nature communications

Article

https://doi.org/10.1038/s41467-024-53943-6

Phylotranscriptomics reveals the phylogeny of Asparagales and the evolution of allium flavor biosynthesis

Received: 23 April 2024

Accepted: 29 October 2024

Published online: 08 November 2024

Check for updates

Xiao-Xiao Wang[1,2,8], Chien-Hsun Huang[3,4,8], Diego F. Morales-Briones[5,8], Xiang-Yu Wang[1], Ying Hu[2], Na Zhang[1], Pu-Guang Zhao[1], Xiao-Mei Wei[2], Kun-Hua Wei[2,6] ✉, Xinya Hemu[1], Ning-Hua Tan[1] ✉, Qing-Feng Wang[7] ✉ & Ling-Yun Chen[1] ✉

Asparagales, the largest monocot order, is renowned for its ecological, economic, and medicinal significance. Here, we leverage transcriptome data from 455 Asparagales species to explore the phylogeny of Asparagales. Moreover, we investigate the evolutionary patterns of the genes involved in allium flavor formation. We not only establish a robust bifurcating phylogeny of Asparagales but also explore their reticulate relationships. Notably, we find that eight genes involved in the biosynthesis of allium flavor compounds underwent expansion in *Allium* species. Furthermore, we observe *Allium*-specific mutations in one amino acid within alliinase and three within lachrymatory factor synthase. Overall, our findings highlight the role of gene expansion, increased expression, and amino acid mutations in driving the evolution of *Allium*-specific compounds. These insights not only deepen our understanding of the phylogeny of Asparagales but also illuminate the genetic mechanisms underpinning specialized compounds.

Asparagales consists of approximately 1030 genera and 39,000 species distributed across 14 families, including Orchidaceae, which is one of the largest families of angiosperms according to the Angiosperm Phylogeny Group [APG] IV[1] and the Plants of World Online (POWO; https://powo.science.kew.org/). Known for their diverse applications, Asparagales species serve as vegetables (onions, chives, garlic, asparagus), spices (vanilla), and ornamentals (orchids) and possess medicinal properties (Gastrodiae Rhizoma [Tianma] and Dendrobii Caulis [Shihu] in China and *Cypripedium* species in North America). Despite several studies having investigated the phylogeny of Asparagales[2,3], there is still uncertainty regarding its evolutionary relationships. One area of debate involves the families Ixioliriaceae, Tecophilaeaceae, and Doryanthaceae. Analyses based only on plastid *rbcL* supported a relationship (Ixioliriaceae, (Tecophilaeaceae,

[1]Department of Resources Science of Traditional Chinese Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, 211198 Nanjing, China. [2]National Center for Traditional Chinese Medicine (TCM) Inheritance and Innovation, Guangxi Botanical Garden of Medicinal Plants, 530023 Nanning, China. [3]State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Key Laboratory of Herbage and Endemic Crop Biology, Ministry of Education, School of Life Sciences, Inner Mongolia University, 010021 Hohhot, China. [4]State Key Laboratory of Genetic Engineering, Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering, Institute of Biodiversity Sciences and Institute of Plant Biology, School of Life Sciences, Fudan University, 200438 Shanghai, China. [5]Systematics, Biodiversity and Evolution of Plants, Ludwig-Maximilians-Universität München, 80638 Munich, Germany. [6]Key Laboratory of State Administration of Traditional Chinese Medicine for Production & Development of Cantonese Medicinal Materials, School of Chinese Materia Medica, Guangdong Pharmaceutical University, 510006 Guangzhou, China. [7]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Sino-Africa Joint Research Center, Chinese Academy of Sciences, 430074 Wuhan, China. [8]These authors contributed equally: Xiao-Xiao Wang, Chien-Hsun Huang, Diego F. Morales-Briones. ✉e-mail: divinekh@163.com; nhtan@cpu.edu.cn; qfwang@wbgcas.cn; lychen@cpu.edu.cn

(Doryanthaceae, others)))[4]. In contrast, analyses using plastid genomes[5] and nuclear genes[6–8] suggested alternative topologies. Transcriptome-based phylogenetics are robust approaches[9,10] that can be used to explore the phylogeny of Asparagales better.

Asparagales has a global distribution spanning all continents, with North America and Asia exhibiting the highest species diversity, according to POWO. Several studies investigated the biogeography of Asparagales[11–14]. For example, the ancestral area for Asteliaceae[11], Blandfordiaceae[11], Boryaceae[11], Iridaceae[13], and Orchidaceae[14] was determined to be Australia. In total, biogeographic origins for 11 out of the 14 families within Asparagales are reported to be from Australia, South Africa, or Gondwana. However, the biogeographic origins of the remaining three families, Amaryllidaceae, Asparagaceae, and Asphodelaceae, remain to be determined. In the current phylogenomic era, it is promising to bridge our understanding of evolution by synthesizing evidence from phylogenomics and biogeographic patterns of plants[15]. Hence, additional analyses are imperative to reassess the biogeography of Asparagales.

The distinctive allium flavor is attributed to a wide variety of sulfur-containing compounds generated from S-Alk(en)ylcysteine sulfoxides (CSOs) found in *Allium*, such as garlic and onions. Major CSOs, such as alliin and isoalliin, serve as primary sources of medicinal and flavor compounds in *Allium* species[16]. The biosynthesis pathway for alliin and isoalliin originates from cysteine and involves at least seven steps[17]. When garlic bulbs are crushed, alliin undergoes successive conversion into allicin through the action of the enzyme alliinase[17–19]. Comparative analyses using genomic data from three *Allium* species (garlic [*A. sativum*], green onion [*A. fistulosum*], and onion [*A. cepa*]), *Arabidopsis thaliana*, and ten monocot species have revealed an expansion of *alliinase* and lachrymatory factor synthase (*LFS*), specifically in the three *Allium* species[17]. The study also revealed that *LFS* exists only in the three *Allium* species and is absent in others[17]. This finding is supported by a recent study[20] in which the same three *Allium* species and 86 other species were used. Transcriptome analysis revealed that the *alliinase*, *ATP-sulfurylase* (*ATPS*), and *O-acetylserine (thiol) lyase* (*OASTL*) expanded in Chinese chive (*A. tuberosum*)[21]. *Allium* includes more than 1000 species, and CSOs are natural products characteristic of the genus[16,22]. It remains unclear whether these genes have widely expanded across *Allium* species. Additionally, although previous reviews hypothesized that CSOs exist in all *Allium* species[16,22], whether the CSOs biosynthesis pathway commonly exists in *Allium* species also remains uncertain.

In this study, we used a dataset comprising transcriptome or genome data from 501 samples, with 196 samples from 169 species generated in this study. These samples represented 464 species, covering 13 of the 14 families within Asparagales, 37 *Allium* species, and nine outgroup species. Our study was designed to achieve three objectives: (1) establishing a robust phylogenetic framework for Asparagales; (2) exploring the biogeography of Asparagales; (3) examining the evolutionary patterns of genes involved in the CSOs biosynthesis pathway using high-resolution mass spectrometry, transcriptome-wide characterization, gene expression analyses with additional transcriptome sequencing, and molecular docking. Our study provides valuable insights into this diverse and ecologically significant order of Asparagales through these integrative approaches.

