the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

*The Plant Journal* (2020) **104,** 864–879 doi: 10.1111/tpj.14992

RESOURCE

# Blue genome: chromosome-scale genome reveals the evolutionary and molecular basis of indigo biosynthesis in *Strobilanthes cusia*

**Wei Xu[1], Libin Zhang[1,3], Anthony B. Cunningham[4], Shan Li[1], Huifu Zhuang[1], Yuhua Wang[1,*] and Aizhong Liu[2,*]**

[1]*Department of Economic Plants and Biotechnology, Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, 132 Lanhei Road, Kunming 650201, China,*

[2]*Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming 650224, China,*

[3]*University of the Chinese Academy of Sciences, Beijing 100049, China, and*

[4]*School of Veterinary and Life Sciences, Murdoch University, 90 South St, Murdoch, WA 6150, Australia*

Received 27 March 2020; accepted 15 September 2020; published online 27 September 2020.
*For correspondence (e-mail wangyuhua@mail.kib.ac.cn; liuaizhong@mail.kib.ac.cn).

## SUMMARY

**Natural plant dyes have been developed and used across many traditional societies worldwide. The blue pigment indigo has seen widespread usage across South America, Egypt, Europe, India and China for thousands of years, mainly extracted from indigo-rich plants. The utilization and genetic engineering of indigo in industries and ethnobotanical studies on the effects of cultural selection on plant domestication are limited due to lack of relevant genetic and genomic information of dye plants. *Strobilanthes cusia* (Acanthaceae) is a typical indigo-rich plant important to diverse ethnic cultures in many regions of Asia. Here we present a chromosome-scale genome for *S. cusia* with a genome size of approximately 865 Mb. About 79% of the sequences were identified as repetitive sequences and 32 148 protein-coding genes were annotated. Metabolic analysis showed that the main indigoid pigments (indican, indigo and indirubin) were mainly synthesized in the leaves and stems of *S. cusia*. Transcriptomic analysis revealed that the expression level of genes encoding metabolic enzymes such as monooxygenase, uridine diphosphate-glycosyltransferase and β-glucosidase were significantly changed in leaves and stems compared with root tissues, implying their participation in indigo biosynthesis. We found that several gene families involved in indigo biosynthesis had undergone an expansion in number, with functional differentiation likely facilitating indigo biosynthesis in *S. cusia*. This study provides insight into the physiological and molecular bases of indigo biosynthesis, as well as providing genomic data that provide the basis for further study of *S. cusia* cultivation by Asia's traditional textile producers.**

**Keywords: genome, indigo, indigo-rich plant, *Strobilanthes cusia*, transcriptome.**

**Linked article:** This paper is the subject of a Research Highlight article. To view this Research Highlight article visit https://doi.org/10.1111/tpj.15016.

## INTRODUCTION

Natural plant dyes have been developed and used across many traditional societies worldwide. Indigo, the oldest blue pigment used by humans, has been prized by textile weavers in diverse ethnic cultures and societies for thousands of years (Balfour-Paul, 2006). Based on anthropological evidence, indigo dye began being used in textiles and clothing in 4000 BC in South America (Splitstoser *et al.*, 2016). The utilization of indigo dye has been recorded in many areas of the world, such as in India (approximately 2600 BC) (Burkhill, 1921), Egypt (2400 BC) (Balfour-Paul, 2006), Europe (at Austria in 1500 BC) (Hartl *et al.*, 2015), China (approximately 900 BC) (Zhang *et al.*, 2008; Kramell *et al.*, 2014) and Asia, as the center of diversity for indigo-producing plant species (Cardon, 2007). However, during the late 19th century, the extremely valuable trade in natural indigo declined due to competition from synthetic indigo, but recently there has been a resurgence of economic interest in natural dyes and pigments from the

"green" fashion industry, in part due to the deleterious effects of environmental pollutants in synthetic indigo production (Muthu and Gardetti, 2016).

Globally, at least 31 plant species across eight different families are indigo sources, including *Persicaria tinctoria* (Polygonaceae), *Isatis indigotica* (Brassicaceae), *Indigofera tinctoria* (Fabaceae) and *Strobilanthes cusia* (Nees) Kuntze (Acanthaceae) (Hadders, 1933; Cardon, 2007; Yusuf and Shahid-ul-islam, 2017). Although the organic compounds of blue pigments were identified as indigo and indirubin, mainly derived from their precursor indoxyl-3-*O*-β-D-glucoside (indican), the independent evolutionary process and molecular mechanisms behind indigo biosynthetic pathways in plants remain largely unclear. Previously, most studies on dye plants focused on resource utilization from ethnobotanical investigations (Cardon, 2007; Han and Quye, 2018; Andriamanantena *et al.*, 2019; Zhang *et al.*, 2019), colorant extraction and compound identification in phytochemistry (Chanayath *et al.*, 2002; Tayade and Adivarekar, 2014; Dutta *et al.*, 2017). Only limited genetic and genomic resources of indigo plants are available (Kang *et al.*, 2020), and the investigation of the molecular mechanism of indigo biosynthesis has been largely ignored. The "green" market is now demanding the development of genetically improved natural indigo varieties with higher indigo content. Genomic information could potentially establish foundational resources for deciphering the molecular basis behind indigo biosynthetic pathways, crucial to the production process in agriculture.

Increasing evidence reveals that *S. cusia* ($2n = 32$) has the highest indigo yields when compared with other indigo sources (Laitonjam and Wangkheirakpam, 2011). Consequently, *S. cusia* was extensively cultivated and used for centuries by ethnic minority groups in many countries, including China, Bangladesh, India and Myanmar. However, unlike food crops the domestication of *S. cusia* was mainly driven by the cultural values of indigo-dyed clothing and local perceptions of the medicinal values of indigo-dyed textiles. In particular, the Landian Yao people worship the color blue and enjoy dressing in indigo-dyed blue clothing in daily life (Figure 1a,b). Owing to the high demand for dyed clothing, leaves and stems of *S. cusia* are commonly sold in local markets. Natural indigo extracted from the leaf and stem of *S. cusia* has been used as traditional Chinese medicine, named *qing dai* or *nanbanlangen* to treat dental ulcers, ulcerative colitis and psoriasis (Lin *et al.*, 2014; Sugimoto *et al.*, 2016; Zhao *et al.*, 2016) due to its antivirus, anti-inflammatory and anti-leukemia activities (Liau *et al.*, 2007; Hu *et al.*, 2015). Since the SARS outbreak in 2003, interest in the antiviral properties of *nanbanlangen* has increased (Ni *et al.*, 2012), with particular recent attention being paid to the chemistry of *S. cusia* and the new indole alkaloids it contains (Lee *et al.*, 2019). Previous studies based on transcriptomic analysis revealed many candidate genes involved in the biosynthesis of indican backbone biosynthesis (Lin *et al.*, 2018, 2019).

Here, we conducted single-molecule sequencing combined with high-throughput chromosome conformation capture (Hi-C) technology to assemble the chromosome-scale genome of *S. cusia*. Furthermore, we identified candidate gene sets involved in the biosynthesis of indigo. This study could provide basic data useful in dissecting the molecular mechanisms behind natural pigment biosynthesis *in vivo* while also paving a critical way forward for deciphering how cultural selection or drivers have affected the domestication and spread of *S. cusia* throughout montane Southeast Asia.

## RESULTS

### Chromosome-scale assembly of *Strobilanthes cusia* genome

Before genome sequencing, the genome size of *S. cusia* was determined based on the *k*-mer method using high-quality reads from Illumina genome shotgun sequencing data. We obtained, in total, 57.98 Gb sequencing data (Table S1), with a *k*-mer number of 47 753 291 639 and K-mer depth of 57 (Figure S1a), which estimated an 826.37 Mb genome size. There was low genome heterozygosity (0.53%) and relatively high repeat content in the *S. cusia* genome (about 68.34%). Estimated genome size was nearly consistent with the observation of flow cytometry (approximately 857 Mb) (Figure S1b). We used MinION single-molecule sequencing to yield about 5 663 926 nanopore long reads, in total, 114.47 Gb with an N50 length of 22.6 kb and a maximum read length of 163 kb, which was about 130× coverage of the *S. cusia* genome (Table S1). Upon using integrative genome assemblers such as Caun, Falcon and WTDGB, the corrected long reads were assembled into an initial genome with a total length of 865.49 Mb, consisting of 1602 contigs with a contig N50 of 4.33 Mb (Table 1).

To generate chromosome-scale genome assembly in *S. cusia*, the Hi-C sequencing was employed for improving the assembly further. In total, we obtained 158 290 970 Hi-C paired-end reads, yielding about 49.18 Gb of clean data (Table S1). As expected, there was a substantially strong interactive signal within intra-chromosomal and at diagonal regions, as shown by HiCplotter at 100-kb resolution (Figure 1c). As a result, 933 contigs with a total length of 848.78 Mb were anchored into 16 pseudochromosomes (Table 1 and Figure 1c,d) ranging from 39.6 to 70.7 Mb (Figure 1d and Table S2). This means that 98.07% of the *S. cusia* genome was accurately assembled. An initial estimation of the completeness of the *S. cusia* genome was performed by searching against the CEGWA database with 458 conserved core eukaryotic genes and the embryophyta_odb10 database with 1440 benchmarking universal single-copy orthologs (BUSCOs). We found that there were 442

