frontiers in Plant Science

ORIGINAL RESEARCH
published: 30 September 2021
doi: 10.3389/fpls.2021.742420

High-Quality Genome of the Medicinal Plant *Strobilanthes cusia* Provides Insights Into the Biosynthesis of Indole Alkaloids

*Yongle Hu[1,2,3†], Dongna Ma[4†], Shuju Ning[5], Qi Ye[1], Xuanxuan Zhao[1], Qiansu Ding[4], Pingping Liang[4], Guoqian Cai[1], Xiaomao Ma[1], Xia Qin[1] and Daozhi Wei[1]**

[1] College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [2] College of Ecology and Resource Engineering, Wuyi University, Wuyishan, China, [3] Fujian Provincial Key Laboratory of Eco-Industrial Green Technology, Wuyishan, China, [4] Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China, [5] College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, China

OPEN ACCESS

Edited by:
Jianwei Zhang, Huazhong Agricultural University, China

Reviewed by:
Zhihua Wu, South-Central University for Nationalities, China
Huan Liu, Beijing Genomics Institute (BGI), China

***Correspondence:**
Daozhi Wei
weidz888@sohu.com

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *16 July 2021*
Accepted: *26 August 2021*
Published: *30 September 2021*

Citation:
Hu Y, Ma D, Ning S, Ye Q, Zhao X, Ding Q, Liang P, Cai G, Ma X, Qin X and Wei D (2021) High-Quality Genome of the Medicinal Plant Strobilanthes cusia Provides Insights Into the Biosynthesis of Indole Alkaloids. Front. Plant Sci. 12:742420. doi: 10.3389/fpls.2021.742420

Strobilanthes cusia (Nees) Kuntze is an important plant used to process the traditional Chinese herbal medicines "Qingdai" and "Nanbanlangen". The key active ingredients are indole alkaloids (IAs) that exert antibacterial, antiviral, and antitumor pharmacological activities and serve as natural dyes. We assembled the *S. cusia* genome at the chromosome level through combined PacBio circular consensus sequencing (CCS) and Hi-C sequencing data. Hi-C data revealed a draft genome size of 913.74 Mb, with 904.18 Mb contigs anchored into 16 pseudo-chromosomes. Contig N50 and scaffold N50 were 35.59 and 68.44 Mb, respectively. Of the 32,974 predicted protein-coding genes, 96.52% were functionally annotated in public databases. We predicted 675.66 Mb repetitive sequences, 47.08% of sequences were long terminal repeat (LTR) retrotransposons. Moreover, 983 *Strobilanthes*-specific genes (SSGs) were identified for the first time, accounting for ∼2.98% of all protein-coding genes. Further, 245 putative centromeric and 29 putative telomeric fragments were identified. The transcriptome analysis identified 2,975 differentially expressed genes (DEGs) enriched in phenylpropanoid, flavonoid, and triterpenoid biosynthesis. This systematic characterization of key enzyme-coding genes associated with the IA pathway and basic helix-loop-helix (bHLH) transcription factor family formed a network from the shikimate pathway to the indole alkaloid synthesis pathway in *S. cusia*. The high-quality *S. cusia* genome presented herein is an essential resource for the traditional Chinese medicine genomics studies and understanding the genetic underpinning of IA biosynthesis.

Keywords: *Strobilanthes cusia*, medicinal plant, whole-genome sequencing, lineage-specific genes, basic helix-loop-helix, indole alkaloid biosynthesis

INTRODUCTION

Strobilanthes cusia (Nees) Kuntze ($2n = 32$) is a perennial dicotyledonous herb of the order Acanthaceae and is broadly distributed from South to East Asia, the countries such as, India, Bangladesh, Thailand, Bhutan, and China (Hu et al., 2011). Generally, *S. cusia* grows in clay or moist soil in mountainous areas and is suitable for transplanting because of its ability to tolerate

different soils. As an important medicinal plant and natural dye, *S. cusia* has been cultivated and processed in China for thousands of years (Lin et al., 2018). For instance, the stems and leaves of *S. cusia* are regularly processed to obtain Qingdai (Indigo Naturalis), whereas the dried rhizomes and roots are known as Nanbanlangen (Rhizoma et Radix Baphicacanthis Cusiae), both of which are listed in the Chinese Pharmacopoeia as traditional medicines (Chinese Pharmacopoeia Committee, 2020). The medicinal uses of Qingdai include the treatment of various inflammatory diseases, oral ulcers, and skin diseases. Nanbanlangen has demonstrated efficacy in preventing and treating influenza A infection, mumps, and other infectious diseases (Tanaka et al., 2004; Yu et al., 2021). The phytochemical analyses showed that *S. cusia* can produce high quantities of biologically active compounds, such as, indole alkaloids (IAs), quinolone alkaloids, phenylethanoid glycosides, lignan glycosides, triterpenoids, steroids, amino acids, and flavonoids (Gu et al., 2014; Xiao et al., 2018; Yu et al., 2021). Among these chemical components, indigo and indirubin are the major medicinal ingredients and are isomers of each other ($C_{16}H_{10}N_2O_2$). The Chinese Pharmacopoeia stipulates that the mass fractions of indigo and indirubin in Qingdai should be higher than 2.0 and 0.13%, respectively, and indigo and indirubin are used as the criteria for identifying Nanbanlangen (Chinese Pharmacopoeia Committee, 2020). Further, clinical and pharmacological evidence suggests that the main alkaloids can treat leukemia (Wang et al., 2008; Wu et al., 2016) and dermatoses (Hsieh et al., 2012); protect against tissue damage (Huang et al., 2017); and exert anti-inflammatory (Sugimoto et al., 2016; Kawai et al., 2017), antibacterial (Chiang et al., 2013; Tsai et al., 2020), and immune-regulatory effects (Zhang et al., 2015; Jie et al., 2017).

Despite the medical importance of *S. cusia*, previous research has mainly focused on its chemistry and pharmacology, whereas research of the active ingredients remains limited. The biosynthetic pathway of the primary medicinal ingredient of *S. cusia*, IAs, remains largely unknown. The previous transcriptomic analyses revealed that cytochrome P450, uridine diphosphate-glycosyltransferase, β-glucosidase, and tryptophan synthase participate in the biosynthesis of indigo and the indole glycoside backbone (Lin et al., 2018, 2019; Xu et al., 2020). Additionally, the enzyme analysis and overexpression experiments confirmed that 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) contributes to the regulation of indigo synthesis (Yu et al., 2019). Additional key enzymes and transcription factors (TFs) involved in synthesizing IAs have not been thoroughly investigated in the *S. cusia* genome.

Owing to its long history of herbal use, wide distribution, and efficacy, *S. cusia* is in high demand in developing countries (Pal and Shukla, 2003; Yu et al., 2021). The Herbs Genome Program is a whole-genome sequencing and post-genomics study of medicinal plants with economically important and characteristic secondary metabolic pathways. This program aimed to characterize the genetic information and regulatory networks of plants used as traditional Chinese medicines through genome sequencing and determine the biosynthetic pathways of the active ingredients in these medicines and the mechanisms underlying the prevention of human diseases (Chen et al., 2015; Hu et al., 2019). To date, genomic data for *Panax ginseng* (Xu et al., 2017), *Scutellaria baicalensis* (Zhao et al., 2019), *Andrographis paniculata* (Sun et al., 2019), and many herbs have been obtained through high-throughput sequencing, and abundant gene information related to the biological evolution, growth and development, stress resistance, and secondary metabolism have been obtained by integrating transcriptome, metabolomic, and proteomics approaches. Thus, the construction of a high-quality *S. cusia* genome will serve to identify, or supplement, the candidate genes associated with the biosynthesis pathway of bioactive compounds and reveal the molecular mechanisms underlying its documented medicinal value.

Although a reference genome sequence was previously published for *S. cusia* (MinION), the assembled long reads were generated using the Oxford Nanopore Technologies platform (Xu et al., 2020). This platform has certain advantages, such as, the generation of longer reads through high-throughput sequencing; however, it also has a moderately high associated error rate, leading to lower assembly accuracy. Thus, in the present study, we assembled high-fidelity reads using the latest PacBio circular consensus sequencing (CCS) technology, which combines both long read length and high accuracy. The CCS sequencing results easily span the shorter repetitive complex regions of the genome while ensuring the precision and completeness of genome assembly (Wenger et al., 2019). In this study, we present the high-quality chromosome-level genome assembly and systematic analysis of species-specific genes, the basic helix-loop-helix (bHLH) family, and the IA pathway in the *S. cusia* genome. Our results provide an important resource for future studies on the mechanisms of active medicinal ingredient synthesis in *S. cusia*.

MATERIALS AND METHODS

Sample Collection and Genome Sequencing

We collected fresh leaves of *S. cusia* for genomic sequencing from our experimental field at the Fujian Agriculture and Forestry University (26°08′ E, 119°23′ N). The leaves were stored at −80°C until DNA extraction using the cetyltrimethylammonium bromide method.

An Illumina paired-end genomic library (average insert size of 350 bp) was constructed according to the standard protocols of Illumina and sequenced on an Illumina HiSeq X Ten platform (San Diego, CA, USA). Library construction (15-kb DNA SMRTbell library) and long-read sequencing were performed using the PacBio CCS technology platform (Menlo Park, CA, USA). We extracted tender leaves to validate the cell fixation and observed the integrity of the nuclei through DAPI staining. In addition, a Hi-C library was created using the *Hin*dIII enzyme according to the BioMarker Technologies Company instructions (Rohnert Park, CA, USA). The sequencing was performed on the Illumina HiSeq X Ten platform. All the sequencing services were provided by Biomarker Technologies Co., Ltd. (Beijing, China).

Genome Survey and Assembly

GenomeScope in conjunction with Jellyfish (version 2.2.3) (Vurture et al., 2017) was used for genome size estimation of *S. cusia*. In total, 49.98 Gb of high-quality paired-end reads were generated by Illumina genomic sequencing (∼52.73 × coverage, **Supplementary Table 1**). To evaluate the genome size, repeat content, and heterozygosity of *S. cusia*, the k-mer distribution with 21 nt was constructed using clean Illumina short-read data. For PacBio sequencing data, low-quality, joints, and short read filtering of the raw data yielded 483.97 Gb of sub-reads (∼532×, **Supplementary Table 1**), which were assembled using hifiasm (Cheng et al., 2021). Burrows-Wheeler Aligner (BWA; version 0.7.10-r789, a software package) (Li, 2013) and SAMtools (version 1.9) (Li et al., 2009) were used for assembly statistics. Finally, we used the BUSCO (version 3.0) (Simão et al., 2015) database embryophyta_odb10 models to assess the genome integrity.

