Kang et al. *Horticulture Research* (2020)7:18
https://doi.org/10.1038/s41438-020-0240-5

Horticulture Research
www.nature.com/hortres

ARTICLE Open Access

A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine

An *Isatis* genome

Minghui Kang[1], Haolin Wu[1], Qiao Yang[1], Li Huang[1], Quanjun Hu[1], Tao Ma[1], Zaiyun Li[2,1] and Jianquan Liu[1,3]

Abstract

Isatis indigotica ($2n = 14$) is an important medicinal plant in China. Its dried leaves and roots (called Isatidis Folium and Isatidis Radix, respectively) are broadly used in traditional Chinese medicine for curing diseases caused by bacteria and viruses such as influenza and viral pneumonia. Various classes of compounds isolated from this species have been identified as effective ingredients. Previous studies based on transcriptomes revealed only a few candidate genes for the biosynthesis of these active compounds in this medicinal plant. Here, we report a high-quality chromosome-scale genome assembly of *I. indigotica* with a total size of 293.88 Mb and scaffold N50 = 36.16 Mb using single-molecule real-time long reads and high-throughput chromosome conformation capture techniques. We annotated 30,323 high-confidence protein-coding genes. Based on homolog searching and functional annotations, we identified many candidate genes involved in the biosynthesis of main active components such as indoles, terpenoids, and phenylpropanoids. In addition, we found that some key enzyme-coding gene families related to the biosynthesis of these components were expanded due to tandem duplications, which likely drove the production of these major active compounds and explained why *I. indigotica* has excellent antibacterial and antiviral activities. Our results highlighted the importance of genome sequencing in identifying candidate genes for metabolite synthesis in medicinal plants.

Introduction

The plant family Brassicaceae (Cruciferae) comprises over 330 genera and ~3700 species with a worldwide distribution[1–5]. Numerous crops are derived from this family, including vegetables (*Brassica* and *Raphanus*), ornamentals (*Matthiola*, *Hesperis*, and *Lobularia*), spices (*Eutrema* and *Armoracia*), and medicines (*Isatis*). Based on sequenced genomes, several model species have been developed for diverse studies, including *Arabidopsis thaliana* for molecular function studies, *Brassica* for polyploidization and whole-genome duplication (WGD) studies, and *Eutrema salsugineum* for abiotic tolerance-related studies. However, genetic biosynthesis of the major active compounds in medicinal plants of this family remains poorly investigated.

Isatis indigotica ($2n = 14$) belongs to tribe Isatideae in lineage II of the family[3,6–10]. This species is widely cultivated in China as an important medicinal plant because its dried leaves and roots are used as a traditional Chinese medicine for curing diseases and viruses[11–13]. The major active compounds isolated from this species comprise terpenoids, lignans, and indole alkaloids[14–17]. These

Correspondence: Zaiyun Li (lizaiyun@mail.hzau.edu.cn) or
Jianquan Liu (liujq@nwipb.ac.cn)
[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education & State Key Laboratory of Hydraulics & Mountain River Engineering, College of Life Sciences, Sichuan University, Chengdu 610065, China
[2]National Key Laboratory of Crop Genetic Improvement, National Center of Oil Crop Improvement (Wuhan), College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China
Full list of author information is available at the end of the article.

Table 1 Statistics for the final genome assembly of *I. indigotica*

	I. indigotica genome (PacBio + Hi-C)
Sequencing platform	PacBio Sequel
Assembly size (bp)	293,875,465
GC %	38.18
Number of scaffolds	810
Scaffold N50 size (bp)	36,165,591
Scaffold N90 size (bp)	87,397
Number of contigs	1199
Contig N50 size (bp)	1,176,212
Contig N90 size (bp)	75,736
Gap %	0.01
Longest sequence length (bp)	38,253,781

compounds were confirmed to have antiviral[18,19], antibacterial[20], anti-inflammatory[21,22], and antileukemia[23,24] functions. Previous studies based on transcriptomes revealed a few candidate genes involved in the biosynthesis of active compounds in this species[25–28]. However, the limitations of transcriptome quality and integrity hinder the identification of all candidate biosynthesis-related genes.

In the present study, we used single-molecule sequencing combined with high-throughput chromosome conformation capture (Hi-C) technology to assemble the genome and construct the pseudochromosomes of *I. indigotica*. Based on homolog searching and functional annotations, we aimed to identify candidate gene sets involved in the biosynthesis of putative active components. The candidate genes and genomic resources recovered here will be critically important for further experimental verification and artificial syntheses of the active compounds of this medicinal plant in the future.

Results

Genome assembly and construction of pseudochromosomes

The genome size, genome repeat size, and heterozygosity rate of *I. indigotica* were estimated using K-mer analysis. The 19-mer frequency of Illumina short reads with the highest peak occurred at a depth of 94. The genome was estimated to be 279.90 Mb in size with 48.99% repeats, and the heterozygosity rate was estimated to be 0.44% (Supplementary Table S3 and Supplementary Fig. S3). In addition, the genome size of *I. indigotica* was estimated to be ~305 Mb based on flow cytometric analyses using *Vigna radiata* as the internal standard (Supplementary Fig. S2).

We sequenced and assembled the genome of *I. indigotica* using single-molecule real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio) and anchored the assembled contigs to seven pseudochromosomes using Hi-C techniques. The final chromosome-scale genome was 293.88 Mb in length with 1199 contigs (contig N50 = 1.18 Mb), a scaffold N50 = 36.17 Mb, and a maximum pseudochromosome length of 38.25 Mb (Table 1, Supplementary Table S4, and Supplementary Fig. S4).

The completeness of the genome assembly was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO)[29]. Of the 1440 plant-specific orthologs, 1416 (98.33%) were identified in the assembly, of which 1400 (97.22%) were considered to be complete (Supplementary Table S5). The assembly base accuracy was also assessed based on Illumina short read mapping. In total, 99.97% of the clean reads were mapped to the genome assembly, and 94.55% of them were properly mapped (Supplementary Table S6). The base error percentage of the genome assembly was estimated to be 0.000081% (Supplementary Table S7). All these evaluations indicate the high completeness, high continuity, and high base accuracy of the present genome assembly.