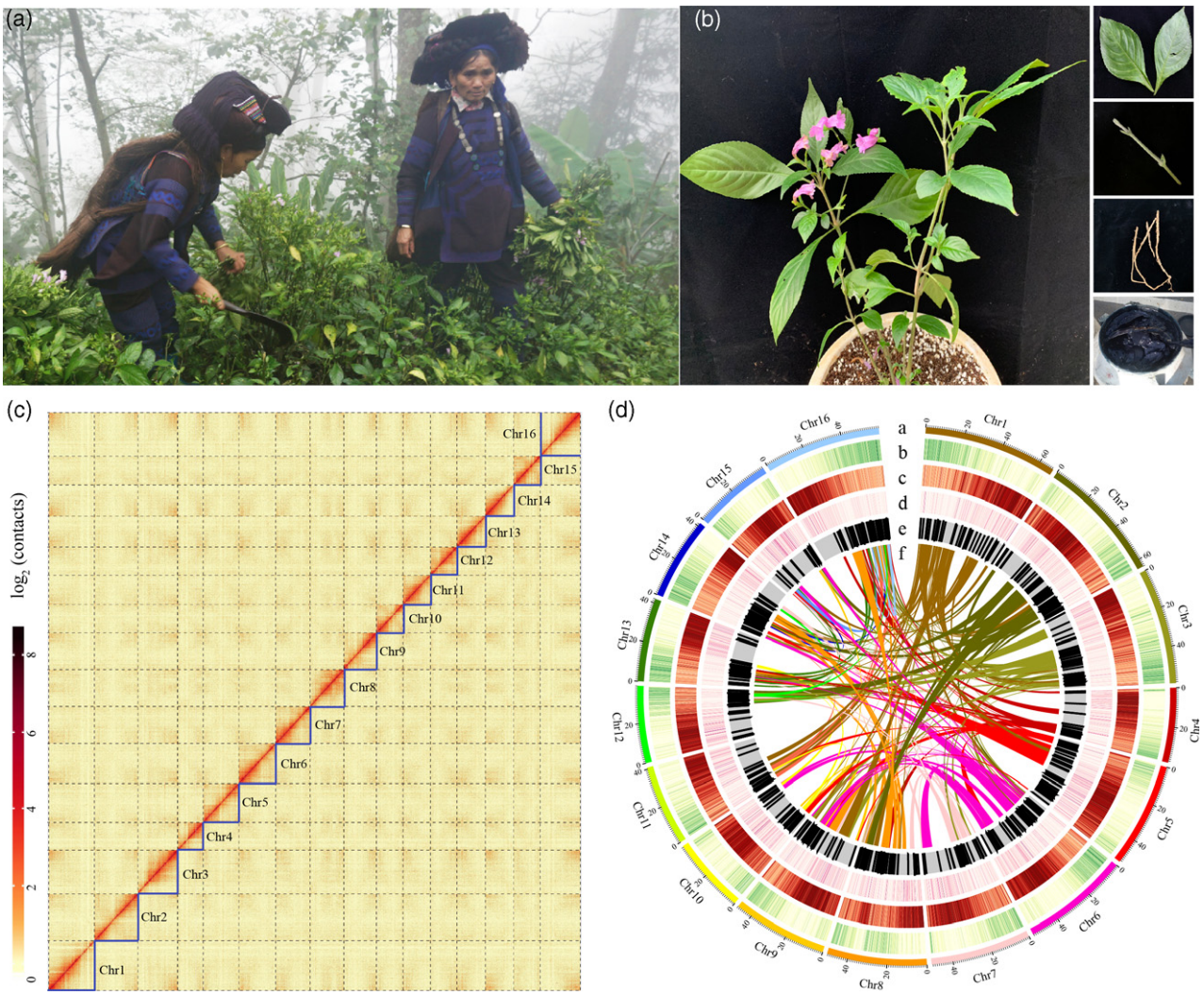


**Figure 1.** The plant *Strobilanthes cusia* and genome assembly.
(a) Landian Yao people dressed in indigo-dyed blue clothing while harvesting the leaf and stem of *S. cusia* for making blue indigo.
(b) The sequenced strain named "LanDian Yao." Morphology of the whole plant (left panel), leaf, stem and root (right panel) was displayed. Extracted indigo from leaf and stem shown in the bottom of the right panel.
(c) High-throughput chromosome conformation capture map of the *S. cusia* genome showing strong interactive signal within intra-chromosome (intra-Chr) and at diagonal regions.
(d) Overview of *S. cusia* genome assembly. Chr (a), gene (b), repeat (c), pseudogene (d), transcripts (e) and colinearity blocks (f) were displayed. The color from light to dark on the circular density scale indicates a density of features from low to high presented in a given track.

(96.5%) complete core eukaryotic genes and 1269 (88.1%) BUSCOs (5.56% of which were duplicated) present in the *S. cusia* genome (Table S3). In addition, the completeness and accuracy of genome sequences were further determined by mapping Illumina reads back to the *S. cusia* genome, which showed that 98.4% of Illumina reads were aligned to the nuclear genome sequences with >99% sequence identity (Table S1). In sum, these benchmarks supported the high-quality assembly of the *S. cusia* genome at chromosomal scale. All genome information, including genome sequences, GFF3 file, functional annotation, coding and protein sequences, and basic BLAST tools are available at the website we developed (http://indigoid-plant.iflora.cn).

## Genome annotation and evolutionary analysis

Genome sequence analysis showed that about 79.02% (683.9 Mb) of the *S. cusia* genome consisted of repetitive sequences (Table 1 and Table S4). Among these repetitive sequences, transposable elements (TEs) were dominant, accounting for about 74.4% of the genome. Further classification showed a substantially higher proportion of retrotransposons (69.93%) than DNA transposon (4.46%) in the genome, and the TE subclass Gypsy (42.89%) was the most abundant followed by LARD (17.39%) in the genome (Table S4).

Based on integrative analysis using *ab initio* and the homology-based method on RNA-sequencing (RNA-seq)

**Table 1** Summary of assembly and annotation of the *Strobilanthes cusia* genome

| | |
|---|---|
| Genome assembly | |
| Number of contig | 1602 |
| Total length of contig (bp) | 865 492 022 |
| N50 of contig (bp) | 4 329 179 |
| N90 of contig (bp) | 606 222 |
| Max. contig (bp) | 20 838 575 |
| GC content | 36.71% |
| Chromosome | 16 |
| Number of anchored contigs | 933 |
| Total length of anchored contig (bp) | 848 778 248 |
| Gene annotation | |
| Number of genes | 32 148 |
| Total length of genes (bp) | 110 914 527 |
| Mean gene length (bp) | 3450.12 |
| Mean exon length (bp) | 244.49 |
| Mean intron length (bp) | 441.85 |
| Total length of repetitive sequence (bp) | 683 920 465 |
| Non-coding RNAs | 1181 |
| Pseudogenes | 4844 |

data obtained from five tissues with three biological replicates (Table S1), we identified 32 148 protein-coding genes with an average length of 3450 bp (Table 1 and Table S5) in the repeat-masked *S. cusia* genome. Of them, 29 489 (91.72%) of protein-coding genes were supported by transcripts from RNA-seq. Further functional prediction revealed 30 417 genes (94.62%) with known functional annotations in public databases (Table S6). These results showed that predicted genes in the *S. cusia* genome are highly reliable. Moreover, we identified 1181 non-coding RNAs (including 685 transfer RNAs, 435 ribosomal RNAs and 61 microRNAs [miRNAs]) and 4844 pseudogenes in the genome (Table 1).

In the *S. cusia* genome, about 82.7% (26 601) of genes were assigned to 14 163 gene families with 1.88 average genes per family, a number comparable with gene families in other plant species such as *Mimulus guttatus* (Phrymaceae, 13 772), *Salvia miltiorrhiza* (Lamiaceae, 13 045) and *Sesamum indicum* (Pedaliaceae, 13 341), but significantly higher than the same family plant *Andrographis paniculata* (Acanthaceae, 12 283) (Table S7). Among these gene families, 562 gene families were unique to *S. cusia* when compared with *A. paniculata*, *S. miltiorrhiza* and *S. indicum* (Figure 2a). The phylogenetic tree and divergence time were analyzed based on the 378 orthologous sing-copy genes among 12 species. The results showed that the divergence between *A. paniculata* and *S. cusia* (in the same family Acanthaceae) occurred approximately 33.7 MYA (Figure 2b), and split from their ancestors *M. guttatus*, *S. miltiorrhiza* and *S. indicum* approximately 51.7 MYA. Inspecting the gains and losses within the gene family, we found that there were 1072 gene families undergoing expansion but 420 undergoing contractions in *S. cusia* (Viterbi $P < 0.05$; Figure 2c). Interestingly, pfam annotation showed that these expanded genes families were abundant in the RNA-dependent DNA polymerase, gag-polypeptide of long terminal repeat (LTR) copia-type, protein kinase domain, NB-ARC domain, glycosyl hydrolases family, cytochrome P450 (CYP), F-box domain, Myb-like DNA-binding domain, uridine diphosphate (UDP)-glucoronosyl and UDP-glycosyltransferase (UGT) (Table S8). Moreover, both Ks and 4DTv analyses showed that the *S. cusia* genome may have only experienced one ancient whole-genome triplication ($\gamma$), and no specific whole-genome duplication occurred in the *S. cusia* genome after divergence from *Vitis vinifera* (Vitaceae) (Figure 2d; Figure S2).

### Putative indigo biosynthetic pathway and content detection in *Strobilanthes cusia*

Studies have shown that three secondary metabolites, including indican, indigo and indirubin are the main active materials in *S. cusia* (Liau *et al.*, 2007; Lin *et al.*, 2018), but little is known about the biosynthesis of these compounds to date. Presently, the putative pathway of indigo biosynthesis was summarized as follows (Warzecha *et al.*, 2017; Hsu *et al.*, 2018): the indole was first oxygenated to the highly reactive indoxyl by monooxygenase, the precursor substance of indigo; subsequently, the indoxyl was immediately glucosylated by UGT to form the colorless and stable material indican *in vivo*. When the plant tissue is damaged, the indican can be reversibly hydrolyzed into the indoxyl by $\beta$-glucosidase (BGL), and indoxyl was spontaneously oxidized into indigo and indirubin, which both exhibited visible coloration (Figure 3a).

Here, we measured the content of these three major secondary metabolisms in different tissues, including root, stem and leaf at four developing stages using high-performance liquid chromatography technology (Figure 3b). The result showed that these three metabolisms were mainly synthesized and accumulated in leaf tissues, followed by the stem and finally root (Figure 3b). In the leaf, indigo was most abundant, up to 2000 $\mu g\ g^{-1}$ at the stage of the leaf 4, while the content of indican exhibited a strongly negative correlation with that of indigo (Pearson's correlation coefficient $r_p = -0.93$). Interestingly, we found that the content of indirubin was substantially lower in all tested tissues, at $\leq 13\ \mu g\ g^{-1}$. The result revealed that indigo is mainly synthesized in the leaf and stem, consistent with our previous ethnobotanical investigation in which local residents used the leaf and stem as the main raw materials for indigo production (see Figure 1a,b) (Li *et al.*, 2019; Zhang *et al.*, 2019). Furthermore, we found that indican was frequently hydrolyzed and converted into the indigo, not the indirubin *in vivo*.

