Journal Pre-proof

Chromosome-level *Alstonia scholaris* genome unveils evolutionary insights into biosynthesis of monoterpenoid indole alkaloids

Haixia Chen, Sunil Kumar Sahu, Shujie Wang, Jia Liu, Jinlong Yang, Le Cheng, Tsan-Yu Chiu, Huan Liu

PII: S2589-0042(24)00821-6

DOI: https://doi.org/10.1016/j.isci.2024.109599

Reference: ISCI 109599

To appear in: *ISCIENCE*

Received Date: 17 November 2023

Revised Date: 25 January 2024

Accepted Date: 25 March 2024

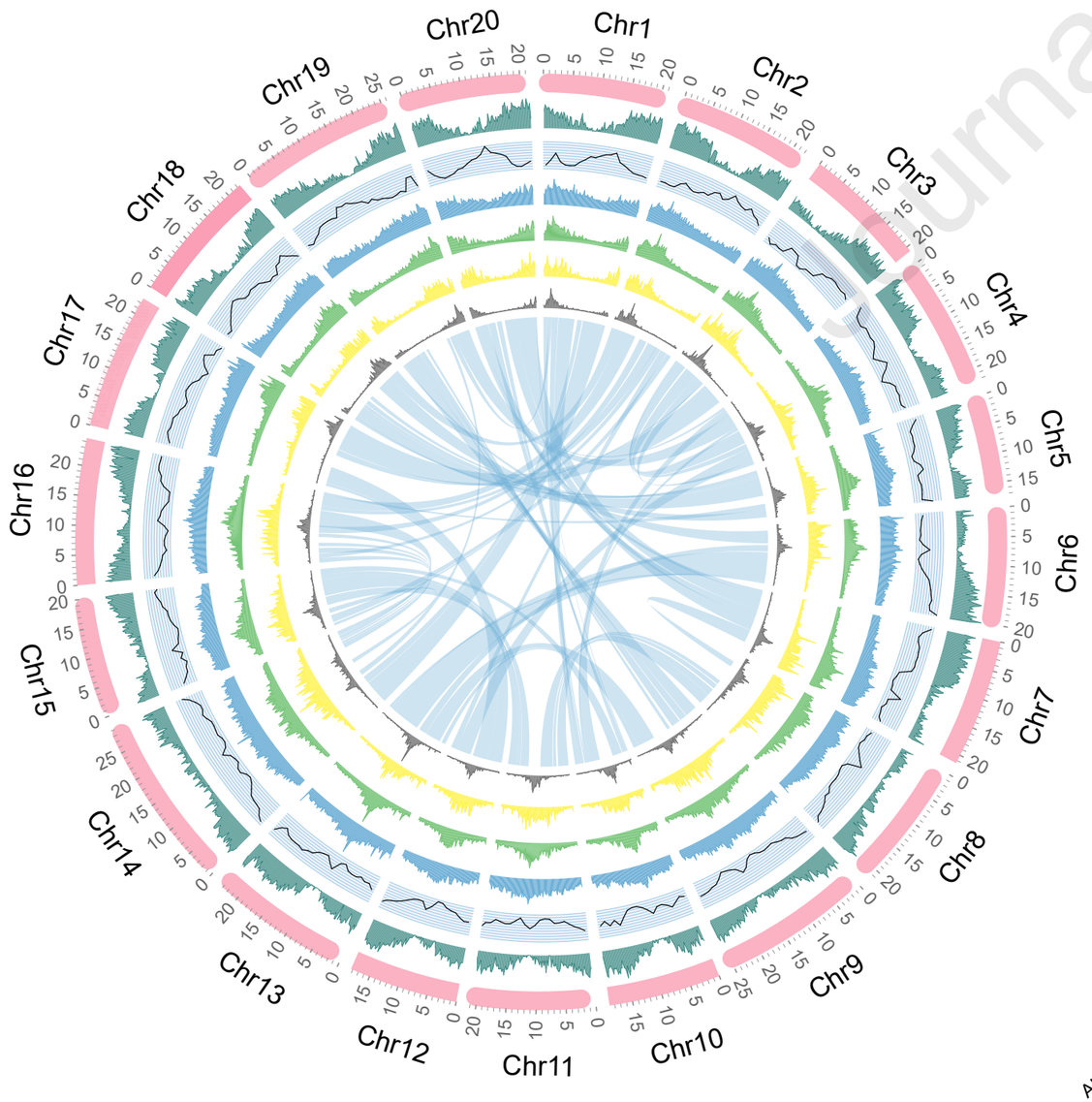
Please cite this article as: Chen, H., Sahu, S.K., Wang, S., Liu, J., Yang, J., Cheng, L., Chiu, T.-Y., Liu, H., Chromosome-level *Alstonia scholaris* genome unveils evolutionary insights into biosynthesis of monoterpenoid indole alkaloids, *ISCIENCE* (2024), doi: https://doi.org/10.1016/j.isci.2024.109599.

Alstonia scholaris

Nanopore and Hi-C sequencing → A chromosome-level genome

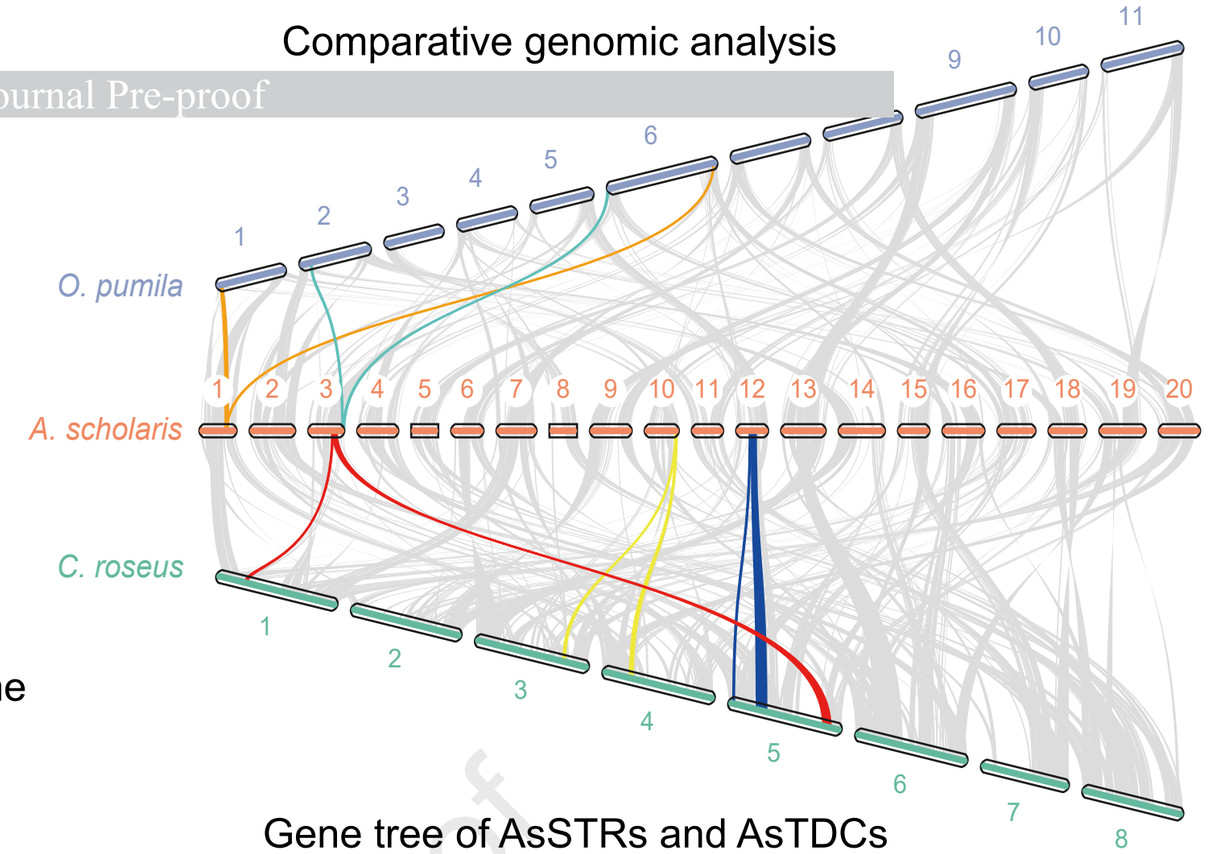
RNA-seq of multi tissues → Different and co-expression analyses

Sequence alignment
Construct gene tree
Identify gene clusters → Candidate genes related to MIAs

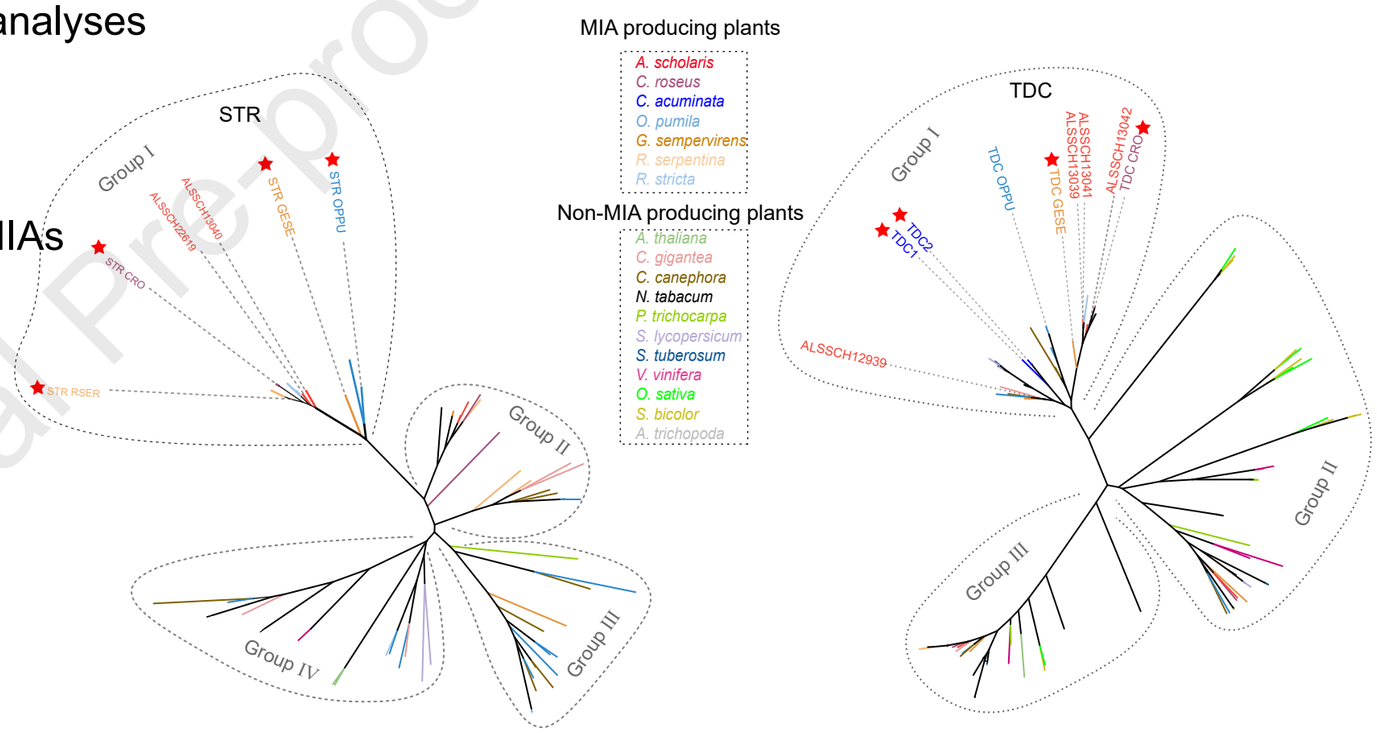


Genomic features

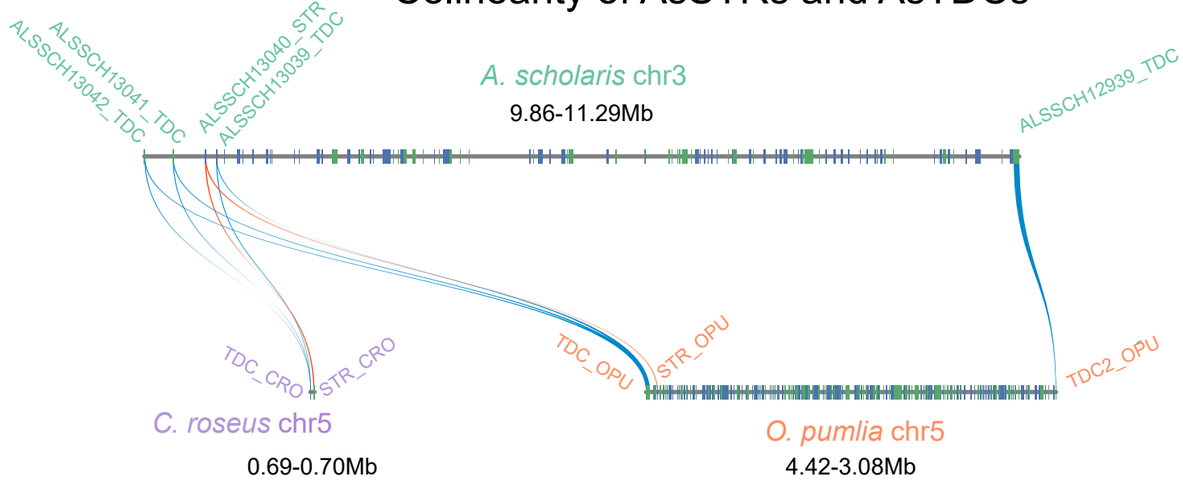
Comparative genomic analysis



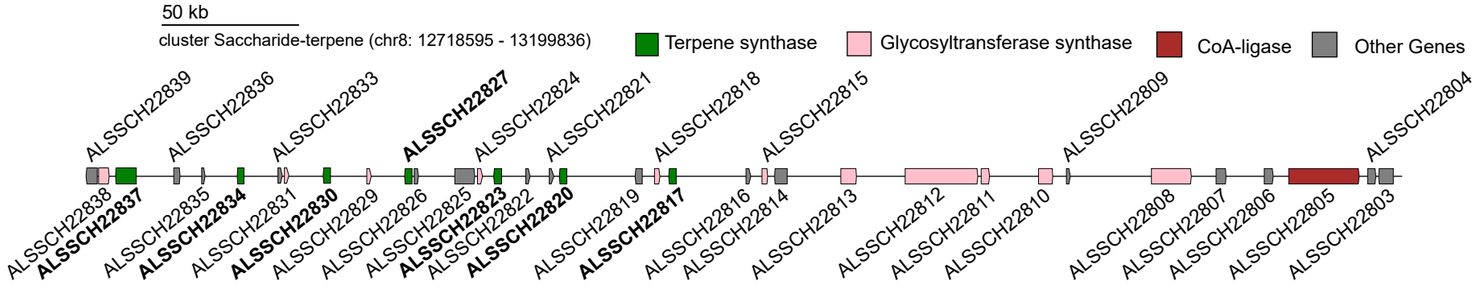
Gene tree of AsSTRs and AsTDCs



Colinearity of AsSTRs and AsTDCs



A saccharide-terpene cluster



Chromosome-level *Alstonia scholaris* genome unveils evolutionary insights into biosynthesis of monoterpenoid indole alkaloids

Haixia Chen[1,2,4], Sunil Kumar Sahu[1,2,4], Shujie Wang[1], Jia Liu[3], Jinlong Yang[1], Le Cheng[1], Tsan-Yu Chiu[1*], Huan Liu[1,5*]

[1]State Key Laboratory of Agricultural Genomics, Key Laboratory of Genomics, Ministry of Agriculture, BGI Research, Shenzhen 518083, China. [2]BGI Research, Wuhan 430074, China. [3]Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin 150081, China.