Repeat and gene annotations

Repetitive sequences were identified using a combination of ab initio and homology-based approaches. In total, we identified 53.27% of the assembled sequences as repetitive sequences, including 34.67% retrotransposons and 7.37% DNA transposons. Long terminal repeat (LTR) retrotransposons were found to account for 30.09% of the genome (Supplementary Table S8). We annotated protein-coding genes by combining transcriptome-based, homology-based, and ab initio predictions. Finally, we predicted a total of 30,323 genes, of which 5973 had alternatively spliced transcripts. The average transcript length and coding sequence size were 2693 and 1387 bp, respectively, with a mean of 5.50 exons and 1.39 transcripts per gene (Table 2). Overall, 29,522 genes (97.36%) were assigned functions, and 76.16% and 91.69% of these genes had homologies and annotated proteins in the Swiss-Prot and TrEMBL databases. Further functional annotations using InterProScan estimated that 95.86% of the genes contained conserved protein domains, and 87.32% of the genes were classified by Gene Ontology (GO) terms, with 29.41% mapped to known plant biological pathways based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Supplementary Table S9).

Chromosome structure of *I. indigotica*

Evolution of chromosome structures in Brassicaceae has been traced and established through comparative chromosome painting techniques using BAC probes of the *A. thaliana* genome[4,30]. Using these techniques, Lysak and

Table 2 Statistics of predicted protein-coding genes in the *I. indigotica* genome

	I. indigotica genome
Number of protein-coding genes	30,323
Number of transcripts	42,061
Average transcript length (bp)	2693.27
Average exon length (bp)	252.24
Average intron length (bp)	215.32
Average number of exons per gene	5.50
Average exon length per gene (bp)	1387.32

his colleagues[31] proposed a model comprising eight chromosomes and 24 genomic blocks (GBs, named from A to X) for comparative genomics and chromosomal analyses based on the Ancestral Crucifer Karyotype (ACK; $n = 8$) concept[31]. This model was further updated to comprise 22 GBs by merging the four GBs to form two new blocks (K-L and M-N), and the GB boundaries defined by *A. thaliana* gene loci were also updated[32] (Fig. 1a). Six tribes (Calepineae, Coluteocarpeae, Conringieae, Eutremeae, Isatideae, and Sisymbrieae) of expanded lineage II were found to derive from a common ancestor with the Proto-Calepineae Karyotype (PCK; $n = 7$). Among these tribes, three (Eutremeae, Isatideae, and Sisymbrieae) displayed an additional whole-arm translocation in the second and seventh chromosomes (translocation PCK, tPCK; $n = 7$)[32–34] (Fig. 1a). ACK and PCK shared five similar chromosomes. Thus, they might descend from a common ancestor; alternatively, PCK may have evolved from ACK.

To determine whether the *I. indigotica* genome sequence also supported tPCK structure in Isatideae, we compared the seven pseudochromosomes of *I. indigotica* with the *A. thaliana* genome by LAST and MCScanX. We determined syntenic relationships and constructed the order and orientation of the 22 GBs along the seven pseudochromosomes of the *I. indigotica* genome (Supplementary Figs. S5, S6). Based on the gene intervals of each GB of *A. thaliana*, we determined the corresponding intervals and boundaries of each block in *I. indigotica* and renamed the pseudochromosomes based on Fig. 1a (Fig. 1b and Supplementary Table S10). The *I. indigotica* genome has good collinearity in each GB compared with the *A. thaliana* genome and is consistent with tPCK structure in both order and orientation (Fig. 1 and Supplementary Figs. S5, S6). Furthermore, we carried out sequence alignments between the genomes of *I. indigotica* and the other three species that might also display tPCK structure (*Sisymbrium irio* for Sisymbrieae, *E. salsugineum* for Eutremeae, and *Schrenkiella parvula* for unassigned genera) using LAST. Our analyses suggested that these four species have similar chromosome structures (Supplementary Figs. S7–S9). However, we found obvious inversions in the *S. parvula* genome and low continuity of sequences in the *E. salsugineum* and *S. irio* genomes. These comparisons suggest that the present *I. indigotica* genome was better assembled in terms of both accuracy and continuity than others with tPCK structure.

Phylogenetic relationships and WGD analyses

We clustered the annotated genes into gene families among *I. indigotica* and eight other Brassicaceae species with *Cleome hassleriana* as the outgroup. A total of 24,382 *I. indigotica* genes (80.41%) clustered into 18,900 gene families, of which 10,826 (57.28%) gene families were shared with nine other species and 896 (4.74%) were *I. indigotica* specific (Fig. 2c and Supplementary Table S11). We selected 822 single-copy gene families among 10 species to construct a phylogenetic tree, which showed that *I. indigotica* was sister to *S. irio*. We further estimated the divergence time between them as 15.86 (12.71–19.20) million years ago (Mya) (Fig. 2a). The relationships of all 10 species are consistent with those from previous phylogenetic analyses[3,6–8,10].

Then, we used synonymous substitution rates (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 with time. A density plot of Ks values for the collinear gene pairs suggested that *I. indigotica* experienced a recent WGD event with a peak value of ~0.76, consistent with At-α-WGD for all Brassicaceae species[8,35,36]. An independent WGD event was identified for *B. rapa* after its divergence from *I. indigotica* at Ks = 0.30–0.34, previously reported as a Brassiceae-specific triplication (Br-α-WGD)[8,37–39] (Fig. 2b). Whole-genome alignment among the *I. indigotica*, *A. thaliana*, and *B. rapa* genomes carried out by LAST also confirmed the collinear relationship and these WGD events. For each genomic region of *I. indigotica*, we typically found one matching region in *A. thaliana* and three matching regions in *B. rapa*. These comparisons suggest that *I. indigotica* did not experience an independent WGD event after At-α-WGD (Supplementary Figs. S5, S10).

The expansion and contraction of gene families play critical roles in driving phenotypic diversification and enhancing special traits in plants. We discovered 1357 expanded and 3074 contracted gene families in *I. indigotica* relative to *S. irio* (Fig. 2a). Tandem duplication was the main contributor to the gene family expansions. GO enrichment analysis of tandem repeat genes suggested that they were enriched in defense response to virus, indole biosynthetic process, lignin biosynthetic process,