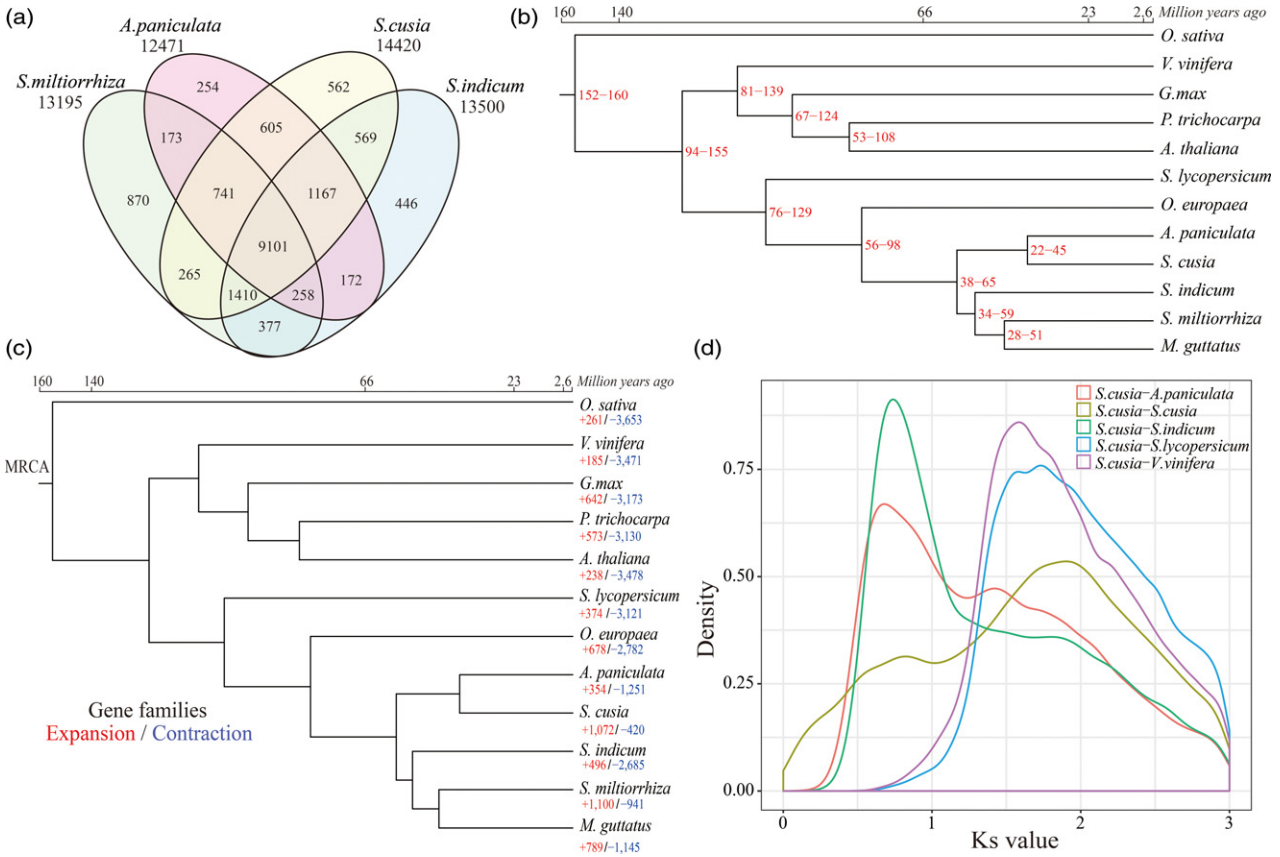


**Figure 2.** Evolutionary analyses of *Strobilanthes cusia* genome.
(a) Comparative analysis of orthologous gene families among four species, including *S. cusia*, *Andrographis paniculata* (in the same family, Acanthaceae), *Salvia miltiorrhiza* and *Sesamum indicum*.
(b) Divergence time estimation among 12 plant species. Phylogenetic tree was constructed based on orthologous single-copy genes.
(c) Expansion and contraction of gene families among 12 plant genomes. Number of gene family expansion and contraction was indicated by red and blue number, respectively. MRCA, most recent common ancestor.
(d) Distribution of Ks value between syntenic gene pairs among *S. cusia*, *A. paniculata*, *Solanum lycopersicum*, *S. indicum* and *Vitis vinfera*.

### Transcriptome analysis

To find out the key genes involved in the production of indican, indigo and indirubin, we employed transcriptome analysis to profile the spatio-temporal expressions of genes across different tissues with three independent biological replicates, including leaves at two developing stages (L1 and L4), stems at two developing stages (S1 and S2) and roots (R). In total, we obtained approximately 8.27 Gb of clean data for each sample, and over 88% reads were able to align on to the *S. cusia* genome uniquely (Table S1). In total, 27 479 genes with expression levels of fragments per kilobase of transcript per million fragments mapped (FPKM) ≧0.5 in at least one sample were detected among these tissues (Table S9). We found that the samples from the same tissue or developing stage were tightly clustered and exhibited a strong correlation with a Pearson's correlation coefficient of $r_p > 0.94$ (Figure 3c). Differential expression analysis revealed that there were 4893, 3778, 4040 and 3068 significantly upregulated and 3928, 4357, 3322 and 3431 significantly downregulated genes (false discovery rate [FDR] <0.05) in L1, L4, S1 and S2, respectively, relative to R (Figure 3d and Tables S10–S13). Among these differentially expressed genes (DEGs), there were 1314 upregulated and 1965 downregulated genes shared by the leaf (L1 and L4) and stem (S1 and S2) (Table S14), which were significantly enriched in Gene Ontology (GO) terms (adjusted $P < 0.01$), including catalytic activity, oxidoreductase activity and transporter activity (Figure S3). These DEGs were used to identify the candidate genes involved with indigo biosynthesis, as the indigo pathway was mainly activated in leaf and stem tissues.

### Monooxygenase in *Strobilanthes cusia* genome

As mentioned above, the first and key step for indigo biosynthesis is the oxidative modification of the indole to form the indoxyl by monooxygenase. The plant CYP family is typically defined as monooxygenase and plays critical

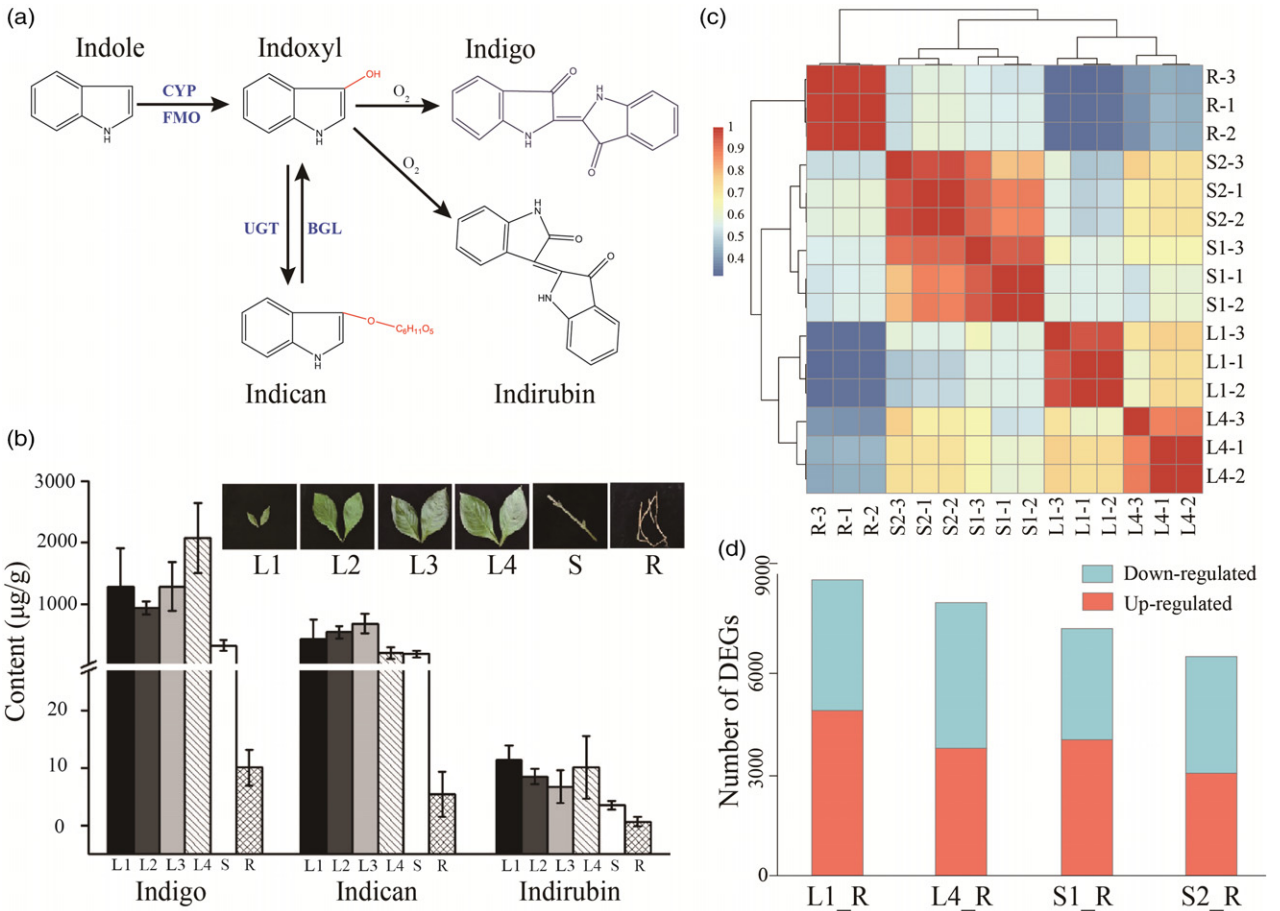


**Figure 3.** Secondary metabolisms detection and transcriptome analysis.
(a) Putative biosynthesis of indican, indigo and indirubin.
(b) Content of indican, indigo and indirubin in different tissues, including leaf from four developing stages (from L1 to L4), root (R) and stem (S) in *Strobilanthes cusia*.
(c) Pearson's correlation coefficient among samples based on gene expression level. Color bar indicates the Pearson correlation coefficient values.
(d) Differentially expressed genes (DEGs) between leaf (L1 and L4), stem (S1 and S2) and root.

roles in the biosynthetic pathways of secondary metabolisms, but they catalyze extremely diverse reactions and have a substantially lower identity in sequences (Bak *et al.*, 2011). Interestingly, some evidence suggests that human CYP enzymes can oxidize indole to form indigo (Gillam *et al.*, 2000; Gillam and Guengerich, 2001), but the function of plant CYP in the formation of indigo has yet to be investigated. In the *S. cusia* genome, we totally identified 342 genes encoding the putative CYPs, but 31 genes were excluded in subsequent analysis due to their short protein length (<200 amino acids) or incomplete p450 domain. Finally, 311 CYPs were classified into two subfamilies: A-type (148 members) and non-A-type (163 CYP71 members) (Figure 4a and Table S15). Among the A-types, they were further grouped into different clans, including CYP51 (two members), CYP74 (six members), CYP85 (48 members), CYP710 (two members), CYP72 (50 members), CYP97 (four members) and CYP86 (36 members). Importantly, we found that 18 gene families related to CYPs were significantly expanded in the *S. cusia* genome compared with other plant species because of tandem duplication (Figure 4a and Table S8).

