Genomics 114 (2022) 110382

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Assembly of a hybrid mangrove, *Bruguiera hainesii*, and its two ancestral contributors, *Bruguiera cylindrica* and *Bruguiera gymnorhiza*

Jeremy R. Shearman [a], Chaiwat Naktang [a], Chutima Sonthirod [a], Wasitthee Kongkachana [a], Sonicha U-thoomporn [a], Nukoon Jomchai [a], Chatree Maknual [b], Suchart Yamprasai [b], Waratthaya Promchoo [b], Panthita Ruang-areerate [a], Wirulda Pootakham [a], Sithichoke Tangphatsornruang [a,*]

[a] National Omics Center, National Science and Technology Development Agency, 111 Thailand Science Park, Paholyothin Road, Khlong Nueng, Khlong Luang, Pathumthani 12120, Thailand
[b] Department of Marine and Coastal Resources, 120 The Government Complex, Chaengwatthana Rd., Thung Song Hong, Bangkok 10210, Thailand

ARTICLE INFO

Keywords:
Bruguiera
Genome assembly
Comparative genomics
Mangrove

ABSTRACT

Mangroves are plants that live in tropical and subtropical coastal regions of the world, they are adapted to high salt environments and cyclic tidal flooding. Mangroves play important ecological roles, including acting as breeding grounds for many fish species and to prevent coastal erosion. The genomes of three mangrove species, *Bruguiera gymnorhiza*, *Bruguiera cylindrica*, and a hybrid of the two, *Bruguiera hainesii* were sequenced, assembled and annotated. The two progenitor species, *B. gymnorhiza* and *B. cylindrica*, were found to be highly similar to each other and sufficiently similar to *B. parviflora* to allow it to be used for reference based scaffolding to generate chromosome level scaffolds. The two subgenomes of *B. hainesii* were independently assembled and scaffolded. Analysis of *B. hainesii* confirms that it is a hybrid and the hybridisation event was estimated at 2.4 to 3.5 million years ago using a Bayesian Relaxed Molecular Clock approach.

1. Introduction

Mangroves are important plants that have adapted to survive high-salt coastal regions. They are found in tropical and subtropical coastal regions and intertidal estuaries, in which they endure hours of salt water submergence and intermittent tidal and storm wave forces [1–3]. Mangroves are a valuable part of biodiversity and serve as breeding grounds for a large number of marine species, including many commercial species of fish and crab [1,3]. Mangroves help maintain water quality by trapping and filtering sediment and they help stabilize coastal shores by preventing erosion and providing a natural barrier to storm surge and flooding [1,3]. Mangrove species have a variety of mechanisms that allow them to survive the harsh salt environment, including salt exclusion by differential absorption, salt excretion from the roots or leaves, or long term salt storage within the cell [4]. Mangroves are found in several genera and even multiple diverse families, consistent with convergent evolution [4].

Bruguiera is a genus of true mangroves, containing five species and three natural hybrids, in the family Rhizophoraceae. *Bruguiera* species have adapted to estuarine and sheltered coastal conditions by having roots that can absorb water while filtering out the salt, referred to as ultrafiltration [5]. *Bruguiera* species can be roughly separated into two groups based on morphology. One group has large leaves and large solitary-flowered inflorescences and consists of *Bruguiera exaristata*, *Bruguiera gymnorhiza*, and *Bruguiera sexangula*. The other group has smaller leaves and inflorescences that have multiple flowers, this group includes *B. cylindrica* and *B. parviflora* [3,6]. *Bruguiera hainesii* has morphological features that represent an intermediate state between these two groups, which led some to consider that the species is a hybrid [7]. Investigations into nuclear and chloroplast loci were consistent with *B. hainesii* being a hybrid and chloroplast sequence shows that *B. cylindrica* is the maternal ancestor, leaving *B. gymnorhiza* as the paternal ancestor [7]. Sequencing and assembly of the complete chloroplast sequence of *B. hainesii*, *B. cylindrica*, and *B. gymnorhiza* confirmed this relationship with the chloroplast sequence of *B. hainesii* identical to the chloroplast sequence of *B. cylindrica* except for a single base difference [8].

The geographic distribution of *B. hainesii* extends from Thailand and

* Corresponding author.
E-mail address: sithichoke.tan@nstda.or.th (S. Tangphatsornruang).

https://doi.org/10.1016/j.ygeno.2022.110382
Received 7 January 2022; Received in revised form 19 April 2022; Accepted 2 May 2022
Available online 6 May 2022

Myanmar across the Malay archipelago to Papua New Guinea [3,6]. However, the species is considered critically endangered on the IUCN red list, and has only a few hundred known mature trees in existence [6,9]. The low number of individuals could be consistent with this species being a hybrid. We sequenced and assembled the nuclear genomes of *B. hainesii*, *B. cylindrica*, and *B. gymnorhiza* to provide high quality reference genomes and to investigate the specific relationship of these three species.

2. Materials and methods

2.1. Sample, DNA/RNA extraction and flow cytometry

The mangrove samples of *B. gymnorhiza* (*Bg*), *B. cylindrica* (*Bc*), and *B. hainesii* (*Bh*) were collected from Ranong Biosphere Reserve, Thailand. Young leaves were collected from a single plant for each species and placed into liquid nitrogen. Frozen leaf was used for DNA and RNA extraction with the standard CTAB method followed by clean-up using a DNeasy Mini spin column from Qiagen for DNA and precipitation of RNA using the standard LiCl method. Poly(A) mRNA was enriched using a Dynabeads mRNA Purification Kit (ThermoFisher Scientific, Waltham, USA). Flow cytometry, according to the protocol of Dolzel and Bartos [10], was used to estimate the genome size of each mangrove species with *Arabidopsis thaliana* as a reference species.