Chromosome Assembly Using Hi-C

The Hi-C data (∼106.27 Gb) generated were mapped back to the draft assembly using BWA software, and a PERL script developed by LACHESIS software was used to obtain clean data for the Hi-C (Burton et al., 2013). Only mapping data were retained for linkage to pseudo-chromosomes using the ALLHiC pipeline (Zhang et al., 2019). The HiC-Pro (version 2.10.0) program was used to determine the Hi-C mapping rate and for quality assessment (Servant et al., 2015).

Protein-Coding Gene Prediction

Protein-coding genes were predicted using the three approaches: *ab initio* gene prediction, transcript evidence, and homologous-based analyses. The MAKER pipeline (Cantarel et al., 2007), which integrates all three approaches, was used. For homology-based analysis, we downloaded proteomes homologous to *S. cusia*, such as, *Olea europaea* var. *sylvestris*, *A. paniculata*, *Mimulus guttatus*, *Antirrhinum majus*, *Salvia splendens*, *Sesamum indicum*, and *Handroanthus impetiginosus*. The protein sequences from each species were then paired with the *S. cusia* genome using TBLASTN software, and the gene structure was predicted using GeneWise software.

For transcript evidence, we used Scallop (version 0.10.4) (Shao and Kingsford, 2017) software and assembled the RNA-seq samples (stems, roots, and leaves). The transcripts obtained were used for training programs of SNAP (version 2006-07-28) (Bromberg and Rost, 2007), GENEMARK (version 4.48_3.60_lic) (Besemer et al., 2001), and AUGUSTUS (version 3.3.3) (Stanke et al., 2004). The MAKER pipeline was used to integrate these layers of coding evidence to generate predictions of high-quality protein-coding genes.

Functional Annotation

Functional annotation of the protein-coding genes in *S. cusia* using BLASTP (*E*-value ≤ 1e-5) were aligned with the public databases NR, Swiss-Prot (Bairoch and Apweiler, 2000), Pfam (Finn et al., 2016), Gene Ontology (GO) (Ashburner et al., 2000), Clusters of Orthologous Groups (COG) (Tatusov et al., 2000), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2013). A Pfam database search, using PfamScan, identified the protein domain. The online platform OmicShare (https://www.omicshare.com/) was used for the GO enrichment and KEGG pathway analyses. The TFs were predicted using iTAK software, and tRNAscan-SE (version 1.3.1) with eukaryotic parameters (Chan and Lowe, 2019) was used to identify transfer RNA (tRNA) genes; ribosomal RNA (rRNA) fragments were predicted using RNAmmer (version 1.2) (Lagesen et al., 2007). For small nuclear RNA (snRNA) and microRNA (miRNA) gene analysis, INFERNAL (version 1.1.3) with default parameters was used for alignment with the Rfam database (Jones et al., 2014).

Analysis of Repetitive Elements and Synteny

For *de novo* prediction, we used the RepeatModeler pipeline (Flynn et al., 2020) to customize the repeated sequence library of the genome, which utilizes RECON and RepeatScout to obtain the consensus repeat library. RepeatMasker was further used in a homology-based method to identify and cluster repetitive elements (Tarailo-Graovac and Chen, 2009). A TEclass software was employed to classify unknown transposable element (TE) types (Abrusán et al., 2009). To identify tandem repeated sequences in the *S. cusia* genome, the Finder package was used to determine higher-order repeated sequences (Benson, 1999). The distribution of tandem repeats on the chromosomes was used to predict centromeres and telomeres in the same manner as in the *Oropetium thomaeum* genome (VanBuren et al., 2015). We downloaded the previously published fasta and hic.gff files of the *S. cusia* genome from http://indigoid-plant.iflora.cn. The syntenic gene pairs were identified and plotted by MCScan (Python version) [https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)].

Phylogenetic and Whole-Genome Duplication (WGD) Analyses

To investigate the evolution of *S. cusia*, we compared its genome with 12 other sequenced plant species. These included one other plant in Acanthaceae (*A. paniculata*), six others in the order Lamiales (*M. guttatus*, *S. splendens*, *H. impetiginosus*, *S. indicum*, *A. majus*, and *O. europaea*), four others in the clade Eudicots (*Solanum tuberosum*, *Solanum lycopersicum*, *Populus trichocarpa*, and *Vitis vinifera*), and *Oryza sativa* as an outgroup. Single-copy orthologous genes were identified by OrthoFinder (v2.3.3) (Emms and Kelly, 2015). MAFFT (v6.864b) with default parameters was used to align each single-copy gene (Katoh and Standley, 2013). The conserved sites were extracted by filtering the alignment with in-house Python scripts. We then strung them together into a unique super sequence. IQ-TREE (v1.7-beta12) (Nguyen et al., 2015) with default parameters was used to predict the best substitution structure models, and maximum likelihood gene trees were constructed by RAxML (Stamatakis, 2014) with 500 bootstraps, with *O. sativa* chosen as an outgroup. For the selected species, we estimated the divergence time. First, we set two fossil constraint divergence times from the TimeTree database (Kumar et al., 2017). *V. vinifera* and *O. sativa* were estimated to have diverged ∼160 million years ago

(Mya), whereas *S. indicum* and *P. trichocarpa* diverged ∼117 Mya. Using these two nodes as fossil times, the divergence time of all plants was evaluated using r8s software (v1.83) (Sanderson, 2003). CAFE (v3.1) (De Bie et al., 2006) was employed to infer the expansion and contraction of the gene family with a *P*-value < 0.01 based on phylogeny.

We performed a BLASTP all-to-all search to identify the homologous genes with an *E*-value $\leq$ 1E-8 by examining WGD in *S. cusia* and *A. paniculata*. Collinearity blocks with the MCScanX (https://github.com/tanghaibao/jcvi/wiki/) were identified. Next, the synonymous substitution rates (*Ks*) of collinear orthologous gene pairs were calculated with Python script synonymous_calc.py (https://github.com/tanghaibao/bio-pipeline/) using the Nei-Gojobori method (Nei and Gojobori, 1986).

Transcriptome Analysis

As a vital cellular regulator, methyl jasmonate (MeJA) plays a role in inducing signal transduction during secondary metabolism and can change the metabolites in plants by promoting or inhibiting gene expression. Previously, we reported that the content of indigo in MeJA-treated leaves and roots of *S. cusia* was higher than that in the control group (Lin et al., 2019). The transcriptome data of *S. cusia* accessions were imported from our previous study and filtered using Trimmomatic (Bolger et al., 2014). The treated groups (BL: treated leaf, BS: treated stem, and BR: treated root) were sprayed with 0.01% (v/v) Tween 20 solution containing 22.29 μM MeJA (Sigma-Aldrich, St. Louis, MO, USA), whereas the control groups (AL: control leaf, AS: control stem, and AR: control root) were treated with 0.01% (v/v) Tween 20 solution without MeJA to the point of runoff. Three independent biological replicates were prepared for RNA sequencing. The predicted coding sequences (CDS) of the *S. cusia* genome were then mapped, and the fragments per kilobase of exon per million mapped reads of each gene were calculated using the RSEM (version 1.3.2) program (https://github.com/deweylab/RSEM). The DESeq2 package was used to identify differentially expressed genes (DEGs). The *p*-values were adjusted using Benjamin-Hochberg multiple testing corrections, and the genes were considered as differentially expressed if the false discovery rate was <0.05 and $|\log2 \text{ fold-change}| > 1$.

Identification and Analysis of Lineage-Specific Genes

The lineage-specific genes (LSGs) have no significant similarity to any sequence in other species (Fischer and Eisenberg, 1999). The plant traits are mainly determined by genes, and adaptation of species to the environment is the main driving force for LSG evolution. We used a comparative genomics approach to identify LSGs in the *S. cusia* genome (Chen et al., 2020). First, we used BLASTP (*E*-value < 1e-3) to compare the proteomic data collected from seven closely related plant species: *O. europaea* var. *sylvestris*, *A. paniculata*, *M. guttatus*, *A. majus*, *S. splendens*, *S. indicum*, and *H. impetiginosus*. Homologous sequences were filtered out, and homology searches were conducted using the Plant-PUTs (http://www.plantgdb.org/prj/ESTCluster/progress.php), Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html), UniProt-KB (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/), and NR databases. Finally, the genes with no homology to any databases were classified as LSGs, whereas, the remaining genes with similarity were classified as evolutionarily conserved genes (ECs). Gene duplication is the primary mechanism of origin. ECs were searched using BLASTP (*E*-value < 1e-3) to infer the paralogs of the LSGs and define hit sequences as gene duplications. The isoelectric points of the LSGs and ECs were calculated using DAMBE software (Xia, 2018), and subcellular localization was evaluated using BUSCO software (Savojardo et al., 2018). The guanine plus cytosine (GC) content, protein length, and exon number of LSGs and ECs were calculated using Python scripts. One-way ANOVA was used to establish any significant differences between the LSGs and ECs.

Gene Family Analysis of bHLH

The bHLH TFs are important members of the plant gene regulatory network. In addition to participating in the growth and development and responding to adversity stress, bHLH TFs can also play a role in the process of biosynthesis (Jia et al., 2021; Singh et al., 2021). To identify and analyze *S. cusia* at the genome-wide level, we downloaded the protein sequence of bHLHs from *Arabidopsis thaliana* (https://www.arabidopsis.org/browse/genefamily/bHLH.jsp). BLASTP (*E*-value < 1e-5) was performed to obtain the potential bHLHs, whereas SMART (http://smart.embl-heidelberg.de/) was used to predict the conserved structural domains and obtain high-quality bHLHs. The essential physicochemical characteristics of bHLHs, such as, their amino acid number, molecular weight, and isoelectric point, were predicted using DAMBE. MEME (https://meme-suite.org/meme/index.html) was utilized for the enrichment analysis of the conserved structural domains. Finally, a phylogenetic tree with 1,000 bootstrap replicates was constructed using the MEGA 7 program with the neighbor-joining method (Kumar et al., 2016).