Expression analysis indicated that 240 CYPs were expressed with FPKM >0.5 in at least one sample, while 71 CYPs had no transcripts in all tested samples (Table S15). Among these expressed CYPs, 24 genes exhibited significantly higher expression levels in the leaf and stem than in the root (FDR <0.01), while there were 58 downregulated CYPs (Figure 4b,c and Table S15). Subsequently, four highly expressed members (including EVM0017260, EVM0027604, EVM0028387, EVM0016194) were selected for quantitative reverse transcription–polymerase chain reaction (qRT-PCR) validation. The result showed that they were significantly and highly expressed in leaf (L1 and/or L4) and stem compared with root (Figure 4d), consistent with the result from RNA-seq.

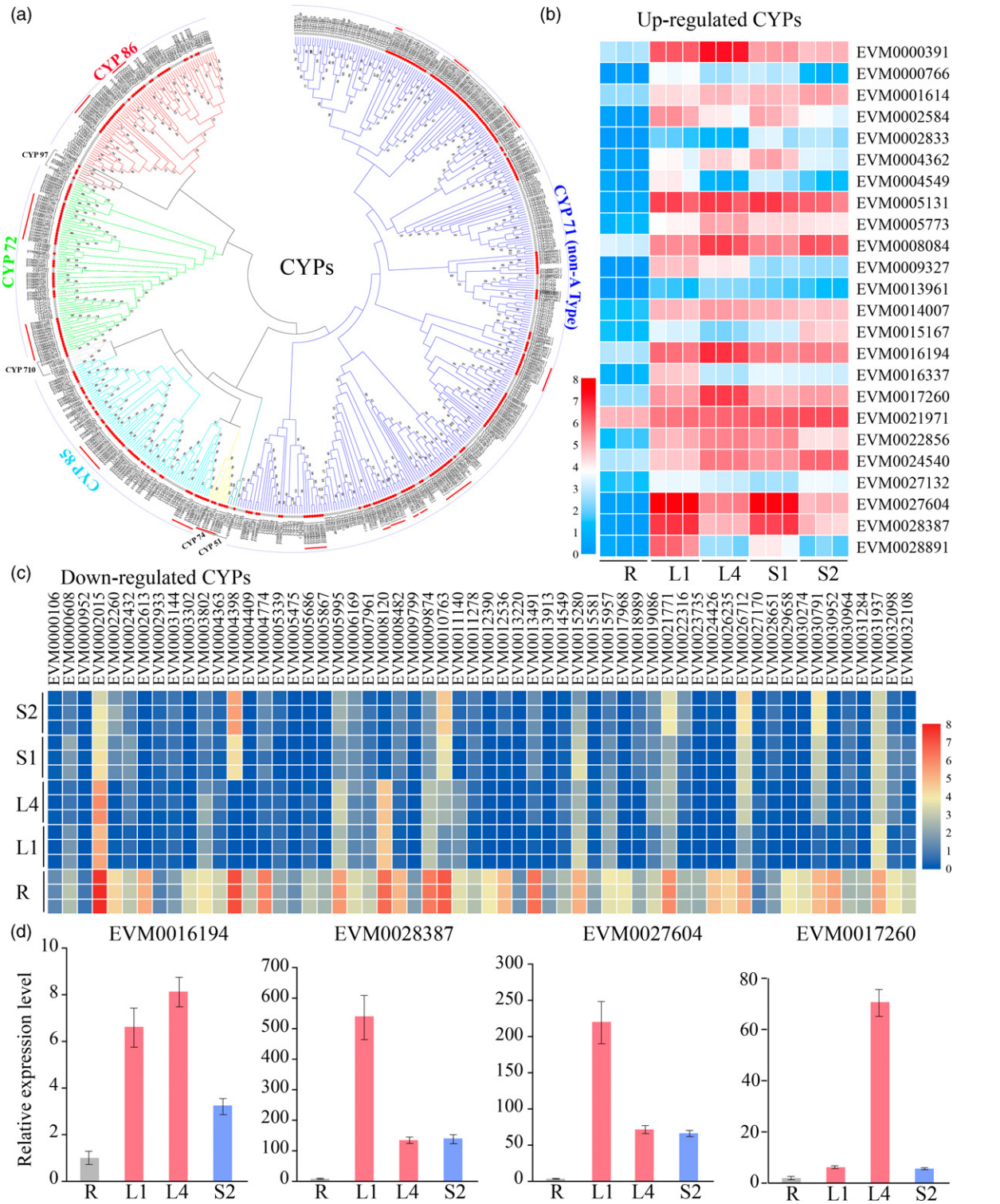


**Figure 4.** Cytochrome P450 (CYPs) in *Strobilanthes cusia* genome.
(a) Phylogenetic tree of CYPs based on the protein sequence alignments from *S. cusia* and Arabidopsis using the MEGA X with the neighbor-joining methods and 10 000 bootstraps. Gray and red dots represented CYP members from Arabidopsis and *S. cusia*, respectively. Red lines indicated expanded CYP genes.
(b) CYP members with significantly upregulated expression in leaf and stem, relative to root. Color bar indicates the genes with different expression levels.
(c) CYP members with significantly downregulated expression in leaf and stem, relative to root. Color bar indicates the genes with different expression levels.
(d) Quantitative reverse transcription–polymerase chain reaction validation of gene expression selected. Error bars showed the standard error across the five biological replicates, and the expression levels of root sample were normalized to 1.

Interestingly, these DEGs included those significantly expanded gene families. For example, the three members EVM0001614, EVM0008084 and EVM0017260, which were highly expressed in the leaf and stem, significantly expanded in the *S. cusia* genome because of tandem duplicates in the *S. cusia* genome.

In addition to CYPs, studies have revealed that other gene families such as flavin-containing monooxygenase (FMO) can also catalyze indole to form indoxyl (Choi *et al.*, 2003). We identified 36 genes encoding FMO in the *S. cusia* genome, and 26 members were detected in at least one sample (FPKM >0.5) (Table S16). Most of the FMOs (26 members) appear to be expressed with a low FPKM value (no more than 10) in all tested tissues (Table S16). Two upregulated members (EVM0030158 and EVM0009245) and eight downregulated members (EVM0021123, EVM0031493, EVM0001389, EVM0007768, EVM0016593, EVM0008637, EVM0019399 and EVM0029964) were identified in the leaf and stem compared with the root (FDR <0.05; Tables S14 and S16). These DEGs in CYP and FMO families were considered candidate genes for further functional inquiry.

### UGT in *Strobilanthes cusia* genome

The highly reactive indoxyl is extremely unstable and immediately glucosylated by UGT, generating the stable molecule indican *in vivo*. A recent study suggests that the glucosyltransferase PtUGT1 from the indigo plant *Polygonum tinctorium* can produce the indican efficiently by heterologous expression in *Escherichia coli* (Hsu *et al.*, 2018). In total, 173 putative genes encoding UGTs were identified in the *S. cusia* genome (Table S17), and phylogenetic studies showed 17 phylogenetic groups designated A–R (Figure 5a). Notably, *S. cusia* UGT members from subfamilies A–Q were tightly clustered with members from *Arabidopsis thaliana*, but the R and Q subfamilies only contain members from *S. cusia*. Sequence alignment of UGT proteins showed the diverse N-terminal residues that mainly interact with the diverse acceptor and conserved C-terminal residues that principally interact with the specific sugar donor, while the linker region between the N- and C-terminal domains exhibited flexibility in the sequences and length (Figure S4) (Osmani *et al.*, 2009). In the C-terminal domains, we found a conserved PSPG motif, which plays a critical role in the direct interaction with the UDP-sugar donor as demonstrated by the crystal structure analysis of UGTs from plants, particularly in the 10 amino acid residues in the PSPG motif as shown in Figure 5(b) (Osmani *et al.*, 2009).

Transcriptome analysis showed that 146 members of UGTs were expressed in at least one tissue with FPKM >0.5, and 27 did not detect transcripts in all tested samples (Table S17). Among these expressed genes, there were 21 upregulated members and 33 downregulated members in the leaf and stem compared with the root (Figure 5c and Table S17). Subsequently, seven genes (including EVM0021179, EVM0005112, EVM0028744, EVM0000484, EVM0022246, EVM0024034 and EVM0013920) were selected for qRT-PCR analysis because of their higher expression levels (average FPKM value >50). Results from qRT-PCR showed that except EVM0000484 gene, all six genes exhibited a relatively high transcript level in the leaf (L1 and/or L4) and stem (S2) when compared with the expression in the root (Figure 5d), in solid agreement with RNA-seq results. In addition, we identified two homologs (EVM0014302 and EVM0020688) of PtUGT1 in the *S. cusia* genome with over 52% sequence identity, but unfortunately, we could not detect their transcripts in various tissues or at developing stages.

Notably, we found that 10 gene families related to UGTs were significantly expanded in the *S. cusia* genome compared with other plant species (Figure 5a and Table S8). These expanded genes were extensively distributed in the subfamilies A, C, E, F, G, I, L, M and, in particular, R, resulting in the substantial increase in copy number within these subfamilies. Interestingly, we found that over half of the genes (21 members as shown in Figure 5b) highly expressed in leaf and stem tissues may be because of gene family expansion, suggesting that gene duplication may be one of the primary driving forces for the generation of new functions. For example, expanded UGT genes include four members (EVM0027831, EVM0016992, EVM0028627 and EVM0028874) in the R subfamily, three members (EVM0011669, EVM0000484 and EVM0009218) in the F subfamily, and four members (EVM0023821, EVM0024034, EVM0013737 and EVM0005112) from the other four subfamilies (A, E, G and L, respectively) that were significantly expressed in the leaf and stem relative to the root, which may indicate involvement in the biosynthesis of indican or other secondary metabolites.