2.2. Genome sequencing and assembly

DNA was used to prepare linked-reads [11,12] using 10× genomics (10× Genomics, Pleasanton, USA) for sequencing on an Illumina HiSeq X Ten following Illumina protocols for 150 bp paired-end sequences (Ilumina, San Diego, USA). The genomes were assembled using supernova v2.1.1 [13] with default settings and the 'pseudohap output' style using an estimated genome size based on the flow cytometry results. Since *Bh* was considered to be a hybrid, it was assembled in two ways, the first was to perform an assembly using all of the reads. A second assembly was performed on subsets of reads grouped by similarity to the ancestral species. The subsetting was performed by mapping the raw reads using Burrows-Wheeler Aligner (BWA) [14] with default settings against the combined assemblies of *Bg* and *Bc* to identify which ancestral species they most closely matched. Contigs were mapped to other assemblies using BWA with relaxed error penalties: -B 1 -O 1 -T 500. These two assemblies were compared by considering the contig N50 size and the gene content.

Each genome was scaffolded using RagTag v2.0.1 [15] with *B. parviflora* [16] as the reference to generate chromosome level scaffolds. The default settings were used with the setting to not split contigs. The reliability of these scaffolding steps were investigated using comparative genomics according to the methods below.

2.3. Annotation and RNA-seq

The RNA from each of the three species was sequenced on an MGISEQ-2000RS using the MGISEQ-2000RS Sequencing Flow Cell V3.0 (MGI Tech, Shenzhen, China) to obtain 150 bp paired-end sequence data.

Annotation was performed for each genome by EvidenceModeler v1.1.1 [17] using the RNA-seq data, protein data from related species, and ab initio gene prediction for each of the three mangrove species. For each species, the RNA-seq data was mapped to the genome assembly using GMAP [18] and a transcriptome assembly was performed using PASA v2.4.1 [17]. Protein sequences were obtained from public databases for *Oryza sativa*, *Mimulus guttatus*, *Sesamum indicum*, *Populus trichocarpa* and *Eucalyptus grandis* and aligned to the genomes using AAT [19]. The ab initio prediction program Augustus v3.3.3 [20] was trained using the PASA2 alignment assembly [17]. All gene predictions were then combined by EvidenceModeler to generate consensus gene models using the following weights for each evidence type: PASA2 - 1, GMAP - 0.5, AAT - 0.3, and Augustus - 0.3. Genome completeness was estimated by comparing each annotation gene set against the embryophyta_odb10 data set using Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.5 [21]. The predicted gene sets for each species were functionally annotated using OmicsBox v2.0.10 (https://www.biobam.com/omicsbox) [22].

Gene expression data for within sample comparison was calculated for *Bh* by mapping the RNA-seq data to the assembled genome and calculating Fragments Per Kilobase of transcript per Million mapped reads (FPKM) using StringTie v2.1.1 [23]. Expression of gene pairs within CD-hit groups were compared within *Bh* to identify which subgenome provided the highest expression based on read counts.

2.4. Repeat sequence annotation

RepeatModeler version 2.0.1 [24] was used to identify and classify de novo repeat families on each assembled genome. The repeat sequences identified in the genomes were aligned to Genbank's non-redundant protein database (using BLASTX with an *E*-value cutoff of $1 \times 10–6$) to exclude repeat sequences that contain large families of protein-coding genes.

2.5. Comparative genomics

All of the scaffolded genomes were compared to *B. parviflora* using the nucmer program of MUMmer v3.23 [25] to determine scaffold quality based on colinearity within contigs. The output of this was plotted in R to identify rearrangements relative to *B. parviflora*. The subgenomes of *Bh* were compared to their respective progenitor species and plotted the same way to identify inversions and translocations. Figures were also generated using Circos [26] based on blast results of peptide sequences for each genome being plotted. The BLASTP alignment results were searched for colinear blocks using MCScanX [27] with regions of at least ten syntenic genes (with no more than six intervening genes allowed) considered as intra- or intergenomic homologous regions.

Genes were grouped using CD-hit [28] with a 95% threshold using the CDS of *Bc*, *Bg* and *Bh* combined. The closest match of each *Bh* gene was identified by blasting each *Bh* gene against a database of *Bc* and *Bg* genes and the output was combined with the CD-hit based gene groupings. This data was then combined with RNA-seq expression data to identify qualitative expression differences between gene homologues from each ancestor species.

A phylogenetic tree was constructed using the RAxML-NG program v1.0.2 [29] based on orthologous groups that were identified by OrthoFinder v2.4.0 [30]. Species included in the phylogenetic tree were *Bc*, *Bg*, the subgenomes of *Bh*, *B. parviflora*, *Ceriops tagal*, *Kendelia obovata*, *Rhizophora apiculata*, *Arabidopsis thaliana*, *Cucumis melo*, *Cucumis sativus*, *Ricinus communis*, *Popoulus trichocarpa*, and *Oryza sativa*. Protein sequences were aligned using MUSCLE [31] and a substitution model was estimated using the ModelTest-NG program v0.1.7 [32]. The divergence times were estimated by the Bayesian Relaxed Molecular Clock approach using MCMCtree v4.0 [33]. Divergence times were anchored by dated fossil records, the root node of the common ancestor of Rhizophoraceae, Euphorbiaceae (*R. communis*) and Salicaceae (*P. trichocarpa*) is dated at 105–120 MYA, fossils of the ancestor of Rhizophoreae are dated at 48–56 MYA, and a fossil recognized as ancestral Rhizophora is dated to 34–38 MYA. Gene family size changes, based on orthogroups, were detected using CAFE5 [34] to identify expansion and contraction of gene numbers within gene families.