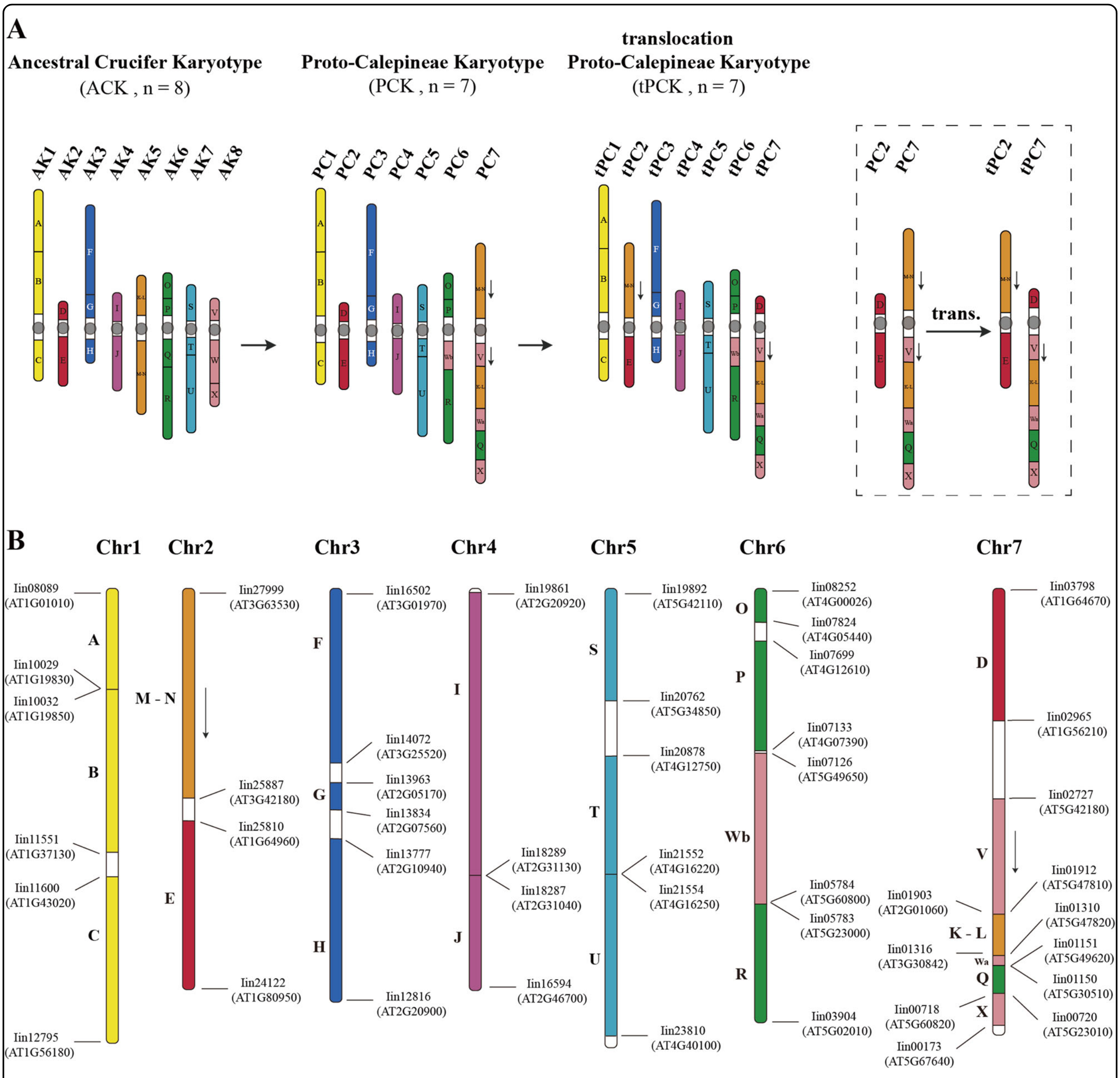

Fig. 1 Ancestral Brassicaceae genomes and the distribution of ancestral GBs along the seven pseudochromosomes of *I. indigotica*. a The ancestral genomes ACK, PCK, and tPCK, each comprising 22 ancestral GBs. Blocks with opposite orientations relative to that of ACK are represented by downward-pointing arrows. M-N and D GBs are translocated in tPCK chromosomes tPC2 and tPC7 compared with PCK chromosomes PC2 and PC7. The structures were drawn based on previous studies[32–34]. **b** Twenty-two GBs and their positions within the *I. indigotica* genome. Genes of each GB boundary are shown beside the chromosomes with the corresponding *A. thaliana* locus IDs within parentheses based on MCScanX results. The *A. thaliana* GB boundaries were derived from a previous study[34]

flavone synthase activity, and glucosyltransferase activity, some of which might be involved in the biosynthesis of active compounds in *I. indigotica* (Supplementary Table S12). We also performed GO enrichment analysis of the contracted gene families, and the results showed that they were enriched in proton export across plasma membrane, proton-exporting ATPase activity, regulation of stomatal movement, and defense response to other organism (Supplementary Table S13), which are probably related to the environmental adaptation of the species.

Identification of genes involved in the biosynthetic pathways of active compounds

Based on the KEGG database, GO classification, and the suggested biosynthesis pathways, we used a combined method of homolog searching and functional annotation