### BGL in the *Strobilanthes cusia* genome

β-glucosidases (BGLs), belonging to the glycoside hydrolase family 1 in plants, is largely involved in various development and stress responses in plants (Xu *et al.*, 2004). Evidence has showed that it can hydrolyze indican to form indoxyl and subsequently indigo (Hsu *et al.*, 2018). Here, we systematically identified the BGLs in the *S. cusia* genome and, in total, 29 genes were discovered to encode putative BGL (Table S18). Phylogenetic analysis of BGLs from *S. cusia* and *A. thaliana* showed 10 distinct subgroups, namely, those from BGL-a to BGL-j, but in subgroups c–g and j there are a lack of members from *S. cusia*. In addition, we found a specific subgroup in *S. cusia* that contains the two members EVM0003108 and EVM0000169 (Figure 6a). Gene family analysis revealed that members from BGL-b underwent significant expansion; members from BGL-b were thought to be involved in

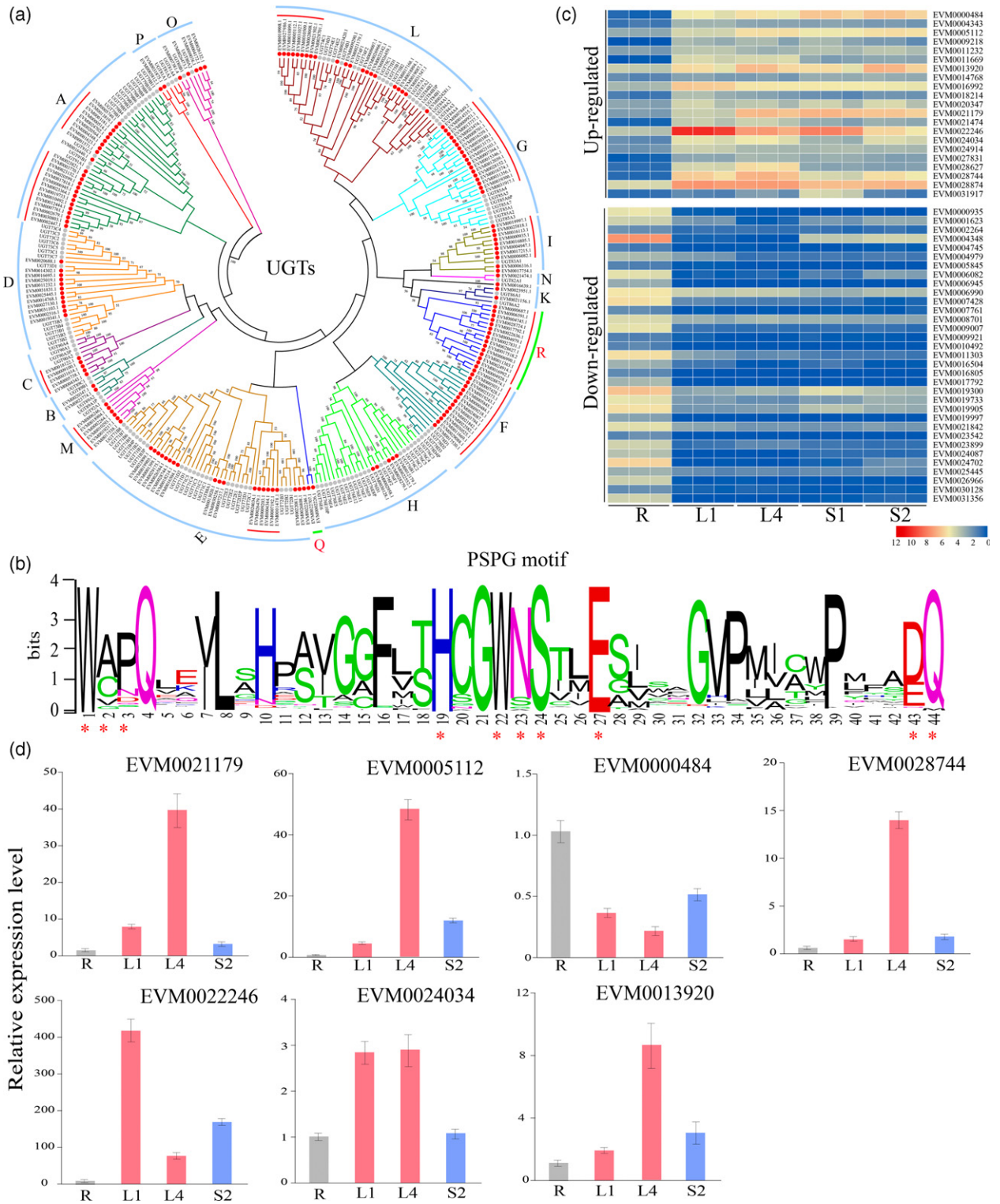


**Figure 5.** UDP-glycosyltransferase (UGTs) in *Strobilanthes cusia* genome.
(a) Phylogenetic tree of UGTs based on the protein sequence alignments from *S. cusia* and Arabidopsis using the MEGA X with the neighbor-joining methods and 10 000 bootstraps. Gray and red dots represented the UGT members from Arabidopsis and *S. cusia*, respectively. Red lines indicated expanded UGT genes.
(b) Logo showed the conserved amino acid residues (marked by asterisk) within PSPG motif.
(c) UGT members with significantly up- and downregulated expression in leaf and stem relative to root.
(d) Quantitative reverse transcription–polymerase chain reaction validation of gene expression selected. Error bars showed the standard error across the five biological replicates, and the expression level of the root sample was normalized to 1.

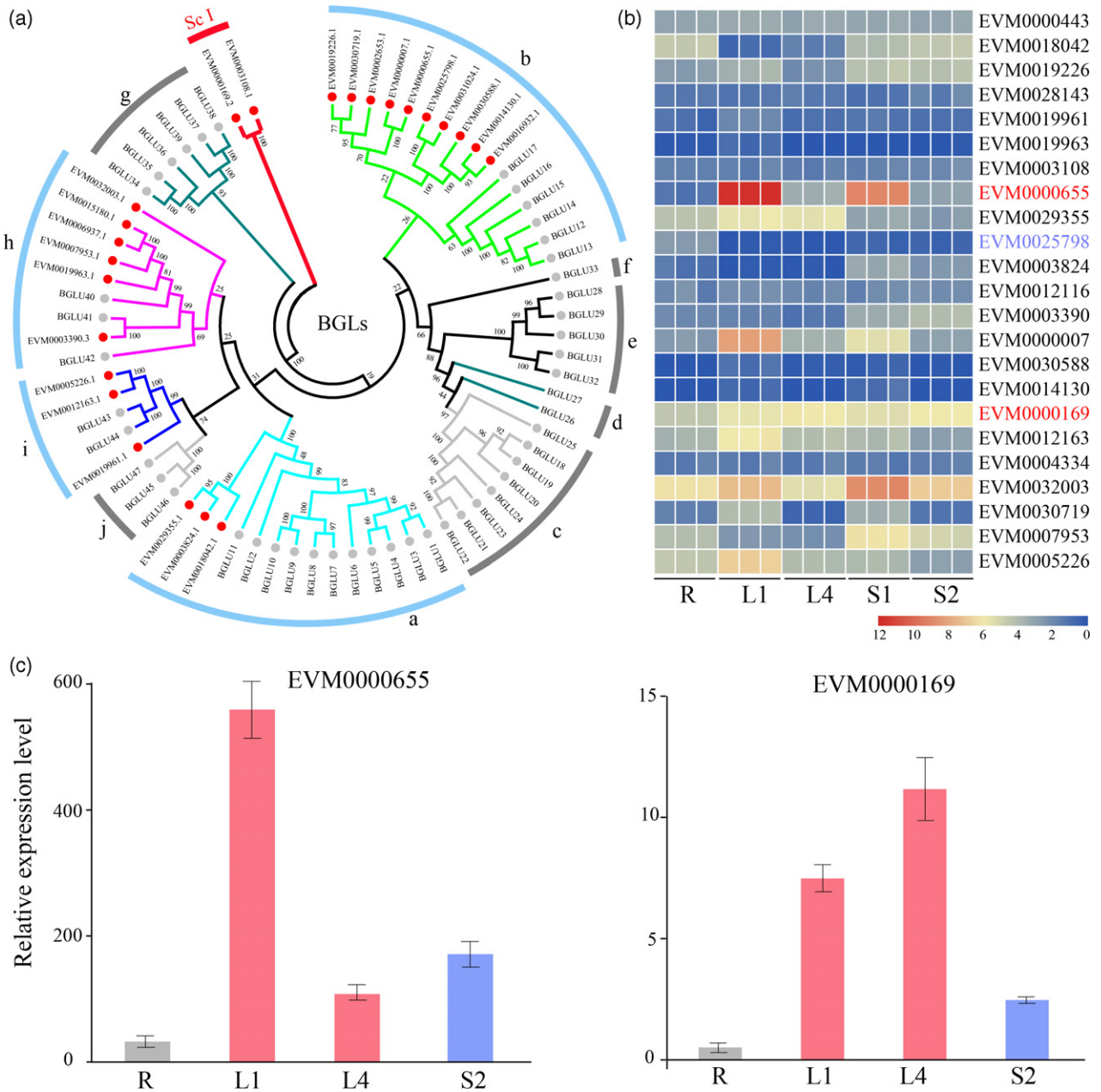


**Figure 6.** β-glucosidases (BGLs) in *Strobilanthes cusia* genome.
(a) Phylogenetic tree of BGLs based on the protein sequence alignments from *S. cusia* and Arabidopsis using the MEGA X with the neighbor-join methods and 10 000 bootstraps. Gray and red dots represented the BGLs members from Arabidopsis and *S. cusia*, respectively. BGL members marked by gray lines indicated those subfamilies that lacked members from *S. cusia*.
(b) Expression patterns of BGL members at different tissue or developing stages. Genes marked by red and blue colors indicate up- and downregulated genes, respectively. Color bar indicates genes with a different expression level. L1 and L4, leaves at two developing stages; R, roots; S1 and S2, stems at two developing stages.
(c) Quantitative reverse transcription–polymerase chain reaction validation of gene expression selected. Error bars showed the standard error across the five biological replicates, and the expression level of the root sample was normalized to 1.

flavonoid utilization (Xu *et al.*, 2004). Expression analysis showed that 23 BGL members were expressed in at least one tissue with FPKM >0.5, while six members exhibited low or no expression levels in all tested samples (Figure 6b and Table S18). Among these expressed genes, two members, EVM0000655 and EVM0000169, were expressed significantly higher in the leaf and stem than the root, which was further confirmed by qRT-PCR (Figure 6c), while only one gene EVM0025798 was significantly downregulated (Figure 6b and Table S18).

## DISCUSSION

The resurgence of natural dyes requires genetically improved natural indigo varieties with higher indigo content. Genomic information would enhance the breeding process of dye plants. To date, the genome of only one indigo-containing plant, *I. indigotica*, has been sequenced, and it was a recent occurrence (Kang *et al.*, 2020), but the potential molecular mechanism of indigo biosynthesis was not investigated in that study. It seems that *S. cusia* represents the most important indigo source for the study on natural dye biosynthesis as well as for examining the interaction between human culture and plant communities. Moreover, *S. cusia* is capable of extreme environmental adaptation and relatively fast growth with clonal propagation, exhibiting its potential as a resurgent natural indigo source.

Here, we present a chromosome-scale assembly of *S. cusia* genome using long reads from Nanopore sequencing and Hi-C technology. The *S. cusia* genome is approximately 865 Mb in size anchored into 16 chromosomes, with 32 148 protein-coding genes. This relatively bigger genome seems to have not undergone lineage-specific whole-genome duplication compared with other phylogenetically related species such as *A. paniculata* (269 Mb) (Sun *et al.*, 2019), *M. guttatus* (322 Mb) (Hellsten *et al.*, 2013) and *S. indicum* (274 Mb) (Wang *et al.*, 2014). The expanded repeat sequences or transposon elements may be major factors in the increase of the *S. cusia* genome size, which make up approximately 79.02% of the *S. cusia* genome, a substantially higher percentage than *A. paniculata* (53.3%) (Sun *et al.*, 2019), *S. indicum* (28.46%) (Wang *et al.*, 2014) and *S. miltiorrhiza* (53.58%) (Zhang *et al.*, 2015). Like other plants, LTR were most abundant in the *S. cusia* genome in which LTR/Gypsy accounted for 42.89% of the genome size, but the gene family of gag-polypeptide of LTR copia-type is significantly expanded in the *S. cusia* genome, which might be partly related to the increase of LTR/copia copies in the *S. cusia* genome.