3. Results and discussion

3.1. Genome assembly

Sequencing the genomes of the three species produced between 111 and 118 Gb of paired-end reads. The genomes assembled into 236 Mb with an N50 of 2 Mb for *Bc*; 289 Mb with an N50 of 280 kb for *Bg*; and 419 Mb with an N50 of 40 kb for *Bh* using all of the reads. These genome sizes are consistent with the estimates that were calculated from flow cytometry, which were 237.4, 285, and 422.2 Mb for *Bc*, *Bg*, and *Bh*, respectively. Repeat analysis (described below) of *Bc* and *Bg* shows that *Bg* has 89.2 Mb (30.9% of the genome) of repeat sequence, while *Bc* has 66 Mb (28% of the genome) of repeat sequence (Table 1). The larger amount of repeat sequence in *Bg* is a likely cause for the assembly being more fragmented compared to the *Bc* assembly. Hybridisation events in plants often result in progeny that have a tetraploid genome that contains a diploid copy of both progenitor genomes [35]. The assembly of *Bh* was significantly larger than the assemblies of *Bc* or *Bg* and is consistent with *Bh* being a tetraploid that consists of diploid copies of both *Bc* and *Bg* genomes.

To investigate the potential of *Bh* consisting of both *Bc* and *Bg* genomes, we mapped the *Bh* contigs using BWA against each of the *Bc* and *Bg* genomes separately using relaxed error penalties to align similar sequences. Approximately 82.6% of the *Bh* genome mapped to *Bc* and covered 94.6% of the *Bc* genome, while 81.6% of the *Bh* genome mapped to *Bg* and covered 84.9% of the *Bg* genome. This showed that a large amount of sequence between *Bc* and *Bg* is highly similar. To investigate this, *Bc* was mapped to *Bg* using the same parameters and it was found that 73.3% of the *Bc* sequence mapped to *Bg*. So, *Bh* is more similar to each of *Bc* and *Bg* than those species are to each other, which is consistent with *Bh* being a hybrid of the two.

The *Bh* reads were subsetted by mapping them against the genomes of *Bc* and *Bg* and separating them according which ancestral species they most closely resembled. This resulted in 48.6% of the reads mapping to *Bg*, 41.5% of the reads mapping to *Bc*, 6.3% of the reads having a partial or overlapping match to both ancestral species, and 3.5% of the reads not returning any mapping alignment. Individual assemblies were performed using reads that mapped to a single ancestral species and GapFiller was used on each resulting assembly using all reads to fill any gaps that might arise from excluding the reads that had no alignment or aligned to both ancestors. The resulting assembly of the *Bc* subgenome consisted of 9305 contigs, totalling 245.5 Mp, the largest contig was 8.5 Mb and the N50 was 1.94 Mb. The subgenome from *Bg* consisted of 29,420 contigs, totalling 279.9 Mp, the largest contig was 7.9 Mp and the N50 was 1.23 Mb. The subgenome assemblies closely resembled their respective ancestral genome assembly in size, plus the N50 of each subgenome was significantly longer than the N50 of the assembly that used all of the reads, suggesting that the subgenome assemblies are superior to the assembly that used all of thereads.

3.2. Genome scaffolding and comparative genomics

We used the assembly of *Bruguiera parviflora*, which has chromosome-level scaffolds through the use of Hi-C, as a reference to perform reference based scaffolding for each of the three genomes [16]. Such an approach relies on the reference species having high sequence similarity and colinearity. Major structural differences in sequence order can introduce errors in the resulting scaffolds. Since the *Bc* assembly had the largest N50, at 2 Mb, it had the highest ability to show discordant mapping within contigs if *B. parviflora* was not a good species to scaffold against. Scaffolding produced chromosome-level scaffolds for *Bc* with 87.2% of the assembly sequence placed into 18 chromosomes (Fig. 1 A). Aligning the resulting scaffolds back to *B. parviflora* showed a high level of colinearity with no translocations and two small inversions on chromosomes 10 and 14 (Supplementary Fig. 1). The resulting alignment would be disorderly if *B. parviflora* was too distantly related to be useful as a reference since the method did not break any contigs prior to placement and the contigs were large to begin with. This shows that *B. parviflora* is a suitable reference to scaffold against for *Bc* and considering the sequence similarity of *Bc* to *Bg*, likely to be suitable for *Bg* also. Scaffolding the *Bg* genome also produced the expected 18 chromosome-level scaffolds with no inversions or translocations, but only used 67.6% of the assembled sequence (Fig. 1 B, Supplementary Fig. 2). The *Bg* chromosome-level scaffolds used less sequence and had larger gaps than the *Bc* scaffolds, likely because of the higher level of fragmentation in the contigs and larger amount of repeat sequence. The resulting chromosome-level scaffolds for both species were numbered according to the chromosome numbers of *B. parviflora* since it was used

Table 1
Assembly statistics and repeat types for the genomes of *B. cylindrica*, *Bruguiera gymnorhiza* and the two subgenomes of *B. hainesii*.