[4]These authors contributed equally

[5]Lead contact

[*]Correspondence:

Huan Liu: liuhuan@genomics.cn; Tsan-Yu Chiu: qiucanyu@genomics.cn

SUMMARY

Alstonia scholaris of the Apocynaceae family is a medicinal plant with a rich source of bioactive monoterpenoid indole alkaloids (MIAs), which possess anti-cancer activity like vinca alkaloids. To gain genomic insights into MIA biosynthesis, we assembled a high-quality chromosome-level genome for *A. scholaris* using Nanopore and Hi-C data. The 444.95 Mb genome contained 35,488 protein-coding genes. A total of 20 chromosomes were assembled with a scaffold N50 of 21.75 Mb. The genome contained a cluster of strictosidine synthases and tryptophan decarboxylases with synteny to other species and a saccharide-terpene cluster involved in the monoterpenoid biosynthesis pathway of the MIA upstream pathway. The multi-omics data of *A. scholaris* provide a valuable resource for understanding the evolutionary origins of MIAs and for discovering biosynthetic pathways and synthetic biology efforts for producing pharmaceutically useful alkaloids.

INTRODUCTION

The *Alstonia scholaris* from Apocynaceae family is commonly known as milkwood pine, blackboard tree, or devil tree, and is widely distributed in the tropical regions of Africa and Asia[1]. Since ancient times, the use of natural products from terrestrial plants has been indispensable to humans in each civilization[2]. For instance, the bark of *A. scholaris* is used in traditional medicine in South and Southeast Asia to treat dysentery and malaria[3]. The traditional system of Indian medicine named Ayurveda uses the bark in numerous compound formulations, including mahatikta ghrita, saptachchhadadi taila, saptaparnaghana vati, and saptachchhadadi kvatha[4] (*A. scholaris* named Saptaparna or Saptaparn in Sanskrit language). The leaves of these plants are used in "Dai" ethnopharmacology to treat chronic respiratory diseases in Yunnan Province of China[5]. Based on traditional utilization, the leaf extract has also been industrialized as an OTC (over the counter) drug in China, popularly known as "Deng-Tai-Ye" tablet[6]. Additionally, it is also used to treat chronic bronchitis cough and was approved by the China Food and Drug Administration (CFDA)[7]. In addition, the extracts of *A. scholaris* have been shown to have anti-diabetic[8], anti-inflammatory[9], anti-tussive, anti-asthmatic[10], and most importantly, anti-tumor activities[11].

Members of the Apocynaceae family serve as the major natural source for procuring monoterpenoid indole alkaloids (MIAs), which can be used to treat various human diseases. For example, catharanthine and vindoline from *Catharanthus roseus* can be used for diabetes treatment. Catharanthine, vinorelbine and vincristine are currently being used for anti-cancer treatment[12]. Ajmalicine from *Rauvolfia verticillate* has neurological function and hypotensive effects[13]. Camptothecin from *Camptotheca acuminate* also has anti-cancer effects[14]. MIAs have a wide range of diverse and important pharmacological properties, some of which have been used clinically. The MIA biosynthesis pathways from *C. roseus* (vinblastine and vincristine)[15-18] and *R. serpentina* (reserpine)[19] were characterized at the molecular level. Several studies have also been carried out on *C. acuminata*[20] and *Ophiorrhiza pumila*[21] to characterize the early steps of camptothecin biosynthesis[22].

Akuammiline alkaloids are a class of MIAs in *A. scholaris*, and more than 300 compounds with various pharmacological activities have been identified in *A. scholaris*[23, 24]. For example, echitamines[2] exhibit both *in vitro* and *in vivo* cytotoxicity[25], while strictamines[26] inhibit the transcription factor NF-kB[27]. Additionally, the renal cortex protein SGLT2 is inhibited by the derivatives of picraline[28-30], whereas aspidophylline A reverses drug resistance in cancerous cell lines[31]. Biogenetically, the akuammiline alkaloids are derived from geissoschizine, a key intermediate in the biosynthetic pathway of MIAs[32]. The formation of strictosidine from secologanine and tryptamine, catalyzed by the enzyme strictosidine synthase (STR)[33], and by the presence of two enzymes function, strictosidine-β-D-glucosidase (SGD)[34-36] and geissoschizine synthase (GS)[18], results in the production of geissoschizine. The intramolecular oxidative coupling between C7 and C16 of geissoschizine results in the formation of the framework of akuammiline[32]. Moreover, this coupling forms the caged indolenine framework of (+)-rhazimal[32].

Elucidating the biosynthetic pathway of bioactive compounds will greatly benefit the development of synthetic biology tools for medicinal plants. The advancement of long-read sequencing has closed the gap in genomic information provided by short-read sequencing[20, 37]. Although there are several existing studies on various medicinal plants [38-42], the biosynthetic pathway of akuammiline alkaloids is still vague due to the lack of valid omics data. Therefore, these high-quality genome and transcriptome data of *A. scholaris* provide a

solid foundation for identifying potential genes involved in the akuammiline alkaloid production pathway and advancing synthetic biology research on anticancer bioactives from *A. scholaris.*

RESULTS

De novo genome assembly and pseudochromosome construction

We used approximately 45 Gb (~90×) of short reads for genome survey analysis, and the estimated genome size was 489 Mb based on k-mer analysis (Figure S1). Using a combination of 295 Gb (~590×) short and 62 Gb (~124×) long nanopore reads, we generated an assembly of 444,958,049 bp with a contig N50 size of 13.24 Mb (Table 1). We anchored the contig-level genome onto 20 pseudochromosomes with 69 Gb (~138×) of Hi-C data (Figure 1). The N50 value increased to 21.75 Mb (Table 1), and the length of the chromosomes ranged from 17.02 Mb to 29.20 Mb. The BUSCO (benchmarking universal single-copy orthologs)[43] results showed that 2286 out of 2326 plant BUSCOs (98.3%) could be found in both the contig and chromosome-level genome assemblies (Figure S2, Table S1).

Protein-coding gene prediction and functional annotation

We found 38.26% repetitive elements in the *A. scholaris* genome. The most abundant type was long terminal repeats (LTRs), accounting for 28.81% of the *A. scholaris* genome. DNA class repeat, LINE and SINE classes accounted for 5.03%, 2.21%, and 0.01%, respectively, of this genome (Table S2). A total of 35,488 genes with 5.64 exons per gene on average were predicted by combining three methods, namely *de novo*, homology, and transcriptome-based methods. The average lengths of the mRNAs, exons and introns were 3,852 bp, 217 bp, and 565 bp, respectively (Table S3). The length distributions of the gene sets of *A. scholaris* and the other seven species (*C. gigantea*, *C. canephora, C. roseus, G. sempervirens, N. tabacum, S. lycopersicum* and *O. pumila*) are shown in Figure S3. The complete and single-copy genes accounted for 92.6% of the predicted gene set (Figure S2, Table S1).

The functional annotation results revealed that approximately 96.70% of the genes had a conserved motif or homolog match in at least one of the public databases, including

SwissProt (78.16%), InterPro (93.53%), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (75.34%). For the non-coding RNAs, we also identified 142 microRNAs, 621 tRNAs, 135 rRNAs and 829 small nuclear RNAs in the *A. scholaris* genome (Table S4).

Comparative Genomic Analysis

We compared the *A. scholaris* genome with 15 other sequenced genomes and identified 14,289 gene families. A total of 107 expanded gene families, and 42 contracted families were significantly differentially expressed in *A. scholaris,* respectively. The 16 species family numbers and gene numbers are summarized in Figure 2a and Table S5. Geno ontology (GO) enrichment analysis of the 107 significantly expanded gene families revealed 25 GO terms. A portion of the expanded genes were enriched in binding terms in the molecular function category, such as ion binding (38), and organic cyclic compound binding (36). The other genes were enriched in several enzyme activity terms, including oxidoreductase activity (25), monooxygenase activity (14) and protein kinase activity (12) (Table S6). A total of seven expanded genes were located in the monoterpenoid biosynthesis pathway; three of these genes were annotated as 10 HGOs (ALSSCH34014, ALSSCH34015, and ALSSCH34016), two of them are candidate G10Hs (ALSSCH04576 and ALSSCH081114) (Table S7). Additionally, these significantly contracted gene families were enriched in 69 GO categories and 35 pathways. For example, a total of 164 genes were enriched in the plant-pathogen interaction pathway, and 83 genes were enriched in the phenylpropanoid biosynthesis pathway (Table S8, S9).

A total of 102 single-copy orthologous groups were used for constructing a phylogenetic tree to estimate the divergence times of 16 plants. We found that *A. scholaris* was clustered with the Apocynaceae species group, which was separated approximately 67 million years ago from *C. gigantea*, while *R. serpentina* and *C. roseus* were closest and diverged almost 39 million years ago. These two species diverged 54 million years ago from *R. stricta* (Figure 2b).

Two whole-genome duplications shaped *A. scholaris* evolution

Ancient whole-genome duplication (WGD) events have contributed to plant adaptation, and

are prevalent in plants[44]. In our study, we used the ks (a synonymous substitution per synonymous site) value to determine whether the *A. scholaris* genome had undergone WGD. We found a peak between ks values of 0.3, indicating that a WGD event occurred approximately 35.1 million years ago (Figure 3a, b), which was later than the divergence time between *A. scholaris* and other Apocynaceae plants. We also performed a synteny analysis of the *A. scholaris* genes using MCScanx to confirm the collinearity relationship. We detected 4,543 syntenic blocks across the whole genome, including 26,485 genes (74.63%). Furthermore, a ks plot of the paralogs of *C. canephora, C. roseus, O. pumila* and *V. vinifera* confirmed that these species underwent WGD in accordance with previous reports (Figure 3a). Synteny analysis revealed 1:2 syntenic depth ratios in both *A. scholaris*-*C. roseus* and *A. scholaris*-*O.pumila* comparisons (Figure 3c; Figure S4), which suggested that two WGD events occurred during the evolution of *A. scholaris*.

Gene clusters involved in MIAs biosynthesis in *A. scholaris*

Secologanin synthase (SLS) and strictosidine synthase (STR), which were identified in Gentianales and catalyze the synthesis of strictosidine, were discovered in a previous study of *O. pumila* and demonstrated significant importance in enabling the evolution of novel enzymes for MIA biosynthesis and diversification[21]. In our study, we used more species from Gentianales and found several STR copies in MIA-producing plants, including *A. scholaris, C. roseus, R. serpentina, C. gigantea, O. pumila, R. stricta,* and *G. sempervirens*. However, no STR was found in *Amborella trichopoda*, *Oryza sativa*, *Solanum tuberosum* or *Sorghum bicolor* genomes. Phylogenetic analysis of the STRs revealed an MIA-specific plant gene family (Group I) that included previously functionally characterized STRs involved in the MIA biosynthesis pathway and two AsSTRs identified in our study. Another two AsSTRs (ALSSCH12919 and ALSSCH22548) clustered with *O. pumila*, *G. sempervirens*, *C. roseus* and *R. serpentina* (Group II) (Figure 4a). Therefore, AsSTRs in Group I are more likely to have true STR activity. Groups III and IV contained homologous gene from *P. trichocarpa*, *S. lycopersicum* and *V. vinifera*, respectively. However, SLS was almost expanded in all the MIA-producing plants and the other ten non-MIA-producing plants (Table S10). Tryptophan decarboxylase (TDC) also plays an essential role in strictosidine biosynthesis, and we found

that TDC was not only expanded in all MIA-specific plants but also in other plants. Phylogenetic analysis of TDC genes showed three branches that included MIA-producing plants. Group I that included four candidate TDC genes of *A. scholaris* and TDC1, TDC2 from *C. acuminata*, one TDC from *O. pumila*, one TDC from *G. sempervirens*, also one TDC from *C. roseus*. In addition, Group I also includes other non-MIA-producing plants, such as *C. canephora* and *S. lycopersicum*. Group II and III consisted of genes from MIA-producing plants (*A. scholaris*, *C. roseus*, *G. sempervirens*, and O. *pumila),* and they included genes from non-MIA-producing plants (Figure 4b).