One of the main objectives was to dissect potential molecular mechanisms underlying indigo biosynthesis and identify involved genes in *S. cuisa*. Based on the processes of indigo or indican biosynthesis (see Figure 3a), the key enzyme genes were supposed to be focused on encoding monooxygenase (including CYP and FMO), UGT and BGLs. Usually, these key enzymes were encoded by multiple genes; thus, identification of the gene families responsible for encoding key enzymes in the *S. cusia* genome would be crucial for understanding the biosynthesis of indigo or indican. Intriguingly, we found that many members within the CYP and UGT gene families were substantially expanded. In total, we identified 342 genes encoding CYP and 173 genes encoding UGT in the *S. cusia* genome, substantially more than that of *A. paniculata* (278 encoding CYP and 120 encoding UGTs) (Sun *et al.*, 2019), a phylogenetically related species in Acanthaceae. Differential or expanded genes encoding CYPs and UGTs in *S. cusia* relative to *A. paniculata* might be one of the bases for indigo biosynthesis. Moreover, the candidate genes identified from transcriptome and metabolome data (in particular, CYPs, FMOs, UGTs and BGLs) might directly participate in indigo or indican biosynthesis. In summary, the increased number of genes that encode key enzymes responsible for indigo or indican biosynthesis, coupled with their transcription during indigo or indican biosynthesis, complicates their functions in the indigo biosynthesis pathway. However, the functions of candidate genes need to be further tested *in vivo*.

In general, substantial phenotypic and genetic variations arise during the cultivation and domestication of crops because of artificial selection (Gaut *et al.*, 2018). Although dye plants such as *S. cusia*, *Justicia spicigera*, *I. indigotica* and *I. tinctoria* have been widely cultivated in many areas throughout global history, the cultivation and domestication of dye plants have mainly been driven by cultural demands. How this cultural selection affects phenotypic and genetic variations remain an open question. In addition, although Asia is the center of diversity for indigo-producing plant species cultivated by diverse indigenous communities (Cardon, 2007), how these indigo-producing plants were spread by cultural drivers also remains uncertain. The indigenous Landian Yao people have cultivated *S. cusia* for thousands of years in southwest China because of its cultural value ascribed to the color blue (Li *et al.*, 2019). Was the cultivation and utilization of *S. cusia* in other regions such as Malaysia, Thailand and Vietnam because of cultural spread or other drivers? The full *S. cusia* genome would provide an opportunity to develop genomic markers to help decipher genetic links among *S. cusia* germplasm cultivated across communities with different cultures in different countries. Further study may provide potential pathways forward for understanding how cultural selection or other drivers affect plant cultivation, domestication or spread.

## CONCLUSION

In conclusion, we have added a high-quality genome dataset to the family Acanthaceae, identified diverse candidate genes responsible for blue pigment biosynthesis in *S. cusia*, and developed a genome website for this species that includes genome and transcriptome data (http://indigoid-plant.iflora.cn). Moreover, this study provided an opportunity for deciphering the effects of cultural selection or drivers on plant cultivation, domestication or spread as well as elucidating the mechanisms behind the independent evolution of the indigo pathway among indigo-rich plant species.

## EXPERIMENTAL PROCEDURES

### Plant materials

*Strobilanthes cusia* (Nees) accession "Landian Yao" (voucher no. 0098327, deposited at KUN) was originally collected from Yao communities in Honghe, southern Yunnan Province, China, where it has been cultivated for thousands of years. The Landian Yao people harvest its leaf and stem to dye textiles and clothing, worship the color blue and enjoy dressing in the indigo-dyed blue clothing every day (Figure 1a,b). Owing to the high demand of dyed clothing, the leaf and stem resources of *S. cusia* are common in local markets. In this study, the whole plant was transplanted to our medicinal garden and stored for approximately 2 years. Young leaves were collected and immediately frozen using liquid nitrogen. All samples were stored at −80°C for future use.

### Genome survey by Illumina sequencing

We predicted the genome size and complexity using the *k*-mer analysis. In short, high-quality genomic DNA was extracted using a Qiagen DNA purification kit (Qiagen, Darmstadt, Germany), and sequencing libraries were constructed following the manufacturer's protocol (Illumina, San Diego, CA, USA). The constructed libraries were subsequently sequenced on the Illumina HiSeq 4000 platform with paired-end 150 bp. Raw reads from Illumina sequencing were subjected to fastp for base quality control (Chen *et al.*, 2018). The genome size was estimated based on the *k*-mer distribution of the short reads following the empirical formula: $G = K\_num/peak\_depth$, where $G$ is the genome size, *K_num* is the total number of *k*-mers and *peak_depth* is the depth of the major peak. Meanwhile, the genome heterozygosity and repeat percentage were estimated. In addition, we collected the young leaves from *S. cusia* and immediately carried out the flow cytometry to determine genome size by using maize genome (B73 with a genome size 2300 Mb) as an internal standard.

### MinION sequencing

The genomic DNA was extracted from about 10 g of leaf. DNA quality and integrity was assessed by agarose gel electrophoresis, and the purity and concentration of DNA were determined on NanoDrop (Thermo Fisher Scientific, Wilmington, DE, USA) and Qubit (Thermo Fisher Scientific), respectively. High-molecular weight genomic DNA were selected using BluePippin DNA Size Selection System. The MinION sequencing library was then constructed using SQK-LSK109 kit according to the procedure by Oxford Nanopore Technologies (Oxford Science Park, Oxford, UK). Finally, the DNA library was added into the flow cell and performed on the platform of PromethION as per the manufacturer's protocol (Oxford Science Park).

### Transcriptome sequencing

To assist the gene prediction and dissect the molecular basis underlying the indigo biosynthesis in *S. cusia*, we performed transcriptome sequencing for different tissues from roots (R), stems at two developing stages (S1 and S2) and leaves at two developing stages (L1 and L4). Three biological replicates were used for each sample. The total RNA was extracted and purified according to the manufacturer's instructions for using the TRNzol reagent (Invitrogen). High-quality RNA was fragmented into 500 bp, which subsequently was used for the synthesis of double-stranded cDNAs. The cDNAs were then end-repaired and ligated with the sequencing adapter. Libraries were sequenced on the Illumina HiSeq 4000 platform with paired-end 150 bp following the manufacturer's protocol (Illumina).

Raw reads obtained from transcriptome sequencing were filtered out to remove the low-quality reads and adapter sequence. The clean data were mapped on to the genome using the HISAT2 package (Kim *et al.*, 2015) and were assembled using STRINGTIE (Pertea *et al.*, 2015). Gene expression levels were normalized to FPKM. DEGs between two samples were identified based on fold-change ($\log_2$(sample1/sample2) $\geq 1$) and adjusted ($P \leq 0.05$) using DESEQ software (Anders and Huber, 2010).

### Genome assembly

The sequenced MinION long reads were subjected into base calling by using MinKNOW, which can convert the MinION's electrical signals (FAST5 binary files) into a sequence of nucleotides (FASTQ files). FASTQ long reads were preprocessed to filter out clipped adapter sequences, low-quality reads and short reads with <2000 bp. Pass reads were used for genome assembly using different assemblers, briefly described as follows: (i) long reads were initially base-corrected by CANU software (Koren *et al.*, 2017) based on the falcon_sense method with the parameter CorrectedErrorRate = 0.025; (ii) corrected-reads from CANU were assembled using Caun and WTDDGB2 (https://github.com/ruanjue/wtdbg) to generate the draft genome, respectively; (iii) Quickmerge (available at https://github.com/mahulchak/quickmerge) was further used to merge different assemblies to obtain a more accurate assembly; and (iv) the draft assembly was polished twice to generate a final genome. The first-round polishing was subjected to Nanopolish (available at https://github.com/jts/nanopolish), and the second polishing was carried out by PILON ver. 1.13 using Illumina data (Walker *et al.*, 2014).

### Hi-C sequencing and chromosome-scale assembly

To assist the genome assembly into chromosome, young *S. cusia* leaves were collected for Hi-C sequencing, which could be used to probe the interacting regions of chromosomes within nuclei, based on paired-end reads sequencing. Briefly, fresh tissues were cross-linked with formaldehyde, and cross-linked DNA was then digested by the *Hin*dIII restriction enzyme. The sticky ends of these fragments were end-repaired, marked with biotin and then blunt-end ligation was performed in close proximity to generate circular molecules. Subsequently, these circular DNA molecules were fragmented into 300–500 bp, DNA ends were sheared, enriched by biotin pull-down and processed to paired-end sequencing (150-bp paired-end) on an Illumina HiSeq X Ten platform. Hi-C read pairs were submitted to LACHESIS software (Burton *et al.*, 2013) to refine the genome assembly, which would result in chromosome-length scaffolds. The completeness and accuracy of genome assembly were quantitatively assessed by the CEGMA (ver. 2.5) and BUSCO (ver. 3.0). Furthermore, genomic Illumina reads and the RNA-seq reads were remapped to the final genome assembly to assess the reads alignment rate.

### Genome annotation

To predict the repeat sequences and TEs, we adopted two main strategies: one *de novo*-based, and the homology-based method. We first produced a custom *de novo* repeat library using REPEATSCOUT ver. 1.0.5 (Price *et al.*, 2005) and LTR-FINDER ver. 1.05 (Xu and Wang, 2007) with default settings to identify TEs. We next employed PASTECLASSIFIER ver. 1.0 (Hoede *et al.*, 2014) to classify TEs in the *S. cusia* library. Then, the constructed TE library was subjected to WU-BLAST against the known Repbase database (ver.

19.06), based on the homology sequence alignment method. Finally, the genomic sequences were repeat-masked by REPEATMASKER ver. 4.0.6 (Tarailo-Graovac and Chen, 2009).