Species	Bc	Bc	BhBc subgenome	BhBg subgenome
Contigs	4546	26,333	8230	26,780
Total length	235,932,177 bp	289,032,877 bp	245,499,040 bp	279,939,462 bp
GC level	34.55%	33.88%	34.56%	34.10%
Bases masked	71,582,842 bp (30.34%) length (percentage)	96,329,235 bp (33.33%) length (percentage)	70,723,610 bp (28.81%) length (percentage)	88,165,204 bp (31.49%) length (percentage)
SINEs	271,758 bp (0.12%)	301,756 bp (0.10%)	380,088 bp (0.15%)	292,847 bp (0.10%)
ALUs	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
MIRs	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
LINEs	1,322,779 bp (0.56%)	1,837,972 bp (0.64%)	1,155,594 bp (0.47%)	1,161,246 bp (0.41%)
LINE1	953,909 bp (0.40%)	1,107,677 bp (0.38%)	746,838 bp (0.30%)	754,611 bp (0.27%)
LINE2	56,557 bp (0.02%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
L3/CR1	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
LTR elements	38,390,106 bp (16.27%)	46,567,071 bp (16.11%)	38,612,951 bp (15.73%)	42,431,825 bp (15.16%)
ERVL	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
ERVL-MaLRs	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
ERV_classI	130,566 bp (0.06%)	95,085 bp (0.03%)	43,663 bp (0.02%)	37,162 bp (0.01%)
ERV_classII	181,005 bp (0.08%)	39,628 bp (0.01%)	0 bp (0.00%)	0 bp (0.00%)
DNA elements	3,312,675 bp (1.40%)	3,705,834 bp (1.28%)	4,315,612 bp (1.76%)	4,370,870 bp (1.56%)
hAT-Charlie	43,030 bp (0.02%)	69 bp (0.00%)	55,218 bp (0.02%)	55,485 bp (0.02%)
TcMar-Tigger	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)	0 bp (0.00%)
Unclassified	22,720,238 bp (9.63%)	36,871,401 bp (12.76%)	20,346,156 bp (8.29%)	33,695,265 bp (12.04%)
Total interspersed repeats	66,017,556 bp (27.98%)	89,284,034 bp (30.89%)	64,810,401 bp (26.40%)	81,952,053 bp (29.27%)
Small RNA	285,677 bp (0.12%)	125,453 bp (0.04%)	232,744 bp (0.09%)	245,320 bp (0.09%)
Satellites	60,331 bp (0.03%)	76,337 bp (0.03%)	107,954 bp (0.04%)	66,139 bp (0.02%)
Simple repeats	4,219,610 bp (1.79%)	5,587,894 bp (1.93%)	4,507,773 bp (1.84%)	4,787,628 bp (1.71%)
Low complexity	1,217,717 bp (0.52%)	1,443,600 bp (0.50%)	1,326,688 bp (0.54%)	1,410,832 bp (0.50%)

as the reference.

Scaffolding against *B. parviflora* was performed with the *Bh* genome that used all reads and each of the *Bh* subgenomes separately. The *Bh* genome from all reads resulted in a single set of 18 scaffolds that used less than half of the assembled sequence, suggesting an incomplete genome. There were no apparent rearrangements, but the highly fragmented nature of the assembly would mask any real rearrangements that may exist. To attempt to generate both subgenomes from these contigs, the contigs were blasted against the *Bc* and *Bg* genomes and independently scaffolded based on which group they matched. This resulted in 41% of the sequence scaffolding into 18 chromosomes consisting of *Bc* contigs and 28% of the sequence scaffolding into 18 chromosomes from *Bg* contigs with 31% of the sequence unscaffolded. Considering that *Bg* is the larger genome, this is further evidence that this assembly method was more erroneous.

The subgenomes of *Bh* from splitting the reads were independently scaffolded. Both subgenomes produced the expected 18 scaffolds, which are likely more reliable than the single *Bh* assembly since they had a much larger N50. The *Bc* subgenome scaffolds contained 199.7 Mb out of the 245.5 Mb, which is 81.3% of the sequence (Fig. 1 C, Supplementary Fig. 3). The *Bg* subgenome scaffolds contained 205.5 Mb out of the 279.9 Mb, which is 73.4% of the sequence (Fig. 1 C, Supplementary Fig. 4). To compare the subgenome assemblies to the *Bh* genome that used all reads, the amount of sequence in each subgenome more closely reflected the ancestral species and more sequence was scaffolded in total. Each chromosome level scaffold tended to have a single disorderly region where sequence was less connected and alternated between direct and inverted sequence, which is likely to represent centromere sequence (Supplementary Figs. 3 and 4). This distribution is more reflective of the progenitor genome sizes and indicates that the split read assembly is superior.

The large contig size allowed for the detection of inversions and translocations within the subgenomes compared to their respective progenitor species. There was one large translocation, one large inversion, and two inverted translocations within the *Bc* subgenome of *Bh* (*BhBc*), each spanning hundreds of kb (Supplementary Fig. 3). Each of these regions have an LTR Gypsy or Copia repeat within a few kb, although none occurred exactly at the junction, so it is unclear if those repeats were the cause of the rearrangements. The translocation involved a segment of scaffold 4 from the progenitor genome getting split into two pieces and inserted in swapped order into scaffold 9 of *BhBc* (Supplementary Fig. 4). The inverted sequence occurred on scaffold 12 and joins to one of the inverted translocated sequences, which is a segment of ancestral scaffold 12 translocated to *BhBc* scaffold 16 (Supplementary Fig. 3). The other inverted translocation is a segment of ancestral scaffold 18 translocated to *BhBc* scaffold 15 (Supplementary Fig. 3). The scaffolding program incorrectly scaffolded small contigs into the gaps from these translocated regions based on the reference, so these were removed from the scaffolds and returned to the unscaffolded contigs list.

The *Bg* subgenome of *Bh* (*BhBg*) had three translocations, two inversions, and six inverted translocations (Supplementary Fig. 5). It is remarkably more rearranged compared to its progenitor genome (*Bg*) than the *BhBc* subgenome is to *Bc*. Only four chromosome scaffolds remained fully contiguous between *BhBg* and the progenitor genome. The translocations are progenitor segments of chromosome 4, 14, and 16 to the *BhBg* subgenome 5, 13, and 1, respectively, relative to the progenitor genome (Supplementary Fig. 5). The inversions are on chromosomes 9 and 10. The inverted translocations are progenitor segments of chromosome 9, 5, 2, 10, and 18 to *BhBg* subgenome chromosomes 8, 14, 14, 8, and 17, respectively (Supplementary Fig. 5). While the *BhBc* subgenome tended to have LTR motifs near the translocations and inversions, the *BhBg* subgenome in many cases had no repeat sequences nearby the rearrangements.