The four candidate AsTDCs exhibited collinearity with TDCs from both *C. roseus* and *O. pumlia*. However, only one AsTDC from *O. pumlia* displayed collinearity with TDC2. Furthermore, a single AsSTR showed collinearity with both STR_CRO and STR_OPU. AsTDCs and AsSTRs were located on chromosome 3, suggesting the possibility of forming a gene cluster (Figure 4c). In addition, a saccharide-terpene cluster on chromosome 8 was identified in our study, this cluster included seven terpene synthases, 11 glycosyltransferase synthases, one CoA-ligase and 16 other genes (Figure 4d).

We compared the expression levels of the various genes in different tissues (Figure S5). The upregulated genes in both the leaf and branch, compared to those in the control (trunk bark), were predominantly enriched in GO terms related to the membrane, oxidoreductase activity, transported activity and transmembrane transporter activity. The genes whose expression was significantly greater in the petiole than in the trunk bark were mostly enriched in metabolic processes, cellular metabolic processes, biosynthetic processes, and organic substance biosynthetic processes. More than two hundred genes were also enriched in the membrane and oxidoreductase activity terms. The results of the KEGG enrichment analysis showed that the metabolic pathways, photosynthesis pathway, and photosynthesis-antenna proteins pathway were more highly expressed in the leaves, branches and petioles than in the trunk bark.

Additionally, compared with those in the branches, the upregulated genes in the trunk bark were enriched in the flavonoid biosynthesis pathway. Moreover, nine genes with higher expression levels in trunk bark than in leaf were enriched in the monoterpenoid biosynthesis pathway. Five (ID: ALSSCH22817, ALSSCH22820, ALSSCH22827, ALSSCH22830, and

ALSSCH22834) of the nine genes were contained in the saccharide-terpene cluster mentioned above. Similarly, the upregulated genes in the petiole (compared to those in the branches) were enriched in the flavonoid biosynthesis pathway, as well as the monoterpenoid biosynthesis pathway. A total of six terpene synthases (ID: ALSSCH22817, ALSSCH22823, ALSSCH22827, ALSSCH22830, ALSSCH22834, and ALSSCH22837) were located on the saccharide-terpene cluster (Table S11). In particular, AsGESs (ALSSCH22817, ALSSCH22830, and ALSSCH22834) act on geranyl-FP to produce geraniol in the monoterpenoid biosynthesis pathway. These enzymes are located on the saccharide-terpene cluster. This cluster is the first terpene gene cluster in MIA-producing species. The results of all the GO and KEGG enrichment analyses are individually summarized in Table S12-S32.

Candidate genes of the monoterpene indole alkaloid biosynthetic pathway

Previous studies have reported the composition and distribution of monoterpene indole alkaloids (MIAs) (picrinine, picralinal, echitamine and akuammidine) in the leaves, flowers, trunk barks and fruits of *A. scolaris*[45]. MIAs are a large group of plant-produced natural products of which more than 3000 have been identified[46], mostly in Gentianale[47]. Additionally, a review paper summarized 444 monoterpene indole alkaloids that were reported from six genera of the Apocynaceae family between 2010 and 2020[48]. In this study, the extracted metabolites of *A. scholaris* leaves, trunk barks and branches were compared with the ionic fragments and separation times of secologanin and tryptamine standards purchased from a certified vendor. The results showed that the same ionic fragments as the standard were found at the same separation time, which indicated the presence of two key precursors required for MIA biosynthesis (Figure S6).

The biosynthesis pathway of alkaloids from *A. scholaris* has not been elucidated. These pathways start with the common precursor strictosidine, which undergoes several steps of reaction to form Rhazimal akuammiline (Figure 5a). Hence, we focused on identifying potential enzyme-coding genes involved in the akuammiline biosynthesis pathway. Initially, we compiled a preliminary gene list by aligning sequences with known genes from MIA biosynthesis pathways, and we filtered out genes with low or no expression. Furthermore,

we screened candidate genes using qualitative protein data of multi-tissues of *A. scholaris,* resulting in the identification of 55 candidate genes in the *A. scholaris* genome (Table S33). All known MIA biosynthesis-related protein sequences and accession numbers are summarized in Table S34. The processed protein group data are presented in Table S35.

The results of the co-expression analysis indicated strong correlations between modules and specific plant parts: the trunk bark (skyblue, darkmagenta, bisque4, darkorange), the petiole (darkorange2 and maroon), and the leaf (darkgreen and red) (Figure S7).

Within the skyblue, darkmagenta, bisque4, and darkorange modules, we identified the presence of AsGES, As10HGO, As7-DLH, AsTDC, AsSLS, and AsGS. While in the darkgreen and red modules, we observed AsAS, AsIGPS, AsTSA, AsTSB, AsGO, AsIS, As7-DLH, and AsSLS. We also found AsIS and AsGO in the darkorange2 and maroon modules. These findings suggest a potential co-expression pattern of alkaloid biosynthesis genes in *A. scholaris*.

In addition, we compared the expression levels of those genes in the leaf, petiole, branch and trunk bark tissues of *A. scholaris*. We found that 7DLGT/UGT6, TSB2, TSB3, TDC1, and TDC2 were more highly expressed in the petioles of *A. scholaris*. AsGOs were highly expressed in trunk bark and petioles. 7-DLH, PAT1, G10H, IGPS, IO, SGD, SLS, TSA and TSB1 exhibited relatively average expression levels in all tissues, exhibiting a co-expression pattern in different modules (Figure 5b).

DISCUSSION

Monoterpenoid Indole Alkaloids (MIAs) are natural compounds derived from secologanin and tryptamine that are subsequently obtained from tryptophan through decarboxylation. One of the most comprehensively elucidated MIA biosynthesis pathways is the vinca alkaloid biosynthesis pathway found in *C. roseus,* which leads to the production of compounds like vincristine, vinblastine, catharanthine, tarbersonine, and vindoline[15, 46, 49-52]. A previous study has also identified various MIAs in different parts of *A. scholaris*[53]. Here, we assembled a chromosome-level genome of *A. scholaris,* an MIA-producing plant with multiple medicinal benefits. We detected the presence of the MIA precursor, secologanin, and tryptamine in

various parts of *A. scholaris*. By performing an alignment with previously identified enzymes from MIA biosynthesis pathways, a series of candidate genes involved in MIA biosynthesis were identified in the *A. scholaris* genome. *A. scholaris* contains several unique alkaloids, such as 19-epi-scholaricine, scholaricine, 19,20-Z-vallesamine and picrinine which are the main medicinal components of the "Deng-Tai-Ye" tablet[54]. Because there is no established commercial standard for these alkaloids, detecting metabolic differences across multiple tissues in *A. scholaris* is challenging.

The biosynthesis pathway of camptothecin is similar to that of vinblastine/vincristine in *C. roseus* and involves the production of loganic acid [55-57]. However, secologanic acid is transformed to strictosidinic acid by STRAS in *C. acuminata,* which has different modifications than *C. roseus*[20]. In *O. pumila*, STR has the same function as in *C. roseus* and similarly produces strictosine[21]. We used additional Gentianales species for comparison with the *A. scholaris* genome and found STR expansion in MIA-specific plants, but not in non-MIA producing plants. These findings indicate that STR is conserved in the MIA biosynthesis pathway. However, SLS was retained in all the species rather than in the MIA-producing plants. Besides, the existence of an alternate pathway for MIA biosynthesis in *C. acuminata* has been proven to occur through strictosidinic acid, which is synthesized by the condensation of secologanic acid with tryptamine by SLAS, an SLS-like enzyme[21, 55], suggesting that SLS is not necessary for all MIA-producing plants. TDC also plays an essential role in strictosidine biosynthesis. and found that TDC was expanded in all MIA-specific plants but also in other plants. The phylogenetic analysis of TDC genes showed a branch that included both MIA-specific plants and non-MIA-producing plants, suggesting an essential role of TDC in amino acid metabolism in plants[21].

A recent study identified an STR-TDC cluster in the *C. roseus* v3 genome[58]. In our study, we discovered a gene cluster on the third chromosome of the *A. scholaris* genome that combines AsTDCs and one AsSTR. These genes exhibited collinearity with the TDC and STR genes from *C. roseus* and *O. pumila*, respectively. In addition, we also found a saccharide-terpene cluster on the eighth chromosome of the *A. scholaris* genome. This terpene cluster included seven terpene synthases that had higher expression levels in the trunk bark and petioles of *A. scholaris*. Taken together, these findings show that the

monoterpenes of the MIA upstream pathway may be synthesized or transferred to the trunk bark and petiole. This cluster also included three AsGESs which catalyze the first step in the monoterpene synthesis pathway.

To maximize the value of our genomic data in the akuammilan alkaloid synthetic pathway analysis of *A. scholaris*, we published the genomic data of short reads in advance. A study reported the discovery of a series of new enzymes involved in akuammilan alkaloid biosynthesis by using our publicly available *A. scholaris* genome data. Among these enzymes, AsRHS and AsGO share a significant sequence identity of 62.4%. Notably, the amino acid residue at position 372 plays a crucial role in regulating the geissoschizine reaction by altering the distance between C-2 and C-7 in relation to the heme. Consequently, one enzyme may predominantly oxidize C2 (GO), while the other may target C7 (RHS) of geissoschizine[59]. We analyzed to investigate the co-expression patterns of these genes, revealing several modules that exhibited strong correlations with the leaf, trunk bark and petiole parts of *A. scholaris*. Among these modules, we identified candidate MIA genes, suggesting a potential co-expression pattern for the alkaloid biosynthesis genes in *A. scholaris*. Overall, our findings contribute to a deeper understanding of *A. scholaris* and pave the way for innovative applications in the field of alkaloid biosynthesis, setting the stage for exciting future research in this domain. In addition, our data are conducive to revealing the mechanism of MIA evolution.

Limitations of the study

While the current study identified putative candidate genes implicated in the MIA biosynthetic pathway, functional characterization of these genes in heterologous hosts such as tobacco or *E. coli* was not performed. The incorporation of metabolomic data could further elucidate the MIA landscape in *A. scholaris*. Despite the lack of reference standards for most *A. scholaris* MIAs, future investigations could employ total ion chromatograms coupled with mass spectrometric analysis to tentatively annotate major alkaloid peaks based on their m/z values.

AUTHOR CONTRIBUTIONS

H.C., L.C. and H.L. designed the study and all the experiments. H.C., S.K.S. and T.-Y.C. performed the data analysis. H.C., S.W. and J.L. collected samples and did metabolome analysis. H.C., T.-Y.C. and S.K.S wrote the manuscript. All the authors have read and agreed to the final version of the manuscript.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - **Lead contact**
  - **Materials availability**
  - **Data and code availability**
- **METHOD DETAILS**
  - **Plant sample collection and sequencing**
  - **Genome assembly and chromosome anchoring**
  - **Identification of repetitive sequences**
  - **RNA-seq analysis**
  - **Gene model prediction and functional annotation**
  - **Candidate MIA gene prediction**
  - **Gene Family analysis and Evolutionary tree construction**
  - **Protein detection**
  - **Metabolite detection**

353

354

355 Table 1 Statistics of the genome assembly.

Assembly		*Alstonia scholaris*
Genome-sequencing depth (~124×)	Nanopore sequencing (Gb)	62
	Hi-C (Gb)	69
Estimated genome size (Mb)		489
Estimated heterozygosity (%)		0.835
Assembly size (Mb)		445
GC content (%)		34.73
Scaffold N50 (Mb)		13.244
BUSCO completeness of assembly (%)		98.3
Total length of pseudochromosome assembly (Mb)		445
Pseudochromosome number		20
Scaffold N50 of pseudochromosome assembly (Mb)		21.753
BUSCO completeness of pseudochromosome assembly (%)		98.3
The rate of pseudochromosome anchored genome (%)		99.9

356

357 Figure1. Genome information and morphological features of *A. scholaris*. a. Characteristics
358 of the 20 chromosomes of *A. scholaris.* The tracks from the outer to the inner regions of the
359 circle individually represent the length of chromosomes (pink), gene numbers (dark green),
360 the content of GC (black line), repeat sequences (blue), LTRs (green), LTR *Copia* (yellow),
361 LTR *Gypsy* (gray) and the links inside the circle shows syntenic collinearity. b. Hi-C plot of
362 the pseudochromosome-level assembly of *A. scholaris* genome. The axis refers to the
363 genome size, and each blue box represents one chromosome.

364

365 Figure 2. Gene family analysis and phylogenetic tree construction. a. Bar chart of the
366 ortholog numbers in these 16 species. b. Phylogenetic tree showing the sizes of significantly
367 expanded and contracted gene families. The branch labels in yellow and blue represent the
368 significantly expanded and contracted gene families (Pvalue < 0.05), respectively, of each node.
369 The right column shows significantly expanded and contracted gene families of individual
370 species. Furthermore, the statistical method of enrichment analysis is $\chi 2$ test. AdjustedPv is
371 a corrected p-value that is obtained by performing false discovery rate (FDR) testing on p-

values. AdjustedPv obtained from significance tests are generally considered to have significant statistical differences when AdjustedPv < 0.05.

Figure 3. The analysis of whole genome duplication in *A. scholaris*. a. The distribution of synonymous substitution rate (ks) distances observed for paralogs from *A. scholaris, C. canephora, C. roseus, O. pumila* and *V. vinifera.* b. The distribution of ks values of orthologs between *A. scholaris* and the previously mentioned species. c. Synteny between genomic regions in *A. scholaris, C. roseus* and *O. pumila*. The gray lines highlight major syntenic blocks spanning the genomes. The colored lines represent examples of syntenic genes found in two species that correspond to one copy in *A. scolaris*, and two in *C. roseus* and *O. pumila*.

