International Journal of
Molecular Sciences

MDPI

Brief Report

Genome Assembly of *Cordia subcordata*, a Coastal Protection Species in Tropical Coral Islands

Yi-Lan Chen [1,2,3], Zheng-Feng Wang [1,4,5,*], Shu-Guang Jian [1], Hai-Min Liao [3] and Dong-Ming Liu [1,*]

[1] Innovation Academy of South China Sea Ecology and Environmental Engineering, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China
[2] Key Laboratory of Karst Georesources and Environment, Ministry of Education, College of Resources and Environmental Engineering, Guizhou University, Guiyang 550025, China
[3] Collaborative Innovation Center for Mountain Ecology & Agro-Bioengineering, College of Life Sciences/Institute of Agro-Bioengineering, Guizhou University, Guiyang 550025, China
[4] Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China
[5] Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems, Key Laboratory of Carbon Sequestration in Terrestrial Ecosystem, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China
* Correspondence: wzf@scib.ac.cn (Z.-F.W.); liudm@scib.ac.cn (D.-M.L.)

Abstract: *Cordia subcordata* trees or shrubs, belonging to the Boraginaceae family, have strong resistance and have adapted to their habitat on a tropical coral island in China, but the lack of genome information regarding its genetic background is unclear. In this study, the genome was assembled using both short/long whole genome sequencing reads and Hi–C reads. The assembled genome was 475.3 Mb, with 468.7 Mb (99.22%) of the sequences assembled into 16 chromosomes. Repeat sequences accounted for 54.41% of the assembled genome. A total of 26,615 genes were predicted, and 25,730 genes were functionally annotated using different annotation databases. Based on its genome and the other 17 species, phylogenetic analysis using 336 single-copy genes obtained from ortholog analysis showed that *C. subcordata* was a sister to *Coffea eugenioides*, and the divergence time was estimated to be 77 MYA between the two species. Gene family evolution analysis indicated that the significantly expanded gene families were functionally related to chemical defenses against diseases. These results can provide a reference to a deeper understanding of the genetic background of *C. subcordata* and can be helpful in exploring its adaptation mechanism on tropical coral islands in the future.

Keywords: *Cordia subcordata*; tropical coral islands; sequencing; genome assembly; gene annotation; phylogenetic analysis

Citation: Chen, Y.-L.; Wang, Z.-F.; Jian, S.-G.; Liao, H.-M.; Liu, D.-M. Genome Assembly of *Cordia subcordata*, a Coastal Protection Species in Tropical Coral Islands. *Int. J. Mol. Sci.* **2023**, *24*, 16273. https://doi.org/10.3390/ijms242216273

Academic Editor: Abir U. Igamberdiev

Received: 21 September 2023
Revised: 4 November 2023
Accepted: 6 November 2023
Published: 13 November 2023

1. Introduction

Cordia subcordata (2n = 2x = 32) trees or shrubs (Figure 1) belong to the Boraginaceae family [1]. Most species of the Boraginaceae family are herbaceous, but all species from the genus *Cordia* are woody; thus, there are special groups for their physical forms [2]. The genus *Cordia* has 250–300 species around the world, mainly distributed on the east coast of Africa, India, Vietnam, and the southern Pacific islands [3]. In China, there are just six species of the genus *Cordia*, namely *C. cochinchinensis*, *C. cumingiana*, *C. dichotoma*, *C. furcans*, *C. myxa*, and *C. subcordata*, which are mostly distributed in southwest and southern China and are mainly found in Hainan province [4]. Most species of the genus *Cordia* have special chemical, biological, and pharmacological properties, which have a high research value in ethnobotany and ethnopharmacology. In addition, their secondary chemicals have a variety of applications in many aspects, such as anti-inflammation, wound healing, de-worming, anti-malaria, diuretics, and treating pulmonary diseases [5].

Figure 1. *Cordia subcordata*. (**A**) Whole tree, (**B**) flowers, and (**C**) fruits. Photographed by D.-M. Liu.

The tropical coral islands of China have a typical tropical ocean monsoon climate, with long sunshine durations, strong radiation, and high temperature all year around [6]. The soils are strongly alkaline and rich in calcium and phosphorus [1,7]. *C. subcordata* is a dominant evergreen small tree found on the tropical coral islands of China. It has a strong resistance to wind and dust because of its developed root systems. A physiological ecology study revealed that its characteristics, including a large leaf area, a high leaf epidermal stomata density, a thick upper epidermis, a low specific leaf area value, a high vessel diameter and density, low leaf water conductivity, and high xylem density, could favor *C. subcordata* in fully utilizing water and adapting to the poor soil, high temperatures, strong sunlight, and drought conditions [8]. In the field, *C. subcordata* is also observed to be pest- and pathogen-resistant (personal observation by D.-M. Liu). This could be attributed to the phytoconstituents in its tissues. In fact, it has been reported that the *Cordia* species contains various chemical components, such as quinones, terpenoid, steroids, alkaloids, flavonoids, and/or saponins [5,9], in their leaves, stem, root, fruits, or seeds. These chemical components not only have medicinal value (see above), but are also widely used as weapons against pests and pathogens in plants in nature. For example, *Cordia sebestena* is a widely planted tree in the area of the tropics [10]. It has long been used as a traditional medicinal plant and has also been found to be pest- and disease-resistant, due to the chemical composition of many parts of the plant (leaves, flower, and bark). Therefore, given its high levels of environmental adaptation, *C. subcordata* could be used

for artificial plant community construction, vegetation restoration, and the improvement of the environmental conditions of tropical coral islands [8,11].

Moreover, *C. subcordata* could have more than medical and ecological value. A pollen morphology study on the genus *Cordia* in Boraginaceae showed that its pollen morphologies were diverse and unique, displaying rather primitive characteristics [2], which exhibited the fact that the *Cordia* species also has a high scientific research value. Specifically, reproductive and pollination biological studies showed that *C. subcordata* was a typical distylous species, and such distyly is rare on oceanic islands [12].

Due to the anthropogenic disturbances of coral islands, the populations of *C. subcordata* have declined and have been threatened. Currently, it is listed as a second-level national key protected wild plant in China (https://www.gov.cn/zhengce/zhengceku/2021-09/09/content_5636409.htm, accessed on 19 September 2023). Considering its low seedling regeneration rate in nature, Xiong et al. [13] successfully established a tissue culture protocol, using young stem sections with apical shoot buds, for the mass propagation of *C. subcordata*, which provided a powerful tool for preserving this endangered plant species and may lead to its future usage in reforestation and coral island preservation.