All three genomes showed gene collinearity evidence indicative of a known ancient genome duplication event (Fig. 1 A, B, and C), which was also found in *B. parviflora* and estimated to have occurred 74.2 mya [2,16,36]. This ancestral duplication is also present in each of the *Bh* subgenomes (*BhBc* and *BhBg*) and the inversions and translocations that were identified in the subgenomes affect only one copy of each duplicated chromosome, consistent with the hybridisation event being more recent than the ancestral duplication.

3.3. Genome annotation and expression analysis

The *Bc* genome was annotated to include 21,191 genes with 20,773 of those contained within chromosome scaffolds. The *Bg* genome had 22,572 genes with 20,763 of those contained within chromosome scaffolds. The *BhBc* subgenome had 26,497 genes with 23,916 of those contained within chromosome scaffolds. The *BhBg* subgenome had 26,510 genes with 24,616 of those contained within chromosome scaffolds. The *Bh* genes were blasted against *Bc* and *Bg* to identify which species each gene most closely matched. For the *Bc* subgenome of *Bh*, 82.5% of the genes with a match were closest to *Bc* with 17.5% of the genes matching to *Bg*. For the *Bg* subgenome of *Bh*, 85.8% of the genes with a match were closest to *Bg* with 14.2% of the genes matching to *Bc*. The gene sets of *Bc*, *Bg*, *BhBc*, and *BhBg* were assessed for completeness using BUSCO and found to be 97.1%, 96.0%, 96.3%, and 92.7%, respectively. It is interesting that the *BhBg* subgenome had the lowest completeness even though it had the highest gene count. When the subgenomes were combined, the completeness was 97.1%, which suggests that each subgenome may have independently lost genes with the other subgenome compensating for that loss. The lack of these genes in the *BhBg* subgenome suggests a faster rate of evolution for the *BhBg* subgenome than the *BhBc* subgenome. However, The *BhBc* subgenome has a higher count of duplicated genes that are single copy in the progenitor genome, showing that some genomic rearrangement has occurred.

A phylogenetic tree was generated to compare between *Bc*, *Bg* and the *BhBc* and *BhBg* subgenomes using single copy orthologues (Fig. 2). The results suggested that *B. gymnorhiza* and *B. cylindrica* separated from each other 9.5 million years ago (MYA) and they both separated from *B. parviflora* 24.95 MYA. The *BhBc* subgenome had an estimated divergence from *B. cylindrica* of 2.44 MYA and the *BhBg* subgenome was estimated at 3.47 MYA, which also suggests that the *BhBg* subgenome is evolving at a faster rate. The gene orthologue set was used to estimate gene expansion or contraction of gene numbers within gene families using a birth and death process to model gene gain and loss across this phylogenetic tree (Fig. 2). The subgenomes of *Bh* were found to have a larger number of expanded gene families and a lower number of contracted gene families than either progenitor. The *BhBc* subgenome had more expanded gene families and fewer contracted gene families than the *BhBg* subgenome. Since the *BhBg* subgenome had a higher gene count, it follows that the expanded gene families include a larger number of gained genes within each expanded family than were lost in the contracted families.

The RNA-seq data was investigated to detect expression differences between subgenomes. Subgenomes often show unequal contribution to the transcriptome following a polyploidy event, a phenomenon known as subgenome dominance [37]. However, it can be difficult to accurately identify the ancestral progenitor species for each gene and high levels of sequence similarity can result in mapping errors being misinterpreted as real expression differences [38]. This suggests that the high similarity between the *BhBg* and *BhBc* subgenomes could be a problem, so thresholds for expression differences were stringent. The *BhBc* subgenome had 55.8% of the mapped RNA-seq reads, despite having fewer annotated genes. However, that is not a particularly large difference and does not indicate any strong subgenome dominance. Gene groups from CD-Hit that contained a single gene from each of the ancestral species, *Bc* and *Bg*, and one gene from each *Bh* subgenome were identified to dissect subgenome expression. There were 413 such groups where only one *Bh* subgenome gene copy had expression, however, each subgenome