The repeat-masked genome was used for gene prediction. We performed protein-coding genes prediction through an integrative analysis of homology-based *ab initio*. We first adopted a RNA-seq-based gene prediction, where we aligned RNA-seq reads to the *S. cusia* genome and assembled them by using HISAT2 (Kim *et al.*, 2015) and STRINGTIE (Pertea *et al.*, 2015). For homology-based predictions, we used protein sequences from five sequenced plant species (including *A. thaliana*, *Oryza sativa*, *Solanum lycopersicum*, *Solanum melongena* and *Solanum tuberosum*) as queries to search against the *S. cusia* genome using tBLASTn with a significant E-value (1e-5). For the *ab initio*-based prediction, AUGUSTUS ver. 3.0.3 (Stanke *et al.*, 2004) and SNAP (Korf, 2004) were used for coding gene prediction. Finally, EVM ver. 1.1.1 (Haas *et al.*, 2008) was used to produce an integrated gene set. In addition, we predicted the pseudogenes in the *S. cusia* genome. We employed GENBLASTA (ver. 1.0.4, She *et al.*, 2009) to identify homologous gene sequences in a protein-coding-gene-masked genome and used GENEWISE (ver. 2.4.1, Birney *et al.*, 2004) to identify those gene sequences with premature stop codons or frameshift mutations as candidate pseudogenes.

The function annotations of genes predicted in the *S. cusia* genome were carried out by running BLAST v2.2.31 (E-value ≤1e-5) against public databases including NCBI non-redundant (Nr), Swiss-Prot, COG/KOG, TrEMBL and KEGG. Meanwhile, the GO term annotations of genes were performed using BLAST2GO software (Conesa *et al.*, 2005) against the GO database. Besides, the motifs and domains within genes were identified using the INTERPROSCAN ver. 5.36.75 (Quevillon *et al.*, 2005) tool by searching against the InterPro databases including PPROSITE, PRINTS, Pfam, ProDOM, SMART, TIGRFAMS, CATH-Gene3D and PANTHER.

For the non-coding RNA annotations, we employed TRNASCAN-SE ver. 1.3.1 (Chan and Lowe, 2019) to predict the transfer RNAs, and infernal ver. 1.1 (Nawrocki and Eddy, 2013) to detect the miRNAs, ribosomal RNAs, small nuclear RNAs and small nucleolar RNAs by searching the RFAM (ver. 12.1) and MIRBASE (ver. 21).

## Comparative genome analysis

To construct the phylogenetic tree and estimate the divergence time among plant species, we first downloaded protein sets from 12 sequenced plants: *O. sativa*, *Populus trichocarpa*, *A. thaliana*, *Glycine max*, *V. vinifera*, *S. lycopersicum*, *Utricularia gibba*, *A. paniculata*, *S. indicum*, *M. guttatus*, *Olea europaea* and *S. miltiorrhiza*. Orthologous genes between plant species were identified using ORTHOMCL ver. 2.0.9 (Li *et al.*, 2003) with the parameters: Pep_length = 10, Stop_coden = 20, PercentageMatchCutoff = 50, EvalueExponentCutoff = 1e-5 and mcl inflation factor = 1.5. Subsequently, multiple sequence alignments with high accuracy were performed based on the coding sequences from single-copy families using MUSCLE, and poorly conserved blocks were removed using Gblocks (Talavera and Castresana, 2007) with default parameters. After that, the processed alignments were merged into large supergenes. Phylogenetic tree was constructed using PHYML ver. 3.0 (Guindon *et al.*, 2010) based on the maximum-likelihood principle with the parameters: model = HKY85 and bootstrap = 1000. Divergence times between plant species were estimated using MCMCtree of PAML based on global and local clock models with the following parameters: burnin = 10 000, nsample = 100 000, sampfreq = 2. Speciation event dates for *O. sativa*/*S. cusia* (152–160 MYA) and *S. cusia*/*S. indicum* (37–64 MYA) obtained from the Timetree web service (http://www.timetree.org/) were further used to calibrate the divergence time.

Gene family expansion and contraction was performed using the CAFE ver. 4.2 (De Bie *et al.*, 2006) with the parameter lambda = 0.002 and the species tree mentioned above. CAFE adopted a random birth and death model to estimate the size of each family at each ancestral node based on a specified phylogenetic tree, and a family-wide *P* value was calculated for each of the gene families to test the significant expansion or contraction ($P < 0.01$). After that, the expanded or contracted gene family was further determined for specific branches and nodes using Viterbi $P < 0.05$ (De Bie *et al.*, 2006). The 4DTV (transversion rate on four-fold degenerated sites) for gene pairs detected by MCSCANX (Wang *et al.*, 2013) was calculated using the HKY substitution model. The synonymous mutation rate (Ks) of gene pairs detected by MCscan was calculated using the YN00 program of the PAML package (Yang, 2007).

## Determination of metabolite concentrations

We collected fresh tissues from the roots, stems and leaves at four developing stages, and detected the content of indican, indigo and indirubin in *S. cusia*. Briefly, these tissues were immediately frozen in liquid nitrogen and about 500 mg of material was extracted with 10 ml of *N,N*-dimethyl formamide in an ultrasonic bath (NS200-6U; Nissei Corporation, Nagoya, Japan) at 25°C for 30 min. Next, the solution was centrifuged for 5 min at 2518 × ***g***. The aqueous layer was filtered through a 0.45 μm filter membrane. For each tissue part, five replicate samples were prepared. The standard compounds (indican, indigo and indirubin) were dissolved with 10 ml of *N,N*-dimethyl formamide in an ultrasonic bath (NS200-6U; Nissei). Reference solutions were prepared with five different concentration references. The contents of these three compounds were determined by the high-performance liquid chromatography system (Agilent 1260; Infinity LC, Agilent Technologies Inc., Palo Alto, CA, USA).

## Gene family analysis of CYP, UGT and BGL

It is suggested that members from the gene family CYP, UGT and BGL are probably involved in the production of indican, indigo or indirubin; thus, we identified all the members for CYP, UGT and BGL at the genome-wide level in *S. cusia*. First, we downloaded the Arabidopsis protein sequences of CYP, UGT and BGL from the website (http://www.p450.kvl.dk/). These proteins were then designed as query sequences against the *S. cusia* protein database by BLAST+ software (ver. 2.2.24+; NCBI) with the E-value of le-5. Subsequently, we obtained protein sequences of CYP, UGT and BGL, respectively, from the *S. cusia* protein database based on the best hit protein ID by a custom Perl script. For further confirmation of the presence of the characterized domain, these sequences were then submitted to SMART (http://smart.embl.de/), and proteins without characterized domains were discarded. Finally, an unrooted phylogenetic tree was constructed using the neighbor-joining criteria in MEGA X with 10 000 bootstrap replicates (Kumar *et al.*, 2018) based on aligned protein sequences from Arabidopsis and *S. cusia*.

## qRT-PCR validation

To validate the expression level of several candidate genes, tissue samples including young leaf, mature leaf (corresponding to L1 and L4, respectively), stem (corresponding to S2) and root were collected, and five independent replicates of each tissue were used. Total RNA was isolated from each sample using the

RNAprep Pure Plant Kit (Tiangen,TIANGEN BiotechCo.,Ltd., Beijing, China) and was subsequently digested by the DNA enzyme. The cDNA was synthesized using the PrimeScript RT Reagent Kit (TaKaRa, Takara Biomedical Technology Co., Ltd., Beijing, China), and then was used for the qRT-PCR reaction using SYBR Green Master Mix (TaKaRa). The ACTIN2 gene was used as an internal reference. The primers used in this study are listed in Table S19.

## ACKNOWLEDGEMENTS

We thanks Service Center for Information Technology, Kunming Institute of Botany, Chinese Academy of Sciences for the help in the data storage and website development. This work was jointly supported by Strategic Priority Research Program of Chinese Academy of Sciences (grant no. XDA20050204), Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), National Natural Science Foundation of China (32000261) and Youth Innovation Promotion Association of CAS (2020389) to WX.

## AUTHOR CONTRIBUTIONS

AL, YW and WX conceived and designed the experiments; LZ, SL, HZ, ABC and YW collected samples and performed the experiments; XW analyzed the data; XW, ABC, YW and AL wrote the article.

## CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

## DATA AVAILABILITY STATEMENT

The *S. cusia* genome information including genome, genes, coding sequences and proteins sequences and related annotation information including genes, repeats, miRNAs and pseudogene are available at the public database http://indigoid-plant.iflora.cn. All clean sequenced data obtained from Illumina HiSeq, MinION, Hi-C and transcriptome sequencing mentioned above are deposited at the public database http://indigoid-plant.iflora.cn.

## SUPPORTING INFORMATION

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

**Figure S1**. Genome survey and size. (a) Assessment of genome size and complexity based on *K*-mers methods. (b) Estimate of genome size by flow cytometry by using *Zea Mays* genome (B73 with a genome size of 2300 Mb) as an internal standard.

**Figure S2**. Distribution of 4DTv distances between syntenic gene pairs among *Strobilanthes cusia*, *Andrographis paniculata*, *Solanum lycopersicum*, *Sesamum indicum* and *Vitis vinfera*.

**Figure S3**. The GO enrichment analysis for differentially expressed genes (DEGs) shared by leaves (L1 and L4) and stems (S1 and S2) compared with root (R). The number on the bar indicates the gene numbers in specific GO term, and the number in the parentheses indicates the adjusted *P* value.

**Figure S4**. Sequence alignment of *Strobilanthes cusia* UGT proteins. Heights of letters in the logo show the frequency of amino acids at that position.

**Table S1**. Summary of sequenced paired-end libraries for *Strobilanthes cusia* genome.

**Table S2**. Hi-C assisted assembly and chromosome length of *Strobilanthes cusia*.

**Table S3**. Genome and gene BUSCO assessment.

**Table S4**. The transposable elements and repeat sequences in *Strobilanthes cusia* genome.

**Table S5**. Prediction of protein-coding genes in *Strobilanthes cusia* genome

**Table S6**. Functional annotation of *Strobilanthes cusia* genome.

**Table S7**. Summary of gene families in related species.

**Table S8**. Pfam annotation of the expanded genes families.

**Table S9**. Expression level of genes (FPKM ≧0.5) in different tissues or developing stages in *Strobilanthes cusia*.

**Table S10**. Differentially expressed genes (DEGs) between leaf1 and root.

**Table S11**. Differentially expressed genes (DEGs) between leaf4 and root.

**Table S12**. Differentially expressed genes (DEGs) between stem1 and root.

**Table S13**. Differentially expressed genes (DEGs) between stem2 and root.

**Table S14**. Up- and downregulated genes shared by leaf and root relative to root in *Strobilanthes cusia*.

**Table S15**. Cytochrome P450 (CYP) identified in *Strobilanthes cusia* genome and their expression level (FPKM).

**Table S16**. Flavin-containing monooxygenase (FMO) identified in *Strobilanthes cusia* genome and their expression level (FPKM).

**Table S17**. UDP-glycosyltransferase (UGT) identified in *Strobilanthes cusia* genome and their expression level (FPKM).