Results and discussion

Samples, phylogenetic analyses, and concordance analyses

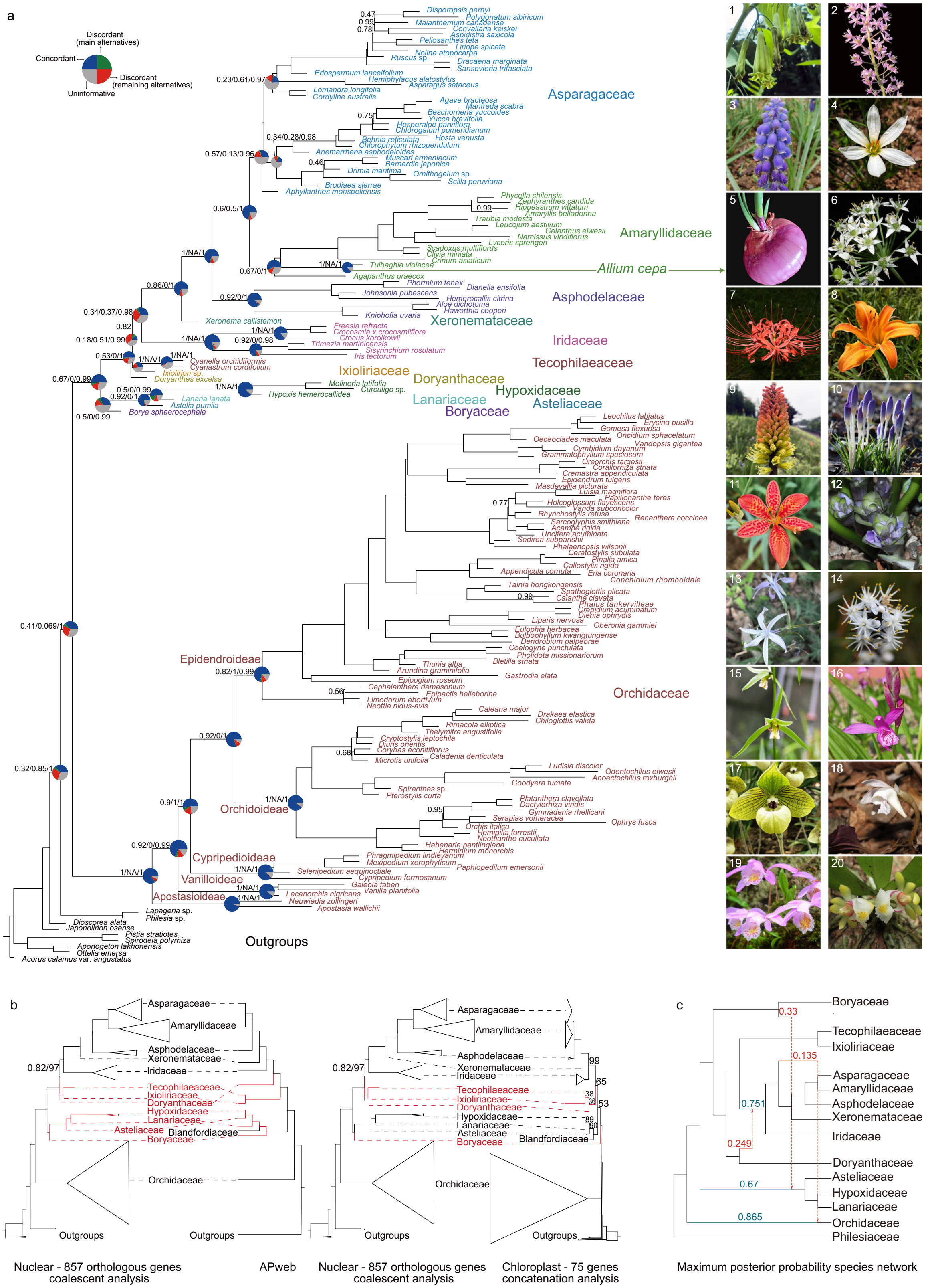
We used 480 de novo assembled transcriptomes and 12 genomes from Asparagales, along with nine outgroups for phylogenetic reconstruction (Supplementary Data 1). We identified 857 nuclear orthologs across the 501 samples using DISCO[23]. Subsequently, coalescence-based (ASTRAL[24]) and concatenation (RAxML[25]) trees were inferred based on these nuclear orthologs. The inter-family relationships depicted in both trees are identical. Most branches in both trees exhibited maximum support [bootstrap support (BS) = 100 or ASTRAL local posterior probability (LPP) = 1] (Fig. 1a; Supplementary Fig. 1). Orchidaceae was resolved as a sister to all other families within Asparagales. The inter-family relationships observed in our trees are similar to those reported in previous studies, such as the Angiosperm Phylogeny Website (APweb)[26], with two major differences. Specifically, our trees recovered a relationship of (Boryaceae, (Asteliaceae, (Lanariaceae, Hypoxidaceae))) (Fig. 1a), whereas APweb proposed a relationship of (Boryaceae, (Lanariaceae, (Asteliaceae, Hypoxidaceae))). Moreover, our trees reveal that (Tecophilaeaceae, Ixioliriaceae) sisters to a subclade comprised Iridaceae, Xeronemataceae, Asphodelaceae, Amaryllidaceae, and Asparagaceae with BS = 97 and LPP = 0.82 (Fig. 1b). Our results differ from the relationship inferred using 353 nuclear genes[8], which suggested a relationship ((Tecophilaeaceae, Ixioliriaceae), Doryanthaceae). This discrepancy was also evident compared to that of APweb and our tree inferred from cpDNA (Fig. 1b and Supplementary Fig. 2).

To assess gene tree discordance, we calculated the internode certainty all (ICA)[27] score and identified the ratio of conflicting/concordant bipartitions through PhyParts[28]. We also calculated the Quartet Concordance (QC), Quartet Differential (QD), and Quartet Informativeness (QI) scores using Quartet Sampling (QS)[29]. The inter-family relationship of Asparagales exhibited varying degrees of support, with two notable exceptions: the clade ((Ixioliriaceae, Tecophilaeaceae), (Iridaceae, others)) and the clade (Iridaceae, others) (Fig. 1a and Supplementary Fig. 3). The discordance in these two clades was evidenced by 65/229 (concordant/discordant genes), 0.16 (ICA), and 0.18/0.51/0.99 (QC/QD/QI) for the former and 143/171, 0.21, and 0.34/0.37/0.98 for the latter (Supplementary Figs. 3 and 4). These findings align with the phylogenetic conflict discussed in the last paragraph, highlighting the inconsistencies among different studies. The monophyly for Orchidaceae, Asphodelaceae, Iridaceae, Hypoxidaceae, and Tecophilaeaceae garnered strong support, with ≥80% of informative gene trees exhibiting concordance (Fig. 1a), ICA ≥ 0.6, and high QS scores (QC/QD/QI ≥ 0.9/NA/1). However, Asparagaceae and Amaryllidaceae had weaker support. For example, Asparagaceae was backed by 136 of the 235 informative gene trees, ICA = 0.37, and QS score = 0.57/0.13/0.96.

The phylogenetic relationships in Asparagales are better explained by both bifurcating trees and networks

We selected 18 clades to evaluate potential reticulation and incomplete lineage sorting (ILS) using PhyloNet[30] and MSCquartets[31], including Asparagales, two clades in Asparagaceae, *Allium* (Amaryllidaceae), Asphodelaceae, two clades in Iridaceae, and 11 clades in Orchidaceae. These clades encompassed most nodes within Asparagales, which displayed either equal support for alternative topologies or clear signals of conflict, indicated by a higher prevalence of discordant genes over concordant genes, coupled with low ICA and QS scores.

For PhyloNet, the maximum posterior probability (MPP) networks with inheritance probability (γ) were analyzed, considering $\gamma < 0.05$ as ILS, $0.05 < \gamma < 0.5$ as introgression or hybridization[32]. Among the 18 clades, ten clades showed reticulation or ILS signals (Supplementary Fig. 5). Specifically, introgression or hybridization was suggested for Asparagales, *Allium* (Amaryllidaceae), Asphodelaceae, *Iris* (Iridaceae), *Dendrobium* (Orchidaceae), and five Orchidaceae clades. Due to introgression and hybridization, the evolutionary relationship of organisms could be explained by combining tree- and network-based inference[33]. The relationship within these ten clades, which have reticulation, should be better explained by phylogenetic networks. However, we recognized that the γ may not accurately reflect historical events due to possible repeated hybridization or horizontal gene transfer[30] during tens of million years of evolution, especially for the family-level relationships of Asparagales. The other eight clades



showed no signals of reticulation (Supplementary Fig. 5). Phylogenetic conflict within these eight clades could be attributed to other factors, such as gene tree estimation error[34]. The relationship within these eight clades could be better represented by “true” bifurcating trees.