Figure 4. Key genes involved in strictosidine biosynthesis. a, b Maximum likelihood phylogenetic tree based on candidate STR and TDC gene families from these 18 species. "*" represents functionally characterized genes. Each species is represented by a different color. c. A gene cluster located on the third chromosome of the *A. scholaris* genome. The blue and orange lines show syntenic blocks of AsTDCs and AsSTRs with TDC and STR from *C. roseus* and *O. pumlia*, respectively. d. The saccharide-terpene cluster. The bolded gene IDs represent genes that are significantly highly expressed in petioles and trunk barks, and are also genes that are enriched in the monoterpene biosynthesis pathway.

Figure 5. Biosynthesis pathway of akuammiline alkaloid and the expression levels of candidate enzymes in the pathway. a. The akuammiline biosynthesis pathway. b. The expression levels of candidate genes in the *A. sholaris* genome. The abscissa of each heatmap indicates the different tissues of *A. scholaris*. B: branch, P: petiole, T: trunk bark, L: leaf. The number represents duplication. Log2GeneCount refers to the normalization of FPKM values by the log method of each row. Different expressed genes match adjusted p value < 0.05 and a $|\log2FoldChange| \geq 2$. Dark-blue color indicates a high expression level, and blue indicates a low expression level.

STAR METHODS

KEY RESOURCES TABLE

RESOURCE AVAILABILITY

Lead contact

Further information and requests can be directed to Prof. Huan Liu (liuhuan@genomics.cn).

Materials availability statements

The study did not generate new unique reagents.

Data and code availability

- The raw data of genome, transcriptome sequencing and assembly data of *A. scholaris* are deposited at CNSA (https://db.cngb.org/cnsa/) under the project accession number CNP0002381, and all datasets are publicly available before the date of publication.
- The DESeq2 and WGCNA analysis R scripts are provided in Supplementary data 1.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Plant sample collection and sequencing

The *Alstonia scholaris* (ID 52822) plant cultivated in the Ruili Botanical Garden of Yunnan Province, China, was used in this study. We collected fresh and young leaves for Nanopore, Hi-C and WGS sequencing. The purity, concentration and integrity of the extracted DNA were tested by Nanodrop, Qubit and Agarose Gel Electrophoresis, respectively. The library was constructed by using SQK-LSK109 kit and the PromethION platform was used for ONT sequencing. For Hi-C experiment, we cut the fresh leaves into fragments with 50 ml of MC buffer and 1.39 ml of 37% methanol to infiltrate those fragments. The methanol-processed tissues were ground to powder in liquid nitrogen for DNA extraction using the CTAB (cetyl

trimethyl ammonium bromide) method[60]. The Hi-C library was constructed and sequenced on the BGISEQ-500 platform, according to their standard protocol.

We collected leaf, petiole, branch, and trunk bark samples to extract total RNA by using the CTAB-βBIOZOL method. The RNA quality was evaluated by Nanodrop, Qubit 2.0 and Agilent 2100 instruments to ensure that the RNA was suitable for library construction and sequencing. Then, 4μL fragmentation buffer was added to the mRNA sample liquid under 150 bp fragmentation conditions for eight minutes. The fragmented samples were mixed with RT buffer to start reverse transcription to obtain the second chain product. After purification, the "A" and adapter bases were added, and after PCR amplification and enzyme cleavage, the RNA-seq library was constructed.

Genome assembly and chromosome anchoring

A total of 295 Gb of short reads and 62 Gb of long reads were generated for the genome assembly. Nextdenovo (v 2.3.0) software (https://github.com/Nextomics/NextDenovo) was used for *de novo* assembly of *A. scholaris* genome via a "correct-then-assemble" strategy. NextPolish (v 1.3.1) (https://github.com/Nextomics/NextPolish) was used to fix base errors in the genome generated by noisy long reads with a combination of short and long-reads data.

A total of 69 Gb data were generated for the Hi-C maps. Juicer[61] (v 1.6) software was used to map the Hi-C data to the assembled genome, and sorting and merging steps generated the input file of the 3D *de novo* assembly (3D-DNA) pipeline[62] which assembles an accurate genome with chromosome-length scaffolds. Juicerbox (https://github.com/aidenlab/juicebox) was used for manual correction, after which 3D-DNA was reanalyzed to generate the final assembled genome. The completeness of the genome assembly was assessed against the eudicot database (odb10) by BUSCO with default settings.

The method for identifying repeat sequences is described in the section titled "Identification of Repetitive Sequences." Gene density was calculated based on the gene positions within each window. A window size of 1 Mb with a step size of 1 Mb was used for sliding windows to calculate the GC content. The colinear regions of chromosomes were obtained using JCVI (https://github.com/tanghaibao/jcvi). Finally, the Circos software[63] was employed to combine all the results and generate the figure.

459

460 **Identification of repetitive sequences**

461 RepeatMasker (v 4.0.6) and RepeatProteinMask (v 4.0.6)[64] were used to search in the
462 Repase[65] database to identify TEs in the DNA and protein leaves. Tandem Repeats Finder (v
463 4.07b)[66] was used to identify tandem repeats. Our strategy for identifying repeat sequences
464 involves a combination of *de novo* and homology-based methods. *De novo* identification was
465 performed with Piler (v1.0)[67] and LTR-FINDER (v 1.06)[68], and RepeatMasker. By using the
466 previously constructed libraries as a database, RepeatMasker was used to identify and
467 classify the final repeats in the *A. scholaris* genome.

468

469 **RNA-seq analysis**

470 Hisat2 (v 2.1.0)[69] was used to map the clean RNA-seq data to the *A. scholaris* genome with
471 the following parameters: hisat2-2.1.0/hisat2-align-s --wrapper basic-0 -t -x index -1
472 clean.read1.fq.gz -2 clean.read2.fq.gz -S clean.sam. Then, we used samtools (v 1.7)[70]
473 software to sort the bam files as an input file of StringTie (v 1.3.3b)[71] to predict each sample's
474 transcript in bulk and integrate them into one nonredundant transcript. The parameters were
475 as follows: stringtie sorted.bam -p 15 -G genome.gtf -o sorted.bam.gtf; stringtie --merge -p
476 20 -G genome.gtf -o merged_stringtie.gtf mergelist.txt. Finally, each sample's gene
477 expression was quantified and integrated by the following procedure: stringtie -e -B -p 8 -G
478 merged_stringtie.gtf -o ballgown/output_merge.gtf sorted.bam; stringtie/prepDE.py -i
479 ballgown.

480 We used DEseq2 R package[72] to perform differential expression analysis with gene
481 count data. A gene matching adjusted p value < 0.05 and a |log2FoldChange| ≥ 2 were
482 considered to indicate differentially expressed genes. The co-expression analysis was
483 conducted using the WGCNA package[73] in R software. All FPKM values data from 19
484 samples were used as the input file, and a power of 8 (soft thresholding power) was used for
485 correlation coefficient analysis to determine the difference between gene correlations.
486 Additionally, we defined clear Pearson correlation coefficient thresholds, such as |r|>0.6 and
487 p<0.05. The DEseq2 and WGCNA Rscript data are supplied in Supplementary Data 1.

488

Gene model prediction and functional annotation

Maker (v 2.31)[74] was used for gene annotation with homology, *de novo* and transcriptome-based prediction evidence. We used the protein sequences of *A. thaliana, C. gigantea, C. roseus, C. canephora, G. sempervirens, N. tabacum, O. pumila, O. sativa, R. serpentina, S. lycopersicum* and known MIA-related genes from the uniport database as homologous species. Genemark-ES (v 4.21)[75] was used for unsupervised self-training of the eukaryote genome with the default criteria. The first round of MAKER analysis was run with EST sequences, homologous species sequences, GeneMark HMMs and Augustus training HMMs of *A. scholaris*. SNAP[76] was subsequently trained with the first round of results. The second round of MAKER was run with the above data and the gff file generated by the first-round analysis.

We aligned the predicted protein sequences against the KEGG[77], COG[78], SwissProt[79], TrEMBL, InterPro, and NR protein databases by BLASTP (E-value ≤1e-05). RNAscan-SE v1.3.1[80] was used for tRNA gene identification. We aligned the assembled genome against the plant rRNA and Rfam[81] databases using BLASTN (E-value ≤1e-05) for rRNA, snRNA and miRNA annotation.

Candidate MIA gene prediction

We downloaded all the identified MIA biosynthesis-related protein sequences (Table S34) as query sequences and performed BLASTP analysis (identity > 40, e-value > 1e-20) with *A. scholaris, A. thaliana, Amborella trichopoda, C. gigantea, C. roseus, C. canephora, G. sempervirens, N. tabacum, O. pumila, O. sativa Japonica Group, Populus trichocarpa, R. serpentina, Rhazya stricta, S. lycopersicum, Solanum tuberosum, Sorghum bicolor,* and *Vitis vinifera* protein sequences. Moreover, the Interpro annotation information (Ipr, Pfam, and GO) was combined to find the best match sequences for the species-specific candidate genes. Gene clusters of *A. scholaris* were predicted by Plantismash (http://plantismash.secondarymetabolites.org/) software with default parameters. The input files include assembled genome with fasta format and the annotation file with gff format.

Gene family analysis and evolutionary tree construction

For the gene family clustering analysis, 15 plant genome sequences, namely those of *C.*

519 *gigantea, C. roseus, R. serpentine*, *A. thaliana, A. trichopoda, R. stricta, C. canephora, G.*
520 *sempervirens, O. pumila, P. trichocarpa, V. vinifera, S. lycopersicum, S. tuberosum*, *S. bicolor*
521 and *O. sativa,* were used with the *A. scholaris* genome. OrthoFinder software (v 2.3.3)[82] was
522 used for gene family cluster identification. The output of OrthoFinder was subsequently
523 passed to identify gene families. If an orthologous group contained more than or equal to
524 eight species then that orthologous gene was considered to be a single-copy ortholog.
525 MAFFT (v 7.310)[83] was used to align single-copy genes in all species. RAxML[84] (v 8.2.4)
526 was used to construct each a gene tree with the PROTCATGTR model. Astral (v 5.5.9)[85] with
527 100 bootstrap replicates was used to construct the species phylogenetic tree.
528 The gene tree about AsSTRs and AsTDCs were also constructed by MAFFT and
529 RAxML softwares, then polished by iTOL (https://itol.embl.de/).
530 MCMCTREE[84] was used to estimate the divergence time between *A. scholaris* and
531 other species with the default parameters. CAFÉ[86] was used to predict the expansion and
532 contraction of gene family numbers by employing a phylogenetic tree and gene family
533 statistics. WGD software[87] was used to perform the Ks distribution analysis.
534
535 **Protein detection**
536 First, proteins from leaf, branch and trunk bark samples were extracted by using the short
537 gradient phenol extraction method[88]. Next, target proteins were detected by using label-free
538 technology with a bottom-up strategy. MaxQuant software was subsequently used to search
539 for proteins in the target database (annotated protein sequence file of *A. scholaris* by
540 ourselves) with the following parameters: Mass accuracy of MS and MS/MS of 20 ppm and
541 0.5 da Orbitrap, respectively.
542
543 **Metabolite detection**
544 The leaf, branch and trunk bark samples were aliquots of 0.5 g each. The samples were
545 triturated with 10 ml of 70% methyl alcohol and incubated under quiescent conditions in an
546 ultrasonic cleaner for 45 mins. Subsequently, the samples were centrifuged at 6000 rpm for
547 15 min, after which the liquid supernatant was collected. The residue was removed, and the
548 previous steps were repeated. The extracted supernatant was combined and drained overnight

549 with a vacuum draining machine. The samples were redissolved in 1 ml of 70% methyl
550 alcohol and placed into an ultrasonic cleaner for 1~2 hours. Next, the samples were
551 transferred to 2 ml centrifuge tubes and centrifuged at 12,000 rpm for 10 minutes. The liquid
552 supernatant was collected and stored at -20°C.

553 We used authentic standards purchased from a certified vendor
554 (https://www.rmuu.com/), including tryptamine and secologan, to carry out targeted
555 metabolomics analysis.

556 Mass spectrometry detection was performed on a quadrupole mass spectrometer, Q-
557 Exactive (Thermo Fisher Scientific) equipped with a heater-electrospray (HESI) source in
558 positive mode for parallel reaction monitoring (PRM) -MS analysis. The conditions of the
559 MS/MS detector were as follows: the flow rate of sheath gas (nitrogen) was 40 arb and the
560 flow rate of auxiliary gas (nitrogen) was 11 arb; the capillary temperature was 320 ℃; the
561 spray voltage was 3.8 kV; the probe heater temperature was 320 ℃ and the S-lens RF level
562 was 50. LC-MS/MS chromatography was performed on a Kinetex® 1.7 μm EVO
563 C18(100×2.1 mm) column (Phenomenex). The column temperature was 30℃. The injection
564 volume was 10 μl. The solvents used were H_2O + 0.1% formic acid as Solvent A and 100%
565 acetonitrile (LC-MS grade) as Solvent B, with a flow rate of 0.3 ml/min. The gradient elution
566 program was as follows: 0 min, 10% B; hold for 2 min; hold for 2 to 5 min, linear gradient
567 to 30 % B; from 5 min to 8 min, linear gradient to 35% B; to 8.5 min, linear gradient to 100%
568 B; hold for 1.5 min; and from 10 min to 11 min, back to 10% B for 2 min to re-equilibrate
569 the column.