However, previous research on *C. subcordata* has mainly focused on its biological properties and morphological structures but does not include genome or related genetic studies. Since the discovery of the DNA double helix structure, many scientists have put a lot of effort into determining the sequence of genomes, and more and more efficient sequencing technologies have been designed to accurately sequence genomes. With the increasing recognition of high-quality genome sequences, the assisted assembly technology for chromosome-level genomes has emerged [14]. In recent years, chromosome-scale genomes have been reported in numerous plant species, which provide insights into their genetic basis and population genetic structure [15–17]. In this study, we sequenced the whole genome and transcriptome of *C. subcordata* with the BGISEQ (BGI-Shenzhen, Shenzheng, China) and OXFORD NANOPORE (Oxford Nanopore Technologies, Oxford, United Kingdom) sequencing platforms and assembled and annotated its whole genome. Additionally, based on this genome, orthologous gene identification and phylogenetic analysis were performed. This deep insight into the genetic background of *C. subcordata* could be helpful in exploring its adaptation mechanisms for high temperature, high salinity, strong alkalinity, and other habitats in tropical coral islands, and in laying a foundation for germplasm conservation and genetic breeding.

2. Results

2.1. Genome Sequencing and Assembly

In this study, the BGISEQ and OXFORD NANOPORE sequencing platform produced approximately 64.3 Gb short whole genome sequencing (WGS) reads, 125.0 Gb long WGS reads, 75.3 Gb Hi-C reads, and 8 Gb RNA-seq reads.

The genome size was initially evaluated to be 524,294,375 bp, using distribution analysis, with a k-mer of 17 and short WGS reads. K-mer analysis also showed that this genome had a heterozygosity of 0.07% and a repetitive sequence content of 42.1%. The primarily assembled genome was 475,321,542 bp, with 92 contigs and a contig N50 length of 16,399,519 bp (Table 1). To obtain high-quality assembly, the contigs were corrected and scaffolded using Hi-C reads. Finally, the assembled genome was 470,353,539 bp in length, with 468,670,648 bp (99.22%) of sequences assembled into 16 chromosomes, and the chromosomes ranged in size from 18,188,694 bp (Chromosome 1) to 42,352,411 bp (Chromosome 16) (Figure 2A, Table S1).

Table 1. The statistical results of the genome assembly.

Type	Contig Length (bp)	Contig Number
N50	16,399,519	10
N60	15,666,772	13
N70	14,697,627	16
N80	12,626,804	19
N90	10,091,180	23
Longest	42,352,411	1
Total	475,321,542	92

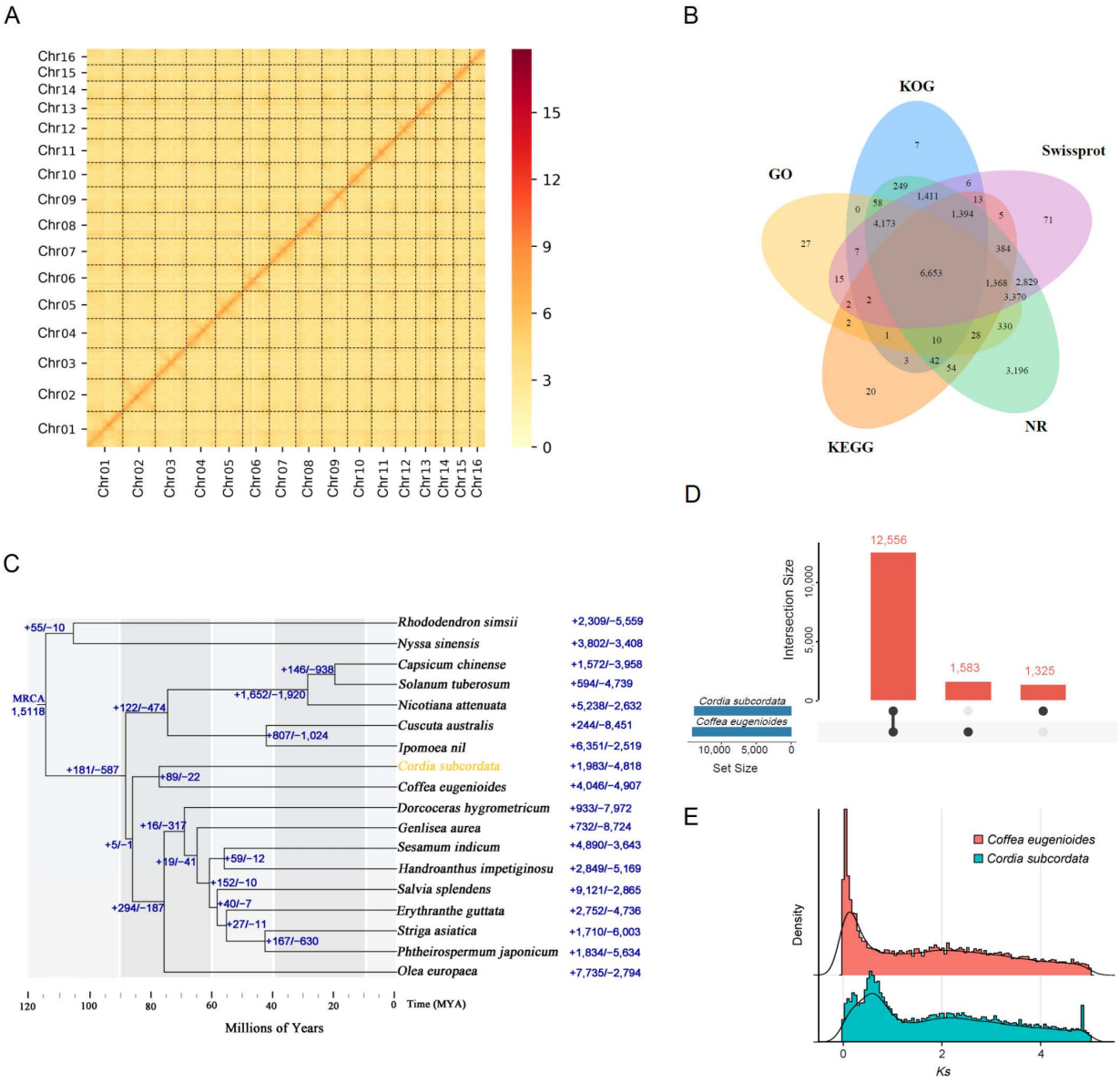

Figure 2. Genome assembly, annotation, and phylogenetic analysis of *Cordia subcordata*. (**A**) Hi-C interaction heat map (bin length 10,000 bp), (**B**) Venn graph of annotation in five databases, (**C**) phylogenetic tree with mapped gene family expansions (+) and contractions (–) in *C. subcordata* and other species, (**D**) shared gene families between *C. subcordata* and *Coffea eugenioides*, (**E**) whole genome duplication analysis with comparisons of the synonymous substitutions per synonymous site (Ks) value in *C. subcordata* and *C. eugenioides*.