Results of the MSCquartets analyses generally aligned with those of PhyloNet, yielding similar conclusions for 17 of the 18 clades examined. For instance, PhyloNet suggested that the ancestor of the clade comprising Asteliaceae, Hypoxidaceae, and Lanariaceae inherited 33% of its genome from an extinct or unsampled taxon, possibly a sister group to Boryaceae, implying a historical hybridization event (Fig. 1c). Meanwhile, MSCquartets revealed that 8.4% of gene trees (indicated by red triangles) in the Asparagales clade rejected the “tree & star” model, with numerous points deviating significantly from the vertices to the centroid (Supplementary Fig. 5), also indicating non-

Fig. 1 | Tree reconstruction of Asparagales. a The species tree of Asparagales with each genus represented by one species. The species tree was inferred using ASTRAL with 857 nuclear genes from 501 samples (Supplementary Fig. 1). Subsequently, one species was selected for each genus from the ASTRAL tree. Pie charts represent gene trees for concordant bipartitions (blue), the most frequent alternative topology (green), remaining alternatives (red), and those uninformative for nodes (BS ≤ 50%) inferred from PhyParts. The numbers next to the branches represent the Quartet Sampling scores (QC, QD, and QI). Pie charts and numbers are displayed only for nodes included in PhyloNet analyses. (1) *Polygonatum cyrtonema* Hua, (2) *Barnardia japonica* (Thunb.) Schult. & Schult.f., (3) *Muscari botryoides* (L.) Mill., (4) *Zephyranthes candida* (Lindl.) Herb., (5) *Allium cepa* L., (6) *Allium tuberosum* Rottler ex Spreng., (7) *Lycoris radiata* (L'Hér.) Herb., (8) *Hemerocallis fulva* (L.) L., (9) *Kniphofia uvaria* (L.) Oken, (10) *Crocus tommasinianus* Herb., (11) *Iris domestica* (L.) Goldblatt & Mabb, (12) *Cyanastrum cordifolium* Oliv., (13) *Ixiolirion tataricum* (Pall.) Schult. & Schult.f., (14) *Borya* sp., (15) *Cymbidium kanran* Makino, (16) *Bletilla striata* (Thunb.) Rchb.f., (17) *Paphiopedilum malipoense* S. C. Chen & Z. H. Tsi., (18) *Changnienia amoena* S.S.Chien, (19) *Pleione formosana* Hayata, (20) *Gastrochilus japonicus* (Makino) Schltr. Photos were taken by Michael L. Moody, Ya-Dong Zhou, Zhong Zhang, Ye-Chun Xu, Xiao-Xiao Wang, Xue-Jia Zhang, Jia-Le Wang, Xiang-Yu Wang, and Ling-Yun Chen. **b** Topological comparison of trees inferred from different datasets or methods. Red lines indicate inconsistent relationships. APweb = Angiosperm Phylogeny Website. **c** The maximum posterior probability species network of Asparagales inferred with PhyloNet. The numbers next to the dashed lines indicate inheritance probabilities (γ). Source data underlying this figure is provided as Source Data file.

tree-like relationships (hybridization and introgression). The only discrepancy between the two methods was observed in the clade Orchidaceae-8 (Supplementary Fig. 5). While PhyloNet detected no reticulation signals within this clade, 11.1% of gene trees in the MSCquartets analysis rejected the "tree & star" model, and approximately 15 points were positioned centrally, suggesting ILS or introgression. This inconsistency, while challenging to discern, is often expected in the detection of ancient ILS or introgression events[30,35].

Overall, the family-level relationships of Asparagales, the species-level relationships of *Allium*, and the genus-level relationships of Asphodelaceae could be explained by networks. The genus-level relationships of Asparagaceae could be explained by bifurcating trees. The species-level relationships of *Iris* (Iridaceae) and relationships of Orchidaceae could be explained by both bifurcating trees and networks (Supplementary Fig. 5). Refer to Supplementary Note 1 for PhyloNet results of other lineages.

The most recent common ancestors of Asparagaceae, Amaryllidaceae, and Asphodelaceae may have originated from Africa during the Late Cretaceous

A time-calibrated tree of Asparagales was constructed using BEAST[36], which incorporated 15 clock-like orthologs and leveraged seven calibration points (Supplementary Data 2). The estimated crown node age of Asparagales was 123.1 million years ago (Ma; 95% highest posterior density [HPD]: 99.2–143.8 Ma; Supplementary Fig. 6 and Supplementary Data 3). This age aligns closely with findings from recent studies, such as 123 Ma[5] and 133 Ma[37]. The estimated crown node age for Orchidaceae was 99.2 Ma, which corresponds well with 101.5 Ma[2] but is older than 83 ± 10 Ma[38].

To explore the biogeographical origins of Asparagales, we carried out an ancestral area reconstruction analysis using BioGeoBEARS[39] with a grafted phylogeny that included 310 taxa representing all 14 families within Asparagales (Supplementary Data 4 and Supplementary Fig. 7). The results indicated that the ancestral regions for Asparagales were likely Asia and Australia (Fig. 2 and Supplementary Data 5). Asparagales comprises two major clades: Clade I, represented by Orchidaceae, and Clade II, formed by the remaining families. Our analyses suggested that Orchidaceae may have originated from the combined regions of Asia and Australia at approximately 99 Ma. A recent study[38], using a broad sampling of orchid lineages, inferred a Laurasian origin of Orchidaceae. Refer to Supplementary Note 2 and Supplementary Fig. 8 for a discussion on the biogeography of Orchidaceae. We did not specifically focus on the origins of Asparagales and Orchidaceae. As recommended[40], further research with additional evidence is necessary to reassess their origins.

For Clade II, the most likely ancestral area of its most recent common ancestor (MRCA) was Australia. Within Clade II, an early dispersal from Australia to Africa occurred for the subclade formed by Asparagaceae, Amaryllidaceae, and Asphodelaceae at approximately 84 Ma (node 1 in Fig. 2). At that time, Australia and Africa were connected via Antarctica[41]. Notably, no previous study has investigated the biogeographic origin of Asparagaceae, Amaryllidaceae, and Asphodelaceae. Our analyses identified Africa and Asia as the ancestral areas for Asparagaceae and Africa as the most likely ancestral area for Amaryllidaceae and Asphodelaceae. Within Asparagaceae, dispersals from Africa or Asia to North America were inferred for the subclade formed by *Agave, Manfreda, Echeandia*, and their relatives (node 2 in Fig. 2; crown age, ca. 20.0 Ma), and the subclade formed by *Bessera*, *Milla*, *Androstephium*, their relatives (node 3 in Fig. 2). From North America, several Asparagaceae genera, such as *Agave* (node 4 in Fig. 2) and *Beaucarnea*, migrated to South America. Following their African origin, Amaryllidaceae and Asphodelaceae underwent dispersal to South America, North America, and Asia. In the case of Amaryllidaceae, the subclade formed by *Eucrosia*, *Caliphruria*, *Griffinia*, etc (node 5 in Fig. 2; crown age, ca. 25.4 Ma) was inferred to be South America. At that time, South America and Africa were already separated by ocean[41], indicating that this separation can be attributed to transoceanic dispersal. From South America, several genera within Amaryllidaceae, such as *Caliphruria*, *Eithea*, and *Haylockia*, migrated to North America. For Asphodelaceae, a dispersal event from Africa back to Australia was inferred for the subclade formed by *Thelionema*, *Herpolirion*, *Xanthorrhoea*, etc (node 6 in Fig. 2). Subsequently, dispersal from Australia to South America (*Eccremis* and *Pasithea*) occurred for several taxa (node 7 in Fig. 2). Refer to Supplementary Data 5 for ancestral areas of all the 14 families.

Genes involved in the CSOs biosynthesis pathway widely exist in Asparagales but have expanded in *Allium*

We employed ultra-high performance liquid chromatography (UPLC) with quadrupole time-of-flight (QTOF) mass spectrometry (MS) to detect eight compounds in the CSOs biosynthesis pathway across nine *Allium* species and seven other species within Asparagales. The results indicated that three compounds upstream of the CSOs biosynthesis pathway—serine, valine, and glutathione—were detected in both *Allium* and non-*Allium* species (Fig. 3a). However, the remaining five compounds were exclusively detected in *Allium* species (Fig. 3a and Supplementary Figs. 9 and 10). Refer to Supplementary Data 6 for details about the compounds. Notably, γ-glutamyl-S-allylcysteine emerged as the most upstream metabolite, specific to *Allium* in the pathway. The gene responsible for synthesizing γ-glutamyl-S-allylcysteine could play a pivotal role in the pathway despite its unreported status.