570 **Quantification and statistical analyses**

571 Bioinformatic analysis was described in the method details section. The standardized
572 thresholds about screening related modules for tissues of co-expression analysis is $|r|>0.6$, p
573 <0.05. The statistical method of enrichment analysis is $\chi 2$ test. AdjustedPv is a corrected p-
574 value that is obtained by performing false discovery rate (FDR) testing on p-values.
575 AdjustedPv obtained from significance tests are generally considered to have significant
576 statistical differences when AdjustedPv < 0.05

577 # ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (No. 2019YFC1711000); Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (Category C) (Ref No: EF038/ICMS-LMY/2021/SZSTIC). This work is part of the 10KP project (https://db.cngb.org/10kp/)[89]. This work is also supported by High-level talent training support plan of Yunnan Province 'to Le Cheng' 2020 and China National GeneBank (CNGB; https://www.cngb.org/).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCE

1. Baliga MS. (2012). Review of the phytochemical, pharmacological and toxicological properties of *Alstonia Scholaris* Linn. R. Br (Saptaparna). Chinese Journal of Integrative Medicine. 10.1007/s11655-011-0947-0
2. Richardson MA, Sanders T, Palmer JL, Greisinger A, Singletary SE. (2000). Complementary/alternative medicine use in a comprehensive cancer center and the implications for oncology. J Clin Oncol 18, 2505-2514. 10.1200/jco.2000.18.13.2505
3. Salim AA, Garson MJ, Craik DJ. (2004). New Indole Alkaloids from the Bark of *Alstonia scholaris*. Journal of Natural Products 67, 1591-1594. 10.1021/np0498612
4. PAGARIYA A JN, MAHAJAN MP. (2020). Saptaparni a Traditional Medicinal Plant-a Concise Review. International Journal of Biology, Pharmacy and Allied Sciences 9. 10.31032/ijbpas/2020/9.4.5038
5. Khyade M, Kasote D, Vaikos N. (2014). *Alstonia scholaris* (L.) R. Br. and Alstonia macrophylla Wall. ex G. Don: A Comparative Review on Traditional Uses, Phytochemistry and Pharmacology. Journal of ethnopharmacology 153, 1-18. 10.1016/j.jep.2014.01.025
6. Cai X-H, Du Z-Z, Luo X-D. (2007). Unique Monoterpenoid Indole Alkaloids from *Alstonia scholaris*. Organic Letters 9, 1817-1820. 10.1021/ol0705301
7. Yang Z, Sun L, Liang C, Xu Y, Cao J, Yang Y, Gu J. (2016). Simultaneous quantitation of the diastereoisomers of scholarisine and 19-epischolarisine, vallesamine, and picrinine in rat plasma by supercritical fluid chromatography with tandem mass spectrometry and its application to a pharmacokinetic study. Journal of separation science 39, 2652-2660. 10.1002/jssc.201600243
8. El-Askary HI, El-Olemy MM, Salama MM, Sleem AA, Amer MH. (2012). Bioguided isolation of pentacyclic triterpenes from the leaves of *Alstonia scholaris* (Linn.) R. Br. growing in Egypt. Nat Prod Res 26, 1755-1758. 10.1080/14786419.2011.608848
9. Shang JH, Cai XH, Feng T, Zhao YL, Wang JK, Zhang LY, Yan M, Luo XD. (2010). Pharmacological evaluation of *Alstonia scholaris*: anti-inflammatory and analgesic effects. J Ethnopharmacol 129, 174-181. 10.1016/j.jep.2010.02.011
10. Shang JH, Cai XH, Zhao YL, Feng T, Luo XD. (2010). Pharmacological evaluation of *Alstonia scholaris*: anti-tussive, anti-asthmatic and expectorant activities. J Ethnopharmacol 129, 293-298. 10.1016/j.jep.2010.03.029
11. Jagetia GC, Baliga MS. (2006). Evaluation of anticancer activity of the alkaloid fraction of *Alstonia scholaris* (Sapthaparna) *in vitro* and *in vivo*. Phytother Research 20, 103-109. 10.1002/ptr.1810
12. Almagro L, Fernández-Pérez F, Pedreño MA. (2015). Indole alkaloids from *Catharanthus roseus*: bioproduction and their effect on human health. Molecules (Basel, Switzerland) 20, 2973-3000. 10.3390/molecules20022973
13. Liu W, Chen R, Chen M, Zhang H, Peng M, Yang C, Ming X, Lan X, Liao Z. (2012). Tryptophan decarboxylase plays an important role in ajmalicine biosynthesis in *Rauvolfia verticillata*. Planta 236, 239-250. 10.1007/s00425-012-1608-z
14. Lorence A, Nessler CL. (2004). Camptothecin, over four decades of surprising findings. Phytochemistry 65, 2735-2749. 10.1016/j.phytochem.2004.09.001
15. Qu Y, Easson M, Simionescu R, Hajicek J, Thamm AMK, Salim V, De Luca V. (2018).

Solution of the multistep pathway for assembly of corynanthean, strychnos, iboga, and aspidosperma monoterpenoid indole alkaloids from 19E-geissoschizine. Proc Natl Acad Sci U S A 115, 3180-3185. 10.1073/pnas.1719979115

16. Caputi L, Franke J, Farrow SC, Chung K, Payne RME, Nguyen T-D, Dang T-TT, Soares Teto Carqueijeiro I, Koudounas K, Dugé de Bernonville T, et al. (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. Science 360, 1235. 10.1126/science.aat4100

17. Levac D, Murata J, Kim WS, De Luca V. (2008). Application of carborundum abrasion for investigating the leaf epidermis: molecular cloning of *Catharanthus roseus* 16-hydroxytabersonine-16-O-methyltransferase. Plant J 53, 225-236. 10.1111/j.1365-313X.2007.03337.x

18. Tatsis EC, Carqueijeiro I, Duge de Bernonville T, Franke J, Dang TT, Oudin A, Lanoue A, Lafontaine F, Stavrinides AK, Clastre M, et al. (2017). A three enzyme system to generate the strychnos alkaloid scaffold from a central biosynthetic intermediate. Nature Communications 8, 316. 10.1038/s41467-017-00154-x

19. Bayer A, Ma X, Stockigt J. (2004). Acetyltransfer in natural product biosynthesis--functional cloning and molecular analysis of vinorine synthase. Bioorg Med Chem 12, 2787-2795. 10.1016/j.bmc.2004.02.029

20. Kang M, Fu R, Zhang P, Lou S, Yang X, Chen Y, Ma T, Zhang Y, Xi Z, Liu J. (2021). A chromosome-level *Camptotheca acuminata* genome assembly provides insights into the evolutionary origin of camptothecin biosynthesis. Nature Communications 12, 3531. 10.1038/s41467-021-23872-9

21. Rai A, Hirakawa H, Nakabayashi R, Kikuchi S, Hayashi K, Rai M, Tsugawa H, Nakaya T, Mori T, Nagasaki H, et al. (2021). Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis. Nature Communications 12, 405. 10.1038/s41467-020-20508-2

22. Luca VD. Monoterpenoid Indole Alkaloid Biosynthesis. In: Hiroshi Ashihara AC, Atsushi Komamine, ed. Plant Metabolism and Biotechnology2011: 263-291.

23. Reddy DS. (2016). PHYTOCHEMICAL ANALYSIS OF ACTIVE CONSTITUENTS OF ALSTONIA SCHOLARIS AND THEIR CYTOTOXICITY IN VITRO. International journal of pharmaceutical sciences and research 7, 3262-3273. 10.13040/ijpsr.0975-8232.7(8).3262-73

24. Zhang L, Zhang C-J, Zhang D-B, Wen J, Zhao X-W, Li Y, Gao K. (2014). An unusual indole alkaloid with anti-adenovirus and anti-HSV activities from *Alstonia scholaris*. Tetrahedron Letters 55, 1815-1817. 10.1016/j.tetlet.2014.01.122

25. Jagetia GC, Baliga MS, Venkatesh P, Ulloor JN, Mantena SK, Genebriera J, Mathuram V. (2005). Evaluation of the cytotoxic effect of the monoterpene indole alkaloid echitamine in-vitro and in tumour-bearing mice. J Pharm Pharmacol 57, 1213-1219. 10.1211/jpp.57.9.0017

26. Schnoes HK, Biemann K, Mokry J, Kompis I, Chatterjee A, Ganguli G. (1966). Strictamine. The Journal of Organic Chemistry 31, 1641-1642. 10.1021/jo01343a507

27. Hou Y, Cao X, Wang L, Cheng B, Dong L, Luo X, Bai G, Gao W. (2012). Microfractionation bioactivity-based ultra performance liquid chromatography/quadrupole time-of-flight mass spectrometry for the identification of nuclear factor-kappaB inhibitors

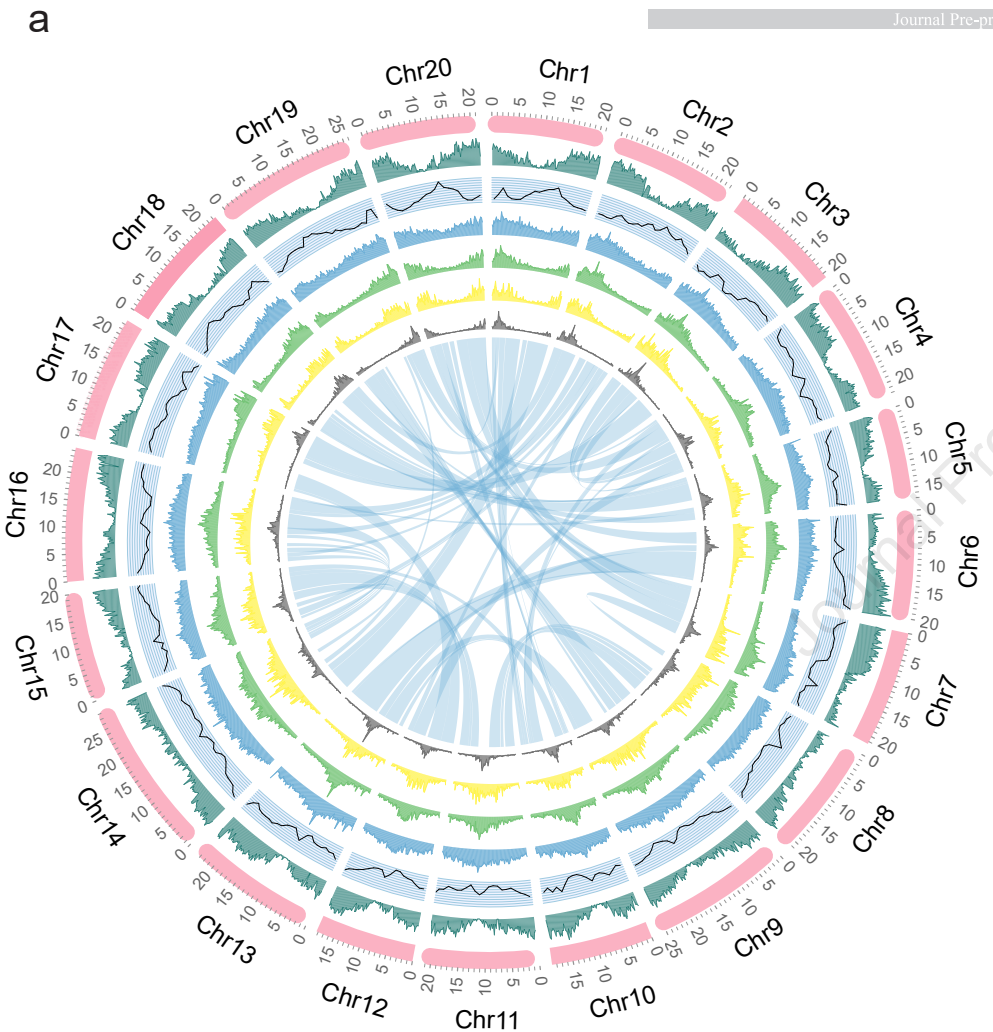
and beta2 adrenergic receptor agonists in an alkaloidal extract of the folk herb *Alstonia scholaris*. J Chromatogr B Analyt Technol Biomed Life Sci 908, 98-104. 10.1016/j.jchromb.2012.10.004
28. Smith AZBaGF. (1963). Akuamma Alkaloids. Part VI.* The Reactions of Picraline. Journal of the Chemical Society(Resumed) 3860. 10.1039/JR9630003850
29. Arai H, Hirasawa Y, Rahman A, Kusumawati I, Zaini NC, Sato S, Aoyama C, Takeo J, Morita H. (2010). Alstiphyllanines E–H, picraline and ajmaline-type alkaloids from Alstonia macrophylla inhibiting sodium glucose cotransporter. Bioorganic & Medicinal Chemistry 18, 2152-2158. 10.1016/j.bmc.2010.01.077
30. Meng W, Ellsworth BA, Nirschl AA, McCann PJ, Patel M, Girotra RN, Wu G, Sher PM, Morrison EP, Biller SA, et al. (2008). Discovery of Dapagliflozin: A Potent, Selective Renal Sodium-Dependent Glucose Cotransporter 2 (SGLT2) Inhibitor for the Treatment of Type 2 Diabetes. Journal of Medicinal Chemistry 51, 1145-1149. 10.1021/jm701272q
31. Subramaniam G, Hiraku O, Hayashi M, Koyano T, Komiyama K, Kam T-S. (2007). Biologically Active Aspidofractinine, Rhazinilam, Akuammiline, and Vincorine Alkaloids from Kopsia. Journal of Natural Products 70, 1783-1789. 10.1021/np0703747
32. Smith JM, Moreno J, Boal BW, Garg NK. (2015). Cascade reactions: a driving force in akuammiline alkaloid total synthesis. Angew Chem Int Ed Engl 54, 400-412. 10.1002/anie.201406866
33. Stöckigt J, Antonchick AP, Wu F, Waldmann H. (2011). The Pictet–Spengler Reaction in Nature and in Organic Chemistry. Angewandte Chemie International Edition 50, 8538-8564. 10.1002/anie.201008071
34. Hemscheidt T, Zenk MH. (1980). Glucosidases involved in indole alkaloid biosynthesis of Catharanthus cell cultures. FEBS Letters 110, 187-191. 10.1016/0014-5793(80)80069-X
35. Luijendijk TJC, Stevens LH, Verpoorte R. (1998). Purification and characterisation of strictosidine β-d-glucosidase from *Catharanthus roseus* cell suspension cultures. Plant Physiology and Biochemistry 36, 419-425. 10.1016/S0981-9428(98)80205-2
36. Barleben L, Ma X, Koepke J, Peng G, Michel H, Stöckigt J. (2005). Expression, purification, crystallization and preliminary X-ray analysis of strictosidine glucosidase, an enzyme initiating biosynthetic pathways to a unique diversity of indole alkaloid skeletons. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1747, 89-92. 10.1016/j.bbapap.2004.09.026
37. Sahu SK, Liu H. (2023). Long-read sequencing (method of the year 2022): the way forward for plant omics research. Molecular Plant 16, 791-793. 10.1016/j.molp.2023.04.007
38. Guo X, Fang D, Sahu SK, Yang S, Guang X, Folk R, Smith SA, Chanderbali AS, Chen S, Liu M. (2021). Chloranthus genome provides insights into the early diversification of angiosperms. Nature communications 12, 1-14. 10.1038/s41467-021-26922-4
39. Wang S, Liang H, Wang H, Li L, Xu Y, Liu Y, Liu M, Wei J, Ma T, Le C. (2022). The chromosome-scale genomes of *Dipterocarpus turbinatus* and *Hopea hainanensis* (Dipterocarpaceae) provide insights into fragrant oleoresin biosynthesis and hard wood formation. Plant biotechnology journal 20(3):538-553. 10.1111/pbi.13735
40. Fan Y, Sahu SK, Yang T, Mu W, Wei J, Cheng L, Yang J, Liu J, Zhao Y, Lisby M.