RESULTS

Genome Assembly

We estimated the genome size of *S. cusia* by 21 k-mer counting's. According to the k-mer distribution (**Supplementary Figure 1**), the length of the genome was ∼947.79 Mb with heterozygosity of 0.45%, repeat content of 78.84%, and GC content of 38.96%. We generated 483.97 Gb (532×) of PacBio long CCS reads and ∼49.98 Gb (53×) of Illumina clean reads (**Supplementary Table 1**) and used hifiasm to assemble the genome with 913.74 Mb that includes 1,357 contigs (N50 size of 35.59 Mb) with the longest contig of 66.79 Mb (**Table 1**). The initial assembly result was ∼8-fold more contiguous than the other released version (MinION), where the N50 size of contigs was 4.33 Mb. We detected genic completeness using BUSCO (97.8%), of which 92.3% were single-copy genes and 5.5% were duplicated (**Supplementary Table 2**), which was slightly higher than that in MinION (88.12%). Additionally, we used Illumina read mapping for assembly with BWA-MEN and showed that ∼49.91 Gb of reads was available for mapping on the genome, with a 99.85% mapping rate (**Supplementary Table 3**).

TABLE 1 | Global statistics for assembly and annotation of *Strobilanthes cusia* genome.

Items	MinION	PacBio CCS
Sequencing assembly		
Estimated genome size (Mb)	826.37	947.79
Contig size (Mb)	865.49	913.74
Number of contigs	1602	1357
Contig N50 (Mb)	4.33	35.59
Longest contig (Mb)	20.84	66.79
BUSCO completeness of assembly (%)	88.12	97.8
Hi-C scaffolding assembly		
Scaffold size (Mb)	865.52	912.62
Number of scaffolds	1354	108
Scaffold N50 (Mb)	50.44	68.44
Average length (Mb)	0.64	8.45
Gene annotation		
Protein-coding genes number	32148	32974
Average gene length (bp)	3450.12	3399.37
Average exon length (bp)	244.48	310.24
Average exon per gene	5.23	6.01
Average intron length (bp)	302.2	306.51
BUSCO completeness of annotation (%)	91.39	98.5

Strobilanthes cusia is a diploid organism with 16 chromosome pairs. We further refined the chromosomal-level assembly based on the generated 106.27 Gb Hi-C clean data. The genome was linked using the ALLHiC pipeline. The final reference assembly included chromosome-scale pseudomolecules, with 16 pseudomolecules >20 Mb in length and covering 98.95% of the 913.74 Mb initial genome (**Figure 1** and **Supplementary Table 4**). The number of scaffolds was lower, with an average length of ∼8-fold >MinION (0.64 Mb) (**Table 1**). We used HiC-Pro to examine the Hi-C data assembly quality, with ∼98.37 Gb on the data mapping, accounting for 92.57%. Unique mapped paired-end reads had ∼7.9 Gb and a lib valid rate of ∼37.93% (**Supplementary Table 5**). Taken together, the genome assembly of *S. cusia* showed high integrity and precision. In addition, our contig assembly shared high collinearity with the corresponding genomes from the previous reports (**Supplementary Figure 2**). Together, these data suggest that the genome assembly of *S. cusia* in this study is more complete compared with that published and of higher quality for subsequent analysis.

Gene Prediction and Annotation

We annotated genes encoding proteins using the MAKER pipeline based on the transcriptome, homology, and *ab initio* prediction. We predicted a total of 32,974 genes (**Table 1**). The average gene length and intron sequence size were 3,399 and 306.51 bp, respectively; we identified 6.01 exons with an average exon length of 310.24 bp per gene. Compared with the MinION version, these gene models had much the longer average gene, exon, and intron lengths (**Table 1**). We evaluated the protein-coding gene completeness using BUSCO (98.5%), revealing 93% single-copy genes and 5.5% duplicated genes, which were slightly higher than those identified by MinION (91.39%) (**Table 1** and **Supplementary Table 6**).

We functionally annotated the predicted protein sequences using NR, Swiss-Prot, and Pfam (**Table 2**); 96.33%, 79.63%, and 82.25% of the genes were homologous, respectively. We used COG, GO, and KEGG for annotation, with 88.74% of genes having COG, 43.9% with GO term classification, and 43% mapping to known biological pathways. These results illustrate the high reliability of the predicted gene model.

Transcription factors regulate gene expression. Herein, we predicted TFs in the *S. cusia* genome using iTAK and showed that of the 32,974 genes encoding 2,666 TFs (**Supplementary Table 7**) were divided into 93 families, and MYB, AP2/ERF-ERF, bHLH, FAR1, and C2H2 were the major TF types.

We used miRDP software to identify miRNA, tRNA, and rRNA genes in the *S. cusia* genome from the miRBase and predicted rRNAs and tRNAs using RNAmmer and tRNAscan-SE, respectively, which revealed 122 miRNAs, 4,034 tRNAs, and 3,403 rRNAs (**Supplementary Table 8**).

Repeat Annotation and Centromere Identification

The *S. cusia* genome has a large number of repeated sequences contained in 675.66 Mb and accounting for 74.04% of the genome (**Supplementary Table 9**). There are 429.62-Mb-long terminal repeat (LTR) retrotransposons, accounting for 47.08%. Non-LTR retrotransposons, comprising of LINE and SINE, represented 8.38% and 0.5%, respectively, whereas 14.23% comprised another type of DNA transposon. We identified 28,229 tandem repeats, constituting 4.22% of the *S. cusia* genome. A similar approach was applied to the *O. thomaeum* genome to identify centromeric repeats; 245 putative centromeric fragments with an average length of 77.93 bp were detected on the *S. cusia* chromosomes (**Supplementary Table 10**). The base centromeric consensus size was 172 bp. Among them, there were 56 on Chr1, forming the largest distribution. We also found 29 putative telomeric fragments. All chromosomes were distributed except for the deletion of Chr14 and Chr15 (**Supplementary Table 11**). The syntenic relationship is presented in a CIRCOS plot showing the TE distribution, gene density, and GC content (**Figure 1**).

Identification and Characterization of Specific Gene Families

Species-specific genes influence the adaptation of species to specific habitats; therefore, a comparative genomic analysis was conducted. Across the six Lamiales species genomes, 99,683 gene family clusters were observed among 231,423 genes (**Supplementary Table 12**). For the *S. cusia* genome, 32,974 genes were clustered into 16,955 gene families, 6,503 of which were single-copy genes and 696 were unique gene families compared with the other five closely related plants (**Figure 2**). GO enrichment analysis indicated that these specific gene families were mainly related to DNA integration, defense response, plant-type hypersensitive response, defense response to fungus, and regulation of salicylic acid biosynthetic process

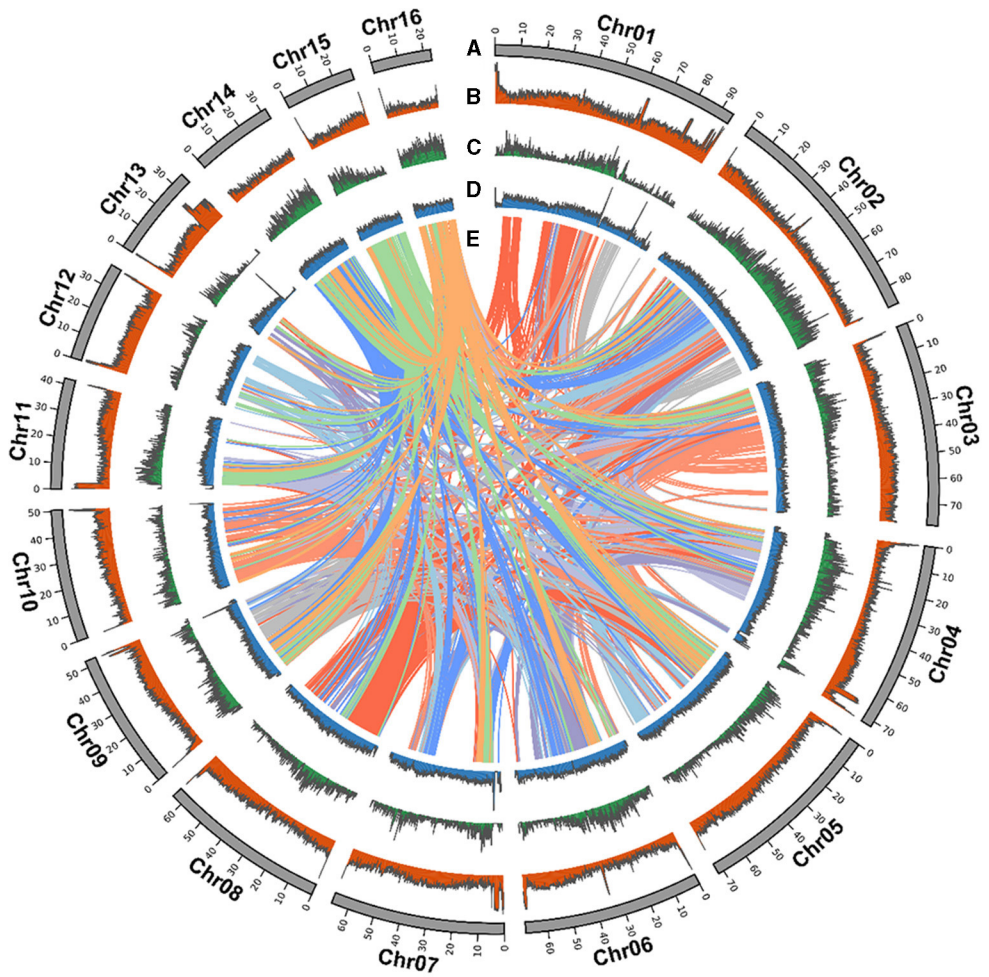

FIGURE 1 | *Strobilanthes cusia* chromosomal elements in global view. **(A)** Chromosome karyotype. **(B)** GC content. **(C)** Gene density. **(D)** DNA transposable elements (TEs). **(E)** Schematic presentation of the major inter-chromosomal relationships in the *S. cusia* genome.

in the "biological process" term and cytochrome-c oxidase activity, and shikimate *O*-hydroxycinnamoyltransferase activity in the "molecular function" term (**Supplementary Table 13**). Notably, the unique gene families were exceedingly enriched in numerous plant defense functions and secondary metabolism. These enriched genes may contribute to the wide distribution and strong environmental adaptability of *S. cusia*.

Evolution of the *S. cusia* Genome

We selected 65 single-copy genes among the 13 species to construct a phylogenetic tree, which showed that *S. cusia* was a sister to *A. paniculata*. Molecular dating using r8s with fossil calibration indicated that *S. cusia* and *A. paniculata* evolved ∼45 Mya (**Figure 3A**). We further estimated the divergence time between these species using the density distribution of *Ks* (**Figure 3B**). The *Ks* values of the orthologues among species pairs revealed a peak of 0.65–0.85 for *S. cusia* and *A. paniculata*, with a corresponding divergence time of 50–65 Mya. Next, we used *Ks* between collinear paralogous genes to identify potential WGD events based on the assumption that the number of silent substitutions per site between two homologous sequences increases in a relatively linear manner over time. A density plot of *Ks* values for the collinear gene pairs suggested that *S.*

TABLE 2 | Annotation statistics for the *S. cusia* genome.