Two upstream sub-pathways, designated as way 1 (glutathione biosynthesis) and way 2 (valine catabolism), as illustrated in Fig. 4a, along with a downstream sub-pathway, comprised a total of 13 genes (Supplementary Data 7), involved in CSOs biosynthesis. Gene trees were constructed, and gene copy numbers for each species were quantified (Supplementary Figs. 11–15). The Mann–Whitney U test indicated that three genes in way 1 (*OASTL*, *GSH1* [encoding γ-glutamylcysteine synthetase], and *GCL* [encoding γ-glutamylcysteine ligase]), two genes in way 2 (*KARI* [encoding ketol-acid reductoisomerase] and *DHAD* [encoding dihydroxy-acid dehydratase]), and three genes in the downstream sub-pathway (*GGT* [encoding γ-glutamyl transpeptidases], *alliinase*, and *LFS*) underwent expansion in *Allium* species compared to non-*Allium* species ($P < 0.05$; Fig. 4b). Refer to Supplementary Data 8 and 9 for the number of gene copies in other species and in previous studies[17,20]. Furthermore, a comparison between early *Allium* species and non-*Allium* species also demonstrated an expansion in *GGT*, *alliinase*, and *LFS* within *Allium* ($P < 0.05$). A previous study[20] suggested that the pathway may have evolved from an ancient, yet uncharacterized, plant defense system, with *alliinase* and *LFS* experiencing *Allium*-

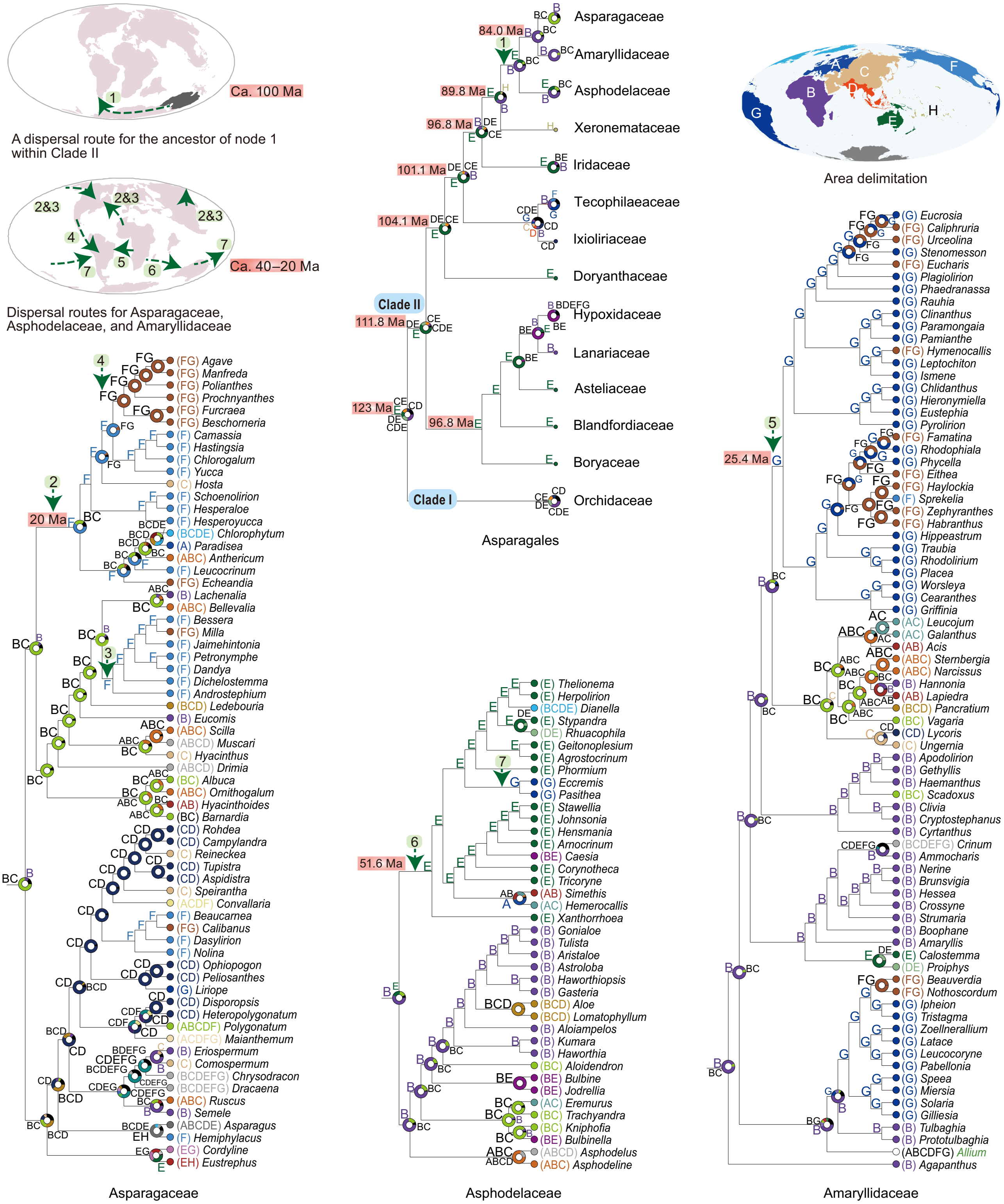


Fig. 2 | Biogeography of Asparagales. Biogeographic analysis was conducted using BioGeoBEARS with a tree that included 310 taxa within Asparagales. The left maps illustrate early major dispersal events within clade II. The arrows and numbers on the maps correspond to the arrows and node numbers on the trees depicted in this figure. The biogeographic patterns of other families are shown in Supplementary Fig. 7. The maps were created using ArcGIS v. 10.8 software. The country boundary data is sourced from the World Geographical Scheme for Recording Plant Distributions (https://github.com/tdwg/wgsrpd). Source data underlying this figure is provided as Source Data file.