2.2. Completeness of the Genome and Quality Evaluation

BUSCO analysis indicated that 95.28% of the core genes were completely captured by the genome assembly, which included 90.49% complete and single-copy and 4.79% complete and duplicated genes, while 0.97% were captured as fragments, and 3.75% were missing (Table 2). Additionally, a simple assessment of genome integrity, by mapping short WGS reads to the assembled genome, indicated that 97.90% of the reads were properly mapped. These evaluations implied a high completeness of the *C. subcordata* genome assembly.

Table 2. BUSCO results.

Type	Number	Percent (%)
Complete BUSCOs (C)	1372	95.28
Complete and single-copy BUSCOs (S)	1303	90.49
Complete and duplicated BUSCOs (D)	69	4.79
Fragmented BUSCOs (F)	14	0.97
Missing BUSCOs (M)	54	3.75
Total BUSCO groups searched	1440	100

2.3. Repeat Sequence Prediction

A total of 256,983,673 bp sequences were identified as repetitive elements by different repeat-identifying programs, accounting for 54.41% of the assembled genome (Table S2). The most abundant repetitive elements were transposable element (TE) sequences, with 241,514,515 bp (51.13%), followed by unknown-type repeats, with 10,733,217 bp (2.27%). Within TE sequences, the most abundant were terminal repeats (LTRs), with 167,223,552 bp (35.4%). The tandem repeats comprised 4,121,922 bp (0.87%), and SSRs comprised 219,915 bp (0.05%) in the assembled genome.

2.4. Gene Prediction and Annotation

A total of 26,615 protein-coding genes were predicted in the *C. subcordata* assembled genome. The average gene length was 4074 bp, and the average coding sequence (CDS) length was 1268 bp. Each predicted gene contained 5.4 exons, with a mean sequence length of 236 bp (Table S3). In addition, 3010 non-coding RNAs, including 928 (0.015%) tRNAs, 282 (0.056%) rRNAs, 90 (0.003%) miRNAs, and 297 (0.004%) snRNAs, were predicted in the *C. subcordata* assembled genome (Table S4). Overall, 25,730 genes (96.67%) were functionally annotated in at least one database (Figure 2B), including 21,703 (81.54%) in the Swiss-Prot protein database, 9981 (37.50%) in the Kyoto Encyclopedia of Gene and Genomes (KEGG) database, 14,029 (52.71%) in the Eukaryotic Orthologous Groups of protein (KOG) database, 16,046 (60.29%) in the Gene Ontology (GO) database, and 25,549 (95.99%) in the Non-Redundant Protein (NR) database (Table 3).

Table 3. The statistics of annotated genes in different databases.

Databases	Number	Percent (%)
Swiss-Prot	21,703	81.54
KEGG	9981	37.50
KOG	14,029	52.71
GO	16,046	60.29
NR	25,549	95.99
Total	25,730	96.67

2.5. Orthologous Gene Identification and Phylogenetic Analysis

A total of 614,770 genes from 18 species were used for orthologous gene group identification, and 30,853 orthologous gene groups were obtained using Orthofinder. For *C. subcordata*, 25,078 genes were assigned to 13,881 orthologous gene groups, and 217 orthologous gene groups contained 737 genes specific to *C. subcordata* (Table S5). The phylogenetic

tree constructed using 336 single-copy genes resolved that *C. subcordata* was a sister to *Coffea eugenioides*, and the divergence time between them was estimated to be 77 MYA (Figure 2C). Both species shared 12,556 gene families, and 1583 were *C. subcordata* specific (Figure 2D).

Gene family evolution analysis indicated that 1983 gene families exhibited expansion, and 4818 families exhibited contraction in *C. subcordata*. For expanded and contracted gene families in *C. subcordata*, 51 gene families, consisting of 639 genes, were significantly expanded ($p < 0.05$), and 44 gene families, consisting of 72 genes, were significantly contracted. Enrichment analysis showed that the significantly expanded gene families were mainly related to the purine, pigment, and lipid metabolic/biosynthetic process in the GO biological process category (Table S6), and to isoquinoline alkaloid biosynthesis, terpenoid backbone biosynthesis, tyrosine metabolism, and ion channels in KEGG (Table S7). The significantly contracted gene families were mainly related to protein phosphorylation in the GO biological process category (Table S8), and to the MAPK signaling pathway, plant–pathogen interaction, and signal transduction in KEGG (Table S9).

2.6. Whole Genome Duplication and Gene Duplication

Whole genome duplication (WGD) analysis indicated that *C. subcordata* underwent two round WGD events (Figure 2E), and the most ancient one was shared with *C. eugenioides*, its sister species in the phylogeny analysis (Figure 2C).

Gene duplications indicated that there were 10,309 WGD-type genes, 2902 tandem duplications (TD)-type genes, 1033 proximal duplications (PD)-type genes, 49 transposed duplications (TRD)-type genes and 8044 dispersed duplications (DD)-type genes in *C. subcordata*. Enrichment analysis indicated that WGD-related genes in *C. subcordata* are mainly associated with monoatomic ion transport, DNA-templated transcription, and cell walls in the GO biological process category (Table S10), and with G protein-coupled receptors, phagosome, and the photosynthesis of antenna proteins in KEGG (Table S11). Enrichment analysis indicated the TD-related genes in *C. subcordata* are mainly associated with defense response and amino sugar catabolic processes in the GO biological process category (Table S12), and with terpenoid biosynthesis, stilbenoid, diarylheptanoid and gingerol biosynthesis, and isoquinoline alkaloid biosynthesis in KEGG (Table S13). Enrichment analysis indicated that PD-related genes in *C. subcordata* are mainly associated with tricarboxylic acid metabolic process in GO biological process category (Table S14), and with the biosynthesis of various plant secondary metabolites and isoquinoline alkaloid biosynthesis in KEGG (Table S15).

