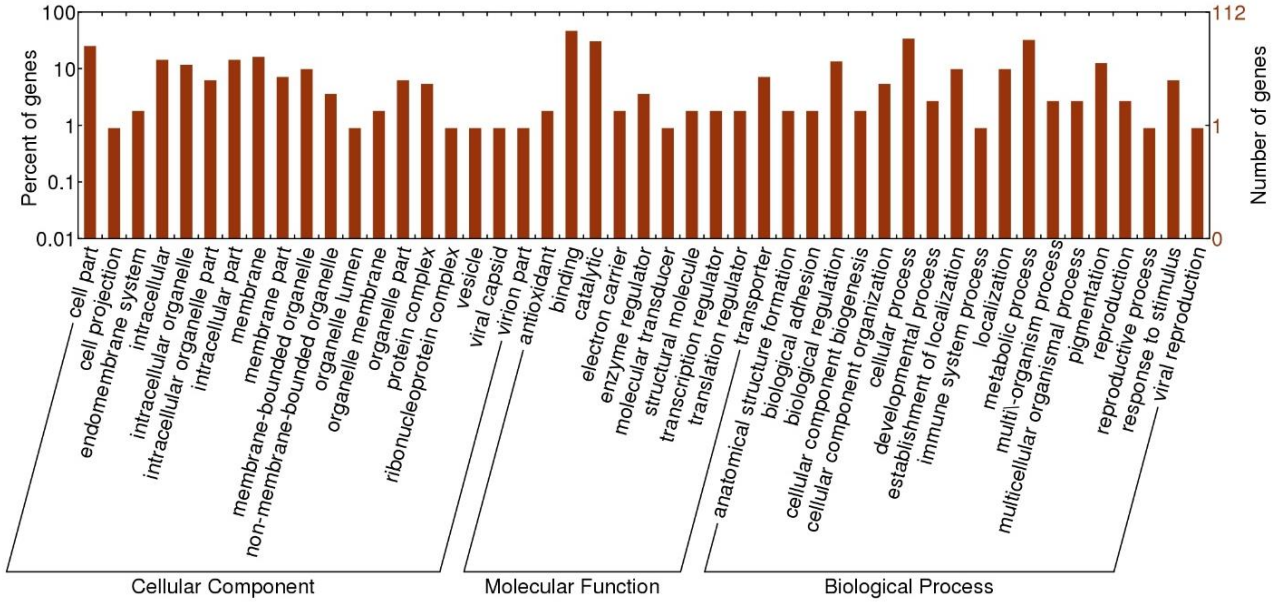
Supplementary Figure 18 | GO terms significantly enriched in differently expressed genes. Each row represents a GO term, and each column represents a pair of treatments. L, leaf. R, root. The GO enrichment analysis was performed using agriGO [98].