Annotation statistics for the genome	Number	Percent (%)
Total proteins	32974	100
NR	31765	96.33
Swiss-Prot	26256	79.63
Pfam	27121	82.25
COG	29260	88.74
GO	14475	43.9
KEGG	14178	43
In all databases	13232	40.13
In at least one database	31828	96.52

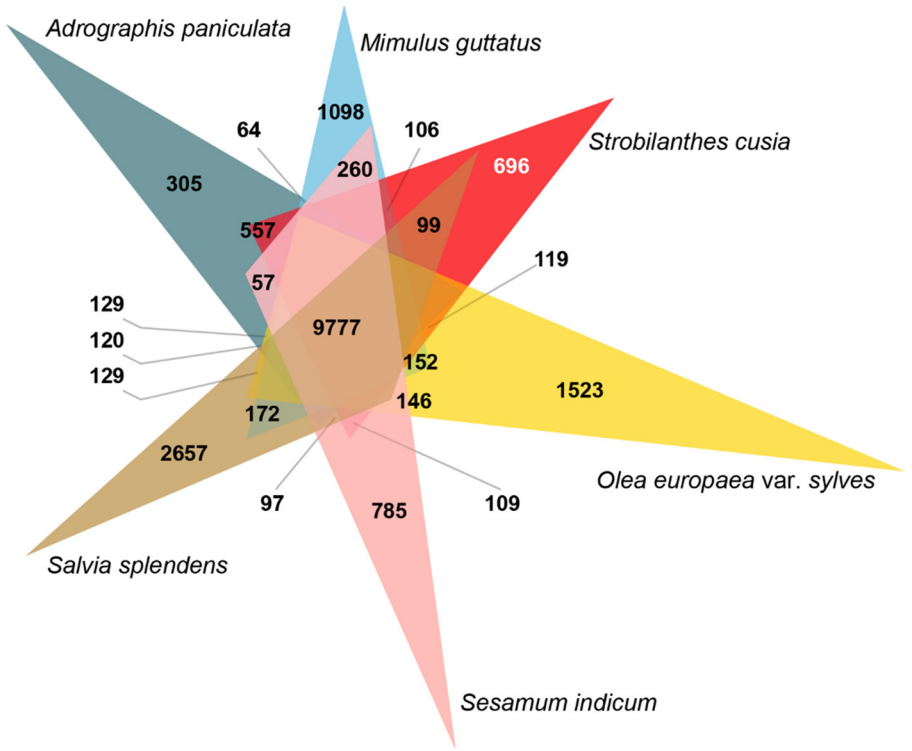

FIGURE 2 | Analysis of Venn diagram of orthologous gene families. Six species (*S. cusia*, *Mimulus guttatus*, *Andrographis paniculata*, *Salvia splendens*, *Sesamum indicum*, and *Olea europaea* var. *sylvestris*) were exploited to generate the Venn diagram based on the gene family cluster analysis. The white number 696 in parentheses represents the gene family specific to *S. cusia* in all six plants considered.

cusia experienced an ancient polyploidization event with a peak value of ∼1.157, and no specific WGD occurred in the *S. cusia* genome after divergence from Eudicots. The WGD data provide a foundation for genome evolution studies of *S. cusia*.

In *S. cusia*, 13,796 gene families were identified, among which 60 and 16 gene families showed rapid expansion and contraction (*P*-value), respectively (**Figure 3A**). Compared with closely related species of *A. paniculata* (24 expansion/20 contraction), *S. cusia* showed distinctly higher gene family expansion than contraction. In total, 60 expanded gene families were annotated to GO terms (**Supplementary Table 14**) and KEGG pathways (**Supplementary Table 15**). The GO analysis showed that the expanded orthogroups were related to spermidine hydroxycinnamate conjugate biosynthetic process, Penta cyclic triterpenoid metabolic process, Penta cyclic triterpenoid biosynthetic process, ethylene-activated signaling pathway, regulation of transcription, DNA-templated, and regulation of gene expression. The KEGG analysis showed that the most expanded genes were clustered in the MAPK signaling pathway, plant-pathogen interaction, diterpenoid biosynthesis, carotenoid biosynthesis, sesquiterpenoid and triterpenoid biosynthesis, and steroid biosynthesis. These expanded gene families were mainly concentrated in the pathways related to environmental adaptation and secondary metabolism, which are important in the biosynthesis of active ingredients and interaction between *S. cusia* and its growth environment.

Transcriptome Analysis

Expression analysis was conducted using transcriptome data and *S. cusia*-annotated gene information. In the control group, 9,264 DEGs (upregulated: 4,704 and downregulated: 4,560) and 6,734 DEGs (upregulated: 2,995 and downregulated: 3,739) were in the leaves and stems, respectively, relative to the roots (**Figure 4A**). Following MeJA treatment, 2,975 DEGs were significantly differentially expressed in 32,974 annotated genes, of which 478 DEGs were upregulated and 164 DEGs were downregulated in MeJA-treated leaves, 545 DEGs were upregulated and 1,182 DEGs were downregulated in MeJA-treated roots, and 361 DEGs were upregulated and 245 DEGs were downregulated in MeJA-treated stems (**Figure 4A**). DEGs in the roots were quite different from those in the stems and leaves, which may be related to tissue specificity. These results indicated that the roots were the most sensitive to MeJA treatment, and the stems were the least sensitive. Through comparison of the data with the KEGG database, the metabolic pathway of *S. cusia* affected by MeJA induction was explored. KEGG enrichment analysis showed that 2,975 DEGs were clustered in the biosynthesis pathways of secondary metabolites, phenylpropanoid biosynthesis, flavonoid biosynthesis, alanine, aspartate and glutamate metabolism, pentose and glucoronate interconversions, sesquiterpenoid, triterpenoid biosynthesis, galactose metabolism, and photosynthesis-antenna proteins (**Figure 4B**). Among them, we identified 27 DEGs in the

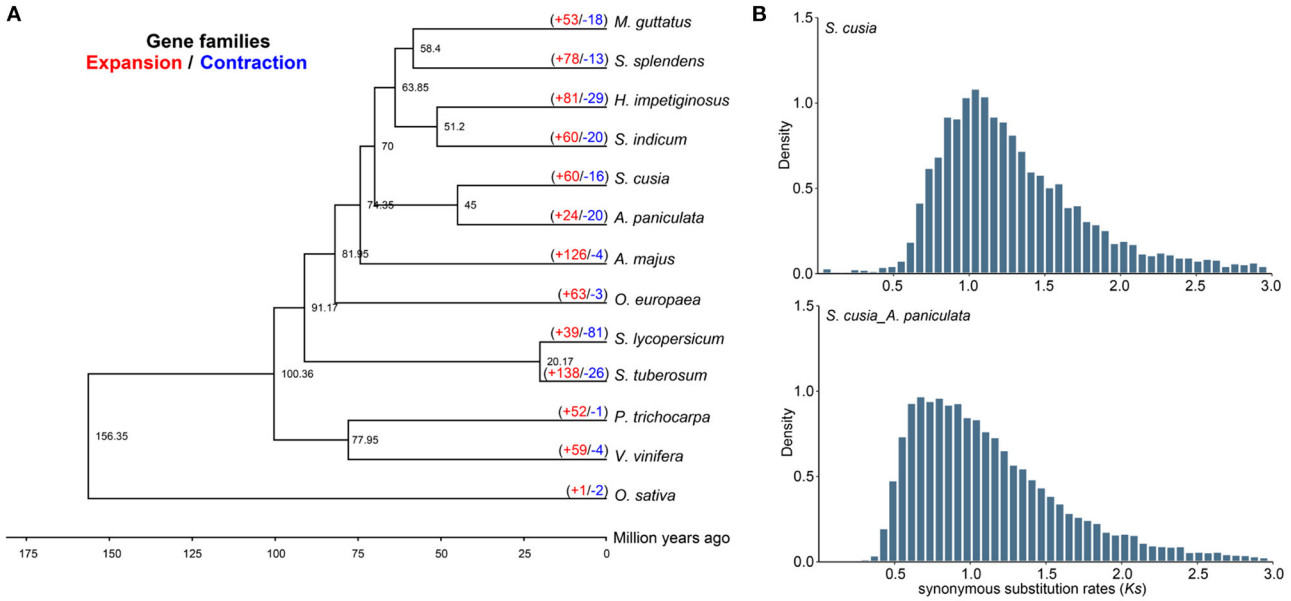

FIGURE 3 | Comparative genomic and phylogenetic relationship analyses. **(A)** Phylogenetic relationship between *S. cusia* and other representative plant species. The divergence times among different plant species are labeled in the bottom. The red number on each branch represents expansion gene families, and blue number indicates the number of contracted gene families. **(B)** Density distribution of *Ks* values between syntenic genes of compared genomes.

roots, stems, and leaves compared with the control group (**Figures 4C,D**). These annotation pathways and significant DEGs provide valuable information for studying the molecular regulatory mechanism of MeJA treatment for effectively enhancing the accumulation of IAs in *S. cusia*.

Identification, Characterization, and Expression Analysis of LSGs

The production of LSGs promotes the evolution and morphological diversity of species (Ma et al., 2020). Some research suggested that the biological functions of LSGs are related to the unique biological characteristics and environmental adaptability of the species. Based on the previous methods for LSG identification (Chen et al., 2020), 983 *Strobilanthes*-specific genes (SSGs) were identified, comprising 2.98% of the entire genome (**Supplementary Table 16**). In total, 31,992 ECs were identified. Gene duplication is the most important mechanism of LSG origin. In this study, we identified 195 SSGs originating from gene duplication, accounting for 19.86% of the SSGs (**Supplementary Table 16**). We compared the structural features of SSGs and ECs; the average EC length (442 aa) was significantly > that of the SSGs (112 aa) by ∼3.65-fold, which was attributed to the presence of fewer exons in SSGs (**Figure 5A**). We also compared the isoelectric point of SSGs (8.64), which was significantly higher than that of ECs (7.54) (**Figure 5A**), and analyzed the SSG distribution on chromosomes based on the annotation information (**Supplementary Table 16**). In total, 980 SSGs were distributed on 16 chromosomes. Chr2 and Chr4 showed the largest number of SSGs. Subcellular localization is an important factor in protein function research and understanding the subcellular localization of proteins can help to determine their biological functions. The 840 SSGs were primarily localized in the nucleus, chloroplast, and extracellular space (400, 239, and 201, respectively), accounting for ∼85.45% of all SSGs (**Figure 5B**). The analyses of the expression patterns of SSGs were conducted using RNA-seq data. Most SSGs showed tissue expression specificity (**Figure 5C**). These LSGs are valuable genetic resources for studying *Strobilanthes*-specific traits; how they participate in the complex biological network and their roles in a very short evolutionary time require further analysis.