3. Discussion

Boraginaceae is the only family in Boraginales that is a core group and is one of the largest groups within the lamiids clade, which is rarely published in genome aspects, so the genome of *C. subcordata* is important for the Boraginaceae family. Generally, Boraginaceae can be classified into four subfamilies: Boraginaceae, Heliotropiaceae, Ehretiaceae, and Cordiaceae [18,19]. However, due to the rapid evolutionary divergence within the Boraginaceae family, the phylogenetic relationships in the family remain elusive [19]. The Cordiaceae subfamily, in addition to *Cordia*, may also include the *Varronia* and *Patagonula* genera [19,20], which means the species in these two genera are the most genetically close to *Cordia*. Cohen [18] conducted both morphological and molecular phylogenetic studies on the Boraginaceae family. Based on combined morphological and molecular markers (including *mat*k, *ndh*F, and *trn*L-*trn*F markers of cpDNA), his results showed that Cordiaceae, Heliotropiaceae, and Ehretiaceae were more phylogenetically closed and formed a clade, but, in the clade, their relationships could not be further resolved [18]. Moreover, his results using only molecular markers (including *mat*k, *ndh*F, and *trnL-trn*F markers of cpDNA, and ITS markers of nrDNA) indicated that Cordiaceae is the most closed to Heliotropiaceae. Nevertheless, further studies using *rbc*L, *ndh*F, and *trn*L-*trn*F markers of cpDNA [19] and whole cpDNA genomes [20] indicated that Cordiaceae is sister to

Ehretiaceae. Since no *Varronia*, *Patagonula*, Ehretiaceae, or Heliotropiaceae genomes are available currently, we could not include them in our phylogeny analysis. Therefore, we expanded our comparative genomic analysis to include more species in Lamiids, as their genomes have been assembled and annotated, which resulted in 15 Lamiids species in 11 families, allowing us to accurately construct the phylogeny of our studied species. Our phylogenetic analysis indicated that *C. subcordata* was sister to *Coffee eugenioides*, when none of the other Boraginaceae species were included. This result was consistent with the results of Alshegaihi et al. [21]. In their study, they assembled the complete cpDNA genome of *C. monoica*. Their phylogeny analysis, using the *C. monoica* cpDNA genome, indicated that Boraginaceae was sister to Rubiaceae (*Coffea arabica* used in their study).

Complete and accurate genome assembly is generally hindered by the large genome size (>1 Gb), high heterozygosity (>0.5%), and repetitive sequences (>50%) in all genomes [22]. The genome features in this study, discovered by the k-mer method with short WGS reads, indicated that *C. subcordata* has an intermediate genome size, low heterozygosity, and repetitive sequences, which facilitated its assembling, which in turn was especially aided with long-read sequencing technology. In this study, the high-quality 470.35 Mb chromosome-level genome of *C. subcordata* was obtained using the Hi-C reads to improve genome assembly. Its genome size was larger than *Echium plantagineum* (351.50 Mb) and *Lithospermum erythrorhizon* (367.41 Mb), which are reported to be in the family Boraginaceae [23,24]. However, compared to the genome assembly of *E. plantagineum* and *L. erythrorhizon*, the predicted protein-coding genes (26,615) in *C. subcordata* were smaller than both of the above, i.e., 42,316 in *E. plantagineum* [23] and 27,720 in *L. erythrorhizon* [24].

A comprehensive genome size estimation conducted for thirty-eight taxa (274 individuals) in the Boraginaceae family from the Czech Republic using flow cytometry revealed that the lowest genome size in the family was about 274 Mb and the largest was 16 Gb, and that most species had a size between 500 Mb and 1.5 Gb [25]. Therefore, *C. subcordata*, as well as the *E. plantagineum* and *L. erythrorhizon* mentioned above, had a relatively small genome size for a member of the Boraginaceae family. A study done by Kobrlova and Hrones [25] also indicated that the genome sizes in the Boraginaceae family were correlated with its life history, ecology, and phylogeny. Perennial plants and species living in natural habitats harbored relatively larger genome sizes in the Boraginaceae family.

Chemical components are accumulated in different parts in different Boraginaceae species [7,8]. For example, alkannin/shikonin are distinctive secondary metabolites and commonly found in the root periderm of *E. plantagineum* and *L. erythrorhizon* [23,24]. The genomes of these two species were used to reveal the alkannin/shikonin pathway and to find some key genes in this pathway. Using CAFE, we detected variations in gene family size, resulting from the gain or loss of genes, which represent the expansion or contraction of the gene family. Expanded gene families are usually created by gene duplication events during selection [26]. Duplicated genes can increase the quantity of protein products, benefiting plant adaptation [27]. We observed both alkaloid- and terpenoid-related gene families which were expanded in *C. subcordata*, and these genes showed TD or PD duplication. TD and PD are central duplication processes related to the biotic and abiotic environmental adaption abilities of plant species [28]. Alkaloid and terpenoid are important chemicals produced by plants. Isoquinoline alkaloids (IQAs) are derived from phenylalanine and tyrosine and constitute one of the largest and most diverse groups of alkaloids. IQAs have only been discovered in a limited number of plant species [29,30] and play a key role in the defense against pathogens and herbivores [31–33]. IQAs also exhibit exceptionally important pharmacological activities [30] and therefore have high pharmaceutical and commercial value. Terpenes are important natural products with a wide range of applications in plants [34–36], but they most often serve as continuously available passive toxic defenses against biological enemies [35]. The *Cordia* species are distributed widely in the tropical regions [9]. The tropics face constant challenges of diverse pests and diseases, and plant defense chemicals play a very important role in their deterrence [37,38]. In prior reports, various chemical components, including sterols, flavonoids, terpenes, alkaloids,

and phenolic acids, have been isolated from members of the *Cordia* species [5,9,39–42]. A recent study, using *C. subcordata* leaf extracts to evaluate its antioxidant properties, identified abundant chemical components, namely polyphenols [43]. From our genome report, further metabolomic analyses are required to identify the chemical constituents of alkaloids and terpenoids in *C. subcordata*.

Interestingly, these expanded gene families are accompanied by the contraction of other biological enemy defense gene families, such as the MAPK signaling pathway. The MAPK signaling pathway has been demonstrated to activate responses and resistance to plant diseases [44–46], and MAPK is also a core mediator for the hypersensitive response and subsequent cell death [45,47]. The contraction of MAPK and the expansion of secondary metabolites gene families could represent a trade-off between active and passive defense, where the utilization of MAPK represents an active approach, while alkaloid and terpenoid, the secondary metabolites, represent a passive approach [48–51]. Therefore, chemicals, such as alkaloids and terpenoids, in *C. subcordata* may be the main mechanism for disease defense on tropical coral islands, and MAPK may have reduced their role in defense. Keeping the balance between gene family expansion and contraction, which maintains the balanced allocation of energy sources, has also been observed in previous studies [52–54]. Considering no relevant studies have been carried out regarding *C. subcordata*, future studies are needed to confirm this in a laboratory setting.