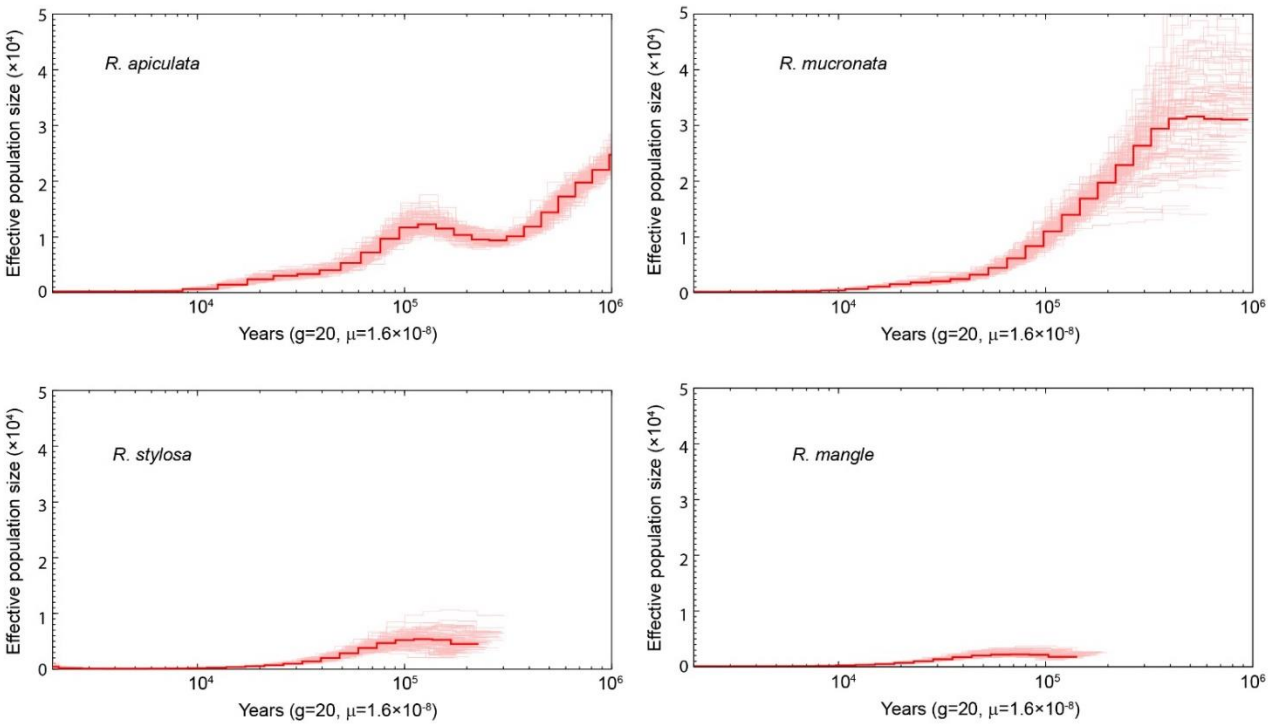
Supplementary Figure 19 | *DFR* (dihydroflavonol reductase) gene in *R. apiculata* and three inland plants. The annotation of each terminal branch consists of a three-letter abbreviation indicates the species and the gene ID. Rap represents *R. apiculata*, Ath represents *A. thaliana*, Rco represents *Ri. communis* and Ptr represents P. *trichocarpa*.



Supplementary Figure 20 | Gene family expansion and contraction in *R. apiculata* and related inland species. The expanded or contracted gene families were detected by a stochastic birth and death process of software CAFE [46]. The computed numbers for the expanded (red) and contracted (green) gene families are shown above each branch.



Supplementary Figure 21 | GO (level-3) enrichment analysis of the expanded gene families in *R. apiculata*. The plot was generated using WEGO [97] (http://wego.genomics.org.cn/cgi-bin/wego/index.pl).



Supplementary Figure 22 | Estimated historical effective population size of *Rhizophora* species using PSMC. The x-axes represent time before the present, and y-axes represent effective population size. The solid red line shows the consensus results, while the pale lines are generated from 100 bootstrap replicates. Generation time (g) is set to 20 years and mutation rate (μ) is 1.6×10^{-8} /bp/generation.

Supplementary Tables

Supplementary Table 1 | Summary of SMRT sequencing data for *R. apiculata*.

Metrics	Pre-Filter	Post-Filter
Polymerase Read Bases (bp)	16,809,960,989	16,229,617,012
Polymerase Reads	2,554,964	1,261,422
Polymerase Read N50 (bp)	17,905	18,074
Polymerase Read Length (bp)	6,579	12,866
Polymerase Read Quality	0.431	0.84

Supplementary Table 2 | Statistics for the final assembly of *R. apiculata*.

	Contigs	Scaffolds
Number	189	142
N50 (bp)	4,319,773	5,420,131
Counts of N50	20	16
N90 (bp)	748,914	1,010,585
Counts of N90	66	48
Shortest (bp)	4,356	5,078
Longest (bp)	10,234,006	13,354,123
Total length (bp)	232,002,840	232,055,149

Supplementary Table 3 | Summary of DNA Illumina libraries and sequencing data for *R. apiculata*.

Libraries	Insert size (bp)	Library number	Read length (bp)	Total data (Gb)
200	180-220	1	100	10
300	280-320	1	100	7.4
400	380-420	1	100	14
600	580-620	1	100	7.9
2 k	500-3,000	2	100	18
5 k	2,000-8,000	2	100	22
10 k	5,000-15,000	2	100	10
Total	-	10	-	89.3

Supplementary Table 4 | Statistics of the assembly of *R. apiculata* using only Illumina short reads.