Identification of bHLHx Family

Among the 93 *S. cusia* TFs identified, bHLH is one of the largest families. In total, 173 bHLH family members were identified in the *S. cusia* genome (**Supplementary Table 17**). ML16673 encodes the highest molecular weight bHLH protein in the bHLH family at 717 amino acids (136,757.33 kDa). In comparison, ML19273, ML14908, and ML19092 encode the bHLH protein with the smallest molecular weight consisting of 93 amino acids (34,371.39 kDa). The isoelectric points of the proteins within the bHLH family recorded in *S. cusia* ranged from 4.5 for ML00529 to 10.78 for ML31426. A phylogenetic tree of the 173 bHLHs in *S. cusia* was constructed using MEGA7 software, and they were classified based on the taxonomy of the subfamily *Arabidopsis*. The results showed that the 173 bHLHs could be classified into 20 subfamilies, with the most distributed gene being subfamily XIII (**Figure 6A**). The gene family analysis showed that the members of bHLH had undergone significant expansion, whereas analysis of the bHLH amino acid conserved motifs using MEME showed that the N-terminal basic region contains the highly conserved H-E-R sequence (His-Glu-Arg),

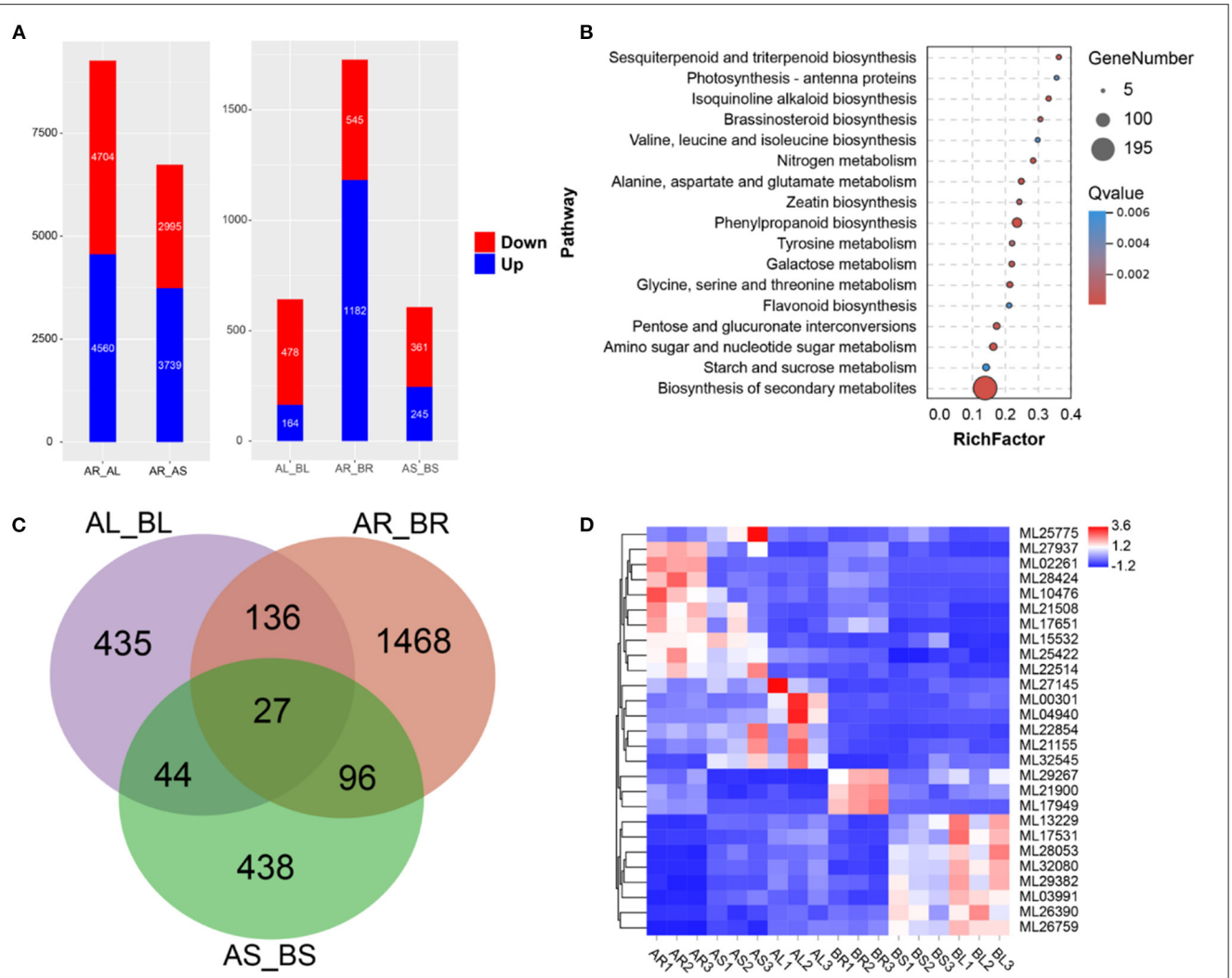

FIGURE 4 | Transcriptome analysis. **(A)** Differentially expressed genes (DEGs) between the leaves, stems, and roots. AR_AL: DEGs in leaves relative to roots; AR_AS: DEGs in stems relative to roots; AL_BL: DEGs in leaves of treated sample relative to control sample. AR_BR: DEGs in roots of treated sample relative to control sample. AS_BS: DEGs in stems of treated sample relative to control sample; **(B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of DEGs in treated samples relative to control samples. **(C)** Venn diagram of DEGs; **(D)** Expression of 27 DEGs in the leaves, stems, and roots.

which is essential for E-box recognition and binding upstream of the target gene promoter. In the C-terminal HLH region, positions 34 and 50 Leu are highly conserved as positions 37 and 38 Leu/Val; these amino acid residues are important for dimer formation and function (**Figure 6B**). We also analyzed the expression of the 173 bHLHs, 17 of which were significantly downregulated in the stems and leaves compared with those in the roots, and five were significantly upregulated in the stems and significantly downregulated in the leaves compared with those in the roots (**Figure 6C**). In the transcriptome data following MeJA treatment, three *bHLH* genes were significantly upregulated and downregulated in the leaves, respectively, compared with those in the control group. In the stem, four genes were significantly downregulated, and one was significantly upregulated, whereas 10 genes were significantly upregulated in the root, with ML17531 significantly upregulated in the roots, stems, and leaves compared with those in the control group (**Figure 6D**).

Genes Involved in the IA Biosynthesis Pathway

Strobilanthes cusia is a natural antibacterial and antiviral raw material or Chinese herbal medicine, and the monomer components or mixtures in its roots, stems, and leaves exhibit significant therapeutic effects. These active substances are mainly derived from the IA biosynthesis pathway (**Figure 7A**). We further evaluated 10 key enzymes associated with the IA synthesis pathway for homology searching and confirmed 18 IA-related coding genes, such as, seven copies of UDP-glucuronosyltransferase (*UGT*), two of indole-3-glycerol phosphate synthase (*IGPS*), two of cytochrome P450

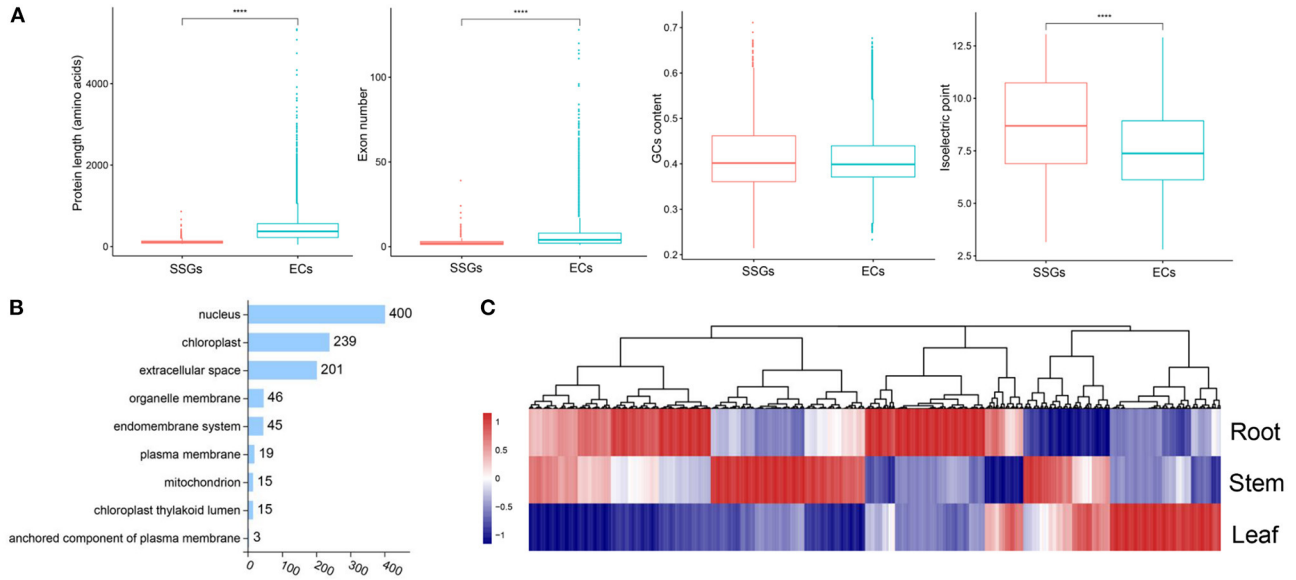

FIGURE 5 | Analysis of lineage-specific genes (LSGs) in *S. cusia*. **(A)** Box-plot comparisons of protein length, exon number, GC content, and isoelectric point for the *Strobilanthes*-specific genes (SSGs) and evolutionarily conserved genes (ECs) in the *S. cusia*. SSGs mean the *S. cusia* LSGs; ECs mean the *S. cusia* evolutionary conserved genes. **(B)** Number of SSGs assigned to different subcellular locations. **(C)** Expression pattern of SSGs in the roots, stems, and leaves at three developmental stages.