4. Materials and Methods

4.1. Sequencing

Leaf samples from one *C. subcordata* individual, planted in the South China Botanical Garden, were collected. The individual was propagated by a shoot cut from a plant from Yongxing Island of the Paracel Islands (16°49′53″ N, 112°20′22″ E; China). After total RNA and genomic DNA were extracted from the samples, the sequencing libraries, including short- and long-whole genome sequencing (WGS), transcriptome, and Hi-C sequencing, were constructed and then sequenced using the BGISEQ and OXFORD NANOPORE sequencing platforms (Table 4).

Table 4. The sequencing information of *C. subcordata* genome.

Sequencing Project	Sequencing Platform	Sequence Type	Instrument
Short whole genome sequencing	BGISEQ	PE150	DNBSEQ-T7
Long whole genome sequencing	OXFORD NANOPORE	Long reads	GridION
Short-length transcriptome	BGISEQ	PE150	DNBSEQ-T7
Full-length transcriptome	BGISEQ	PE150	DNBSEQ-T7
Hi-C	BGISEQ	PE150	DNBSEQ-T7

4.2. Assembly

Before genome assembly, the genome size of *C. subcordata* was estimated using k-mer method, under a k-mer of 17 and short WGS reads. NextDenovo v1.0 [55] was used to correct the long WGS reads with the parameter of "read_cuoff = 1 k, seed_cutoff = 25 k". The corrected long WGS reads were then inputted into Smartdenovo v1.0.0 [56], to perform assembly, with the parameters of "-k 21, -j 3000". After assembly, NextPolish v1.0.1 [57] was used to polish the assembly. LACHESIS software (https://github.com/shendurelab/LACHESIS accessed on 5 November 2023) [58] was further used to cluster the assembled contigs into chromosome groups with Hi-C reads. The evaluated assembly was performed by BUSCO v3.0.1 [59], as well as the mapping of the short WGS against the assembly.

4.3. Annotation

Tandem repeats in the assembly were searched using GMATA v 2.2 [60] and the Tandem Repeats Finder (TRF) v 4.07b [61], with default parameters. The transposable elements (TEs) were identified using MITE-hunter [62], LTR_finder [63], ltr_harverst [64],

LTR_retriver [65], and RepeatModeler v1.0.11 [66]. These inferred results were then combined and inputted into RepeatMasker v1.331 [67], to obtain whole genome repeat sequences.

Gene structure prediction was combined with ab initio-, homology-, and RNA-seq-based methods. Augustus v 3.3.1 [68] was used for ab initio-based gene prediction, and GeMoMa v1.6.1 [69] was used for homology-based gene prediction. Hisat2 v2.1.0 [70], Stringtie v1.3.4d [71], and PASA v2.3.3 [72] were used for RNA-seq-based gene prediction. Finally, EVM [73] and TransposonPSI [74] were applied to integrate gene prediction results and obtain a consensus gene set. For the predicted genes, the annotation of gene functions was carried out by comparing them with the protein sequences of the NR [75], KEGG [76], KOG [77], Swissport [78], and GO [79] databases. Noncoding RNAs, including rRNA, small RNA, cis-regulatory elements, and tRNA, were identified by combined Infernal [80], tRNAscan-SE [81], and RNAmmer [82] programs and the Rfam [83] database.

4.4. Orthologous Gene Group Identification and Whole Genome Duplication Analysis

OrthoFinder 2.4.0 [84,85] was used to identify the orthologous gene groups (i.e., gene families) among *C. subcordata* and the other 17 species (Table S16). These species included 15 Lamiids species in 11 families and two outgroup species (*Rhododendron simsii* and *Nyssa sinensis*). After ortholog group identification, single-copy ortholog sequences were selected by OrthFinder, to perform phylogeny analysis among the species. Based on phylogeny results, species divergence times were estimated by TreePL [86,87] with 5 species pairs (Table 5) used as time calibration points and their estimated divergence time (million years ago, MYA) derived from http://timetree.org/. Then, using the phylogenic tree, containing species divergence time, expansions and contractions of the orthologous gene families were estimated using CAFE v5 [88]. For significantly expanded and contracted gene families, enrichment analyses by GO and the KEGG database were performed using TBtools v1.115 [89]. Ancient whole genome duplication events in *C. subcordata* and its sister species *C. eugenioides* were detected using wgd v1.2 [90]. Gene duplications in *C. subcordata* were examined by Doubletrouble v0.99.1 [91]. These duplications were classified into WGD, tandem duplications (TD), proximal duplications (PD), transposed duplications (TRD), and dispersed duplications (DD) [21]. In the doubletrouble analysis, *Coffea eugenioides* was used as the outgroup species. Finally, for WGD-, TD-, and PD-derived genes, enrichment analysis, according to the GO and KEGG databases, was applied using TBtools.

Table 5. The reference species' differentiation times from Timetree.

Species Pair	Taxon 1	Taxon 2	Estimated Divergent Time (Million Years Ago, MYA)
1	*Phtheirospermum japonicum*	*Striga asiatica*	39–64
2	*Handroanthus impetiginosus*	*Sesamum indicum*	56–73
3	*Coffea eugenioides*	*Cordia subcordata*	75–96
4	*Capsicum chinense*	*Solanum tuberosum*	18.2–31.4
5	*Rhododendron simsii*	*Nyssa sinensis*	105–119

5. Conclusions

In summary, we applied both short- and long-read sequencing technologies to assemble the *C. subcordata* genome. The assembled genome was 470,353,539 bp in length, with 468,670,648 bp (99.22%) of sequences assembled into 16 chromosomes. A total of 26,615 genes were predicted and functionally annotated by various databases. The first reported *Cordia* genome provides important information regarding the phylogeny evolution of this genus and helps explore the adaptive mechanism of *C. subcordata* on tropical coral islands. Moreover, with the genomes of the other Boraginaceae species currently sequenced and available, comparative genome analyses including these other genomes will more

clearly address these mechanisms, particularly in elucidating the biosynthesis related to the various chemical components (such as alkaloid and terpenoid) of *C. subcordata*.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms242216273/s1.