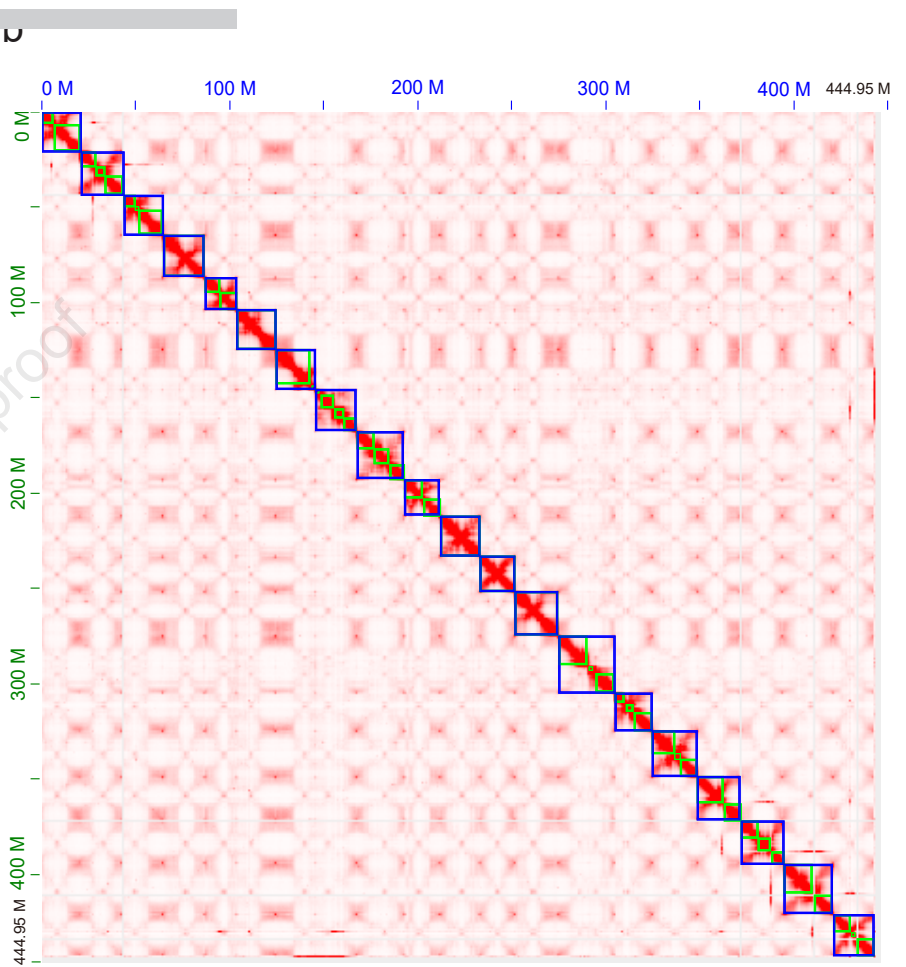
(2021). The *Clausena lansium* (Wampee) genome reveal new insights into the carbazole alkaloids biosynthesis pathway. Genomics 113, 3696-3704. 10.1016/j.ygeno.2021.09.007
41. Fan Y, Sahu SK, Yang T, Mu W, Wei J, Cheng L, Yang J, Mu R, Liu J, Zhao J. (2020). Dissecting the genome of star fruit (*Averrhoa carambola* L.). Horticulture research 7, 1-10. 10.1038/s41438-020-0306-4
42. Sahu SK, Liu M, Yssel A, Kariba R, Muthemba S, Jiang S, Song B, Hendre PS, Muchugi A, Jamnadass R. (2020). Draft Genomes of two Artocarpus plants, Jackfruit (*A. heterophyllus*) and Breadfruit (*A. altilis*). Genes 11, 27. 10.3390/genes11010027
43. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212. 10.1093/bioinformatics/btv351
44. Wu S, Han B, Jiao Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. Molecular Plant 13, 59-71. 10.1016/j.molp.2019.10.012
45. Mahar R, Manivel N, Kanojiya S, Mishra DK, Shukla SK. (2022). Assessment of Tissue Specific Distribution and Seasonal Variation of Alkaloids in *Alstonia scholaris*. Metabolites 12. 10.3390/metabo12070607
46. Pan Q, Mustafa NR, Tang K, Choi YH, Verpoorte R. (2016). Monoterpenoid indole alkaloids biosynthesis and its regulation in *Catharanthus roseus*: a literature review from genes to metabolites. Phytochemistry Reviews 15, 221-250. 10.1007/s11101-015-9406-4
47. De Luca V, Salim V, Levac D, Atsumi SM, Yu F. (2012). Discovery and functional analysis of monoterpenoid indole alkaloid pathways in plants. Methods Enzymol 515, 207-229. 10.1016/B978-0-12-394290-6.00010-0
48. Mohammed AE, Abdul-Hameed ZH, Alotaibi MO, Bawakid NO, Sobahi TR, Abdel-Lateff A, Alarif WM. (2021). Chemical Diversity and Bioactivities of Monoterpene Indole Alkaloids (MIAs) from Six Apocynaceae Genera. Molecules 26. 10.3390/molecules26020488
49. Miettinen K, Dong L, Navrot N, Schneider T, Burlat V, Pollier J, Woittiez L, van der Krol S, Lugan R, Ilc T, et al. (2014). The seco-iridoid pathway from *Catharanthus roseus*. Nature Communications 5, 3606. 10.1038/ncomms4606
50. Franke J, Kim J, Hamilton JP, Zhao D, Pham GM, Wiegert-Rininger K, Crisovan E, Newton L, Vaillancourt B, Tatsis E, et al. (2019). Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis. Chembiochem 20, 83-87. 10.1002/cbic.201800592
51. Nakabayashi R, Mori T, Takeda N, Toyooka K, Sudo H, Tsugawa H, Saito K. (2020). Metabolomics with 15N Labeling for Characterizing Missing Monoterpene Indole Alkaloids in Plants. Analytical Chemistry 92, 5670-5675. 10.1021/acs.analchem.9b03860
52. Sharma A, Amin D, Sankaranarayanan A, Arora R, Mathur AK. (2020). Present status of *Catharanthus roseus* monoterpenoid indole alkaloids engineering in homo- and heterologous systems. Biotechnology Letters 42, 11-23. 10.1007/s10529-019-02757-4
53. Kaushik D, Rana AC, Kaushik P, Sharma N. (2011). *Alstonia scholaris*: It's Phytochemistry and pharmacology. Chronicles of Young Scientists 2. 10.4103/2229-5186.82970
54. Pandey K, Shevkar C, Bairwa K, Kate AS. (2020). Pharmaceutical perspective on bioactives from *Alstonia scholaris*: ethnomedicinal knowledge, phytochemistry, clinical

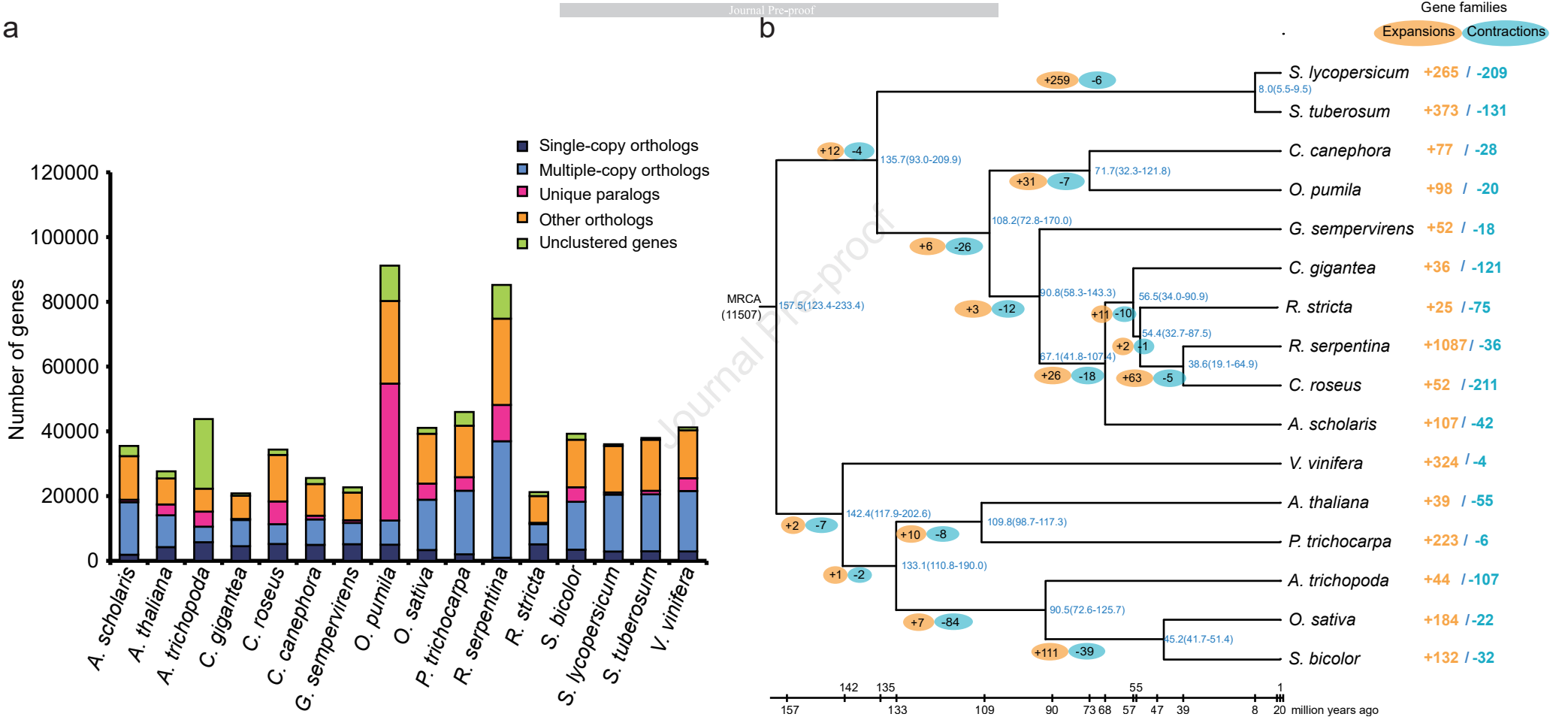
status, patent space, and future directions. Phytochemistry Reviews 19, 191-233. 10.1007/s11101-020-09662-z
55. Sadre R, Magallanes-Lundback M, Pradhan S, Salim V, Mesberg A, Jones AD, DellaPenna D. (2016). Metabolite diversity in alkaloid biosynthesis: a multilane (diastereomer) highway for camptothecin synthesis in *Camptotheca acuminata*. The Plant Cell 28, 1926-1944. 10.1105/tpc.16.00193
56. Murata J, Roepke J, Gordon H, De Luca V. (2008). The leaf epidermome of *Catharanthus roseus* reveals its biochemical specialization. Plant Cell 20, 524-542. 10.1105/tpc.107.056630
57. Salim V, Yu F, Altarejos J, De Luca V. (2013). Virus-induced gene silencing identifies *Catharanthus roseus* 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. Plant J 76, 754-765. 10.1111/tpj.12330
58. Li C, Wood JC, Vu AH, Hamilton JP, Rodriguez Lopez CE, Payne RME, Serna Guerrero DA, Gase K, Yamamoto K, Vaillancourt B, et al. (2023). Single-cell multi-omics in the medicinal plant *Catharanthus roseus*. Nature Chemical Biology. 10.1038/s41589-023-01327-0
59. Wang Z, Xiao Y, Wu S, Chen J, Li A, Tatsis EC. (2022). Deciphering and reprogramming the cyclization regioselectivity in bifurcation of indole alkaloid biosynthesis. Chemical science 13, 12389-12395. 10.1039/d2sc03612f
60. Sahu SK, Thangaraj M, Kathiresan K. (2012). DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. ISRN molecular biology 2012, 205049. 10.5402/2012/205049
61. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems 3, 95-98. 10.1016/j.cels.2016.07.002
62. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. (2017). De novo assembly of the genome using Hi-C yields chromosome-length scaffolds. Science 356, 92. 10.1126/science.aal3327
63. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. (2009). Circos: An information aesthetic for comparative genomics. Genome Research. 10.1101/gr.092759.109
64. Chen N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics Chapter 4, Unit 4.10. 10.1002/0471250953.bi0410s05
65. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462-467. 10.1159/000084979
66. Benson G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573-580. 10.1093/nar/27.2.573
67. Edgar R, Myers E. (2005). PILER: identification and classification of genomic repeats. Bioinformatics (Oxford, England) 21 Suppl 1, i152-158. 10.1093/bioinformatics/bti1003
68. Xu Z, Wang H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research 35, W265-W268. 10.1093/nar/gkm286
69. Kim D, Langmead B, Salzberg SL. (2015). HISAT: a fast spliced aligner with low

memory requirements. Nature methods 12, 357-360. 10.1038/nmeth.3317
70. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. 10.1093/gigascience/giab008
71. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33, 290-295. 10.1038/nbt.3122
72. Love MI, Huber W, Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550-550. 10.1186/s13059-014-0550-8
73. Langfelder P, Horvath S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. 10.1186/1471-2105-9-559
74. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188-196. 10.1101/gr.6743907
75. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33, 6494-6506. 10.1093/nar/gki937
76. Korf I. (2004). Gene finding in novel genomes. BMC Bioinformatics 5, 59. 10.1186/1471-2105-5-59
77. Kanehisa M, Goto S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 27-30. 10.1093/nar/28.1.27
78. Tatusov RL, Galperin MY, Natale DA, Koonin EV. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28, 33-36. 10.1093/nar/28.1.33
79. Bairoch A, Apweiler R. (2000). The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research 28, 45-48. 10.1093/nar/28.1.45
80. Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research 25, 955-964. 10.1093/nar/25.5.955
81. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43, D130-137. 10.1093/nar/gku1063
82. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research 36, 3420-3435. 10.1093/nar/gkn176
83. Katoh K, Standley DM. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30, 772-780. 10.1093/molbev/mst010
84. Alexandros S. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313. 10.1093/bioinformatics/btu033
85. Mirarab S, Reaz R, Bayzid M, Zimmermann T, Swenson M, Warnow T. (2014).