	Contigs (bp)	Scaffolds (bp)
Number	32,160	3,316
N50	9,726	1,461,497
N90	2,587	151,294
Mean	5,983	67,132
Median	3,807	2,054
Shortest	1,000	1,000
Longest	91,252	5,869,168
Total length	192,443,154	222,608,739

Supplementary Table 5 | Summary of sequencing data for *R. mangle, R. stylosa* and *R. mucronata*.

Species	Insert-size (bp)	Total data (Gb)	Mapping rate (%)
R. mangle	300	15.21	96.85%
R. stylosa	300	15.79	98.95%
R. mucronata	300	3.10	98.29%

Supplementary Table 6 | Data sets used in this study.

Name	Data type	Data size (Gb)	Reference	Accession number
Rhizophora apiculata	genome	105.5	This study	PRJEB8423
Rhizophora mucronata	genome	3.10	This study	PRJEB20990
Rhizophora stylosa	genome	15.79	This study	PRJEB20992
Rhizophora mangle	genome	15.21	This study	PRJEB21001
Populus trichocarpa	genome	--	Tuskan *et al.*, 2006 [84]	--
Ricinus communis	genome	--	Chan *et al.*, 2010 [83]	--
Carallia brachiata	transcriptome	2.46	Guo *et al.*, 2017 [15]	SRP093193
Pellacalyx yunnanensis	transcriptome	4.01	Yang *et al.*, 2015 [14]	SRP056405
Bruguiera gymnorrhiza	transcriptome	2.46	Guo *et al.*, 2017 [15]	SRP093193
Kandelia obovata	transcriptome	2.33	Guo *et al.*, 2017 [15]	SRP093193
Ceriops tagal	transcriptome	4.31	Yang *et al.*, 2015 [14]	SRP056405

Supplementary Table 7 | Summary of genome completeness assessment.

Species		*R. apiculata*
Transcript mapping	**Number of transcripts**	151,828
	Mapped transcripts	147,205 (96.95%)
Core eukaryotic gene mapping	**Number of core eukaryotic genes**	458
	Mapped core eukaryotic genes	428 (93.45%)
Sanger sequence mapping	**Number of genes**	79
	Genes with unique mapped position	78 (98.73%)

Supplementary Table 8 | Repeat element statistics of the *R. apiculata* genome.

Type	Number of elements	% of genome
DNA elements	4,845	0.93
LTR elements	44,591	18.02
SINEs	1,696	0.14
LINEs	3,649	0.53
Unclassified	66,502	10.08
Total	121,283	29.69

Supplementary Table 9 | Predicted protein-coding gene statistics in *R. apiculata*.

Species	***R. apiculata***
Gene number	26,640
Average gene length (bp)	2,838
Average CDS length (bp)	1,179
Average exon number	5.3
Average exon length (bp)	221
Average intron length (bp)	383

Supplementary Table 10 | Non-coding RNA identification.

Categories of ncRNA	**Number of genes**
rRNA	1,908
miRNA	221
snoRNA	173
tRNA	451
snRNA	51
Intron	132
SRP	8
others	11

Supplementary Table 11 | Protein-coding gene annotation statistics.

Species	***R. apiculata***
Protein number	26,640
Mean length (AA)	393
Min length (AA)	49
Median length (AA)	318
Max length (AA)	5,345
Total length (AA)	10,468,087
Annotated number	25,271
Number with KEGG annotation	17,869
Number with GO annotation	16,145

AA, amino acid.

Supplementary Table 12 | Transcription factors in *R. apiculata*.