Author Contributions: D.-M.L. and Z.-F.W. designed and conceived the project, D.-M.L. and Z.-F.W. collected the plant materials for sequencing, Z.-F.W. and Y.-L.C. performed the analyses, Y.-L.C. wrote the manuscript draft, and Z.-F.W., S.-G.J., H.-M.L. and D.-M.L. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Key Research and Development Program of China (2022YFC3103802), the Key Research Program of the Chinese Academy of Sciences Grant (KGFZD-135-19-08), the Guangdong Science and Technology Program (No. 2019B121201005), and the National Key Research & Development Program of China (2022YFC3103700).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the genome and raw sequencing reads described in this article are publicly available in the National Center for Biotechnology Information (NCBI) database under project PRJNA909849. The BGISEQ and OXFORD NANOPORE raw sequencing data are deposited under the accession number SRR23354649, SRR23354648, and SRR23354594, and the genome under the accession number ASM2855475v1. Genome assembly and gene annotation data are available at https://doi.org/10.6084/m9.figshare.21286518 accessed on 5 November 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xing, F.W.; Deng, S.W. *Flora of the South China Sea Island*; China Forestry Press: Beijing, China, 2018.
2. Liu, J.X.; Xi, Y.Z.; Ning, J.C.; Zhang, J.M.; Li, Y.X.; Zhao, Y.Y.; Sun, X.H. Pollen morphology and exine ultrastructure of genus *Cordia* in Boraginaceae in China. *Acta. Bot. Sin.* **2001**, *43*, 893–898.
3. Bouby, L.; Bouchette, A.; Figueiral, I. Sebesten fruits (*Cordia myxa* L.) in Gallia Narbonensis (Southern France): A trade item from the Eastern Mediterranean? *Veget. Hist. Archaeobot.* **2011**, *20*, 397–404. [CrossRef]
4. Wang, W. An enumeration of the Boraginaceous plants collected by H.smith from China during 1921–22, 1924 and 1934. *Bull. Bot. Res.* **1993**, *13*, 1–10.
5. Matias, E.F.F.; Alves, E.F.; Silva, M.K.D.N.; Carvalho, V.R.D.A.; Coutinho, H.D.M.; Costa, J.G.M. The genus *Cordia*: Botanists, ethno, chemical and pharmacological aspects. *Rev. Bras. Farmacogn.* **2015**, *25*, 542–552. [CrossRef]
6. Huang, Z.G.; Zhang, W.Q. A discussion on the quaternary climate record from the coral reef in tropical China. *Trop. Geogr.* **2008**, *2*, 11–15.
7. Jian, S.G. Vegetation of tropical coral islands in China. *Guihaia* **2020**, *40*, 443.
8. Wu, S.H.; Chen, H.W.; Jian, S.G.; Liu, H.; Zhang, W.; Ren, H. The biological characteristics of *Cordia subcordata* on tropical coral island in China. *Ecol. Sci.* **2017**, *36*, 57–63.
9. Oza, M.J.; Kulkarni, Y.A. Traditional uses, phytochemistry and pharmacology of the medicinal species of the genus *Cordia* (Boraginaceae). *J. Pharm. Pharmacol.* **2017**, *69*, 755–789. [CrossRef]
10. Nisha, M.; Chandru, N.; Pradeep, P.; Selvarasu, P.; Surendra Kumar, M.; Latha, S.T.; Astalakshmi, N. A review on phytochemical and pharmacological activity of *Cordia sebestena*. *Int. J. Pharm. Sci. Rev. Res.* **2022**, *73*, 156–161.
11. Liu, D.M.; Chen, H.F.; Wang, F.G.; Yi, Q.F.; Xing, F.W. Investigation of introduced plants in Nansha islands and reefs, China. *J. Trop. Subtrop. Bot.* **2015**, *23*, 167–175.
12. Wang, X.P.; Wen, M.H.; Wu, M.S.; Zhang, D.X. *Cordia subcordata* (Boraginaceae), a distylous species on oceanic coral islands, is self-compatible and pollinated by a passerine bird. *Plant Ecol. Evol.* **2020**, *153*, 361–372. [CrossRef]
13. Xiong, Y.P.; Chen, X.H.; Wu, K.L.; da Silva, J.A.T.; Zeng, S.J.; Ma, G.H. Shoot organogenesis and plant regeneration in *Cordia subcordata* Lam. *In Vitro Cell. Dev. Biol. Plant* **2022**, *58*, 392–398. [CrossRef]
14. Giani, A.M.; Gallo, G.R.; Gianfranceschi, L.; Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 9–19. [CrossRef] [PubMed]
15. Sahu, S.K.; Liu, M.; Chen, Y.W.; Gui, J.S.; Fang, D.M.; Chen, X.L.; Yang, T.; He, C.Z.; Cheng, L.; Yang, J.L.; et al. Chromosome-scale genomes of commercial timber trees (*Ochroma pyramidale*, *Mesua ferrea*, and *Tectona grandis*). *Sci. Data* **2023**, *10*, 512. [CrossRef]
16. Wang, Z.F.; Zhang, X.; Lei, W.X.; Zhu, H.; Wu, S.D.; Liu, B.B.; Ru, D.F. Chromosome-level genome assembly and population genomics of *Robinia pseudoacacia* reveal the genetic basis for its wide cultivation. *Commun. Biol.* **2023**, *6*, 797. [CrossRef]