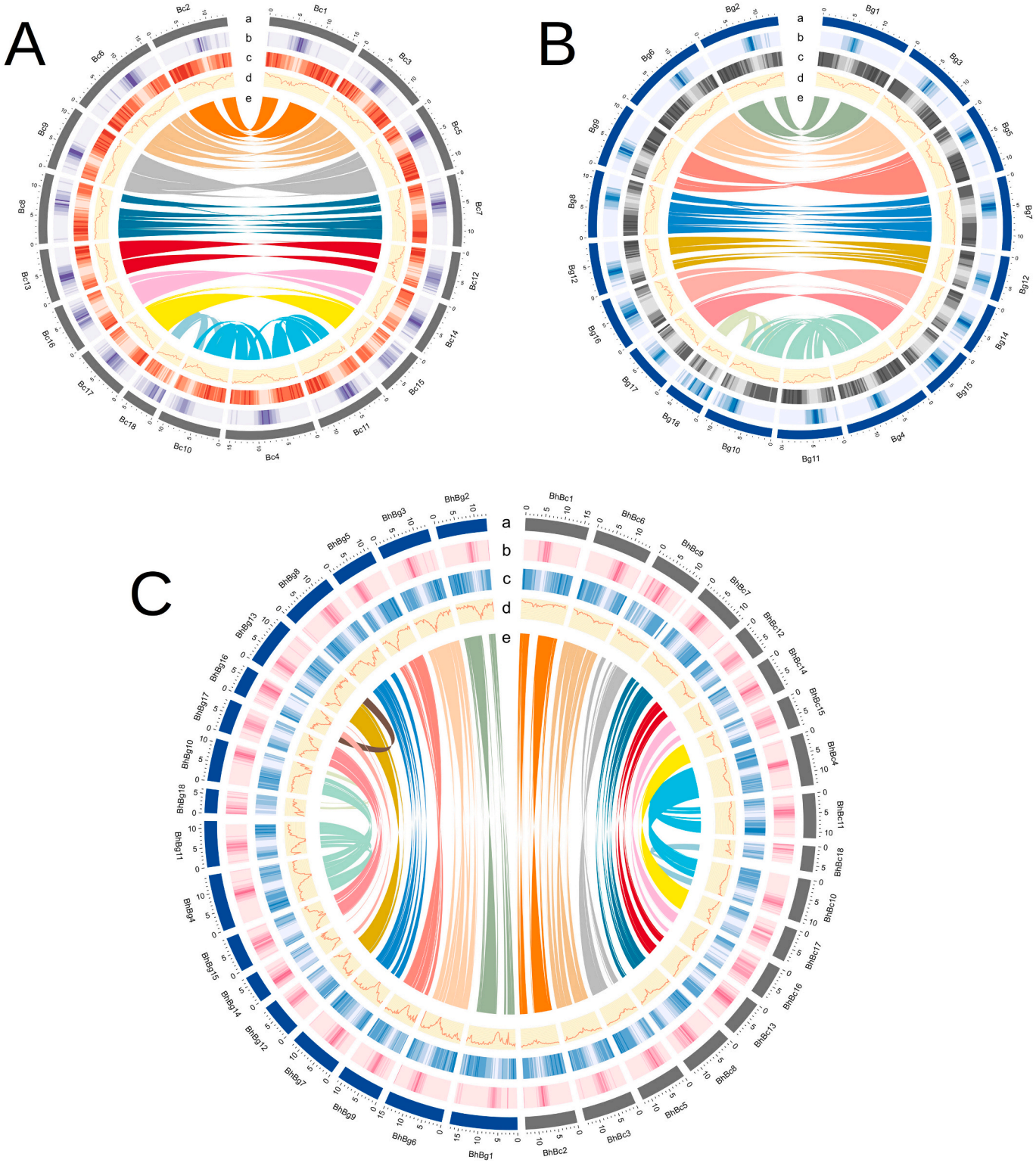

Fig. 1. Chromosomes of the mangrove species *Bruguiera cylindrica*, *Bruguiera gymnorhiza*, and *Bruguiera hainesii*.
The chromosome structure for *B. cylindrica* (A) and *B. gymnorhiza* (B) depicting the 18 chromosomes for each species named according to the species; Bc for *B. cylindrica* and Bg for *B. gymnorhiza*. The structure for *B. hainesii* (C) consists of two subgenomes each containing 18 chromosomes that are named according the ancestral species: BhBc for chromosomes that match *B. cylindrica* and BhBg for chromosomes that match *B. gymnorhiza*. Individual tracks are layered in concentric circles with: (a) representing the chromosomes, which are numbered according to synteny with *B. parviflora*; (b) showing repeat density with darker colours representing higher repeat density; (c) showing gene density with darker colours representing higher gene density; (d) showing the GC content along the length of each chromosome; and (e) showing syntenic blocks depicted by connected lines indicating that each chromosome has a syntenic chromosome within each genome or subgenome.