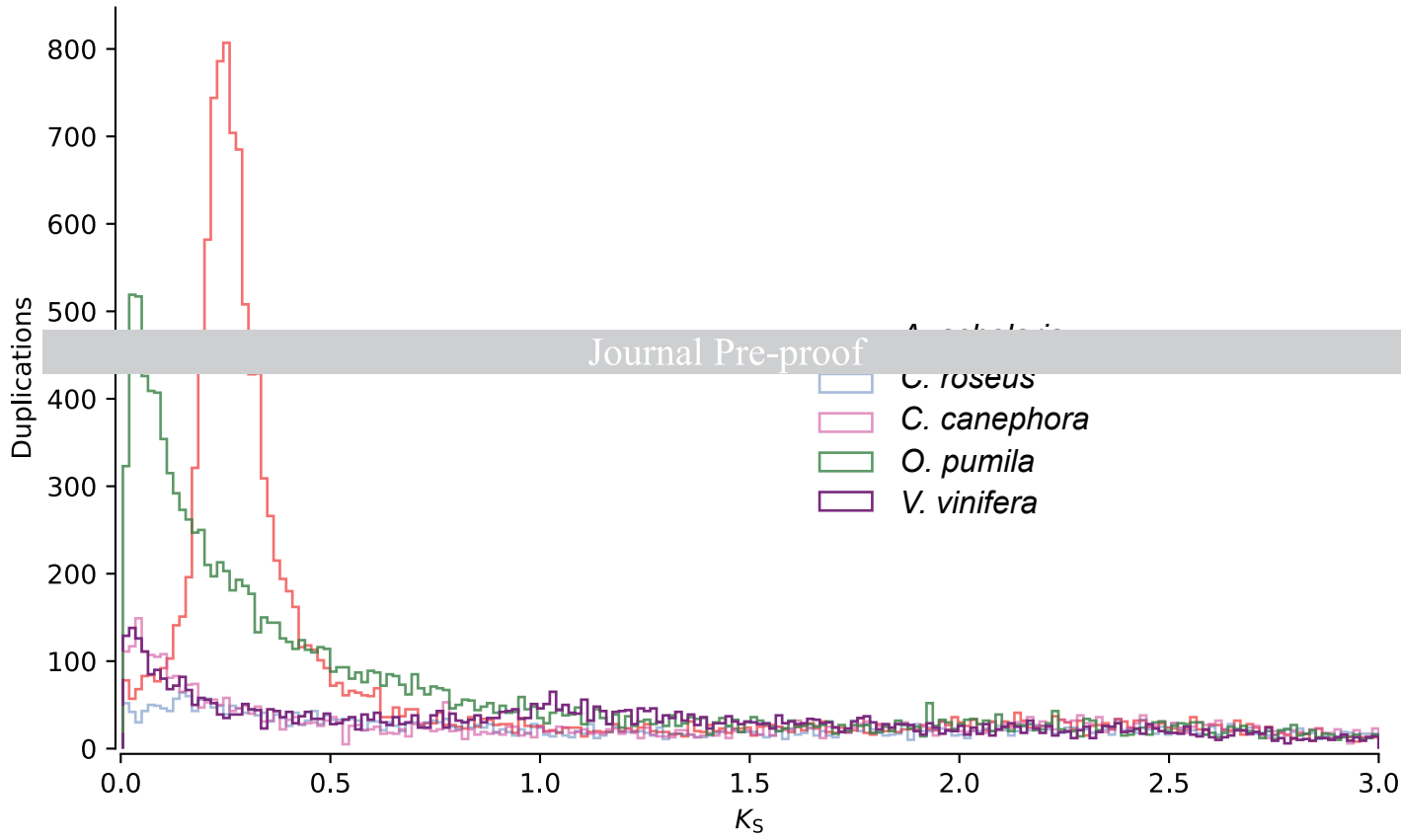
ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics (Oxford, England) 30, i541-i548. 10.1093/bioinformatics/btu462
86. De Bie T, Cristianini N, Demuth JP, Hahn MW. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269-1271. 10.1093/bioinformatics/btl097
87. Zwaenepoel A, Van de Peer Y. (2018). WGD—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35, 2153-2155. 10.1093/bioinformatics/bty915
88. Niu L, Yuan H, Gong F, Wu X, Wang W. (2018). Protein Extraction Methods Shape Much of the Extracted Proteomes. Front Plant Sci 9, 802. 10.3389/fpls.2018.00802
89. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, Li FW, Melkonian B, Mavrodiev EV, Sun W, et al. (2018). 10KP: A phylodiverse genome sequencing plan. Gigascience 7, 1-9. 10.1093/gigascience/giy013