17. Qin, Y.; Zhao, H.; Han, H.W.; Zhu, G.P.; Wang, Z.S.; Li, F.D. Chromosome-level genome assembly and population genomic analyses reveal geographic variation and population genetic structure of *Prunus tenella*. *Int. J. Mol. Sci.* **2023**, *24*, 11735. [CrossRef]
18. Cohen, J.I. A phylogenetic analysis of morphological and molecular characters of Boraginaceae: Evolutionary relationships, taxonomy, and patterns of character evolution. *Cladistics* **2014**, *30*, 139–169. [CrossRef]
19. Hasenstab-Lehman, K.E. Phylogenetics of the Borage family: Delimiting Boraginales and assessing closest relatives. *Aliso A J. Syst. Florist. Bot.* **2017**, *35*, 41–49. [CrossRef]
20. Alawfi, M.S.; Albokhari, E.J. Comparative chloroplast genomics reveals a unique gene inversion in two *Cordia* trees (Cordiaceae). *Forests* **2023**, *14*, 1778. [CrossRef]
21. Alshegaihi, R.M.; Mansour, H.; Alrobaish, S.A.; Al Shaye, N.A.; Abd El-Moneim, D. The first complete chloroplast genome of *Cordia monoica*: Structure and comparative Analysis. *Genes* **2023**, *14*, 976. [CrossRef]
22. Sun, X.; Wang, G.; Yang, J.; Yu, W.; Xu, J.; Tang, B.; Ding, G.; Zhang, D. Whole genome evaluation analysis and preliminary assembly of *Oratosquilla oratoria* (Stomatopoda: Squillidae). *Mol. Biol. Rep.* **2023**, *50*, 4165–4173. [CrossRef]
23. Tang, C.Y.; Li, S.; Wang, Y.T.; Wang, X. Comparative genome/transcriptome analysis probes Boraginales' phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. *Mol. Ecol. Resour.* **2020**, *20*, 228–241. [CrossRef]
24. Auber, R.P.; Suttiyut, T.; McCoy, R.M.; Ghaste, M.; Crook, J.W.; Pendleton, A.L.; Widhalm, J.R.; Wisecaver, J.H. Hybrid de novo genome assembly of red gromwell (*Lithospermum erythrorhizon*) reveals evolutionary insight into shikonin biosynthesis. *Hortic. Res.* **2020**, *7*, 82–97. [CrossRef] [PubMed]
25. Kobrlova, L.; Hrones, M. First insights into the evolution of genome size in the borage family: A complete data set for Boraginaceae from the Czech Republic Bot. *J. Linn. Soc.* **2019**, *189*, 115–131. [CrossRef]
26. Fang, Y.; Jiang, J.; Hou, X.; Guo, J.; Li, X.; Zhao, D.; Xie, X. Plant protein-coding gene families: Their origin and evolution. *Front. Plant Sci.* **2022**, *13*, 995746. [CrossRef]
27. Tao, W.; Li, R.; Li, T.; Li, Z.; Li, Y.; Cui, L. The evolutionary patterns, expression profiles, and genetic diversity of expanded genes in barley. *Front. Plant Sci.* **2023**, *14*, 1168124. [CrossRef]
28. Qiao, X.; Li, Q.H.; Yin, H.; Qi, K.; Li, L.; Wang, R.; Zhang, S.; Paterson, A.H. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **2019**, *20*, 38. [CrossRef] [PubMed]
29. Yamada, Y.; Kokabu, Y.; Chaki, K.; Yoshimoto, T.; Ohgaki, M.; Yoshida, S.; Kato, N.; Koyama, T.; Sato, F. Isoquinoline alkaloid biosynthesis is regulated by a unique bHLH-Type transcription factor in *Coptis japonica*. *Plant Cell Physiol.* **2011**, *52*, 1131–1141. [CrossRef]
30. Singh, S.; Pathak, N.; Fatima, E.; Negi, A.S. Plant isoquinoline alkaloids: Advances in the chemistry and biology of berberine. *Eur. J. Med. Chem.* **2021**, *226*, 113839. [CrossRef] [PubMed]
31. Zeng, W.Y.; Sun, Z.D.; Lai, Z.G.; Yang, S.Z.; Chen, H.Z.; Yang, X.H.; Tao, J.R.; Tang, X.M. Determination of the miRNAs related to bean pyralid larvae resistance in soybean using small RNA and transcriptome sequencing. *Int. J. Mol. Sci.* **2019**, *20*, 2966. [CrossRef]
32. Ren, Z.Y.; Fang, M.K.; Muhae-Ud-Din, G.; Gao, H.F.; Yang, Y.Z.; Liu, T.G.; Chen, W.Q.; Gao, L. Metabolomics analysis of grains of wheat infected and noninfected with *Tilletia controversa* Kuhn. *Sci. Rep.* **2021**, *11*, 18876. [CrossRef] [PubMed]
33. Yang, C.C.; Wu, P.F.; Yao, X.H.; Sheng, Y.; Zhang, C.C.; Lin, P.; Wang, K.L. Integrated transcriptome and metabolome analysis reveals key metabolites involved in *Camellia oleifera* defense against anthracnose. *Int. J. Mol. Sci.* **2022**, *23*, 536. [CrossRef] [PubMed]
34. Chen, F.; Tholl, D.; Bohlmann, J.; Pichersky, E. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **2011**, *66*, 212–229. [CrossRef] [PubMed]
35. Pichersky, E.; Raguso, R.A. Why do plants produce so many terpenoid compounds? *New Phytol.* **2016**, *220*, 655–658. [CrossRef]
36. Jiang, S.Y.; Jin, J.J.; Sarojam, R.; Ramachandran, S. A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* **2019**, *11*, 2078–2098. [CrossRef] [PubMed]
37. Salazar, D.; Lokvam, J.; Mesones, I.; Pilco, M.V.; Zuñiga, J.M.A.; de Valpine, P.; Fine, P.V.A. Origin and maintenance of chemical diversity in a species-rich tropical tree lineage. *Nat. Ecol. Evol.* **2018**, *2*, 983–990. [CrossRef]
38. Hallam, A.; Read, J. Do tropical species invest more in anti-herbivore defence than temperate species? A test in Eucryphia (Cunoniaceae) in eastern Australia. *J. Trop. Ecol.* **2006**, *22*, 41–51. [CrossRef]
39. Abdel-Aleem, E.R.; Attia, E.Z.; Farag, F.F.; Samy, M.N.; Desoukey, S.Y. Total phenolic and flavonoid contents and antioxidant, anti-inflammatory, analgesic, antipyretic and antidiabetic activities of *Cordia myxa* L. leaves. *Clin. Phytosci.* **2019**, *5*, 29. [CrossRef]
40. Ma, Z.Q.; Lu, C.Q.; Wang, Y.; Wang, Q. Phenolpropane compounds of *Cordia dichotoma* and their anticomplementary activities. *Chem. Nat. Compd.* **2021**, *57*, 169–170. [CrossRef]
41. Raghuvanshi, D.; Sharma, K.; Verma, R.; Kumar, D.; Kumar, H.; Khan, A.; Valko, M.; Alomar, S.Y.; Alwasel, S.H.; Nepovimova, E.; et al. Phytochemistry, and pharmacological efficacy of *Cordia dichotoma* G. Forst. (Lashuda): A therapeutic medicinal plant of Himachal Pradesh. *Biomed. Pharmacother.* **2022**, *153*, 113400. [CrossRef]
42. Ahmed Bokhari, S.W.; Sharif, H.; Umer Gilani, S.M.; Ali, S.T.; Ahmed, S.; Ahmed Siddiqui, M.U.; Mohtasheemul Hasan, M. Pharmacognostic and phytochemical study of the flowers of *Cordia sebestena* L. *Pak. J. Pharm. Sci.* **2022**, *35*, 69–76. [PubMed]