TF family	*R. apiculata*	*A. thaliana*	*P. trichocarpa*	*Ri. communis*
MYB	143	206	295	160
NAC	132	112	170	95
bHLH	127	134	171	103
C2H2	118	110	154	97
HB	104	93	127	69
AP2-EREBP	89	145	210	115
bZIP	84	73	91	50
WRKY	77	72	102	58
C3H	74	68	93	52
GRAS	63	34	106	48
LOB	53	43	58	34
ABI3VP1	48	66	120	40
G2-like	46	41	63	32
CCAAT	45	46	62	41
Trihelix	38	26	54	27
MADS	35	108	102	39
mTERF	34	35	55	35
C2C2-GATA	32	30	39	19
FAR1	31	17	51	20
C2C2-Dof	30	36	45	23
TCP	29	24	37	22
ARF	28	23	36	18
HSF	26	24	29	18

OFP	23	17	28	16
SBP	21	17	30	15
FHA	19	17	23	17
zf-HD	18	17	21	11
Tify	17	15	17	11
PLATZ	14	12	21	11
GRF	13	9	19	9
BSD	12	11	14	9
C2C2-YABBY	12	6	12	6
TUB	12	11	12	7
RWP-RK	12	14	20	10
ARR-B	12	14	17	11
GeBP	9	22	6	4
BES1	9	8	14	7
Alfin-like	8	7	9	5
C2C2-CO-like	8	15	11	7
E2F-DP	8	8	9	6
CPP	8	8	12	6
LIM	6	6	12	6
Sigma70-like	6	6	9	5
BBR/BPC	6	7	16	5
SRS	6	11	10	5
CSD	6	4	6	5
TAZ	5	8	7	3
CAMTA	5	6	7	4
DBP	5	2	4	2
EIL	4	6	7	4
ULT	3	2	2	2
PBF-2-like	2	3	3	2
VOZ	2	2	4	1
LFY	2	1	1	1
S1Fa-like	1	3	2	1
HRT	1	2	1	1
NOZZLE	1	1	3	1
SAP	1	1	1	1
TIG	0	0	0	1
Total	1,783	1,433	1,865	2,660

Supplementary Table 13 | Divergence time and credible intervals of each node in the Rhizophoreae group using *MCMCTREE*.

Node	codon123	codon12
1	118.1 [111.6,120.8]	118.0 [111.3,120.8]
2	104.9 [89.0,117.7]	102.8[84.9,117.4]
3	54.6 [50.7,59.4]	55.5 [51.0,61.3]
4	42.4 [37.0,47.6]	41.5 [35.6,47.9]
5	40.7 [39.3,43.2]	40.6 [39.2,43.3]
6	29.9 [27.0,32.9]	30.1 [26.6,33.5]
7	38.6 [37.5, 41.0]	38.7 [37.5, 41.2]
8	10.8 [9.2,12.8]	10.1 [8.2,12.3]
9	6.4 [5.2,7.8]	5.8 [4.6,7.3]
10	2.9 [2.2,3.8]	2.4 [1.8,3.1]

Divergence time and 95% credible interval of each node is shown. Time units: million years. Nodes were marked in Supplementary Fig. 13.
Codon123, the dataset includes all three codon positions. Codon12, the dataset includes the first and second codon positions.

Supplementary Table 14 | Divergence time of each node of the Rhizophoreae group using *r8s*.

Node	**Results of *MCMCTREE***	**HKY85+G in *r8s***			**GTR+I+G in *r8s***		
		PL	LF	NPRS	PL	LF	NPRS
1	118.1 [111.6,120.8]	120.00	120.00	120.00	120.00	120.00	120.00
2	104.9 [89.0,117.7]	87.36	88.84	85.34	87.44	88.85	85.35
3	54.6 [50.7,59.4]	48.11	47.38	53.13	48.09	47.39	53.15
4	42.4 [37.0,47.6]	34.98	33.92	42.34	34.95	33.93	42.37
5	40.7 [39.3,43.2]	38.81	38.72	39.73	38.81	38.72	39.73
6	29.9 [27.0,32.9]	27.72	26.37	29.91	27.66	26.38	29.91
7	38.6 [37.5, 41.0]	38.00	38.00	38.00	38.00	38.00	38.00
8	10.8 [9.2,12.8]	9.22	8.53	11.42	9.19	8.53	11.41
9	6.4 [5.2,7.8]	5.07	4.59	6.76	5.05	4.59	6.76
10	2.9 [2.2,3.8]	2.06	1.84	2.72	2.05	1.84	2.72

Nodes were marked in Supplementary Fig. 13. Time units: 1 million years.