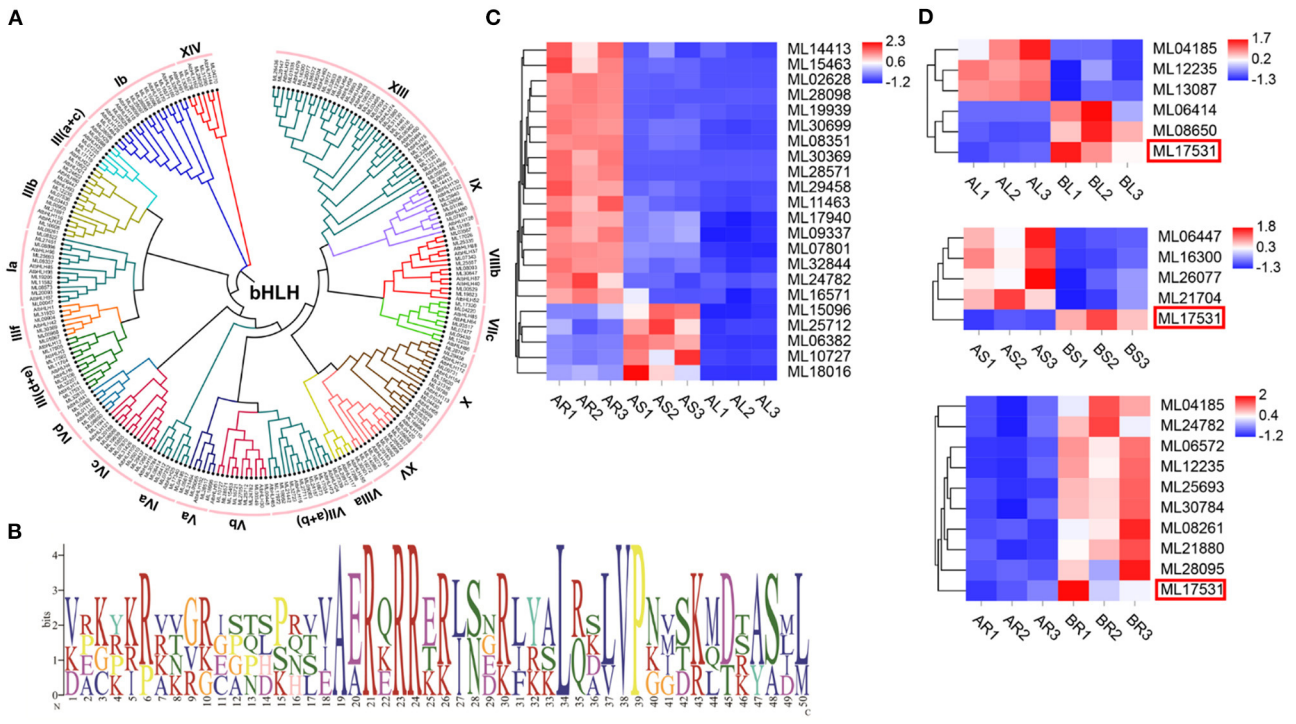

FIGURE 6 | The basic helix-loop-helix (bHLH) in *S. cusia* genome. **(A)** Phylogenetic tree of bHLH based on the protein sequence alignments from *S. cusia* and *Arabidopsis* using MEGA X with the neighbor-joining methods. **(B)** Logo showed the conserved amino acid residues within motif. **(C)** *bHLH* gene members with significantly up and downregulated expression in leaf and stem relative to root. **(D)** *bHLH* gene members with significantly upregulated and downregulated expression in leaf and stem relative to control sample.

monooxygenase (*CPY450*), *EPSPS*, chorismate synthase (*CS*), anthranilate synthase a-subunit (ASA), anthranilate synthase β-subunit *(ASB)*, β-glucosidase *(BGL)*, tryptophan synthase α-subunit *(TSA)*, and tryptophan synthase β-subunit (*TSB*) in the assembled genome of *S. cusia* (**Figure 7B**). Compared with the stems and leaves, the expression of three *UGT* (ML00663/ML00664/ML20757) and two *CYP450* (ML07148/ML07149) genes in the roots decreased significantly, whereas the expression of two *UGT* (ML20753/ML20754), one *ASA* (ML21291), one *TSB* (ML12970), one *BGL* (ML32220), one *CS* (ML21020), and one *EPSPS* (ML13587) gene increased significantly. The expression levels of four genes (ML21291/ML32220/ML12970/ML07148) were greatly increased in the roots treated with MeJA. Two genes (ML13587/ML20757) were significantly downregulated and two *UGT* genes (ML00663/ML00664) were significantly upregulated in MeJA-treated leaves. One *UGT* gene (ML29269) was significantly upregulated in MeJA-treated stems (**Figure 7C**). These DEGs may be closely related to the synthesis of IAs, such as, indigo and indirubin and can provide valuable information on the metabolic regulation of the active ingredients in *S. cusia*.

DISCUSSION

Strobilanthes cusia is a characteristic medicinal plant with a typical metabolic pathway of IAs. The market demand is very high for this Chinese medicine with great development potential. Owing to the low content of natural products in the plants, extraction steps are cumbersome, and the yield is low, making it necessary to obtain sufficient medicinal natural products using new approaches. It is particularly important to analyze the genetic background and secondary metabolic pathways of medicinal plants. In this study, we present a high-quality reference genome assembly of *S. cusia* determined using Illumina, PacBio CCS, and Hi-C sequencing data. The resulting total genome size was 913.74 Mb (the estimated size was 826.37 Mb for MinION versions based on k-mer composition) (Xu et al., 2020). A remarkable increase in the size of the contig and scaffold N50 was achieved using the new "PacBio CCS" assembly; particularly, the contig N50 was 8-fold larger than the previously reported (**Table 1**). The assembly precision based on the Hi-C contact map was improved compared with that of the previous assemblies. These data provide an important reference for improving the assembly quality of the current medicinal plant genome. The content of pharmacodynamics components in Chinese herbal medicine is typically related to various factors, such as, the variety, place of origin, medicinal parts, harvest season, growth period, and other factors. Our assembled genetic information provides a reliable molecular basis for explaining these differences. After enrichment analysis of species-specific genes and expansion gene families, we found that the gene families were concentrated in pathways involved in the plant defense response and secondary metabolism. This phenomenon is related to the secondary metabolites for adaptation to the living environment and is consistent with the substances produced by resisting the negative conditions of the natural world. Additionally, we used comparative genomics to identify LSGs in the *S. cusia* genome and subsequently analyzed their origin, sequence structure, and expression patterns. As LSGs represent important drivers for the generation of new functions and phenotypic changes in species, the LSGs identified in the current study are thought to be important for *S. cusia* breeding.

Currently, the synthesis pathway of IAs, a key medicinal ingredient in *S. cusia*, has been primarily derived based on the presumption of microbial metabolism and the studies on other indigo source plants (Jin et al., 2016; Ma et al., 2016; Kang et al., 2020). It is generally accepted that the synthesis of indigo, indirubin, and indole glycosides in *S. cusia* involves the shikimate and indole pathways. Regulation of the biosynthetic pathway of target secondary metabolites through exogenous elicitors is regarded as an important method for greatly increasing the metabolite content. High-performance liquid chromatography (HPLC) showed that the content of indigo in the leaves and roots of *S. cusia* was significantly increased after treatment with exogenous MeJA (Lin et al., 2019). MeJA is thought to activate or inhibit the activity of corresponding TFs through signal transduction, thereby regulating the expression of key enzyme genes in the secondary metabolic pathway and ultimately affecting the synthesis of secondary metabolites (Wang et al., 2019; Zhou et al., 2021). Therefore, we performed a homology search, high-quality genome functional annotation, and transcriptome analysis of *S. cusia* before and after MeJA treatment to identify IA-related coding genes to better understand the biological characteristics of *S. cusia* and *in vivo* biosynthesis pathways for indigo, indirubin, and indole glycosides. We identified one *EPSPS* and one *CS*, which encode two enzymes responsible for catalyzing the chorismic acid synthesis from precursor substances, as well as key enzyme genes in the shikimate pathway (Wang et al., 2014; Yu et al., 2019). One *ASA* gene and one *ASB* gene were involved in catalyzing the synthesis of anthranilic acid from chorismic acid, whereas ASA is considered to have a rate-limiting effect on indigo synthesis. We also identified two *IGPS* and two *TSA* genes in the *S. cusia* genome. Anthranilic acid synthesizes indole under the combined action of IGPS and TSA. ^{13}C-NMR and mass spectrometry revealed that indole serves as the precursor of indoxyl derivatives, rather than L-tryptophan in plants (Xia and Zenk, 1992). Our identification of genes related to the shikimate pathway in the *S. cusia* genome supports that indole heterocycles in the backbone structures of indigo and indirubin are derived from the shikimate pathway. Additionally, two *CYP450* and seven *UGT* genes were identified: CYP450 oxidizes indole to indoxyl, which is then glucosylated by UGT to form indoxyl-3-*O*-β-D-glucoside (indican) (Marcinek et al., 2000). One *BGL* gene was identified. When exposed to external influences or leaf senescence, indican was reversibly hydrolyzed under BGL to decompose into indoxyl. Subsequently, indigo and indirubin can be synthesized by the polymerization of indophenol under aerobic conditions. However, the functionalities of the obtained candidate genes require further validation *in vivo*. Analyzing the genome of *S. cusia* after MeJA treatment to obtaining candidate genes related to the synthesis of IAs provides valuable data resources for studying the medicinal properties of *S. cusia*.