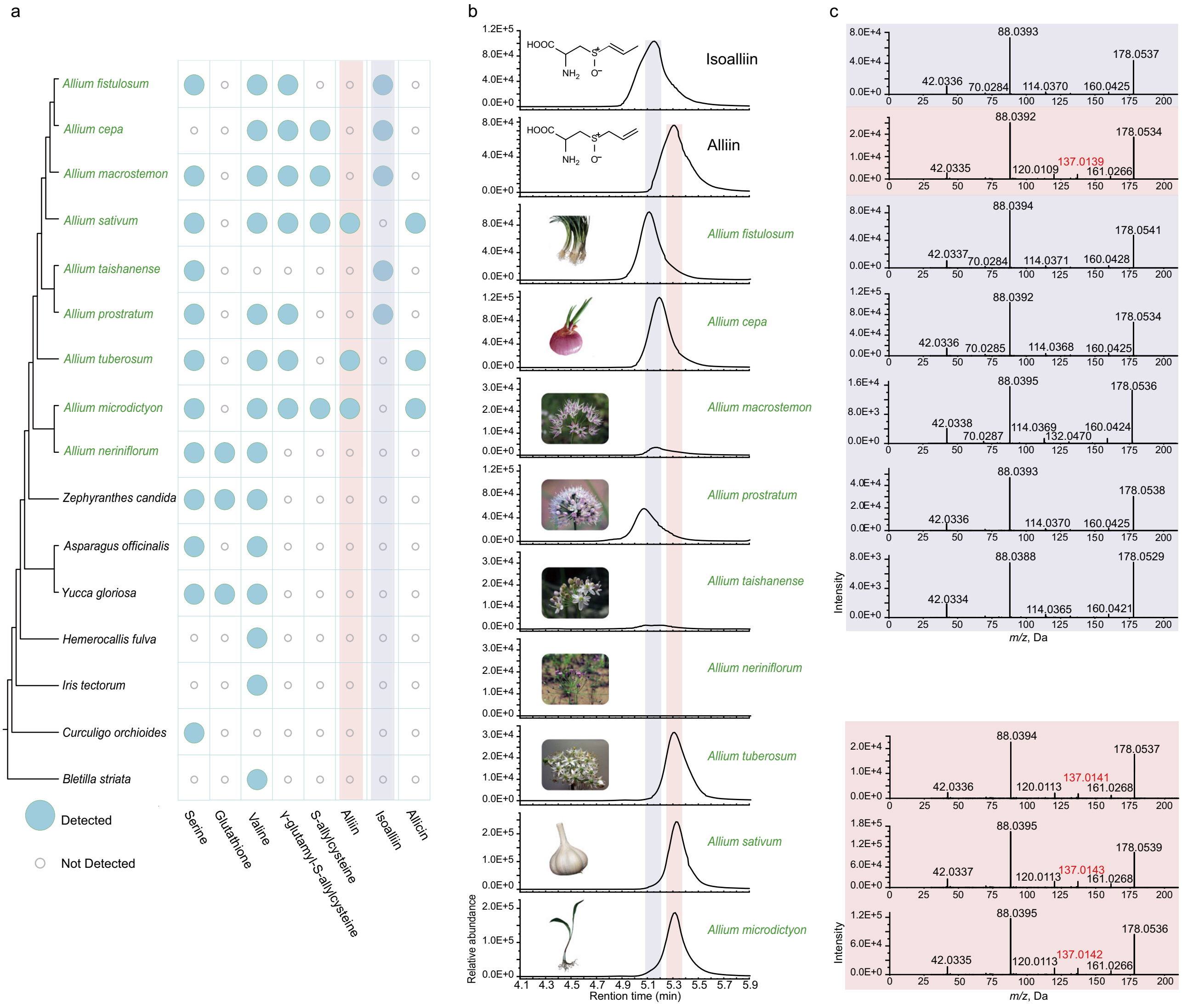
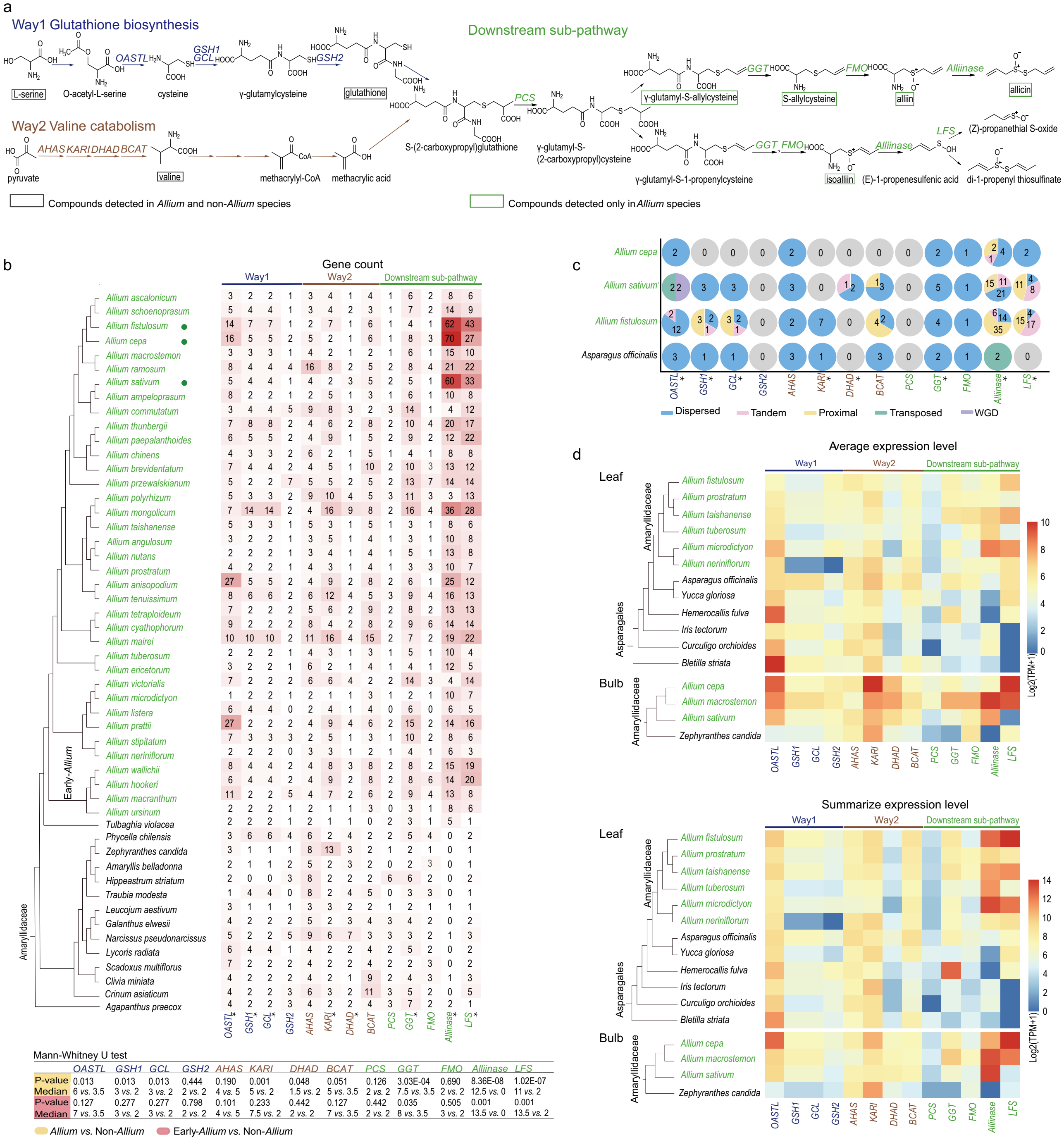


Fig. 3 | Identification of S-Alk(en)ylcysteine sulfoxides (CSOs) using UPLC-QTOF-MS. a A summary of metabolites identified for 16 Asparagales species. **b** Extracted-ion chromatograms of isoalliin and alliin from nine *Allium* species. Plant photos were taken by Bing Liu, Xiao-Wei Xin, Xiao-Xiao Wang, and Ling-Yun Chen. **c** Secondary ion fragments for alliin and isoalliin are depicted as shaded areas on the chromatograms. Red numbers indicate the characteristic secondary fragment ion of alliin.

specific expansion. However, this hypothesis was based on genomic data from only three closely related *Allium* species. By analyzing metabolite data from 16 species and transcriptomes from 110 species, our findings revealed that homologs of genes involved in CSOs biosynthesis are widespread in Asparagales, with eight out of the 13 genes undergoing significant expansion in *Allium*.

Gene trees revealed that six of the eight genes known to have undergone expansion in *Allium*, namely, *GSH1*, *GCL*, *KAR1*, *GGT*, *alliinase*, and *LFS*, duplicated at the MRCA of *Allium* (Supplementary Figs. 11–15), with *alliinase* and *LFS* experiencing at least three times of duplications (Supplementary Fig. 15). In contrast, the five non-expanded genes did not. Moreover, the *alliinase* and *LFS* formed species-specific clusters (Supplementary Fig. 15), indicating their independent expansions, consistent with Liao et al.[17]. Whole-genome duplication (WGD) analyses using Tree2GD[42] and the methods of Yang et al.[43] (*map_dups_mrca.py*) supported a WGD event that occurred at the MRCA of *Allium* (Supplementary Data 10). Three (*OASTL*, *alliinase*, and *LFS*) of the eight expanded genes are indeed included within the 1029 AABB gene clusters that supported the WGD at the MRCA of *Allium*. Analyses using DupGen_finder[44] indicated that the 13 genes derived from dispersed, proximal, tandem duplications with two *OASTL* copies in *A. sativum* derived from WGD (Fig. 4c). In conclusion, the expansion of genes in the CSOs biosynthesis pathway could be attributed to WGD, dispersed, proximal, and tandem duplications.