Supplementary Table 15 | Genes under positive selection in *R. apiculata* and involved in embryo development, ABA biosynthesis and signaling, Ethylene biosynthesis and signaling, which are supposed to relate with vivipary.

Gene name	ID in *A. thaliana*	ID in *R. apiculata*	P value	FDR
EMB2279	AT1G30610.1	evm.model.scaffold1.42	0.00E+00	0.00E+00
XBAT32	AT5G57740.1	evm.model.scaffold9.546	0.00E+00	0.00E+00
TIF3H1	AT1G10840.1	evm.model.scaffold12.888	1.11E-16	2.66E-14
FTA	AT3G59380.1	evm.model.scaffold77.23	1.22E-15	1.99E-13
EMB2768	AT3G02660.1	evm.model.scaffold1.590	5.62E-14	6.74E-12
EIN2	AT5G03280.1	evm.model.scaffold4.466	4.01E-13	4.30E-11
RCE1	AT4G36800.1	evm.model.scaffold7.804	8.87E-10	5.65E-08
DCP2	AT5G13570.2	evm.model.scaffold11.565	3.21E-08	1.64E-06
RRP5	AT3G11964.1	evm.model.scaffold23.64	7.26E-06	2.31E-04
OVA4	AT2G25840.2	evm.model.scaffold9.431	1.07E-05	3.39E-04
LEA family protein	AT3G62580.1	evm.model.scaffold3.530	2.13E-05	6.48E-04
EMB3137	AT5G14320.1	evm.model.scaffold13.99	4.01E-05	1.16E-03
HDC1	AT5G08450.1	evm.model.scaffold5.154	5.14E-05	1.46E-03
GCT	AT1G55325.2	evm.model.scaffold13.252	5.62E-05	1.58E-03
RID3	AT3G49180.1	evm.model.scaffold35.25	1.97E-04	4.91E-03
PYR4	AT4G22930.1	evm.model.scaffold3.860	2.19E-04	5.34E-03
PP2C family protein	AT2G40860.1	evm.model.scaffold13.441	6.77E-04	1.51E-02
ISS1	AT1G80360.1	evm.model.scaffold9.136	1.42E-03	3.04E-02
ABA3	AT1G16540.1	evm.model.scaffold20.63	1.68E-03	3.54E-02

Supplementary Table 16 | Gene Ontology terms with two-fold or larger increase in median dN/dS in *R. apiculata* compared to *Ca. brachiata*.

GO term	Description	dN/dS in *Ca. brachiata*	dN/dS in *R. apiculata*
GO:0016458	gene silencing	0.108	0.250
GO:0044419	competition with other organisms	0.089	0.204
GO:0045454	cell redox homeostasis	0.146	0.333
GO:0016585	chromatin remodeling	0.099	0.220
GO:0048193	Golgi vesicle transport	0.107	0.238
GO:0019725	cellular homeostasis	0.128	0.264
GO:0006397	mRNA processing	0.097	0.197
GO:0016071	mRNA metabolic process	0.097	0.197

Supplementary Table 17 | Genes in the Seedgene database positively selected in the ancestral branch of Rhizophoreae mangroves.

Gene ID in *A. thaliana*	Gene ID in R. *apiculata*	Description	p-value
AT5G39980.1	evm.model.scaffold38.185	Tetratricopeptide repeat (PPR) protein	7.12E-51
AT4G29860.1	evm.model.scaffold7.738	Embryo defective 2757 (EMB2757)	9.83E-05
AT2G37560.1	evm.model.scaffold4.160	Origin Recognition Complex Subunit 2 (ORC2)	7.37E-04
AT4G32260.1	evm.model.scaffold15.236	Pigment defective 334 (PDE334)	4.93E-03
AT2G21470.2	evm.model.scaffold17.155	SUMO activating enzyme 2 (SAE2)	5.01E-03

Supplementary Table 18 | Copy number of genes encoding key enzymes of the flavonoid biosynthesis pathway. For each enzyme, we used the orthologous genes annotated in *A. thaliana* as query to search against other three genomes. For F3’5’H and LAR, which are missing in *A. thaliana*, the copies in *P. trhichocarpa* were used to search against other two genomes.