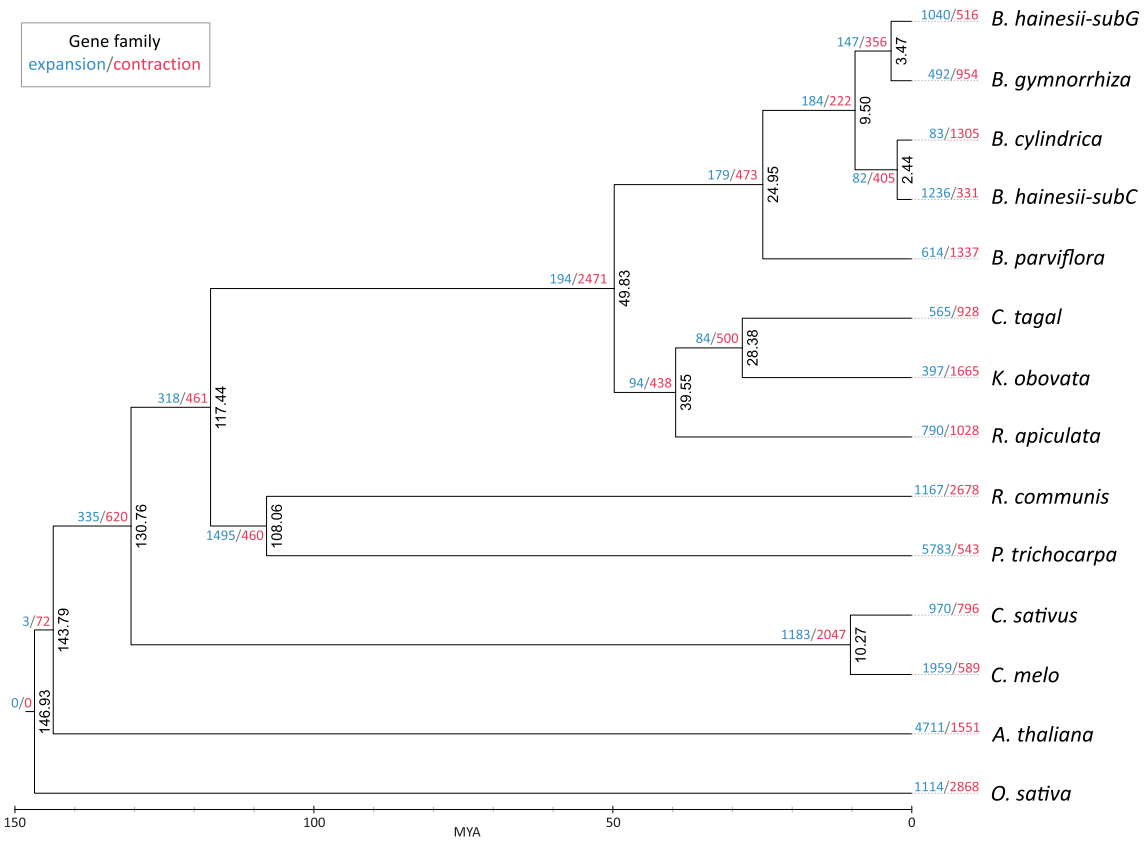

Fig. 2. Maximum-likelihood phylogenetic tree of the mangrove species *Bruguiera cylindrica*, *Bruguiera gymnorhiza*, and each subgenome of *Bruguiera hainesii*. Divergence times (million years ago) are indicated at each node. The number of gene families that have gained or lost genes are indicated as blue and red numbers, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

contained approximately half of the expressed genes and they were almost exclusively genes with the lowest expression suggesting that the majority are unreliable (Supplementary Table 1). There were 19 genes from this group where the expression was at or above the median *Bh* expression level, suggesting that one copy was silenced. From these 19 genes, there were 13 where the expression came only from the *BhBc* subgenome and 6 where expression came only from the *BhBg* subgenome. The two most highly expressed genes are annotated as *DELLA protein GAI1-like* and *ribophorin II family protein* (Supplementary Table 1) and both are expressed only from the *BhBc* subgenome, revealing a slight bias toward the *Bc* subgenome. However, extending the analysis to genes where more than 80% of the expression belonged to a single subgenome identified 1851 gene groups with a slight bias toward the *BhBg* subgenome (Supplementary Table 2). Counting which subgenome contributed the dominant gene for each group reveals 959 genes from the *BhBg* subgenome and 892 genes from the *BhBc* subgenome with 51% of total expression from within this group coming from the *BhBg* subgenome. There did not appear to be any significant enrichment in pathway or GO terms for any group. This shows that neither subgenome is particularly dominant, but does indicate that evolution might be in the process of leading to one copy of each homoeologous gene becoming silenced.

3.4. Repeat sequence analysis

All three species contained 26–31% repeat sequence and the most common repeat type was LTR elements (Table 1). The *Bc* genome had the lowest number of repeats as a percentage of total genome (27.98%) and *Bg* had the most repeat sequence at 30.89%, which translates to a much larger amount of repeat sequence considering that the *Bg* genome is 1.2 times the size of the *Bc* genome. The distribution and proportion of repeat types was highly similar between *Bc* and *Bg*, which is unsurprising considering how closely related the species are. The majority of repeats are LTR elements and account for approximate 16% of each genome followed by unclassified repeats accounting for 9–13% of each genome. Simple repeats and DNA elements each accounted for 1–2% of the genome, and SINE and LINE each accounted for less than 1% of the genomes. Comparing the subgenomes of *Bh* to the genomes of their respective progenitor species revealed a slight decrease in the amount of repeat sequence, 26.4% for *BhBc* and 29.27% for the *BhBg* subgenome compared to 27.98% and 30.89% for their ancestral genomes, respectively. The distributions of repeat types in the *Bh* subgenomes are largely conserved, with most of the repeat loss coming from the unclassified repeat category.

Genes annotated as transposon or retrotransposon were investigated in the two progenitor species and the subgenomes of *Bh*. The *Bc* genome had 236 genes, and the *Bg* genome had 114 genes, that were identified as a transposon or retrotransposon. The subgenomes of *Bh* show significant loss of these genes, particularly from the *BhBc* subgenome, which had 89 transposon genes. The *BhBg* subgenome retained most of the transposon genes that are found in the *Bg* genome with a gene count of 100. This suggests that genome remodelling is purging repeat sequence.

4. Conclusions

The genomes of *B. gymnorhiza* and *B. cylindrica* were each assembled into 18 chromosome-level scaffolds and found to be highly similar to each other. The subgenomes of *B. hainesii* were independently assembled into 18 chromosome-level scaffolds, which were similar to their respective progenitor genomes with some inversions and translocations. Both subgenomes show some evidence of change and remodelling compared to the progenitor genomes, but there is no evidence of genomic shock causing substantial rearrangement or any subgenome

dominance in gene expression. The amount of genomic shock causing rearrangement and gene silencing or loss in an allopolyploidy event is likely correlated with the degree of difference between the progenitor genomes [39]. So it is not particularly surprising that the subgenomes appear relatively harmonious since they are so similar at the genome and gene level.

The hybridisation event was estimated at somewhere between 2.44 and 3.47 million years ago by comparing each subgenome to its progenitor genome. This is a relatively recent event on an evolutionary time scale, which is the most probable reason that so few structural differences were found between the subgenomes and their progenitor genomes. One particularly interesting question raised, but not answered, by this research is why this species has such a widespread geographic distribution, yet so few individuals in existence. There seem to be two possible occurrences that could lead to this situation, a single hybridisation event followed by propagule dispersal aided by ocean currents and human activity, or multiple independent hybridisation events followed by the same types of dispersal. The inversions and translocations identified here would be unlikely to also occur in hybrids that may have occurred from separate hybridisation events, so studies of this species from multiple different locations could help answer this question.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110382.

Competing financial interests

The authors declare that there are no competing financial interests.

Accession codes

B. cylindrica genome and reads: PRJNA725957.
B. gymnorhiza genome and reads: PRJNA725949.
B. hainesii Bc subgenome: JAHOYR000000000.
B. hainesii Bg subgenome: JAJAGF000000000.
B. hainesii raw read data: PRJNA794248.