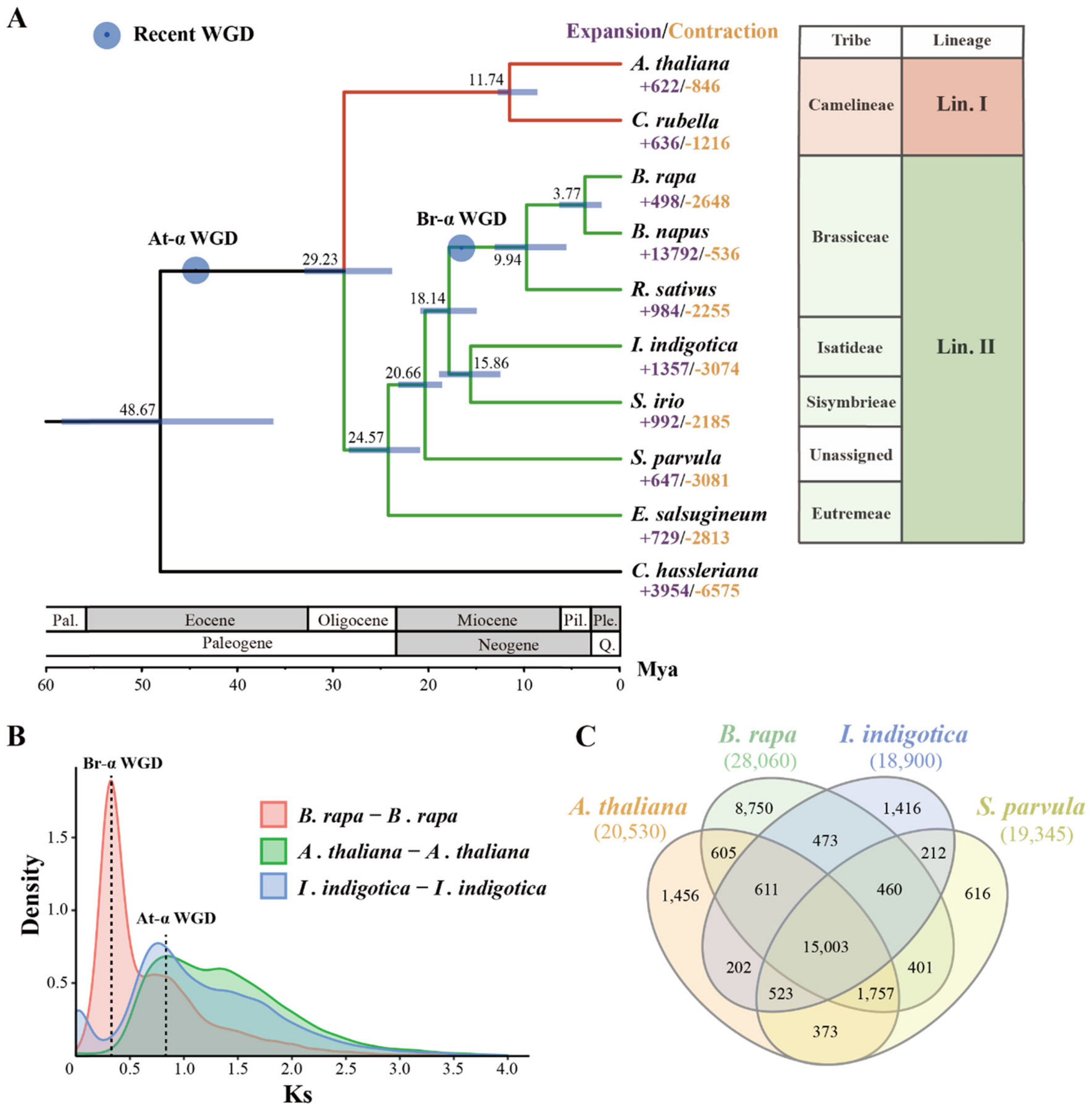

Fig. 2 Evolutionary and comparative genomic analyses. a The phylogenetic tree of *I. indigotica* and eight other Brassicaceae species with *C. hassleriana* as the outgroup. All branch bootstrap values are 100. Gene family expansions are indicated in purple, while gene family contractions are indicated in light brown. The estimated divergence times (million years ago, Mya) are indicated at each node; bars are the 95% highest probability densities (HPDs). Circles in blue represent recent whole-genome duplication events. **b** Ks value distributions between *B. rapa*, *I. indigotica*, and *A. thaliana*. **c** Orthogroups shared by selected species

to identify candidate genes for the biosynthesis of three types of active compounds, namely, terpenoids, phenylpropanoids, and indoles, in *I. indigotica*[14–16,25,40,41]. Sterols are the major terpenoids in *I. indigotica*, mainly comprising β-sitosterol and daucosterol[42]. β-Sitosterol was reported to play a critical role in curing lung inflammation[43], while daucosterol can inhibit cancer cell proliferation[44]. A total of 59 genes in the present genome, which encoded 31 enzymes, were identified to be involved in terpenoid and sterol biosynthesis (Supplementary Table S14). Based on the functional annotations of these genes, the biosynthesis pathway of β-sitosterol is nearly complete and daucosterol can be further synthesized from β-sitosterol by glucosyltransferases (Fig. 3a). In addition, the intermediate product geranyl diphosphate can be used not only to synthesize sterols but also to produce secologanin for monoterpene indole alkaloids in numerous medicinal plants such as *Catharanthus roseus*[45]. However, we annotated genes only with geraniol 10-hydroxylase activity (GO: 0102811). The lack of other related genes may account for the absence of secologanin and other related monoterpene indole alkaloids in *I. indigotica*.

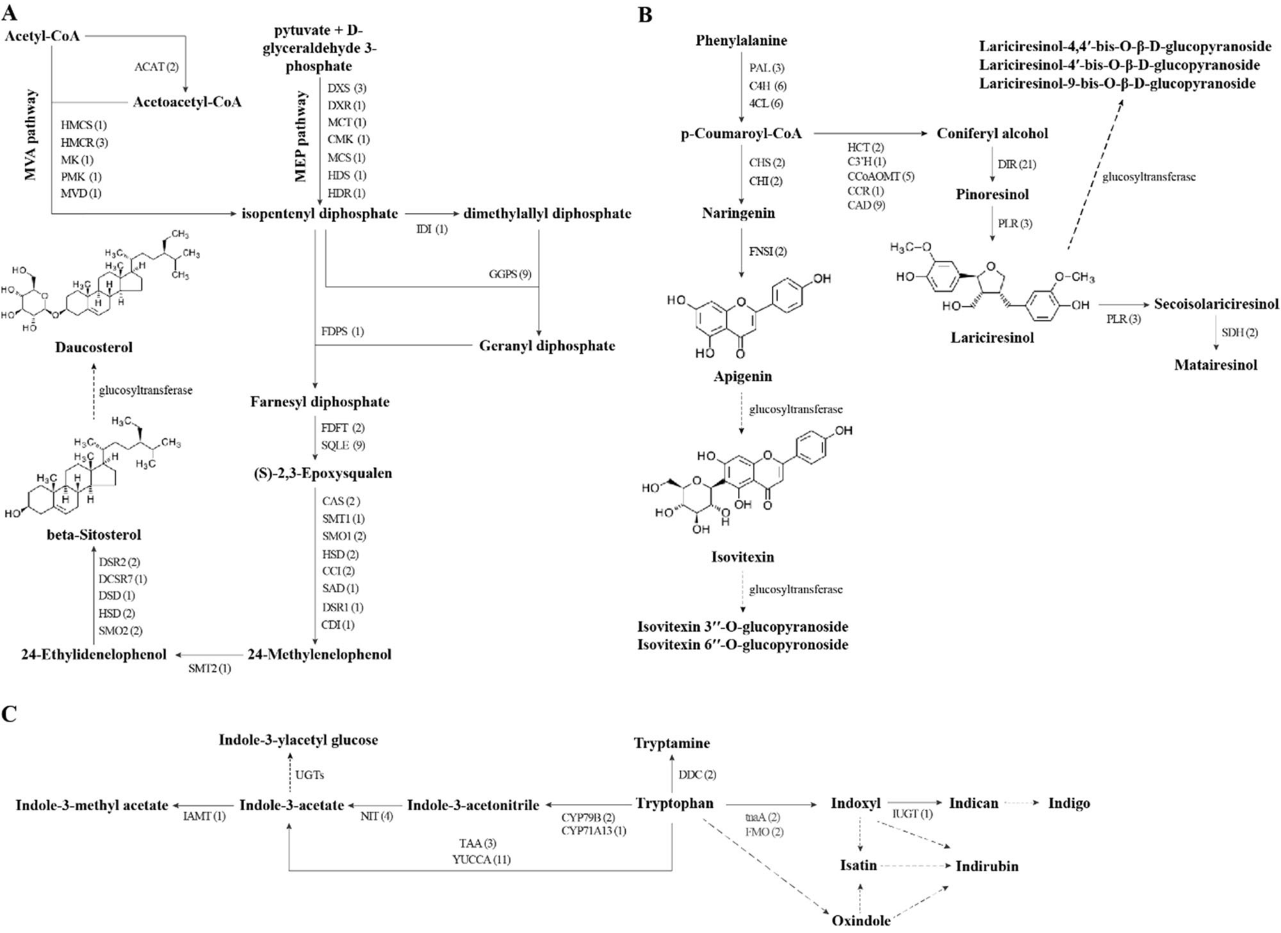

Fig. 3 Putative biosynthetic pathways of three main class active compounds in *I. indigotica*. The putative biosynthetic pathways of terpenoids (**a**), phenylpropanoids (**b**), and indole alkaloids (**c**) of active compounds in *I. indigotica*. Values within brackets indicate the numbers of gene copies corresponding to the catalytic genes in the pathways