The gene expansion of *alliinase* and *LFS* may aid *Allium* species in responding to external stimuli[17]; however, the types of stimuli remain unknown. Our divergence time estimation revealed that the diversification of the two genes in *Allium* occurred in recent 10 Ma with a median age of approximately 5 Ma (Supplementary Fig. 16), much younger than the MRCA of *Allium* (40 Ma; Supplementary Fig. 6). Notably, there has been a significant increase in the global insect population over the past 10 million years[45]. Several insect larvae, such as *Delia antiqua*[46], and leaf miners, such as *Phytomyza gymnostoma*, feed on *Allium* species. Considering the concurrent diversification of insects[47,48], it is plausible that the expansion of these two genes serves as a defense mechanism against insect predation, as proposed[49,50]. To investigate whether there is a gene with a timescale of duplication similar to *alliinase* and *LFS*, we explored research on the evolution of metabolites and their corresponding genes across various studies, such as those on steroids[51], benzylisoquinoline alkaloid[52], and



Gene count

	Way1				Way2				Downstream sub-pathway				
	OASTL*	GSH1*	GCL*	GSH2	AHAS	KARI*	DHAD*	BCAT	PCS	GGT*	FMO	Alliinase*	LFS*
Allium ascalonicum	3	2	2	1	3	4	1	4	1	6	2	8	6
Allium schoenoprasum	5	4	4	1	3	4	2	3	1	7	2	14	9
Allium fistulosum ●	14	7	7	1	2	7	1	6	1	4	1	62	43
Allium cepa ●	16	5	5	2	5	2	1	6	1	8	3	70	27
Allium macrostemon	3	3	3	1	4	2	1	2	1	6	1	15	10
Allium ramosum	8	4	4	4	16	8	2	5	2	8	4	21	22
Allium sativum ●	5	4	4	1	4	2	3	5	2	5	1	60	33
Allium ampeloprasum	8	2	2	1	2	5	1	3	1	6	1	10	8
Allium commutatum	3	4	4	5	9	8	3	2	3	14	1	4	12
Allium thunbergii	7	8	8	2	4	6	2	8	2	10	4	20	17
Allium paepalanthoides	6	5	5	2	4	9	1	5	2	9	2	12	22
Allium chinens	4	3	3	2	6	2	1	5	1	4	1	8	8
Allium brevidentatum	7	4	4	2	4	5	1	10	2	10	3	13	12
Allium przewalskianum	5	2	2	7	5	5	2	5	2	13	7	14	14
Allium polyrhizum	5	3	3	2	9	10	4	5	3	11	3	3	13
Allium mongolicum	7	14	14	2	4	16	9	8	2	16	4	36	28
Allium taishanense	5	3	3	1	5	3	1	5	1	3	1	8	6
Allium angulosum	3	2	2	1	3	5	1	4	1	3	1	10	8
Allium nutans	2	2	2	1	3	4	1	4	1	5	1	13	8
Allium prostratum	4	2	2	1	4	3	1	4	1	3	3	10	7
Allium anisopodium	27	5	5	2	4	9	2	8	2	6	1	25	12
Allium tenuissimum	8	6	6	2	6	12	4	5	3	9	4	16	13
Allium tetraploideum	7	2	2	2	5	6	2	9	2	8	2	13	13
Allium cyathophorum	9	2	2	2	8	4	2	8	2	9	6	14	14
Allium mairei	10	10	10	2	11	16	4	15	2	7	2	19	22
Allium tuberosum	2	2	2	1	4	3	1	3	1	5	1	10	4
Allium ericetorum	3	2	2	1	6	2	1	4	1	4	1	12	5
Allium victorialis	7	2	2	2	6	6	4	2	2	14	3	4	14
Allium microdictyon	1	2	2	1	2	1	1	3	1	2	1	10	7
Allium listera	6	4	4	0	2	4	1	4	0	6	0	6	5
Allium prattii	27	2	2	2	4	9	4	6	2	15	2	14	16
Allium stipitatum	7	3	3	3	2	5	1	3	1	10	2	8	6
Allium neriniflorum	2	2	2	0	3	3	1	2	1	4	1	6	3
Allium wallichii	8	4	4	2	4	8	3	8	2	8	2	15	19
Allium hookeri	6	4	4	2	4	9	2	8	2	8	6	14	20
Allium macranthum	11	2	2	5	4	7	2	6	2	9	4	13	8
Allium ursinum	2	2	2	1	2	1	1	3	0	3	1	8	6
Tulbaghia violacea	2	2	2	1	2	1	1	0	2	3	1	5	1
Phycella chilensis	3	6	6	4	6	2	4	2	2	5	4	0	2
Zephyranthes candida	3	1	1	1	8	13	3	2	1	2	1	0	1
Amaryllis belladonna	2	1	1	2	5	2	3	2	0	2	3	0	1
Hippeastrum striatum	2	0	0	3	8	2	2	2	6	6	2	0	2
Traubia modesta	1	4	4	0	8	2	4	5	0	3	3	0	1
Leucojum aestivum	3	1	1	1	3	3	2	2	1	2	1	1	1
Galanthus elwesii	4	2	2	2	5	2	3	4	3	4	2	0	2
Narcissus pseudonarcissus	5	2	2	5	9	6	7	3	3	3	2	1	3
Lycoris radiata	6	4	4	1	4	2	2	4	2	5	2	0	2
Scadoxus multiflorus	7	2	2	2	4	2	2	4	2	2	3	0	2
Clivia miniata	4	2	2	2	4	2	1	9	2	4	3	1	3
Crinum asiaticum	4	2	2	3	6	3	2	11	3	5	2	0	5
Agapanthus praecox	4	2	2	3	4	2	2	4	3	7	2	2	1

Mann-Whitney U test

	OASTL	*GSH1*	*GCL*	*GSH2*	*AHAS*	*KARI*	*DHAD*	*BCAT*	*PCS*	*GGT*	*FMO*	*Alliinase*	*LFS*
P-value	0.013	0.013	0.013	0.444	0.190	0.001	0.048	0.051	0.126	3.03E-04	0.690	8.36E-08	1.02E-07
Median	6 *vs.* 3.5	3 *vs.* 2	3 *vs.* 2	2 *vs.* 2	4 *vs.* 5	5 *vs.* 2	1.5 *vs.* 2	5 *vs.* 3.5	2 *vs.* 2	7.5 *vs.* 3.5	2 *vs.* 2	12.5 *vs.* 0	11 *vs.* 2
P-value	0.127	0.277	0.277	0.798	0.101	0.233	0.442	0.127	0.442	0.035	0.505	0.001	0.001
Median	7 *vs.* 3.5	3 *vs.* 2	3 *vs.* 2	2 *vs.* 2	4 *vs.* 5	7.5 *vs.* 2	2 *vs.* 2	7 *vs.* 3.5	2 *vs.* 2	8 *vs.* 3.5	3 *vs.* 2	13.5 *vs.* 0	13.5 *vs.* 2

🟨 *Allium vs.* Non-*Allium* 🟥 Early-*Allium vs.* Non-*Allium*

Fig. 4 | Copy numbers and expression levels of the genes related to the CSOs biosynthesis. a The CSOs biosynthetic pathway in *Allium*. **b** Number of gene copies for representative species within Amaryllidaceae and results of the two-sided Mann–Whitney *U* test. Green pies next to species names indicate the three species that utilized genomic data, while other species utilized de novo assembled transcriptomes. The Mann–Whitney *U* test was conducted based on gene copy numbers inferred from de novo assembled transcriptomes of 34 *Allium* species and 14 non-*Allium* species. The medians and *P*-values are shown in the table. For example, 0.013 indicates the *P*-value for *Allium* vs. non-*Allium* of gene *OASTL*, while 6 vs. 3.5 indicates the medians for the two groups. **c** Duplication types. Asterisks indicate the genes significantly expanded in *Allium* species. **d** Gene expression level. The expression level was calculated by averaging the expression of all copies(y) within a gene for each species and summarizing the expression of all copies(y) of a gene separately for each species. Source data underlying panel (**d**) is provided as Source Data file.

terpenoids[53]. However, there was no common pattern regarding the timescale of gene family evolution related to plant metabolites. This variability could be explained by the different environments in which plants live and the diverse ecological functions of plant metabolites, such as attracting pollinators and resisting biotic and abiotic environmental stressors[54]. Furthermore, no similar timescale phenomena were observed for genes associated with plant metabolite resistance to insects; however, such a phenomenon was noted for rye *Pm3*- and wheat *Pm8*- like genes, which are related to pathogen resistance[55].