Author contributions

JRS, WPo and ST conceived of the study. JRS, CS, CN, WK, and PA performed bioinformatics work and prepared figures. WPo, WPr, SU, CM, SY and NJ performed sample collection, laboratory work and library construction. JRS wrote the manuscript and all authors reviewed it.

Acknowledgements

The authors would like to acknowledge funding from the National Science and Technology Development Agency, Thailand.

References

[1] F. Göltenboth, S. Schoppe, 10 - Mangroves, in: F. Göltenboth, K.H. Timotius, P. P. Milan, J. Margraf (Eds.), Ecology of Insular Southeast Asia, Elsevier, 2006, pp. 187–214, https://doi.org/10.1016/B978-044452739-4/50011-5.
[2] W. Guo, et al., Comparative analysis of transcriptomes in Rhizophoraceae provides insights into the origin and adaptive evolution of mangrove plants in intertidal environments, Front. Plant Sci. 8 (2017) 795.
[3] P.B. Tomlinson, The Botany of Mangroves, Cambridge University Press, 2016, https://doi.org/10.1017/CBO9781139946575.
[4] R. Reef, C.E. Lovelock, Regulation of water balance in mangroves, Ann. Bot. 115 (2015) 385–395.
[5] Water Relations and Salt Balance, in: P.B. Tomlinson (Ed.), The Botany of Mangroves, Cambridge University Press, 2016, pp. 109–122, https://doi.org/10.1017/CBO9781139946575.009.
[6] C.-R. Sheue, J. Yong, Y.-P. Yang, The Bruguiera (Rhizophoraceae) species in the mangroves of Singapore, especially on the new record and the rediscovery, Taiwania 50 (2005) 251–260.
[7] J. Ono, et al., Bruguiera hainesii, a critically endangered mangrove species, is a hybrid between B. cylindrica and B. gymnorhiza (Rhizophoraceae), Conserv. Genet. 17 (2016) 1137–1144.
[8] P. Ruang-Areerate, et al., Complete chloroplast genome sequences of five Bruguiera species (Rhizophoraceae): comparative analysis and phylogenetic relationships, PeerJ 9 (2021), e12268.
[9] B.A. Polidoro, et al., The loss of species: mangrove extinction risk and geographic areas of global concern, PLoS One 5 (2010), e10095.
[10] J. Dolezel, J. Bartos, Plant DNA flow cytometry and estimation of nuclear genome size, Ann. Bot. 95 (2005) 99–110.
[11] G.X.Y. Zheng, et al., Haplotyping germline and cancer genomes using high-throughput linked-read sequencing, Nat. Biotechnol. 34 (2016) 303–311.
[12] P. Marks, et al., Resolving the full spectrum of human genome variation using linked-reads, Genome Res. 29 (2019) 635–645.
[13] N.I. Weisenfeld, V. Kumar, P. Shah, D.M. Church, D.B. Jaffe, Direct determination of diploid genome sequences, Genome Res. 27 (2017) 757–767.
[14] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, Bioinform. Oxf. Engl. 25 (2009) 1754–1760.
[15] M. Alonge, et al., RaGOO: fast and accurate reference-guided scaffolding of draft genomes, Genome Biol. 20 (2019) 224.
[16] W. Pootakham, et al., A chromosome-scale reference genome assembly of yellow mangrove (Bruguiera parviflora) reveals a whole genome duplication event associated with the Rhizophoraceae lineage, Mol. Ecol. Resour. (2022), https://doi.org/10.1111/1755-0998.13587.
[17] B.J. Haas, et al., Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, Genome Biol. 9 (2008) R7.
[18] T.D. Wu, C.K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics 21 (2005) 1859–1875.
[19] X. Huang, M.D. Adams, H. Zhou, A.R. Kerlavage, A tool for analyzing and annotating genomic sequences, Genomics 46 (1997) 37–45.
[20] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, Bioinformatics 24 (2008) 637–644.
[21] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212.
[22] S. Götz, et al., High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res. 36 (2008) 3420–3435.
[23] S. Kovaka, et al., Transcriptome assembly from long-read RNA-seq alignments with StringTie2, Genome Biol. 20 (2019) 278.
[24] J.M. Flynn, et al., RepeatModeler2 for automated genomic discovery of transposable element families, Proc. Natl. Acad. Sci. 117 (2020) 9451–9457.
[25] S. Kurtz, et al., Versatile and open software for comparing large genomes, Genome Biol. 5 (2004) R12.
[26] M.I. Krzywinski, et al., Circos: an information aesthetic for comparative genomics, Genome Res. (2009), https://doi.org/10.1101/gr.092759.109.
[27] Y. Wang, et al., MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, Nucleic Acids Res. 40 (2012), e49.
[28] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinform. Oxf. Engl. 28 (2012) 3150–3152.
[29] A.M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, Bioinformatics 35 (2019) 4453–4455.
[30] D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, Genome Biol. 20 (2019) 238.
[31] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[32] D. Darriba, et al., ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models, Mol. Biol. Evol. 37 (2020) 291–294.
[33] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.
[34] F.K. Mendes, D. Vanderpool, B. Fulton, M.W. Hahn, CAFE 5 models variation in evolutionary rates among gene families, Bioinformatics 36 (2020) 5516–5518.
[35] Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy, Nat. Rev. Genet. 18 (2017) 411–424.
[36] S. Xu, et al., The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole-genome sequencing, Natl. Sci. Rev. 4 (2017) 721–734.
[37] K.L. Adams, R. Cronn, R. Percifield, J.F. Wendel, Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 4649–4654.
[38] G. Hu, et al., Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids, Brief. Bioinform. 22 (2021) 1819–1835.
[39] K.A. Bird, R. VanBuren, J.R. Puzey, P.P. Edger, The causes and consequences of subgenome dominance in hybrids and recent polyploids, New Phytol. 220 (2018) 87–93.