Phenylpropanoids comprise lignans and flavonoids, with critical roles in anti-inflammatory response, anti-oxidant activity[46], and attenuation of mammary tumor growth[47]. We annotated 66 genes involved in the biosynthesis of lignans and flavonoids (Supplementary Table S15). The identified putative pathway mainly comprises the biosynthesis of isovitexin and lariciresinol, while their glycosides were further synthesized by glucosyltransferases (Fig. 3b). Indole alkaloids comprise another active component of *I. indigotica*[48–50] with important anti-influenza, anti-inflammatory, and leukocyte inhibition effects[11,23,51–53]. Based on the KEGG maps and previously suggested pathways[25,54,55], we identified 32 genes that encoded 11 enzymes involved in the biosynthesis of indole alkaloids (Fig. 3c and Supplementary Table S16). Because of the lack of downstream pathways, other genes for indole alkaloid biosynthesis in *I. indigotica* need further identification. It should be noted that numerous genes involved in the biosynthesis of the three major types of active compounds increased in copy number because of tandem duplication, for example, geranylgeranyl diphosphate synthase, cinnamate 4-hydroxylase, 4-coumarate-CoA ligase, and indole-3-pyruvate monooxygenase (Fig. 3 and Supplementary Tables S14–S16).

Discussion

Continuity and completeness are important indicators of genome assembly. PacBio-based genome assembly plus error corrections based on Illumina data could greatly improve continuity and completeness[56–59]. Our genome assembly of *I. indigotica* by this strategy showed a highly resolved result with an N50 = 1.22 Mb and longest contig length = 8.99 Mb. In addition, we used Hi-C data to cluster the contigs into seven pseudochromosomes with a final scaffold N50 = 36.17 Mb and longest chromosome length = 38.25 Mb. The completeness and high quality of the present *I. indigotica* genome were further confirmed by BUSCO and comparative chromosome analyses[32]. A total of 97.22% of the genes examined by BUSCO were complete, and the chromosome structure of *I. indigotica* was consistent with the tPCK type.

We constructed the phylogenetic relationships of *I. indigotica* based on genomic data and found that *I. indigotica* of Isatideae is sister to *S. irio* of Sisymbrieae among the sampled species, consistent with the results of previously published phylogenetic analyses[3,6–8,10]. Based on the phylogenetic results, we identified expanded and contracted gene families in *I. indigotica*. The expanded genes in this species were mainly derived from tandem duplications and were obviously enriched in some secondary metabolite pathways. Based on homolog searching and functional annotation in our high-quality genome, we further identified candidate genes for the biosynthesis of three main classes of active compounds in *I. indigotica*: terpenoids, phenylpropanoids, and indole alkaloids. These candidate genes complete or replenish gene sets for biosynthetic pathways of these compounds concentrated in *I. indigotica*[25–28] (Fig. 3). In addition, we found that in some synthesis steps, the copy number of enzyme-coding genes increased to two or more because of tandem duplications. The increase in copy number may drive the production of major active compounds in *I. indigotica* and account for its excellent antibacterial and antiviral activities because gene expansions are responsible for enhancing a special trait or the origin of a new trait[60–62].

Overall, in this study, we present a high-quality genome for *I. indigotica*. We further identify or replenish candidate genes for biosynthesis pathways of the active compounds in this medicinal plant. These genes and genomic resources will provide a solid basis for future biosynthesis-related studies.

Materials and methods

DNA extraction and genome sequencing

We initially extracted high-quality total DNA from fresh young leaves of a 2-month-old plant artificially cultivated in the greenhouse using the cetyltrimethylammonium bromide method. We used a SMRTbell Template Prep Kit 1.0 (PacBio, Menlo Park, CA, USA) to construct the DNA libraries for PacBio long-read sequencing and sequenced them on a PacBio Sequel system. We obtained a total of four SMRT cells with 39.94 Gb of sequencing data (coverage of 142.71×) from the PacBio Sequel platform and generated a total of 4.30 million subreads with an N50 read length of 14.9 kb (Supplementary Table S1 and Supplementary Fig. S1). We also prepared paired-end Illumina libraries using an Illumina Genomic DNA Sample Preparation Kit and sequenced them on an Illumina HiSeq X Ten system for error correction and K-mer analysis and generated a total of 37.50 Gb of data and 31.79 Gb of clean data (Supplementary Table S1).

Genome assembly and pseudochromosome construction

We initially estimated the genome size of *I. indigotica* by flow cytometry with *Vigna radiata* as the reference[63]. We then used clean Illumina short reads to calculate K-mers (Illumina DNA short read size of 19 bp) by Jellyfish v.2.2.9[64] to confirm the genome size. The sequencing depth was estimated by determining the highest peak value of the frequency curve of the K-mer occurrence distribution. We used SMRT Link pipeline v.5.1.0.26412 to process the polymerase reads into subreads with readScore = 0.75 and minSubReadLength = 500 and used Canu v.1.6[65] to correct errors of the PacBio subreads and assemble the corrected reads into contigs after trimming low-quality bases using WTDBG (https://github.com/ruanjue/wtdbg). We corrected the assembled contigs by using 270 bp PE Illumina data by Pilon v.1.13[66] and finally obtained a 293.83 Mb contig-scale assembly with a contig N50 of 1.22 Mb. The genome contained 1162 contigs, and the longest contig was 8.99 Mb with a 38.18% GC content. These contigs were further anchored to chromosomes by the Hi-C technique.