The gene expression levels for each of the 13 genes in the pathway were assessed across nine *Allium* species and seven other species from Asparagales with 48 transcriptomes. Expression was measured using two strategies: averaging the expression of all copies(y) within a gene in a species and summarizing the expression of all copies(y) within a gene in a species (Fig. 4d). Both strategies indicated that the upstream genes in the pathway exhibited relatively low expression levels. Conversely, the downstream genes exhibited relatively high expression levels in *Allium* species, in contrast to those in non-*Allium* species. Interestingly, analyses of genes involved in the biosynthesis of terpenoids, particularly within Lamiaceae, a family renowned for terpenoid richness, showed an opposite pattern, with upstream genes being highly expressed and downstream genes exhibiting low expression levels[56].

A comparison across nine *Allium* species revealed that, except *AHAS* and *BCAT*, 11 of the 13 genes exhibited higher expression levels in the bulb than in the leaf (Supplementary Fig. 17). A previous study reported that garlic bulbs contain a higher concentration of alliin compared to leaves[57]. Our findings revealed that the *FMO* gene, which involves the last step of alliin synthesis, exhibits a higher expression in the bulb than in the leaf, aligning with the results of Yang et al.[57] and Yoshimoto et al.[58]. However, compared to our results, the *FMO* exhibited an opposite expression pattern in two studies[17,59]. This inconsistency could be explained by the factor that the expression of genes in CSOs biosynthesis varied during different growth stages[58]. Further research is needed to explore the relationship among developing stages, organs, and CSOs biosynthesis.

Mutations occurring in the substrate-binding pockets play crucial roles in the CSOs biosynthesis pathway

To uncover gene motif(s) potentially linked to CSOs biosynthesis, motif analyses were conducted on the 13 genes involved in the pathway (Supplementary Data 11). The findings revealed that PCS and LFS exhibited *Allium*-specific gene features. Specifically, the PCS in *Allium* lacks motif 8, which is conserved in other species, and the LFS possesses *Allium*-specific motifs 12 and 14 (Fig. 5a and Supplementary Figs. 18–21). Conversely, the remaining genes did not show motif-level differences between *Allium* and non-*Allium* species.

We then performed multiple sequence alignments for the 13 genes in the CSOs biosynthesis pathway. The results unveiled 34 *Allium*-specific mutations in GSH2, PCS, GGT, FMO, alliinase, and LFS (Supplementary Data 12). Among these mutations, one site in alliinase (Q388) and three sites (F84, F104, and W155) in LFS (Fig. 5b) are likely linked to CSOs biosynthesis. Specifically, these sites in alliinase and LFS are situated within the substrate-binding pockets of *A. sativum* or *A. cepa*, as reported[60,61] (Supplementary Data 13).

Our molecular docking analyses demonstrated that the site Q388 in *A. sativum* alliinase formed hydrophobic (or non-bonded) interactions with the substrate alliin, whereas the corresponding site in non-*Allium* alliinase did not engage in such interactions (Fig. 5c). The sites F84, F104, and W155 in the *Allium cepa* LFS (*Ac*LFS) engaged in hydrophobic (or non-bonded) interactions with the substrate (E)−1-propenesulfenic acid (1-PSA) (Fig. 5d). However, these three sites in non-*Allium* LFSs did not form hydrophobic interactions with the substrate 1-PSA. Previous research utilizing site-directed mutagenesis, protein expression, and activity assays found that mutagenesis at F104 comparatively reduced the activity of *Ac*LFS[61]. The side chain of F84, adjacent to E88—a validated active site—serves as an indicative residue for the binding state of *Ac*LFS[61]. The research[61] also identified E88, Y102, and Y114 as active sites. Our molecular docking analyses confirmed that E88, Y102, and Y114 form hydrogen bonds with 1-PSA. Interestingly, although the three sites E88, Y102, and Y114 are conserved across *Allium* and non-*Allium* LFSs, neither Y102 nor Y114 exhibited hydrogen bonds with the substrate 1-PSA in non-*Allium* species (Fig. 5d). Overall, our findings suggest that mutations at the four identified sites (Q388 in alliinase and F84, F104, and W155 in LFS) may impact protein substrate recognition, consequently influencing metabolite production. Since alliinase and LFS are positioned downstream in the CSOs biosynthesis pathway (Fig. 4a), the four sites are unlikely to be the key enzymes determining whether a plant produces CSOs. Nevertheless, the four mutations in alliinase and LFS might contribute to the diversity of CSOs, such as the formation of alliin and isoalliin. The functions of these four sites require further verification through wet lab experiments.

In summary, our study generated a robust phylogenetic tree of Asparagales, shedding light on the African origins of Amaryllidaceae, Asparagaceae, and Asphodelaceae. We demonstrated that gene expansion, increased expression, and particularly amino acid mutations play pivotal roles in the biosynthesis of allium flavor compounds. The transcriptome dataset generated in this study promises to propel future research in multiple fields significantly. Overall, this study contributes valuable insights into the phylogeny of Asparagales and the evolution of the CSOs biosynthesis pathway.

Methods

A workflow representing the methodological steps employed in this study is presented in Supplementary Fig. 22.

Sampling, sequencing, and transcriptome processing

We collected 196 samples from China between 2018 and 2022 (Supplementary Data 1). Most of these samples were collected from natural populations, with a small subset originating from botanical gardens (Supplementary Data 1). For samples collected during natural populations, we obtained proper permission from the land managers. For samples collected from botanical gardens, we obtained permission from these gardens. RNA-seq reads (2 × 150 bp) were generated for these samples. Sequencing was carried out using the Nova HiSeq 4000 platform or the Beijing Genomic T7 platform. Additionally, 16 samples were sourced from published whole-genome sequencing data, while 289 were derived from RNA-seq reads obtained from NCBI SRA. In total, 501 samples across 464 species were sampled (Supplementary Data 1). Among the 501 samples, 492 samples from 455 species are from Asparagales, representing 160 of the 1144 genera and 13 of the 14 families encompassed by Asparagales. Among the 501 samples, one species from Acorales, four from Alismatales, one from Petrosaviales, one from Dioscoreales, and two from Liliales were outgroups according to the APG IV[1] classification.

The processing of raw reads, assembly, and translation followed the pipeline outlined by Y. Yang and S.A. Smith (https://bitbucket.org/yanglab/phylogenomic_dataset_construction/). Specifically, sequencing errors in raw reads were corrected using Rcorrector v.1.0.4[62]. Adapters and low-quality bases were removed using Trimmomatic v.0.38 with parameters SLIDINGWINDOW:4:15 LEADING:5 TRAILING:4 MINLEN:80[63]. Then, organelle reads were filtered using Bowtie2 v.2.3.5[64] by mapping to Magnoliophyta organelle genomes obtained from the NCBI Organelle Genome Resources database (accessed October 17, 2018). Over-represented reads were detected by FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and removed. Transcriptomes were assembled using Trinity v.2.3.2[65]. Then, the longest transcript within each Trinity “gene” was extracted using the script *get_longest_isoform_seq_per_trinity_gene.pl* in Trinity, followed by translated into coding sequences (CDS) and peptides (PEP) using TransDecoder v.5.5.0[66].