43. Chambon, M.; Ho, R.; Baghdikian, B.; Herbette, G.; Bun-Llopet, S.-S.; Garayev, E.; Raharivelomanana, P. Identification of atioxidant metabolites from five plants (*Calophyllum inophyllum*, *Gardenia taitensis*, *Curcuma longa*, *Cordia subcordata*, *Ficus prolixa*) of the polynesian pharmacopoeia and cosmetopoeia for skin care. *Antioxidants* **2023**, *12*, 1870. [CrossRef]
44. Zhang, S.Q.; Klessig, D.F. MAPK cascades in plant defense signaling. *Trends Plant Sci.* **2001**, *6*, 520–527. [CrossRef] [PubMed]
45. Meng, X.Z.; Zhang, S.Q. MAPK cascades in plant disease resistance signaling. *Annu. Rev. Phytopathol.* **2013**, *51*, 245–266. [CrossRef]
46. Bigeard, J.; Hirt, H. Nuclear signaling of plant MAPKs. *Front. Plant Sci.* **2018**, *9*, 469. [CrossRef] [PubMed]
47. Huang, X.X.; Zhu, G.Q.; Liu, Q.; Chen, L.; Li, Y.J.; Hou, B.K. Modulation of plant salicylic acid-associated immune iesponses via glycosylation of dihydroxybenzoic acids. *Plant Physiol.* **2018**, *176*, 3103–3119. [CrossRef]
48. Pedley, K.F.; Martin, G.B. Role of mitogen-activated protein kinases in plant immunity. *Curr. Opin. Plant Biol.* **2005**, *8*, 541–547. [CrossRef]
49. Mithöfer, A.; Maffei, M.E. General mechanisms of plant defense and plant Toxins. In *Plant Toxins*; Carlini, C., Ligabue-Braun, R., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 3–24.
50. Sood, M.; Kapoor, D.; Kumar, V.; Kalia, N.; Bhardwaj, R.; Sidhu, G.P.S.; Sharma, A. Mechanisms of plant defense under pathogen stress: A review. *Curr. Protein Pept. Sci.* **2021**, *22*, 376–395. [CrossRef]
51. Kaur, S.; Samota, M.K.; Choudhary, M.; Choudhary, M.; Pandey, A.K.; Sharma, A.; Thakur, J. How do plants defend themselves against pathogens-Biochemical mechanisms and genetic interventions. *Physiol. Mol. Biol. Plants* **2022**, *28*, 485–504. [CrossRef]
52. Liang, P.; Liu, S.; Xu, F.; Jiang, S.; Yan, J.; He, Q.; Liu, W.; Lin, C.; Zheng, F.; Wang, X.; et al. Powdery Mildews are characterized by contracted carbohydrate metabolism and diverse effectors to adapt to obligate biotrophic lifestyle. *Front. Microbiol.* **2018**, *9*, 3160. [CrossRef]
53. Sharma, V.; Hecker, N.; Roscito, J.G.; Foerster, L.; Langer, B.E.; Hiller, M. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **2018**, *9*, 1215. [CrossRef]
54. Zhang, T.; Qiao, Q.; Novikova, Y.P.; Wang, Q.; Yue, J.; Guan, Y.; Ming, S.; Liu, T.; De, J.; Liu, Y.; et al. Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 7137–7146. [CrossRef] [PubMed]
55. NextDenovo v1.0. Available online: https://github.com/Nextomics/NextDenovo (accessed on 28 May 2020).
56. Smartdenovo: Ultra-Fast de Novo Assembler Using Long Noisy Reads. Available online: https://github.com/ruanjue/smartdenovo (accessed on 28 May 2020).
57. Hu, J.; Fan, J.; Sun, Z.; Liu, S. NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **2020**, *36*, 2253–2255. [CrossRef]
58. Burton, J.N.; Adey, A.; Patwardhan, R.P.; Qiu, R.; Kitzman, J.O.; Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **2013**, *31*, 1119–1125. [CrossRef]
59. Simao, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
60. Wang, X.W.; Wang, L. GMATA: An integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **2016**, *7*, 1350. [CrossRef]
61. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic. Acids Res.* **1999**, *27*, 573–580. [CrossRef] [PubMed]
62. Han, Y.J.; Wessler, S.R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic. Acids Res.* **2010**, *38*, e199. [CrossRef]
63. Xu, Z.; Wang, H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic. Acids Res.* **2007**, *35*, W265–W268. [CrossRef] [PubMed]
64. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinf.* **2008**, *9*, 18. [CrossRef] [PubMed]
65. Ou, S.; Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long Terminal Repeat Retrotransposons. *Plant Physiol.* **2018**, *176*, 1410–1422. [CrossRef]
66. Bedell, J.A.; Korf, I.; Gish, W. MaskerAid: A performance enhancement to RepeatMasker. *Bioinformatics* **2000**, *16*, 1040–1041. [CrossRef]
67. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker. Available online: http://www.repeatmasker.org (accessed on 28 May 2020).
68. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **2008**, *24*, 637–644. [CrossRef]
69. Keilwagen, J.; Hartung, F.; Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **2019**, *1962*, 161–177.
70. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [CrossRef]
71. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295. [CrossRef] [PubMed]

72. Haas, B.J.; Delcher, A.L.; Mount, S.M.; Wortman, J.R.; Smith, R.K., Jr.; Hannick, L.I.; Maiti, R.; Ronning, C.M.; Rusch, D.B.; Town, C.D.; et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **2003**, *31*, 5654–5666. [CrossRef]
73. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E.; Orvis, J.; White, O.; Buell, C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **2008**, *9*, R7. [CrossRef] [PubMed]
74. Haas, B. TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies. Available online: https://transposonpsi.sourceforge.net (accessed on 28 May 2020).
75. NR: Non-Redundant Protein Sequence Database. Available online: https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins (accessed on 28 June 2020).
76. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
77. Galperin, M.Y.; Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **2015**, *43*, D261–D269. [CrossRef]
78. Stanke, M.; Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **2004**, *19*, II215–II225. [CrossRef]
79. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
80. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef] [PubMed]
81. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **1997**, *25*, 955–964. [CrossRef] [PubMed]
82. Lagesen, K.; Hallin, P.; Rodland, E.A.; Staerfeldt, H.H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [CrossRef] [PubMed]
83. Griffiths-Jones, S.; Moxon, S.; Marshall, M.; Khanna, A.; Eddy, S.R.; Bateman, A. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **2005**, *33*, D121–D124. [CrossRef]
84. Emms, D.M.; Kelly, S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, 157. [CrossRef]
85. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 238. [CrossRef]
86. Sanderson, M.J. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.* **2012**, *19*, 101–109. [CrossRef]
87. Smith, S.A.; O'Meara, B.C. treePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **2012**, *28*, 2689–2690. [CrossRef]
88. De Bie, T.; Cristianini, N.; Demuth, J.P.; Hahn, M.W. CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* **2006**, *22*, 1269–1271. [CrossRef] [PubMed]
89. Chen, C.J.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.H.; Xia, R. TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **2020**, *13*, 1194–1202. [CrossRef] [PubMed]
90. Zwaenepoel, A.; Van de Peer, Y. WGD-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **2019**, *35*, 2153–2155. [CrossRef] [PubMed]
91. Almeida-Silva, F.; Van de Peer, Y. Doubletrouble: Identification and Classification of Duplicated Genes. R Package Version 0.99.1. 2022. Available online: https://github.com/almeidasilvaf/doubletrouble (accessed on 21 July 2023).