We grounded ~3 g of fresh young leaf tissue into powder in liquid nitrogen for Hi-C experiments and constructed a Hi-C library following Louwers et al.[67] with chromatin extraction and digestion and DNA ligation, purification, and fragmentation. Finally, we obtained a total of 79.43 Gb of clean reads for Hi-C analyses by the Illumina HiSeq X Ten platform. We first carried out a preliminary assembly by splitting contigs into segments of 100 kb on average and mapping the Hi-C data to the contigs using BWA v.0.7.10-r789[68] in order to correct contig errors. We then used LACHESIS software[69] with the parameters CLUSTER MIN RE SITES = 22, CLUSTER MAX LINK DENSITY = 2, CLUSTER NON-INFORMATIVE RATIO = 2, ORDER MIN N RES IN TRUN = 10, and ORDER MIN N RES IN SHREDS = 10 to cluster and reorder all corrected contigs into pseudochromosomes. We finally adjusted the order and direction of the contigs on the pseudochromosomes by examining their interactions in the Hi-C heatmap. We evaluated the completeness and quality of the final assembled genome through BUSCO v.3.0[29] tests using gene content from the Embryophyta_odb9 database[29].

Repeat annotation

We identified repetitive elements through both RepeatModeler v.1.0.10 and RepeatMasker v.4.0.7[70,71]. RepeatModeler employed RECON and RepeatScout to predict interspersed repeats and then obtained the consensus repeat library. RepeatMasker recovered the repeats in the *I. indigotica* genome through a homology-based repeat search using the ab initio repeat database and Repbase. The overlapping repeats belonging to the same repeat class were combined according to their coordination in the genome. The overlapping repeats belonging to different repeat classes were then split into different types.

Gene prediction and functional annotation

To improve gene prediction, we further obtained transcriptomes by sequencing high-quality RNA from mixed fresh leaf, flower, and stem tissues and sequenced them by the Illumina HiSeq X Ten platform. We removed adapters and discarded reads with >10% N bases or reads having more than 20% bases of low quality (below 5) using NGS QC Toolkit v.2.3.3[72] and finally generated 19.87 Gb of clean data. We assembled the de novo and genome-guided transcriptomes with clean reads by Trinity v.2.4.0[73]. We also mapped the RNA-sequencing (RNA-seq) reads to the assembled genome to obtain the mapping rate through HISAT2 v.2.1.0[74] to evaluate the completeness of the genome.

We run PASA pipeline v.2.1.0[75] to align the transcripts to the assembled genome to carry out ORF prediction and gene prediction. To train the HMM model for Augustus, we extracted complete, multiexon genes, removed redundant high-identity genes (cut-off all-to-all identity of 70%), and finally generated the best candidate and low-identity gene models for training. We aligned the RNA-seq data to the hard-masked genome assembly by HISAT2[74] and used bam2hints in Augustus to generate the intron hint file. We used this hint file to carry out ab initio gene prediction by Augustus v.3.2.2[76]. For homologous prediction, the reference protein sequences of *Brassica rapa*, *Brassica napus*, *Raphanus sativus*, *Brassica juncea*, and *Brassica nigra* were downloaded and aligned against the *I. indigotica* genome using TBLASTN v.2.2.31[77] and searched with an *e* value of 1e^{-5}. After filtering low-quality results, gene structure was predicted using GeneWise v.2.4.1[78]. We combined the results from PASA, Augustus and GeneWise to generate the final protein-coding gene set using EVidenceModeler v.1.1.1[75]. To obtain the untranslated regions and alternatively spliced isoforms, we used PASA to update the gff3 file for two rounds and obtain the final gene models.

We annotated the functions of the predicated genes against public databases by NCBI BLAST+ v.2.2.31[77] with a cut-off *e* value of 1e^{-5} and maximum number of target sequences of 20, including the Swiss-Prot and TrEMBL databases[79]. Best-hit BLAST results were then used to define gene functions. We used InterProScan v.5.25-64.0[80] to identify motifs and domains by matching against public databases. We identified GO annotations by using Blast2GO v.4.1[81] according to the blast results and combined them with InterPro GO entries. We mapped the existing GO terms to enzyme codes by Blast2GO and submitted the predicted proteins to the KEGG (Kyoto Encyclopedia of Genes and Genomes) Automatic Annotation Server (KAAS)[82] to obtain KO numbers for KEGG pathway annotation.

Gene family and phylogenetic analyses

We used protein sequences of *I. indigotica* and eight other Brassicaceae species (*Arabidopsis thaliana*, *Capsella rubella*, *Brassica rapa*, *Brassica napus*, *Raphanus sativus*, *Schrenkiella parvula*, *Sisymbrium irio*, and *Eutrema salsugineum*) with the outgroup species *Cleome hassleriana* for same-family gene clustering. For genes with alternative splicing variants, the longest transcript was selected to represent the gene. Similarities between sequence pairs were calculated using BLASTP v.2.2.31[77] with a cut-off *e* value of 1e^{-5}. Additionally, OrthoMCL v.2.0.9 was used with default parameters to assess gene family membership based on overall gene similarity combined with Markov Chain Clustering (MCL) v.14-137[83].

We extracted single-copy orthologous genes from the ten species by OrthoMCL and aligned the resulting protein sequences by MAFFT v.7.313[84]. Then, we used Gblocks v.0.91b[85] to extract the conserved sites of multiple sequence alignments and constructed a phylogenetic tree by RAxML v.8.2.11[86]. We used *C. hassleriana* as an outgroup and performed 1000 bootstrap analyses to test the robustness of each branch. We used the Bayesian relaxed molecular clock approach in MCMCTREE of PAML v.4.9e[87] to estimate divergence time. We calibrated this tree based on the estimated divergence times in the TimeTree database[88] for *C. hassleriana–A. thaliana* (35–59 Mya), *A. thaliana–C. rubella* (7.4–12.8 Mya), *B. rapa–S. parvula* (19.3–28.6 Mya), and *B. rapa–A. thaliana* (23.4–33.5 Mya).

Gene families that had undergone expansion or contraction were identified in the eight sequenced species using CAFE[89]. The CAFE parameters included a *p* value threshold = 0.05 and automatic searching for the λ value. The algorithm in CAFE takes a matrix of gene family sizes in extant species as input and uses a probabilistic graphical model to ascertain the rate and direction of changes in gene family size across a given phylogenetic tree.

WGD analysis and identification of tandemly repeated genes

To examine WGD in *I. indigotica* and *B. rapa*, we extracted all homologous proteins between these two species and *A. thaliana* using an all-to-all search in BLASTP v.2.2.31[77] with an *e* value cut-off of 1e^{-9}. We used MCScanX[90] with default parameters to identify collinear blocks, each containing at least five collinear gene pairs. To infer WGD events, we used the downstream MCScanX script add_ka_and_ks_to_collinearity.pl to calculate the Ks values between collinear genes among these three genomes. We further performed whole-genome alignment of the three species by LAST v.946[91] and constructed a dot plot by the downstream program last-dotplot.