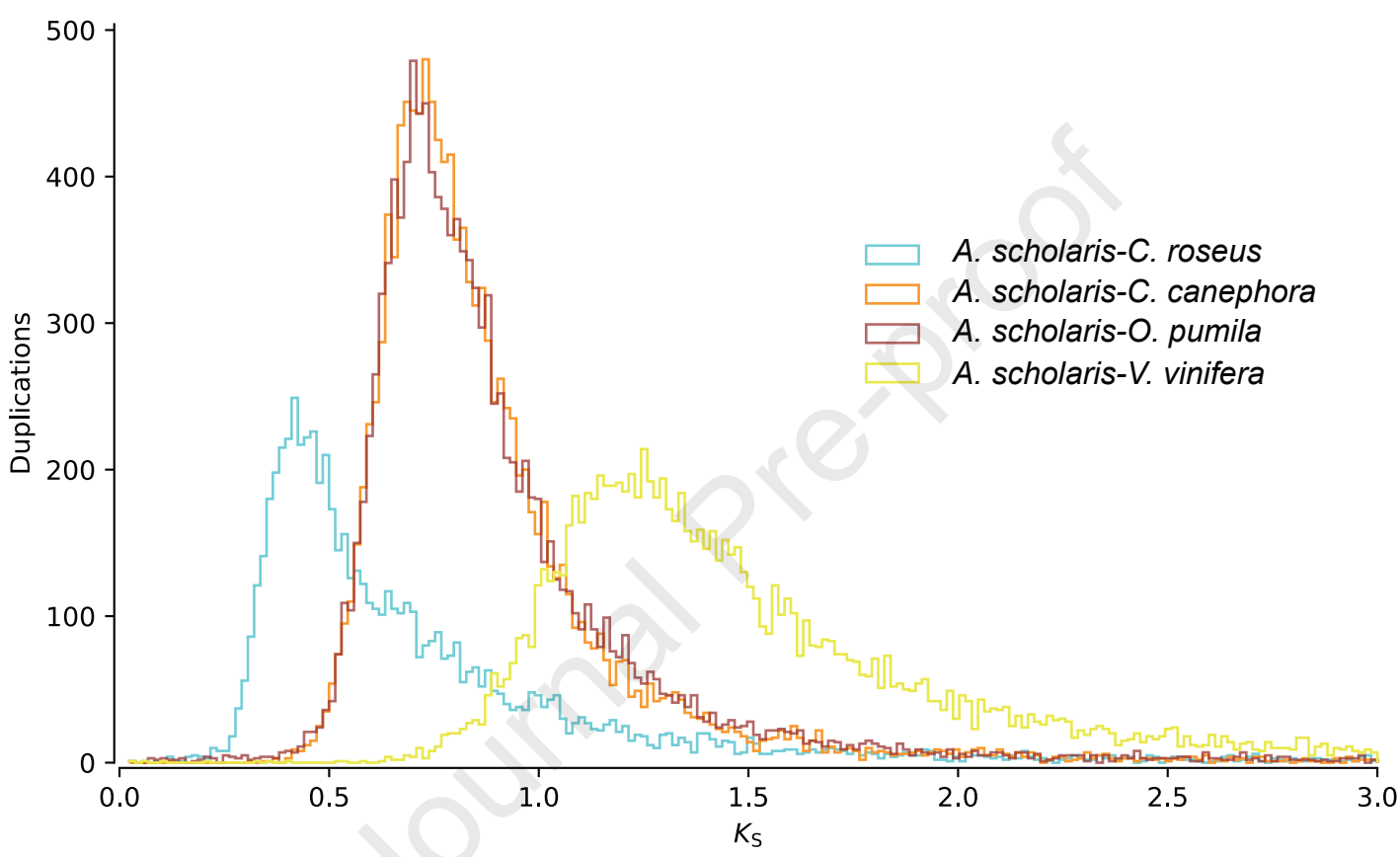




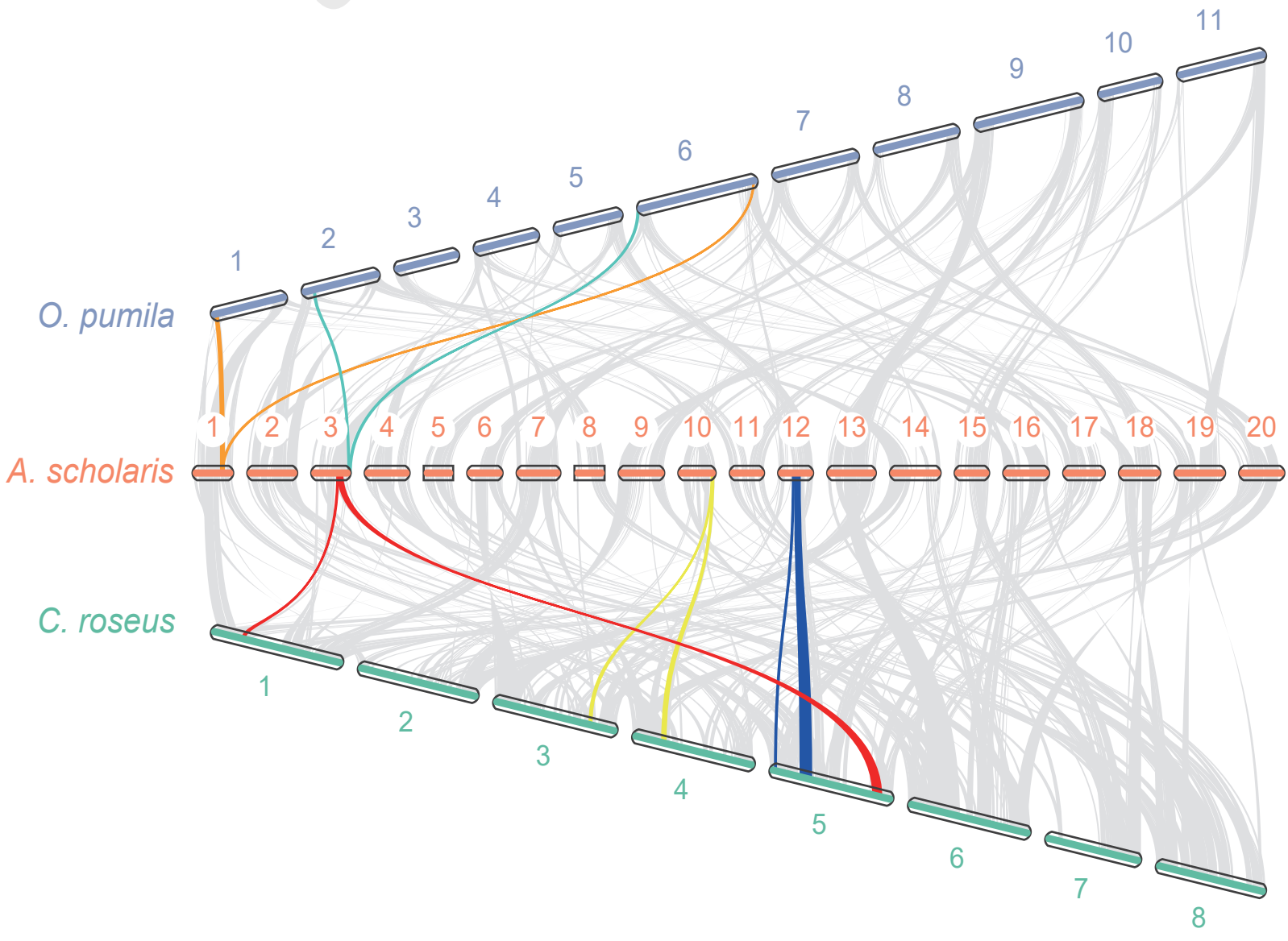
a

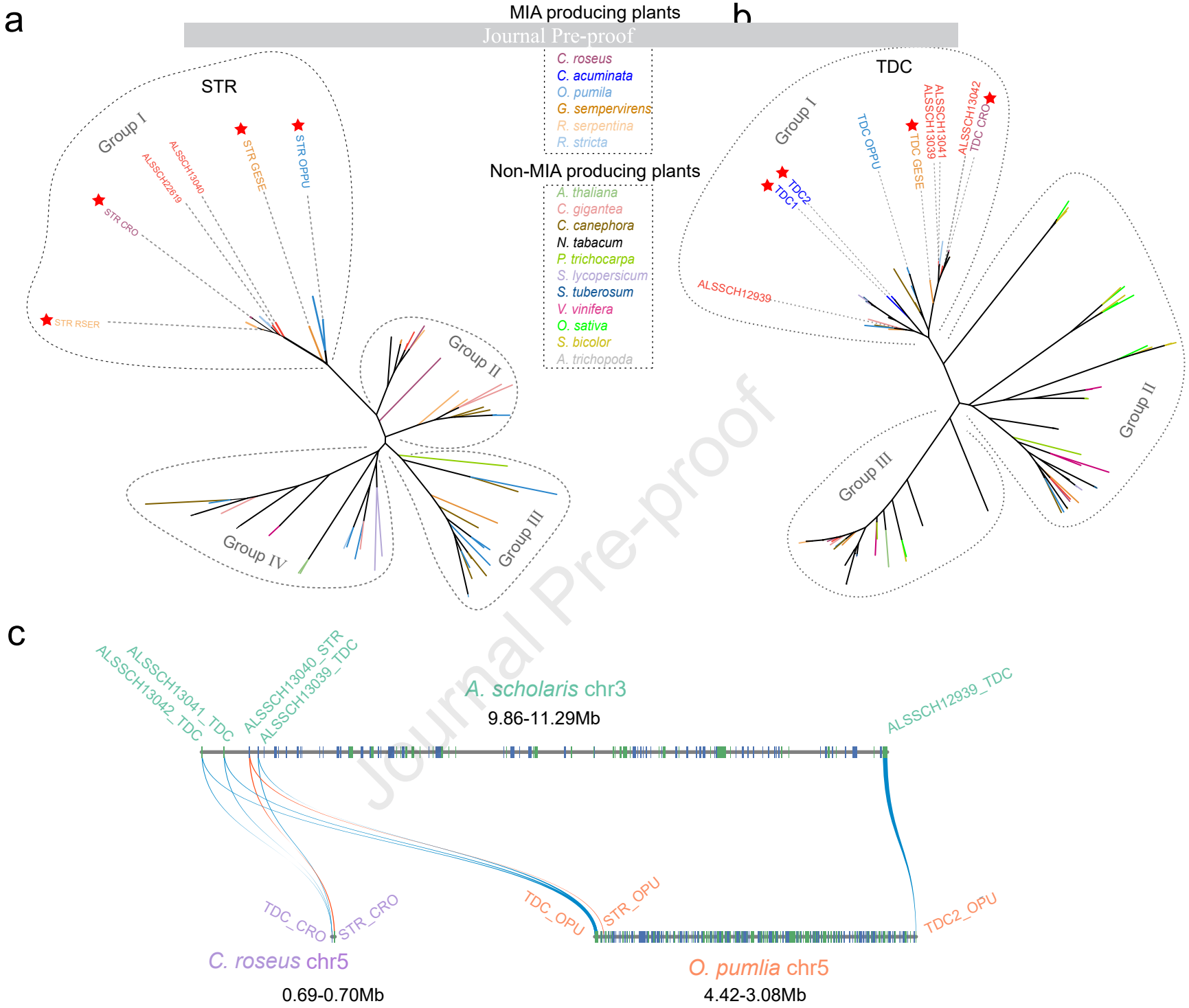


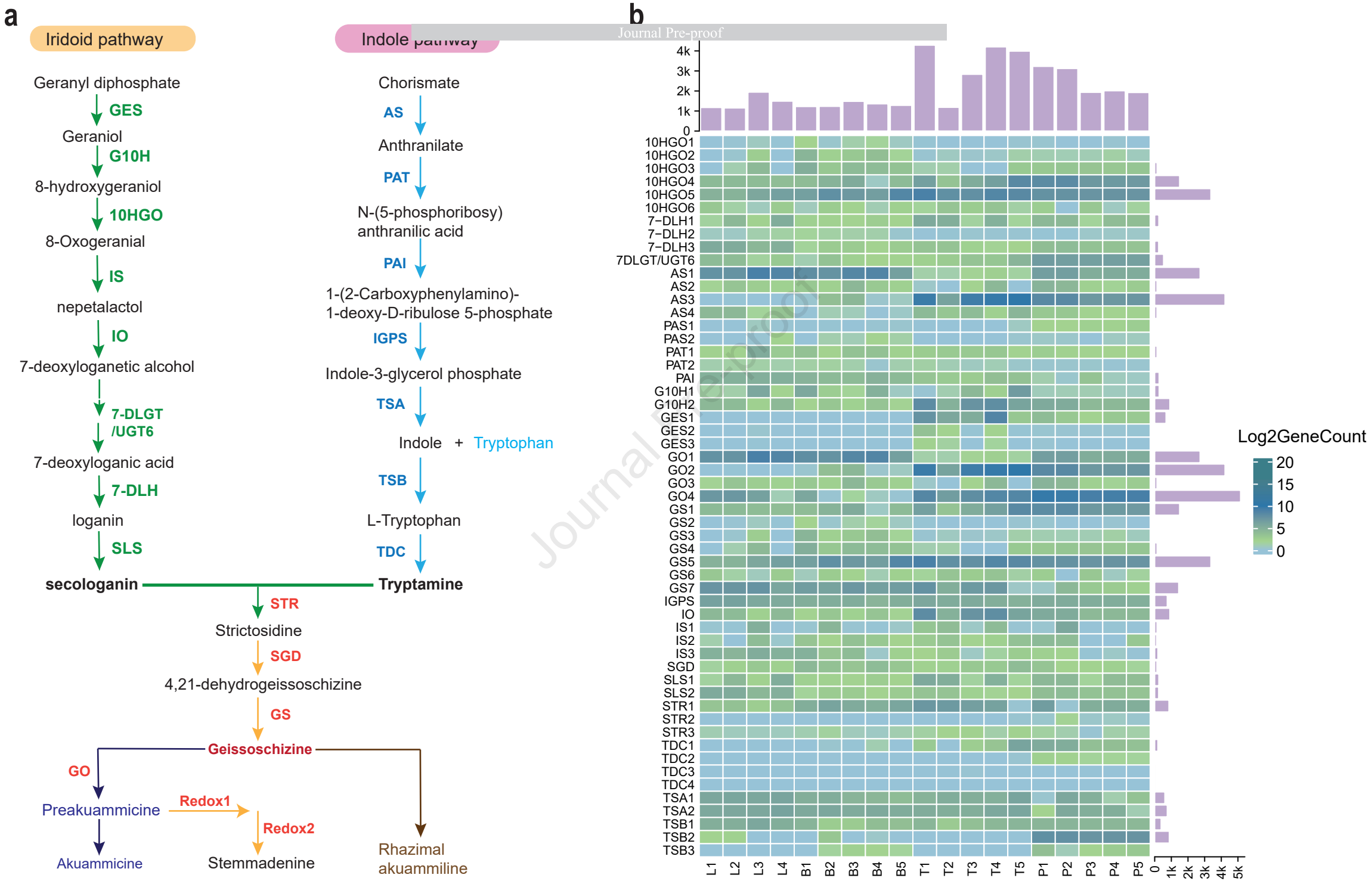
b



c







Highlights

- *Alstonia scholaris* genome illuminates MIA biosynthesis evolution.
- Assembled 444.95 Mb genome into 20 chromosomes with high scaffold quality.
- Identified clusters for key enzymes in MIA biosynthetic pathway.
- Multi-omics data aids understanding of co-expression pattern of MIA genes.

KEY RESOURCES TABLE

The table highlights the reagents, genetically modified organisms and strains, cell lines, software, instrumentation, and source data **essential** to reproduce results presented in the manuscript. Depending on the nature of the study, this may include standard laboratory materials (i.e., food chow for metabolism studies, support material for catalysis studies), but the table is **not** meant to be a comprehensive list of all materials and resources used (e.g., essential chemicals such as standard solvents, SDS, sucrose, or standard culture media do not need to be listed in the table). **However, please note that items in the table must also be reported in the method details section within the context of their use.**

ALL references cited in the key resources table must be included in the main references list. Citations should be formatted as "Author name et al.[#]" (e.g., Smith et al.[1]), with the citation number matching that in the main references list.

Please report the information as follows:

- **REAGENT or RESOURCE:** Provide the full descriptive name of the item so that it can be identified and linked with its description in the manuscript (e.g., provide version number for software, host source for antibody, strain name). See the sample tables at the end of this document for examples of how to report reagents.
    - In the **experimental models sections** (applicable only to experimental life science studies), please include all models used in the paper and describe each line/strain as model organism: name used for strain/line in paper: genotype (e.g., Mouse: OXTR$^{fl/fl}$: B6.129(SJL)-Oxtr$^{tm1.1Wsy/J}$).
    - The **Biological samples section** (applicable only to experimental life science studies) should list all samples obtained in this study or from commercial sources or biological repositories.
    - **You may list a maximum of 10 oligonucleotides or RNA sequences** in the table. If there are more than 10 to report, please provide this information as a supplemental document and reference the file (e.g., See Table S1 for XX) in the key resources table.
    - **Deposited data** should include both newly deposited data from this manuscript and existing datasets that were used in the manuscript.
    - Please include software and code mentioned in the method details or data and code availability section under **software and algorithms**.
    - Any item that does not fit the existing subheadings should be added to the "other" section. **Please do not add your own subheadings.**
- **SOURCE:** Report the company, manufacturer, or individual that provided the item or where the item can be obtained (e.g., stock center or repository).
    - For materials distributed by Addgene, please cite the article describing the plasmid and include "Addgene" as part of the identifier.
    - If an item is from another lab, please include the name of the principal investigator and a citation if it has been previously published.
    - If the material is being reported for the first time in the current paper, please indicate as "this paper."
    - For software, please provide the company name if it is commercially available or cite the paper in which it has been initially described.
- **IDENTIFIER:** Include catalog numbers (entered in the column as "Cat#" followed by the number, e.g., Cat#3879S). Where available, please include unique entities such as RRIDs, Model Organism Database numbers, and accession numbers preceded by database abbreviations such as PDB or

CCDC). Please ensure the accuracy of the identifiers, as they are essential for generation of hyperlinks to external sources when available. For more information about data sharing policies and a list of recommended data repositories for abbreviations, please see the Cell Press Author's guide to data sharing.

- For antibodies, if applicable and available, please also include the lot number or clone identity.
- For software or data resources, please include the URL where the resource can be downloaded.
- When listing more than one identifier for the same item, use semicolons to separate them (e.g., Cat#3879S; RRID: AB_2255011).
- If an identifier is not available, please enter "N/A" in the column.
- ***A NOTE ABOUT RRIDs:*** we highly recommend using RRIDs as the identifier (in particular for antibodies and organisms but also for software tools and databases). For more details on how to obtain or generate an RRID for existing or newly generated resources, please visit the RII or search for RRIDs.

Please use the empty table that follows to organize the information under the provided subheadings and skip sections that are not relevant to your study. To add a row, place the cursor at the end of the row above where you would like to add the row, just outside the right border of the table. Then press the ENTER key to add the row. Alternatively, you can right-click on your mouse and choose Insert > Insert rows above or Insert rows below. Please delete empty rows. Each entry must be on a separate row; do not list multiple items in a single table cell. Please see the sample tables at the end of this document for relevant examples in the life and physical sciences of how reagents and instrumentation should be cited.

TABLE FOR AUTHOR TO COMPLETE

Please do not add custom subheadings. *If you wish to make an entry that does not fall into one of the subheadings below, please contact your handling editor or add it under the "other"* subheading. ***Any subheadings not relevant to your study can be skipped.*** *(**NOTE:** references should be in numbered style, e.g., Smith et al.[1])*

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Bacterial and virus strains		
Biological samples		
Leaf, petiole, branch and Trunk bark of *Alstonia scholaris*.	Ruili Botanical Garden in Ruili, Yunnan Province, China	N/A
Chemicals, peptides, and recombinant proteins		
Critical commercial assays		
Deposited data		
Genome sequencing and assembly data	This study	CNSA with project accession CNP0002381. (https://db.cngb.org/search/?q=CNP0002381)

Experimental models: Cell lines		
Experimental models: Organisms/strains		
Alstonia scholaris ecotype f Yunnan Province, China	Ruili Botanical Garden	N/A
Oligonucleotides		
Recombinant DNA		
Software and algorithms		
Nextdenovo (v 2.3.0)	N/A	https://github.com/Nextomics/NextDenovo
NextPolish (v 1.3.1)	N/A	https://github.com/Nextomics/NextPolish
Juicer (v 1.6)	Durand *et al.*,2016[69]	N/A
Juicerbox	N/A	https://github.com/aidenlab/juicebox
JCVI	N/A	https://github.com/tanghaibao/jcvi
RepeatMasker (v 4.0.6)	Chen et al., 2004[72]	N/A
RepeatProteinMask (v 4.0.6)	Chen et al., 2004[72]	N/A
Tandem Repeats Finder (v 4.07b)	Benson et al., 1999[74]	N/A
Piler (v1.0)	Edgar & Myers, 2005[75]	N/A
LTR-FINDER (v 1.06)	Xu et al., 2007[76]	N/A

Hisat2 (v 2.1.0)	Kim et al., 2015[77]	N/A
StringTie (v 1.3.3b)	Pertea et al., 2015[79]	N/A
WGCNA	Langfelder & Horvath, 2008[81]	N/A
Maker (v 2.31)	Cantarel et al., 2008[82]	N/A
OrthoFinder (v 2.3.3)	Götz et al., 2008[90]	N/A
WGD	Zwaenepoel et al., 2018[95]	N/A
Other		