Gene name	Enzyme	Gene’s copy number of flavonoid synthesis			
		A. thaliana	*P. trichocarpa*	*R. apiculata*	*Ri. communis*
PAL	Phenylalanone Ammonia-lyase	4	5	4	5
C4H	Cinnamate 4-hydroxylase	1	3	3	2
4CL	Coumarate-4-CoA ligase	4	6	4	3
CHS/STSY	Chalcone synthase	1	8	4	3
CHI	Chalcone isomerase	1	1	1	1
F3H	Flavanone 3-hydroxylase	1	3	2	1
F3’H	Flavanone 3’-hydroxylase	1	1	2	1
F3’5’H	Flavanone 3’5’-hydroxylase	0	2	5	0
FLS	Flavonol synthase	1	5	2	2
DFR	Dihydroflavonol-4-reductase	1	2	1	1
LDOX/ANS	Leucoanthocyanidin dioxygenase	1	2	2	1
LAR	Leucoanthocyanidin reductase	0	2	2	1
ANR	Anthocyanidin reductase	1	2	1	1
UFGT	UDP-glucose: flavanoid 3-O-glucosyltransferase	11	1	4	1

Supplementary Table 19 | Differentially expressed genes in the flavonoid biosynthesis pathway. The transcript levels were determined by fragments per kilobase of exon per million fragments mapped (FPKM).

Gene ID in *R. apiculata*	Gene name	Tissue	Salt concentration change (mM/L) (treatment 1-2)	FPKM value in treatment 1	FPKM value in treatment 2	p-value	q-value
evm.model.scaffold6.336	4CL	Leaf	0-250	4.12	1.06	8.28E-03	4.91E-02
evm.model.scaffold7.924	F3H	Leaf	0-250	25.01	7.39	5.64E-04	5.37E-03
evm.model.scaffold36.172	LAR	Leaf	250-500	0.39	3.01	6.20E-03	3.90E-02
evm.model.scaffold38.385	CHS/STSY	Leaf	250-500	10.71	124.19	5.23E-08	1.46E-06
evm.model.scaffold1.577	CHS/STSY	Root	0-250	17.89	41.99	1.80E-03	1.41E-02
evm.model.scaffold11.178	F3'H	Root	0-250	59.85	164.03	6.33E-04	5.91E-03
evm.model.scaffold20.251	DFR	Root	0-250	3.18	12.49	2.68E-04	2.84E-03
evm.model.scaffold32.254	LAR	Root	0-250	27.57	114.42	9.44E-09	3.07E-07
evm.model.scaffold7.924	F3H	Root	0-250	232.46	486.02	6.61E-03	4.10E-02
evm.model.scaffold7.227	CHS/STSY	Root	0-250	304.25	698.97	5.25E-03	3.41E-02
evm.model.scaffold11.512	CHI	Root	250-500	224.12	106.22	3.78E-03	2.61E-02
evm.model.scaffold38.385	CHS/STSY	Root	250-500	2843.70	780.15	3.80E-03	2.62E-02

Supplementary Table 20 | The number of expanded families that involved in plant-pathogen interaction and biosynthesis of secondary metabolites. The p-values were calculated using Fisher's exact test.

Pathway	# in expanded gene families	# in all gene families	p-value
plant-pathogen interaction	7	220	4.8×10^{-4}
Plant-pathogen interaction (ko04626)			
biosynthesis of secondary metabolites	5	149	2.5×10^{-3}
Isoflavonoid biosynthesis (ko00943)			
Flavonoid biosynthesis (ko00941)			
Phenylpropanoid biosynthesis (ko00940)			

Supplementary Table 21 | Gene Ontology enrichment of tandemly duplicated genes in *R. apiculata*. The p-values were calculated using Fisher's exact test.