Identification of tandem repeat genes in the *I. indigotica* genome was based on three criteria: (1) two or more genes had more than 70% identity and 70% coverage according to BLASTP; (2) the pairwise gene distance was <100 kb; and (3) there were no more than 10 genes lying between the repeat genes on a single scaffold[92]. The genes identified in this way were subjected to functional analysis using GO enrichment.

Acknowledgements
This work was supported by the National Key Research and Development Program of China (2017YFC0505203), National Natural Science Foundation of China (31590821 and 31561123001), Fundamental Research Funds for the Central Universities (2018CDDY-S02-SCU and SCU2019D013), National High-Level Talents Special Support Plan (10 Thousand of People Plan), and 985 and 211 Projects of Sichuan University.

Author details
[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education & State Key Laboratory of Hydraulics & Mountain River Engineering, College of Life Sciences, Sichuan University, Chengdu 610065, China. [2]National Key Laboratory of Crop Genetic Improvement, National Center of Oil Crop Improvement (Wuhan), College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China. [3]State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology, Lanzhou University, Lanzhou 730000, China

Data availability
The PacBio long reads and Illumina short reads were uploaded to the NCBI SRA database under BioProject PRJNA549758. The final chromosome-scale genome assembly was submitted to the NCBI with accession number VHIU00000000. The genome fasta and gff3 files were uploaded to Figshare.

Conflict of interest
The authors declare that they have no conflict of interest.

Supplementary Information accompanies this paper at (https://doi.org/10.1038/s41438-020-0240-5).

Received: 16 July 2019 Revised: 24 December 2019 Accepted: 2 January 2020
Published online: 01 February 2020

References

1. Rollins, R. C. *The Cruciferae of Continental North America: Systematics of the Mustard Family from the Arctic to Panama* (Stanford University Press, 1993).
2. Al-Shehbaz, I. A. & Mummenhoff, K. Transfer of the South African genera Brachycarpaea, Cycloptychis, Schlechteria, Silicularia, and Thlaspeocarpa to Heliophila (Brassicaceae). *Novon* **15**, 385–389 (2005).
3. Al-Shehbaz, I., Beilstein, M. A. & Kellogg, E. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol.* **259**, 89–120 (2006).
4. Lysak, M. A. & Koch, M. A. Phylogeny, genome, and karyotype evolution of crucifers (Brassicaceae). In Schmidt, R. & Bancroft, I. (eds) *Genetics and Genomics of the Brassicaceae* 1–31 (Springer, New York, New York, USA, 2011).
5. Warwick, S. I., Francis, A. & Al-Shehbaz, I. A. Brassicaceae: species checklist and database on CD-Rom. *Plant Syst. Evol.* **259**, 249–258 (2006).
6. Beilstein, M., Al-Shehbaz, I. & Kellogg, E. Brassicaceae phylogeny and trichome evolution. *Am. J. Bot.* **93**, 607–619 (2006).
7. Beilstein, M. A., Al-Shehbaz, I. A., Sarah, M. & Kellogg, E. A. Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *Am. J. Bot.* **95**, 1307–1327 (2008).
8. Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A. & Mummenhoff, K. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* **16**, 108–116 (2011).
9. Guo, X. et al. Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* **18**, 176 (2017).
10. Nikolov, L. A. et al. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *N. Phytol.* **222**, 1638–1651 (2019).
11. Chang, S. J., Chang, Y. C., Lu, K. Z., Tsou, Y. Y. & Lin, C. W. Antiviral activity of *Isatis indigotica* extract and its derived indirubin against Japanese encephalitis virus. *Evid. Based Complement Alternat. Med.* **2012**, 925830 (2012).
12. Liu, S. et al. Antiviral action of Radix Isatidis and Folium Isatidis from different germplasm against influenza A virus. *Acad. J. Second Mil. Med. Univ.* **21**, 204–206 (2000).
13. Du, Z., Liu, H., Zhang, Z. & Li, P. Antioxidant and anti-inflammatory activities of Radix Isatidis polysaccharide in murine alveolar macrophages. *Int. J. Biol. Macromol.* **58**, 329–335 (2013).
14. Chen, M. et al. Alkaloids from the root of *Isatis indigotica*. *J. Nat. Prod.* **75**, 1167–1176 (2012).
15. Deng, X., Gao, G., Zheng, S. & Li, F. Qualitative and quantitative analysis of flavonoids in the leaves of *Isatis indigatica* Fort. by ultra-performance liquid chromatography with PDA and electrospray ionization tandem mass spectrometry detection. *J. Pharm. Biomed.* **48**, 562–567 (2008).
16. Li, B. et al. Phenylpropanoids isolated from tetraploid roots of *Isatis indigotica*. *Chin. Tradit. Herbal Drugs* **36**, 326–328 (2005).
17. Zhou, W. & Zhang, X. Y. Research progress of Chinese herbal medicine *Radix isatidis* (banlangen). *Am. J. Chin. Med.* **41**, 743–764 (2013).
18. Lin, C. W. et al. Anti-SARS coronavirus 3C-like protease effects of *Isatis indigotica* root and plant-derived phenolic compounds. *Antivir. Res.* **68**, 36–42 (2005).
19. Sun, D. D., Dong, W. W., Li, X. & Zhang, H. Q. Indole alkaloids from the roots of *Isatis ingigotica* and their antiherpes simplex virus type 2 (HSV-2) activity in vitro. *Chem. Nat. Compd.* **46**, 763–766 (2010).
20. Xia, X., Xiao, J., Shi, G. & W., D. Function research of resistance to Salmonella Typhimurium infection using Banlangen polysaccharide. *Med. J. Wuhan Univ.* **28**, 348–350 (2007).
21. Ho, Y. L. & Chang, Y. S. Studies on the antinociceptive, anti-inflammatory and antipyretic effects of *Isatis indigotica* root. *Phytomedicine* **9**, 419–424 (2002).
22. During, A., Debouche, C., Raas, T. & Larondelle, Y. Among plant lignans, pinoresinol has the strongest antiinflammatory properties in human intestinal Caco-2 cells. *J. Nutr.* **142**, 1798–1805 (2012).
23. Hoessel, R. et al. Indirubin, the active constituent of a Chinese antileukaemia medicine, inhibits cyclin-dependent kinases. *Nat. Cell Biol.* **1**, 60–67 (1999).
24. Molina, P. et al. Inhibition of leukocyte functions by the alkaloid isaindigotone from *Isatis indigotica* and some new synthetic derivatives. *J. Nat. Prod.* **64**, 1297–1300 (2001).
25. Chen, J. et al. Biosynthesis of the active compounds of *Isatis indigotica* based on transcriptome sequencing and metabolites profiling. *BMC Genomics* **14**, 857–857 (2013).
26. Gai, Q. Y. et al. Elicitation of *Isatis tinctoria* L. hairy root cultures by salicylic acid and methyl jasmonate for the enhanced production of pharmacologically active alkaloids and flavonoids. *Plant Cell* **137**, 77–86 (2019).
27. Lu, B. B. et al. Cloning and characterization of a differentially expressed phenylalanine ammonialyase gene (liPAL) after genome duplication from tetraploid *Isatis indigotica* Fort. *J. Integr. Plant Biol.* **48**, 1439–1449 (2006).
28. Hu, Y. et al. Isolation and characterization of a gene encoding cinnamoyl-CoA reductase from *Isatis indigotica* Fort. *Mol. Biol. Rep.* **38**, 2075–2083 (2011).
29. SimãO, F. A., Waterhouse, R. M., Panagiotis, I., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
30. Mandáková, T. & Lysak, M. A. Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. *Curr. Protocols Plant Biol.* **1**, 359–371 (2016).
31. Schranz, M. E., Lysak, M. A. & Mitchell-Olds, T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542 (2006).
32. Lysak, M. A., Mandáková, T. & Schranz, M. E. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Curr. Opin. Plant Biol.* **30**, 108–115 (2016).
33. Terezie, M. & Lysak, M. A. Chromosomal phylogeny and karyotype evolution in $x = 7$ crucifer species (Brassicaceae). *Plant Cell* **20**, 2559–2570 (2008).
34. Cheng, F. et al. Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *Plant Cell* **25**, 1541–1554 (2013).
35. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
36. Barker, M. S., Heiko, V. & Eric, S. M. Paleopolyploidy in the Brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1**, 391–399 (2009).
37. Lagercrantz, U. & Lydiate, D. J. Comparative genome mapping in Brassica. *Genetics* **144**, 1903–1910 (1996).

38. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
39. Tank, D. C. et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *N. Phytol.* **207**, 454–467 (2015).
40. Mohn, T., Plitzko, I. & Hamburger, M. A comprehensive metabolite profiling of *Isatis tinctoria* leaf extracts. *Phytochemistry* **70**, 924–934 (2009).
41. Yang, L. et al. Indole alkaloids from the roots of *Isatis indigotica* and their inhibitory effects on nitric oxide production. *Fitoterapia* **95**, 175–181 (2014).
42. He, L. W., Li, X. & Chen, J. W. Research progress of antiviral active components of *Radix Isatidis*. *Inform. Trad. Chin. Med.* **22**, 37–40 (2005).
43. Yuk, J. E. et al. Effects of lactose-β-sitosterol and β-sitosterol on ovalbumin-induced lung inflammation in actively sensitized mice. *Int. Immunopharmacol.* **7**, 1517–1527 (2007).
44. Zhao, C. et al. Daucosterol inhibits cancer cell proliferation by inducing autophagy through reactive oxygen species-dependent manner. *Life Sci.* **137**, 37–43 (2015).
45. Kellner, F. et al. Genome-guided investigation of plant natural product biosynthesis. *Plant J.* **82**, 680–692 (2015).
46. Lv, H. et al. Isovitexin exerts anti-inflammatory and anti-oxidant activities on lipopolysaccharide-induced acute lung injury by inhibiting MAPK and NF-κB and activating HO-1/Nrf2 pathways. *Int. J. Biol. Sci.* **12**, 72–86 (2016).
47. Saarinen, N. M. et al. Dietary lariciresinol attenuates mammary tumor growth and reduces blood vessel density in human MCF-7 breast cancer xenografts and carcinogen-induced mammary tumors in rats. *Int. J. Cancer* **123**, 1196–1204 (2010).
48. Meng, L. J. et al. Diglycosidic indole alkaloid derivatives from an aqueous extract of *Isatis indigotica* roots. *J. Asian Nat. Prod. Res.* **19**, 529 (2017).
49. Wu, Y. et al. Novel indole C-glycosides from *Isatis indigotica* and their potential cytotoxic activity. *Fitoterapia* **82**, 288–292 (2011).
50. Liu, Y. F. et al. Antiviral glycosidic bisindole alkaloids from the roots of *Isatis indigotica*. *J. Asian Nat. Prod. Res.* **17**, 689–704 (2015).
51. Mak, N. K. et al. Inhibition of RANTES expression by indirubin in influenza virus-infected human bronchial epithelial cells. *Biochem. Pharmacol.* **67**, 167–174 (2004).
52. Kunikata, T. et al. Indirubin inhibits inflammatory reactions in delayed-type hypersensitivity. *Eur. J. Pharmacol.* **410**, 93–100 (2000).
53. Salvini, M. et al. Alpha-tryptophan synthase of *Isatis tinctoria*: gene cloning and expression. *Plant Physiol. Biochem.* **46**, 715–723 (2008).
54. Hsu, T. M. et al. Employing a biochemical protecting group for a sustainable indigo dyeing strategy. *Nat. Chem. Biol.* **14**, 256 (2018).
55. Sheng, H. et al. Altering regioselectivity cytochrome P450 BM-3 saturation mutagenes is for the biosynthesis of indirubin. *J. Mol. Catal. B* **67**, 29–35 (2010).
56. Korlach, J. et al. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* **6**, gix085 (2017).
57. Zhang, L. et al. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Horticult. Res.* **5**, 50 (2018).
58. Chen, J. et al. Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat. Plants* **5**, 18 (2019).
59. Hu, Y. et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **51**, 739–748 (2019).
60. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344**, 510–513 (2014).
61. Kliebenstein, D. J. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PLoS ONE* **3**, e1838 (2008).
62. Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J. & Mitchell-Olds, T. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell* **13**, 681–693 (2001).
63. Kang, Y. J. et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **5**, 5443 (2014).
64. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764 (2011).
65. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722 (2017).
66. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
67. Louwers, M., Splinter, E., Driel, R. V., Laat, W. D. & Stam, M. Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). *Nat. Protoc.* **4**, 1216–1229 (2009).
68. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
69. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119 (2013).
70. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4–10 (2004).
71. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl. 1), i351 (2005).
72. Patel, R. K. & Mukesh, J. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLos ONE* **7**, e30619 (2012).
73. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
74. Daehwan, K., Ben, L. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
75. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
76. Mario, S., Rasmus, S., Stephan, W. & Burkhard, M. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, 309–312 (2004).
77. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
78. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
79. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
80. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
81. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, 619832 (2008).
82. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
83. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
84. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
85. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
86. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
87. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
88. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
89. Han, M. V., Thomas, G. W. C., Jose, L. M. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).
90. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
91. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
92. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S. H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003 (2008).