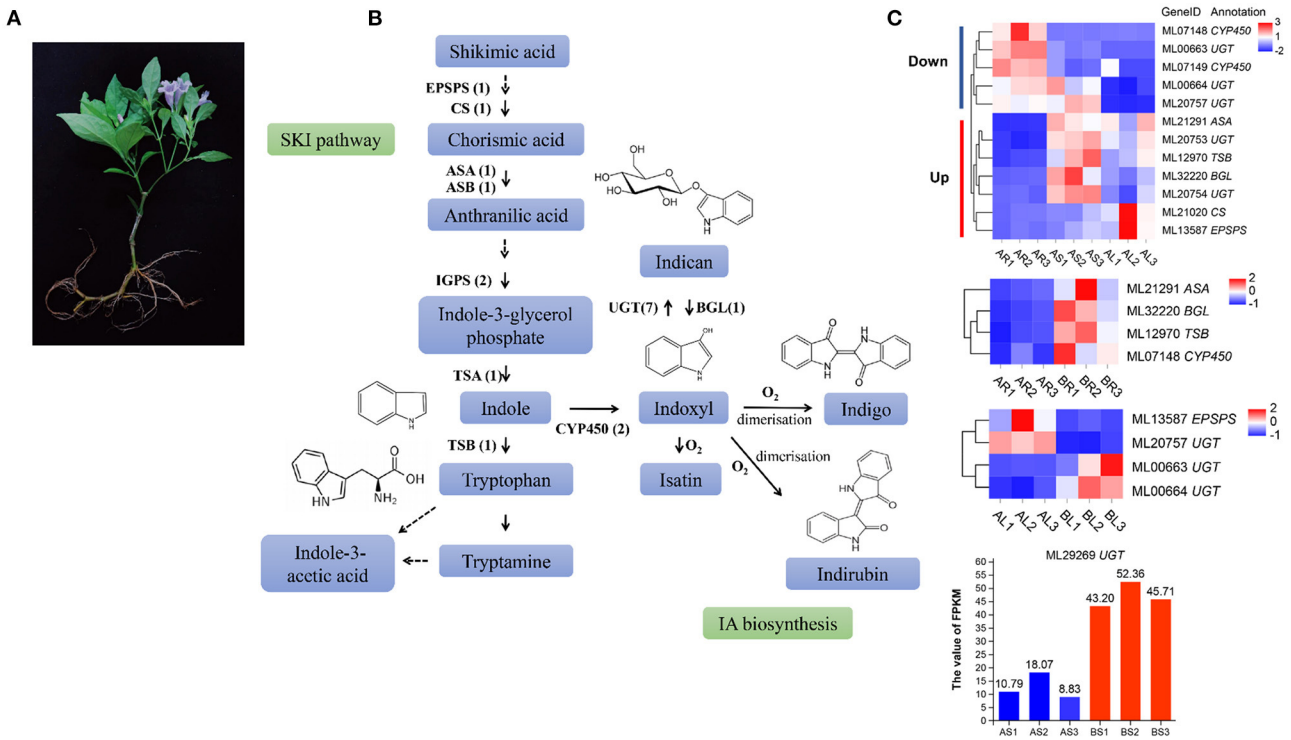

FIGURE 7 | Analysis of indole alkaloid (IA) biosynthesis in *S. cusia*. **(A)** *S. cusia* whole plant display. **(B)** Putative biosynthesis pathway of IAs in *S. cusia*. **(C)** Heat map showing the expression pattern of genes associated with IA biosynthesis in different tissues including the roots, stems, and leaves.

In addition to enzyme genes, TFs are thought to play an important role in secondary metabolism by activating or inhibiting gene transcription through binding with *cis*-acting DNA elements. In recent years, more and more bHLH TFs have been addressed to be closely related to the alkaloid metabolism in medicinal plants such as, *Catharanthus roseus*, *Taxus cuspidata*, and *Coptis japonica* (Yamada et al., 2011; Lenka et al., 2015; Patra et al., 2018). MeJA induction can significantly increase the content of indigo in *S. cusia* (Lin et al., 2019), and bHLHs, especially MYC proteins, can play a key role in the signal transduction process after JA perception (Fernández-Calvo et al., 2011). To discover the candidate bHLHs involved in JA-mediated regulation of secondary metabolism in *S. cusia*. Based on the sequence similarity, evolutionary relationships, and motif diversity, 173 bHLH proteins were identified and characterized. Phylogenetic analysis and the domain distribution of bHLH reflected the conservatism of the functional structure of the family (Li et al., 2006; Carretero-Paulet et al., 2010; Pires and Dolan, 2010). The tissue-specific expression profiles of the bHLH TF family in different tissues (roots, stems, and leaves) were analyzed using RNA-seq data. The expression patterns of *bHLH* differed significantly among the various tissues; however, most *bHLH* levels in the stems and leaves were significantly lower than those in the roots. Moreover, compared with that in the control group, ML17531 was significantly upregulated in the roots, stems, and leaves following MeJA treatment, suggesting that this transcription factor is involved in synthesizing IAs in *S. cusia*. Our research identified and supplemented candidate enzyme genes and TFs involved in the biosynthetic pathways of indigo, indirubin, and indole glycosides in *S. cusia*. These results provide a selectable target for the regulation of IA metabolic flow and valuable information for cultivating *S. cusia* varieties with high medicinal ingredient contents through metabolic engineering.

CONCLUSION

We propose a high-quality genome for the medicinal plant *S. cusia*, which we sequenced using the PacBio CCS sequencing platform, with an assembled genome size of ∼913.74 Mb. From the assembled genomes, we identified 675.66 Mb of repetitive sequences were identified, representing 74.04% of the genome. The chromosome-level genome, resulting from Hi-C data, yielded a contig N50 size of 35.59 Mb and scaffold N50 size of 68.44 Mb. We also predicted 32,974 protein-coding genes, of which 96.52% had annotations in public databases. Further, we identified 245 putative centromeric fragments, 29 putative telomeric fragments, and 983 SSGs in *S. cusia*. We identified the key enzyme genes and candidate bHLH family TFs associated with the IA pathway in *S. cusia*. In conclusion, high-quality genome sequencing of *S. cusia* provided insight into the systematic study of IA synthesis in this medicinally and economically important Chinese herbal medicine. Our study provides a theoretical basis for the breeding of *S. cusia* varieties

and the production of IAs through new methods such as, synthetic biology.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: the data that support the findings of this study have been deposited into CNGB Sequence Archive of CNGBdb with accession number CNP0001632 (https://db.cngb.org/). The genome assembly, annotation (gff), protein, coding sequences (CDS) and cDNA sequences can be found under assembly accession number CNA0019301.

AUTHOR CONTRIBUTIONS

YH, DM, and DW designed and coordinated the entire project. YH and DM led, performed the entire project together, performed the data analysis, and wrote the manuscript. DW supervised the project. GC and XM collected the samples for sequencing. XZ and XQ submitted data to the database. SN, QY, QD, and PL participated in manuscript revision. All authors read and approved the final manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (No. 81573517), Natural Science Foundation of Fujian Province (No. 2019J01827), Science and Technology Innovation Project of Fujian Agriculture and Forestry University (No. CXZX2020011A), and the Central Special Project for Fujian Local Science and Technology Development (No. 2020L3025).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.742420/full#supplementary-material

REFERENCES

Abrusán, G., Grundmann, N., DeMester, L., and Makalowski, W. (2009). TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25, 1329–1330. doi: 10.1093/bioinformatics/btp084

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 29, 2607–2618. doi: 10.1093/nar/29.12.2607

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35, 3823–3835. doi: 10.1093/nar/gkm238

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2007). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907

Carretero-Paulet, L., Galstyan, A., Roig-Villanova, I., Martínez-García, J. F., Bilbao-Castro, J. R., and Robertson, D. L. (2010). Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. *Plant Physiol.* 153, 1398–1412. doi: 10.1104/pp.110.153593

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chen, K., Tian, Z., Chen, P., He, H., Jiang, F., and Long, C. (2020). Genome-wide identification, characterization and expression analysis of lineage-specific genes within *Hanseniaspora* yeasts. *FEMS Microbiol. Lett.* 367:fnaa077. doi: 10.1093/femsle/fnaa077

Chen, S., Song, J., Sun, C., Xu, J., Zhu, Y., Verpoorte, R., et al. (2015). Herbal genomics: examining the biology of traditional medicines. *Science* 347, 27–29. doi: 10.17660/ActaHortic.2015.1089.62

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Chiang, Y., Li, A., Leu, Y., Fang, J., and Lin, Y. (2013). An *in vitro* study of the antimicrobial effects of indigo naturalis prepared from *Strobilanthes formosanus* Moore. *Molecules* 18, 14381–14396. doi: 10.3390/molecules181114381

Chinese Pharmacopoeia Committee. (2020). *The pharmacopoeia of the People's Republic of China*. Beijing: China Medical Science and Technology Press.

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Fernández-Calvo, P., Chini, A., Fernández-Barbero, G., Chico, J., Gimenez-Ibanez, S., Geerinck, J., et al. (2011). The *Arabidopsis* bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *Plant Cell* 23, 701–715. doi: 10.1105/tpc.110.080788

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: toward a more sustainable future. *Nucleic Acids Res.* 44, 279–285. doi: 10.1093/nar/gkv1344

Fischer, D., and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics* 15, 759–762. doi: 10.1093/bioinformatics/15.9.759

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Gu, W., Zhang, Y., Hao, X., Yang, F., Sun, Q., Morris-Natschke, S. L., et al. (2014). Indole alkaloid glycosides from the aerial parts of *Strobilanthes cusia*. *J. Nat. Prod.* 77, 2590–2594. doi: 10.1021/np5003274

Hsieh, W., Lin, Y., Tsai, C., Wang, T., Chen, T., and Pang, J. S. (2012). Indirubin, an acting component of indigo naturalis, inhibits EGFR activation and EGF-induced CDC25B gene expression in epidermal keratinocytes. *J. Dermatol. Sci.* 67, 140–146. doi: 10.1016/j.jdermsci.2012.05.008

Hu, H., Shen, X., Liao, B., Luo, L., Xu, J., and Chen, S. (2019). Herbgenomics: a stepping stone for research into herbal medicine. *Sci. China Life Sci.* 62, 913–920. doi: 10.1007/s11427-018-9472-y

Hu, J., Deng, Y., John, R. I. W., and Thomas, F. D. (2011). *Flora of China: Acanthaceae*. Beijing: Science Press; St. Louis, MO: Missouri Botanical Garden Press.