**Table S18**. β-glucosidase (BGL) identified in *Strobilanthes cusia* genome and their expression level (FPKM).

**Table S19**. The information of primers used in this study.

## REFERENCES

**Anders, S. and Huber, W.** (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

**Andriamanantena, M., Danthu, P., Cardon, D., Fawbush, F.R., Raonizafinimanana, B., Razafintsalama, V.E., Rakotonandrasana, S.R., Ethève, A., Petit, T. and Caro, Y.** (2019) Malagasy dye plant species: a promising source of novel natural colorants with potential applications - a review. *Chem. Biodivers.* **16**, e1900442.

**Bak, S., Beisson, F., Bishop, G., Hamberger, B., Hofer, R., Paquette, S. and Werck-Reichhart, D.** (2011) Cytochrome P450. *Arabidopsis Book*, **9**, e0144.

**Balfour-Paul, J.** (2006) *Indigo*. London: Archetype Publications.

**Birney, E., Clamp, M. and Durbin, R.** (2004) GeneWise and genomewise. *Genome Res.* **14**, 988–995.

**Burkhill, I.H.** (1921) A note upon plants grown for blue dyes in the north of the Malay Penninsula. *Gard. Bull. Straits Settlem.* **2**, 426–429.

**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**, 11191.

**Cardon, D.** (2007) Cocaigne to cowboys: indigo plants, indigo blues. In *Natural Dyes Sources, Tradition, Technology and Science* (Cardon, D., ed). London: Archetype Publications Ltd, pp. 335–408.

**Chan, P.P. and Lowe, T.M.** (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14.

**Chanayath, N., Lhieochaiphant, S. and Phutrakul, S.** (2002) Pigment extraction techniques from the leaves of *Indigofera tinctoria* Linn. and *Baphicacanthus cusia* Brem. and chemical structure analysis of their major components. *Chiang Mai Univ. J.* **1**, 149–160.

**Chen, S., Zhou, Y., Chen, Y. and Gu, J.** (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

**Choi, H.S., Kim, J.K., Cho, E.H., Kim, Y.C., Kim, J.I. and Kim, S.W.** (2003) A novel flavin-containing monooxygenase from *Methylophaga* sp strain

SK1 and its indigo synthesis in *Escherichia coli*. *Biochem. Biophys. Res. Comm.* **306**, 930–936.

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

**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.

**Dutta, S., Roychoudhary, S. and Sarangi, B.K.** (2017) Effect of different physico-chemical parameters for natural indigo production during fermentation of Indigofera plant biomass. *3 Biotech*, **7**, 322.

**Gaut, B.S., Seymour, D.K., Liu, Q. and Zhou, Y.** (2018) Demography and its effects on genomic variation in crop domestication. *Nat. Plants*, **4**, 512–520.

**Gillam, E.M. and Guengerich, F.P.** (2001) Exploiting the versatility of human cytochrome P450 enzymes: the promise of blue roses from biotechnology. *IUBMB Life*, **52**, 271–277.

**Gillam, E.M., Notley, L.M., Cai, H., De Voss, J.J. and Guengerich, F.P.** (2000) Oxidation of indole by cytochrome P450 enzymes. *Biochemistry*, **39**, 13817–13824.

**Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O.** (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.

**Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J, White, O., Buell, C.R. and Wortman, J.R.** (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7.

**Hadders, M.** (1933) Systematische verbreitung und vorkommen der indoxylglucoside. In *Handbuch der Pflanzenanalyse* (Klein, G., ed). Wien: Springer, pp. 1062–1063.

**Han, J. and Quye, A.** (2018) Dyes and dyeing in the Ming and Qing Dynasties in China: preliminary evidence based on primary sources of documented recipes. *J. Text. Hist.* **49**, 44–70.

**Hartl, A., Proaño Gaibor, A.N., van Bommel, M.R. and Hofmann-de Keijzer, R.** (2015) Searching for blue: experiments with woad fermentation vats and an explanation of the colours through dye analysis. *J. Archaeol. Sci. Rep.* **2**, 9–39.

**Hellsten, U., Wright, K.M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S.R., Schmutz, J., Willis, J.H. and Rokhsar, D.S.** (2013) Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc. Natl Acad. Sci. USA*, **110**, 19478–19482.

**Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. and Quesneville, H.** (2014) PASTEC: an automatic transposable element classification tool. *PLoS One*, **9**, e91929.

**Hsu, T.M., Welner, D.H., Russ, Z.N., Cervantes, B., Prathuri, R.L., Adams, P.D. and Dueber, J.E.** (2018) Employing a biochemical protecting group for a sustainable indigo dyeing strategy. *Nat. Chem. Biol.* **14**, 256–261.

**Hu, Z., Tu, Y., Xia, Y. *et al.*** (2015) Rapid identification and verification of indirubin-containing medicinal plants. *Evid. Based Complement. Alternat. Med.* **2015**, 484670.

**Kang, M., Wu, H., Yang, Q., Huang, L., Hu, Q., Ma, T., Li, Z. and Liu, J.** (2020) A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine: An Isatis genome. *Hortic. Res.* **7**, 18.

**Kim, D., Langmead, B. and Salzberg, S.L.** (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357.

**Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M.** (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736.

**Korf, I.** (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

**Kramell, A., Li, X., Csuk, R., Wagner, M., Goslar, T., Tarasov, P.E., Kreusel, N., Kluge, R. and Wunderlich, C.-H.** (2014) Dyes of late Bronze Age textile clothes and accessories from the Yanghai archaeological site, Turfan, China: determination of the fibers, color analysis and dating. *Quatern. Int.* **348**, 214–223.

**Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K.** (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549.

**Laitonjam, W.S. and Wangkheirakpam, S.D.** (2011) Comparative study of the major components of the indigo dye obtained from *Strobilanthes flaccidifolius* Nees. and *Indigofera tinctoria* Linn. *Int. J. Plant Physiol. Biochem.* **3**, 108–116.

**Lee, C.L., Wang, C.M., Hu, H.C., Yen, H.-R., Song, Y.-C., Yu, S.-J., Chen, C.-J., Li, W.-C. and Wu, Y.-C.** (2019) Indole alkaloids indigodoles A-C from aerial parts of *Strobilanthes cusia* in the traditional Chinese medicine Qing Dai have anti-IL-17 properties. *Phytochemistry*, **162**, 39–46.

**Li, L., Stoeckert, C.J. Jr and Roo, D.S.** (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

**Li, S., Cunningham, A.B., Fan, R. and Wang, Y.** (2019) Identity blues: the ethnobotany of the indigo dyeing by Landian Yao (Iu Mien) in Yunnan, Southwest China. *J. Ethnobiol. Ethnomed.* **15**, 13.

**Liau, B.C., Jong, T.T., Lee, M.R. and Chen, S.S.** (2007) LC-APCI-MS method for detection and analysis of tryptanthrin, indigo, and indirubin in daqingye and banlangen. *J. Pharm. Biomed. Anal.* **43**, 34643.

**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**, e0212863.

**Lin, W., Huang, W., Ning, S., Wang, X., Ye, Q. and Wei, D.** (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**, e0199788.

**Lin, Y.K., See, L.C., Huang, Y.H., Chang, Y.-C., Tsou, T.-C., Lin, T.-Y. and Lin, N.-L.** (2014) Efficacy and safety of Indigo naturalis extract in oil (Lindioil) in treating nail psoriasis: a randomized, observer-blind, vehicle-controlled trial. *Phytomedicine*, **21**, 1015–1020.

**Muthu, S.S. and Gardetti, M.A.** (2016) *Green Fashion*. vol **2**, Singapore: Springer.

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

**Ni, Y., Song, R. and Kokot, S.** (2012) Discrimination of *Radix Isatidis* and *Rhizoma et Radix Baphicacanthis Cusia* samples by near infrared spectroscopy with the aid of chemometrics. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **96**, 252–258.

**Osmani, S.A., Bak, S. and Møller, B.L.** (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry*, **70**, 325–347.

**Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L.** (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290.

**Price, A.L., Jones, N.C. and Pevzner, P.A.** (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**(Suppl 1), i351–i358.

**Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R.** (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**(Web Server), W116–W120.

**She, R., Chu, J.S., Wang, K., Pei, J. and Chen, N.** (2009) GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149.

**Splitstoser, J.C., Dillehay, T.D., Wouters, J. and Claro, A.** (2016) Early pre-hispanic use of indigo blue in Peru. *Sci. Adv.* **2**, e1501623.

**Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B.** (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312.

**Sugimoto, S., Naganuma, M., Kiyohara, H. *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.

**Sun, W., Leng, L., Yin, Q. *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.

**Talavera, G. and Castresana, J.** (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577.

**Tarailo-Graovac, M. and Chen, N.** (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, Chapter 4, Unit, 4, 10. https://doi.org/10.1002/0471250953.bi0410s05

**Tayade, P.B. and Adivarekar, R.V.** (2014) Extraction of Indigo dye from *Couroupita guianensis* and its application on cotton fabric. *Fash. Text.* **1**, 16.

**Walker, B.J., Abeel, T., Shea, T. *et al.*** (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

**Wang, L., Yu, S., Tong, C. *et al.*** (2014) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* **15**, R39.

**Wang, Y., Li, J. and Paterson, A.H.** (2013) MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics*, **29**, 1458–1460.

**Warzecha, H., Frank, A., Peer, M., Gillam, E.M., Guengerich, F.P. and Unger, M.** (2017) Formation of the indigo precursor indican in genetically engineered tobacco plants and cell cultures. *Plant Biotechnol. J.* **5**, 185–191.

**Xu, Z., Escamilla-Treviño, L., Zeng, L. *et al.*** (2004) Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 1. *Plant Mol. Biol.* **55**, 343–367.

**Xu, Z. and Wang, H.** (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**(suppl_2), W265–W268.

**Yang, Z.** (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

**Yusuf, M. and Shahid-ul-islam**. (2017) Natural dyes from indigoid-rich plants: an overview. In *Plant-Based Natural Products* (Shahid-ul-islam, ed). New Jersey: Wiley, pp. 27–46.

**Zhang, G., Tian, Y., Zhang, J., Shu, L., Yang, S., Wang, W., Sheng, J., Dong, Y. and Chen, W.** (2015) Hybrid de novo genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *Gigascience*, **4**, 62.

**Zhang, L., Wang, L., Cunningham, A.B., Shi, Y. and Wang, Y.** (2019) Island blues: indigenous knowledge of indigo-yielding plant species used by Hainan Miao and Li dyers on Hainan Island, China. *J. Ethnobiol. Ethnomed.* **15**, 31.

**Zhang, X., Good, I. and Laursen, R.** (2008) Characterization of dyestuffs in ancient textiles from Xinjiang. *J. Archaeol. Sci.* **35**, 1095–1103.

**Zhao, X., He, X. and Zhong, X.** (2016) Anti-inflammatory and in-vitro antibacterial activities of Traditional Chinese Medicine Formula Qingdaisan. *BMC Complement Altern. Med.* **16**, 503.