GO term	Ontology	Description	Gene number	p-value	FDR
GO:0055114	P	oxidation reduction	280	2.80E-15	5.60E-12
GO:0009875	P	pollen-pistil interaction	22	1.40E-11	5.70E-09
GO:0008037	P	cell recognition	22	1.40E-11	5.70E-09
GO:0048544	P	recognition of pollen	22	1.40E-11	5.70E-09
GO:0009856	P	pollination	22	1.40E-11	5.70E-09
GO:0071554	P	cell wall organization or biogenesis	31	1.10E-07	3.80E-05
GO:0006855	P	multidrug transport	23	1.90E-07	4.30E-05
GO:0042493	P	response to drug	23	1.90E-07	4.30E-05
GO:0015893	P	drug transport	23	1.90E-07	4.30E-05
GO:0030259	P	lipid glycosylation	13	1.80E-06	0.00036
GO:0071555	P	cell wall organization	24	3.80E-06	0.0007
GO:0042545	P	cell wall modification	19	1.20E-05	0.0021
GO:0005975	P	carbohydrate metabolic process	140	1.50E-05	0.0024
GO:0051313	P	attachment of spindle microtubules to chromosome	14	3.30E-05	0.0044
GO:0008608	P	attachment of spindle microtubules to kinetochore	14	3.30E-05	0.0044
GO:0034453	P	microtubule anchoring	14	4.70E-05	0.006
GO:0007154	P	cell communication	25	5.00E-05	0.006
GO:0006633	P	fatty acid biosynthetic process	24	9.60E-05	0.011
GO:0030258	P	lipid modification	14	0.00018	0.019
GO:0042221	P	response to chemical stimulus	54	0.00021	0.021
GO:0055085	P	transmembrane transport	128	0.00027	0.026

73. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002; **12**:656-664.
74. Jones P, Binns D, Chang H-Y *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014; **30**:1236-1240.
75. Grabherr MG, Haas BJ, Yassour M *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011; **29**:644-652.
76. Altschul SF, Madden TL, Schäffer AA *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**:3389-3402.
77. Suyama M, Torrents D and Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 2006; **34**:W609-W612.
78. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004; **5**:113.
79. Yim H-S, Cho YS, Guang X *et al*. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet* 2014; **46**:88-92.
80. Frantz L, Schraiber JG, Madsen O *et al*. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome Biology* 2013; **14**:R107.
81. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 2003; **19**:301-302.
82. Kaul S, Koo HL, Jenkins J *et al*. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000; **408**:796-815.
83. Chan AP, Crabtree J, Zhao Q *et al*. Draft genome sequence of the oilseed species Ricinus communis. *Nat Biotechnol* 2010; **28**:951-956.
84. Tuskan GA, DiFazio S, Jansson S *et al*. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 2006; **313**:1596-1604.
85. Farnsworth E and Farrant J. Reductions in abscisic acid are linked with viviparous reproduction in mangroves. *Am J Bot* 1998; **85**:760-760.
86. Lois LM, Lima CD and Chua N-H. Small ubiquitin-like modifier modulates abscisic acid signaling in Arabidopsis. *The Plant Cell* 2003; **15**:1347-1359.
87. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995; **57**:289-300.
88. Yang Z and Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 1998; **46**:409-418.
89. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 1998; **15**:568-573.
90. Hoagland DR and Arnon DI: The water-culture method for growing plants without soil: The College of Agriculture, University of California, Berkeley; 1950.
91. Prado-Martinez J, Sudmant PH, Kidd JM *et al*. Great ape genetic diversity and population history. *Nature* 2013; **499**:471-475.
92. Zhao S, Zheng P, Dong S *et al*. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat Genet* 2013; **45**:67-71.
93. Hung C-M, Shaner P-JL, Zink RM *et al*. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proc Natl Acad Sci USA* 2014; **111**:10636-10641.
94. Groenen MA, Archibald AL, Uenishi H *et al*. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 2012; **491**:393-398.
95. Kelley JL, Peyton JT, Fiston-Lavier A-S *et al*. Compact genome of the Antarctic midge is likely an adaptation to an

extreme environment. *Nat Commun* 2014; **5**.

96. Albert VA, Barbazuk WB, Der JP *et al*. The Amborella genome and the evolution of flowering plants. *Science* 2013; **342**:1241089.
97. Ye J, Fang L, Zheng H *et al*. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 2006; **34**:W293-W297.
98. Du Z, Zhou X, Ling Y *et al*. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* 2010; **38**:W64-W70.