Huang, M., Wang, L., Zeng, S., Qiu, Q., Zou, Y., Shi, M., et al. (2017). Indirubin inhibits the migration, invasion, and activation of fibroblast-like synoviocytes from rheumatoid arthritis patients. *Inflamm. Res.* 66, 433–440. doi: 10.1007/s00011-017-1027-5

Jia, N., Wang, J., Liu, J., Jiang, J., Sun, J., Yan, P., et al. (2021). DcTT8, a bHLH transcription factor, regulates anthocyanin biosynthesis in *Dendrobium candidum*. *Plant Physiol. Bioch.* 162, 603–612. doi: 10.1016/j.plaphy.2021.03.006

Jie, C., Luo, Z., Chen, H., Wang, M., Yan, C., Mao, Z., et al. (2017). Indirubin, a bisindole alkaloid from *Isatis indigotica*, reduces H1N1 susceptibility in stressed mice by regulating MAVS signaling. *Oncotarget* 8, 105615–105629. doi: 10.18632/oncotarget.22350

Jin, Z., Kim, J., Park, S. U., and Kim, S. (2016). Cloning and characterization of indole synthase (INS) and a putative tryptophan synthase α-subunit (TSA) genes from *Polygonum tinctorium*. *Plant Cell Rep.* 35, 2449–2459. doi: 10.1007/s00299-016-2046-3

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, 199–205. doi: 10.1093/nar/gkt1076

Kang, M., Wu, H., Yang, Q., Huang, L., Hu, Q., Ma, T., et al. (2020). A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine. *Hortic. Res.* 7:18. doi: 10.1038/s41438-020-0240-5

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kawai, S., Iijima, H., Shinzaki, S., Hiyama, S., Yamaguchi, T., Araki, M., et al. (2017). Indigo naturalis ameliorates murine dextran sodium sulfate-induced colitis via aryl hydrocarbon receptor activation. *J. Gastroenterol.* 52, 904–919. doi: 10.1007/s00535-016-1292-z

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Lenka, S. K., Nims, N. E., Vongpaseuth, K., Boshar, R. A., Roberts, S. C., and Walker, E. L. (2015). Jasmonate-responsive expression of paclitaxel biosynthesis genes in *Taxus cuspidata* cultured cells is negatively regulated by the bHLH transcription factors TcJAMYC1, TcJAMYC2, and TcJAMYC4. *Front. Plant Sci.* 6:115. doi: 10.3389/fpls.2015.00115

Li, H. (2013). *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. Available online at: https://arxiv.org/abs/1303.3997v2 (accessed May 26, 2013).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., et al. (2006). Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. *Plant Physiol.* 141, 1167–1184. doi: 10.1104/pp.106.080580

Lin, W., Huang, W., Ning, S., Gong, X., Ye, Q., and Wei, D. (2019). Comparative transcriptome analyses revealed differential strategies of roots and leaves from methyl jasmonate treatment *Baphicacanthus cusia* (Nees) Bremek and differentially expressed genes involved in tryptophan biosynthesis. *PLoS ONE* 14:e212863. doi: 10.1371/journal.pone.0212863

Lin, W., Huang, W., Ning, S., Wang, X., Ye, Q., Wei, D., et al. (2018). *De novo* characterization of the *Baphicacanthus cusia* (Nees) Bremek transcriptome and analysis of candidate genes involved in indican biosynthesis and metabolism. *PLoS ONE* 13:e199788. doi: 10.1371/journal.pone.0199788

Ma, R. F., Liu, Q. Z., Xiao, Y., Zhang, L., Li, Q., Yin, J., et al. (2016). The phenylalanine ammonia-lyase gene family in *Isatis indigotica* Fort.: molecular cloning, characterization, and expression analysis. *Chin. J. Nat. Med.* 14, 801–812. doi: 10.1016/S1875-5364(16)30097-8

Ma, S., Yuan, Y., Tao, Y., Jia, H., and Ma, Z. (2020). Identification, characterization and expression analysis of lineage-specific genes within *Triticeae*. *Genomics* 112, 1343–1350. doi: 10.1016/j.ygeno.2019.08.003

Marcinek, H., Weyler, W., Deus-Neumann, B., and Zenk, M. H. (2000). Indoxyl-UDPG-glucosyltransferase from *Baphicacanthus cusia*. *Phytochemistry (Oxford)* 53, 201–207. doi: 10.1016/S0031-9422(99)00430-6

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Pal, S. K., and Shukla, Y. (2003). Herbal medicine: current status and the future. *Asian Pac. J. Cancer Prev.* 4, 281–288.

Patra, B., Pattanaik, S., Schluttenhofer, C., and Yuan, L. (2018). A network of jasmonate-responsive bHLH factors modulate monoterpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *New Phytol.* 217, 1566–1581. doi: 10.1111/nph.14910

Pires, N., and Dolan, L. (2010). Origin and diversification of basic-helix-loop-helix proteins in plants. *Mol. Biol. Evol.* 27, 862–874. doi: 10.1093/molbev/msp288

Sanderson, M. J. (2003). R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, 459–466. doi: 10.1093/nar/gky320

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C., Vert, J., et al. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16:259. doi: 10.1186/s13059-015-0831-x

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35, 1167–1169. doi: 10.1038/nbt.4020

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Singh, S. K., Patra, B., Paul, P., Liu, Y., Pattanaik, S., and Yuan, L. (2021). *BHLH IRIDOID SYNTHESIS 3* is a member of a bHLH gene cluster regulating terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Direct.* 5:e00305. doi: 10.1002/pld3.305

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32, 309–312. doi: 10.1093/nar/gkh379

Sugimoto, S., Naganuma, M., Kiyohara, H., Arai, M., Ono, K., Mori, K., et al. (2016). Clinical efficacy and safety of oral Qing-Dai in patients with ulcerative colitis: a single-center open-label prospective study. *Digestion* 93, 193–201. doi: 10.1159/000444217

Sun, W., Leng, L., Yin, Q., Xu, M., Huang, M., Xu, Z., et al. (2019). The genome of the medicinal plant *Andrographis paniculata* provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide. *Plant J.* 97, 841–857. doi: 10.1111/tpj.14162

Tanaka, T., Ikeda, T., Kaku, M., Zhu, X., Okawa, M., Yokomizo, K., et al. (2004). A new lignan glycoside and phenylethanoid glycosides from *Strobilanthes cusia* Bremek. *Chem. Pharm. Bull.* 52, 1242–1245. doi: 10.1248/cpb.52.1242

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 10, 1–14. doi: 10.1002/0471250953.bi04 10s25

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Tsai, Y., Lee, C., Yen, H., Chang, Y., Lin, Y., Huang, S., et al. (2020). Antiviral action of tryptanthrin isolated from *Strobilanthes cusia* leaf against human coronavirus NL63. *Biomolecules* 10:366. doi: 10.3390/biom10030366

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 527, 508–511. doi: 10.1038/nature15714

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wang, L., Zhou, G., Liu, P., Song, J., Liang, Y., Yan, X., et al. (2008). Dissection of mechanisms of Chinese medicinal formula Realgar-*Indigo naturalis* as an effective treatment for promyelocytic leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4826–4831. doi: 10.1073/pnas.0712365105

Wang, W., Xia, H., Yang, X., Xu, T., Si, H. J., Cai, X. X., et al. (2014). A novel 5-enolpyruvoylshikimate-3-phosphate (EPSP) synthase transgene for glyphosate resistance stimulates growth and fecundity in weedy rice (*Oryza sativa*) without herbicide. *New Phytol.* 202, 679–688. doi: 10.1111/nph.12428

Wang, Y., Liu, W., Jiang, H., Mao, Z., Wang, N., Jiang, S., et al. (2019). The R2R3-MYB transcription factor MdMYB24-like is involved in methyl jasmonate-induced anthocyanin biosynthesis in apple. *Plant Physiol. Bioch.* 139, 273–282. doi: 10.1016/j.plaphy.2019.03.031

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. doi: 10.1038/s41587-019-0217-9

Wu, X., Chen, X., Jia, D., Cao, Y., Gao, S., Guo, Z., et al. (2016). Characterization of anti-leukemia components from *Indigo naturalis* using comprehensive two-dimensional K562/cell membrane chromatography and *in silico* target identification. *Sci. Rep.* 6:30103. doi: 10.1038/srep30103

Xia, X. (2018). DAMBE7: new and improved tools for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* 35, 1550–1552. doi: 10.1093/molbev/msy073

Xia, Z. Q., and Zenk, M. H. (1992). Biosynthesis of indigo precursors in higher plants. *Phytochemistry* 31, 2695–2697. doi: 10.1016/0031-9422(92)83613-4

Xiao, C., Yang, W., Tu, J., Yuan, C., Huang, L., Hu, Y., et al. (2018). Determination analysis of six bioactive constituents different parts in different habitats of *Baphicacanthus cusia* (Nees) Bremek by RP-HPLC. *Nat. Prod. Res. Dev.* 30, 1188–1194. doi: 10.16333/j.1001-6880.2018.7.016

Xu, J., Chu, Y., Liao, B., Xiao, S., Yin, Q., Bai, R., et al. (2017). *Panax ginseng* genome examination for ginsenoside biosynthesis. *Gigascience* 6, 1–15. doi: 10.1093/gigascience/gix093

Xu, W., Zhang, L., Cunningham, A. B., Li, S., Zhuang, H., Wang, Y., et al. (2020). Blue genome: chromosome-scale genome reveals the evolutionary and molecular basis of indigo biosynthesis in *Strobilanthes cusia*. *Plant J.* 104, 864–879. doi: 10.1111/tpj.14992

Yamada, Y., Kokabu, Y., Chaki, K., Yoshimoto, T., Ohgaki, M., and Yoshida, S., et al. (2011). Isoquinoline alkaloid biosynthesis is regulated by a unique bHLH-type transcription factor in *Coptis japonica*. *Plant Cell Physiol.* 52, 1131–1141. doi: 10.1093/pcp/pcr062

Yu, H., Li, T., Ran, Q., Huang, Q., and Wang, J. (2021). *Strobilanthes cusia* (Nees) Kuntze, a multifunctional traditional Chinese medicinal plant, and its herbal medicines: a comprehensive review. *J. Ethnopharmacol.* 265:113325. doi: 10.1016/j.jep.2020.113325

Yu, J., Zhang, Y., Ning, S., Ye, Q., Tan, H., Chen, R., et al. (2019). Molecular cloning and metabolomic characterization of the 5-enolpyruvylshikimate-3-phosphate synthase gene from *Baphicacanthus cusia*. *BMC Plant Biol.* 19:485. doi: 10.1186/s12870-019-2035-0

Zhang, A., Ning, B., Sun, N., Wei, J., and Ju, X. (2015). Indirubin increases $CD4^+CD25^+foxp3^+$ regulatory T cells to prevent immune thrombocytopenia in mice. *PLoS ONE* 10:e142634. doi: 10.1371/journal.pone.0142634

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* 5, 833–845. doi: 10.1038/s41477-019-0487-8

Zhao, Q., Yang, J., Cui, M., Liu, J., Fang, Y., Yan, M., et al. (2019). The reference genome sequence of *Scutellaria baicalensis* provides insights into the evolution of wogonin biosynthesis. *Mol. Plant* 12, 935–950. doi: 10.1016/j.molp.2019.04.002

Zhou, W., Shi, M., Deng, C., Lu, S., Huang, F., Wang, Y., et al. (2021). The methyl jasmonate-responsive transcription factor SmMYB1 promotes phenolic acid biosynthesis in *Salvia miltiorrhiza*. *Hortic. Res.* 8:10. doi: 10.1038/s41438-020-00443-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.