Phylogenomic analyses and discordance assessment

To identify orthologs, we first conducted an all-by-all BLASTN search using NCBI BLAST v.2.9.0+[67] for the CDSs across the 501 samples. Putative homolog groups were clustered using MCL v.1.37[68]. Homolog groups were aligned using MAFFT v.7.407[69] with the settings “--genafpair --maxiterate 1000”, and low occupancy columns were trimmed using Phyutility[70] (all sequences in this study were aligned using MAFFT and trimmed using Phyutility unless otherwise noted). A maximum likelihood (ML) gene tree was constructed for each homolog group with RAxML v.8.2.12[25] (all ML trees in this study were inferred using RAxML with the GTRCAT model and 100 bootstrap replicates unless otherwise

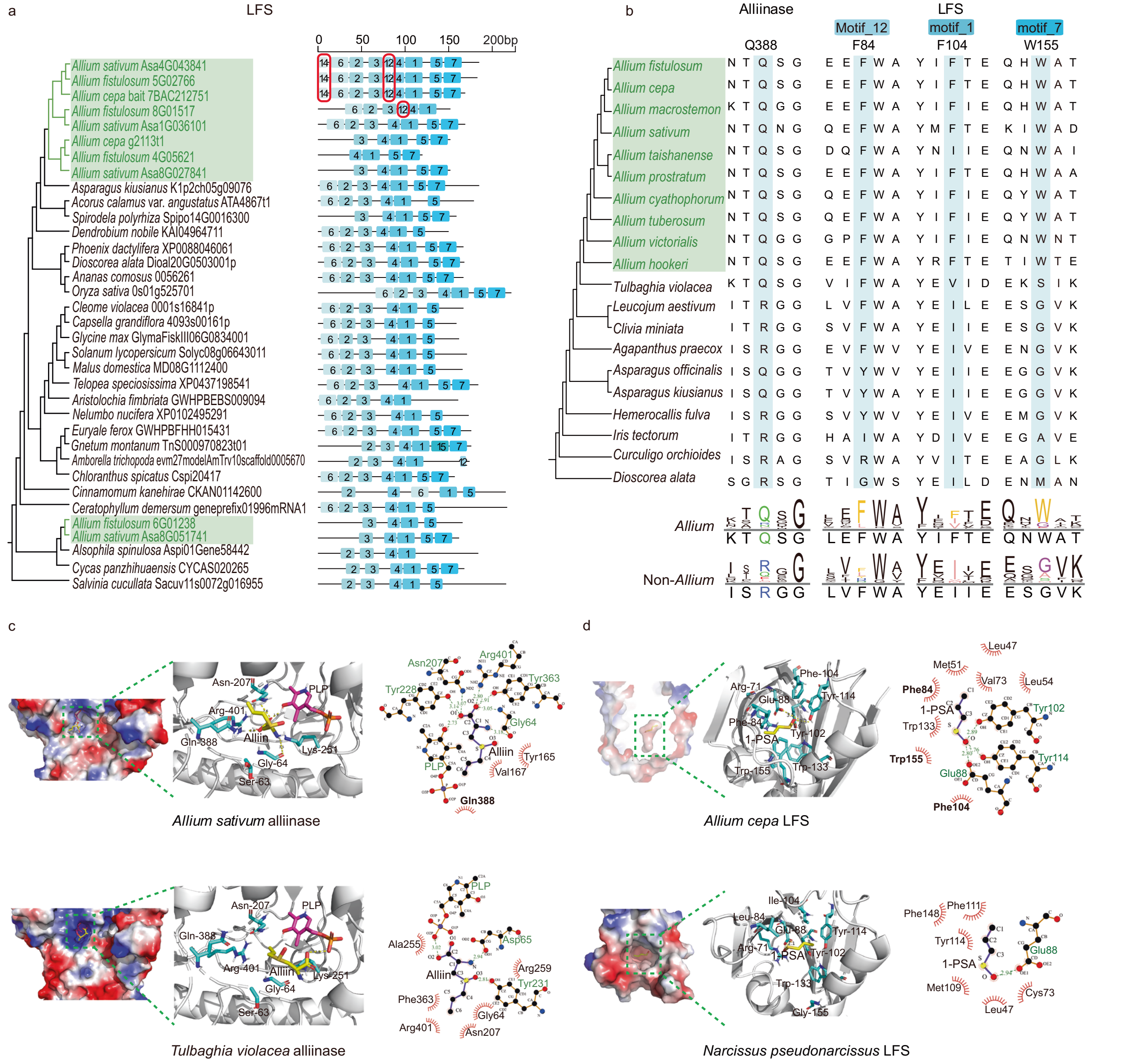

a
LFS
0 50 100 150 200bp
Allium sativum Asa4G043841
Allium fistulosum 5G02766
Allium cepa bait 7BAC212751
Allium fistulosum 8G01517
Allium sativum Asa1G036101
Allium cepa g2113t1
Allium fistulosum 4G05621
Allium sativum Asa8G027841
Asparagus kiusianus K1p2ch05g09076
Acorus calamus var. angustatus ATA4867t1
Spirodela polyrhiza Spipo14G0016300
Dendrobium nobile KAI04964711
Phoenix dactylifera XP0088046061
Dioscorea alata Dioal20G0503001p
Ananas comosus 0056261
Oryza sativa 0s01g525701
Cleome violacea 0001s16841p
Capsella grandiflora 4093s00161p
Glycine max GlymaFiskIII06G0834001
Solanum lycopersicum Solyc08g06643011
Malus domestica MD08G1112400
Telopea speciosissima XP0437198541
Aristolochia fimbriata GWHPBEBS009094
Nelumbo nucifera XP0102495291
Euryale ferox GWHPBFHH015431
Gnetum montanum TnS000970823t01
Amborella trichopoda evm27modelAmTrv10scaffold0005670
Chloranthus spicatus Cspi20417
Cinnamomum kanehirae CKAN01142600
Ceratophyllum demersum geneprefix01996mRNA1
Allium fistulosum 6G01238
Allium sativum Asa8G051741
Alsophila spinulosa Aspi01Gene58442
Cycas panzhihuaensis CYCAS020265
Salvinia cucullata Sacuv11s0072g016955
14 6 2 3 12 4 1 5 7 15
b
Alliinase
LFS
Motif_12
motif_1
motif_7
Q388 F84 F104 W155
Allium fistulosum N T Q S G E E F W A Y I F T E Q H W A T
Allium cepa N T Q S G E E F W A Y I F T E Q H W A T
Allium macrostemon K T Q G G E E F W A Y I F I E Q H W A T
Allium sativum N T Q N G Q E F W A Y M F T E K I W A D
Allium taishanense N T Q S G D Q F W A Y N I I E Q N W A I
Allium prostratum N T Q S G E E F W A Y I F T E Q H W A A
Allium cyathophorum N T Q S G Q E F W A Y I F I E Q Y W A T
Allium tuberosum N T Q S G Q E F W A Y I F I E Q Y W A T
Allium victorialis N T Q G G G P F W A Y I F I E Q N W N T
Allium hookeri N T Q S G E E F W A Y R F T E T I W T E
Tulbaghia violacea K T Q S G V I F W A Y E V I D E K S I K
Leucojum aestivum I T R G G L V F W A Y E I L E E S G V K
Clivia miniata I T R G G S V F W A Y E I I E E S G V K
Agapanthus praecox I S R G G E V F W V Y E I V E E N G V K
Asparagus officinalis I S Q G G T V Y W V Y E I I E E G G V K
Asparagus kiusianus I S Q G G T V Y W A Y E I I E E G G V K
Hemerocallis fulva I S R G G S V Y W V Y E I V E E M G V K
Iris tectorum I T R G G H A I W A Y D I V E E G A V E
Curculigo orchioides I S R A G S V R W A Y V I T E E A G L K
Dioscorea alata S G R S G T I G W S Y E I L D E N M A N
Allium K T Q S G L E F W A Y I F T E Q N W A T
Non-Allium I S R G G L V F W A Y E I I E E S G V K
c
Asn-207 PLP Arg-401 Gln-388 Alliin Lys-251 Gly-64 Ser-63
Asn207 Arg401 Tyr363 Tyr228 Gly64 Alliin Tyr165 PLP Val167 Gln388
Allium sativum alliinase
Asn-207 PLP Gln-388 Arg-401 Alliin Lys-251 Gly-64 Ser-63
PLP Asp65 Ala255 Alliin Arg259 Tyr231 Phe363 Gly64 Arg401 Asn207
Tulbaghia violacea alliinase
d
Phe-104 Arg-71 Glu-88 Tyr-114 Phe-84 1-PSA Tyr-102 Trp-155 Trp-133
Leu47 Met51 Val73 Leu54 Phe84 1-PSA Tyr102 Trp133 Trp155 Tyr114 Glu88 Phe104
Allium cepa LFS
Ile-104 Leu-84 Glu-88 Tyr-114 Arg-71 1-PSA Tyr-102 Trp-133 Gly-155
Phe148 Phe111 Tyr114 Glu88 1-PSA Met109 Cys73 Leu47
Narcissus pseudonarcissus LFS

Fig. 5 | Motifs and amino acid mutations of proteins related to the CSOs biosynthesis. a Schematic representation of the motifs for the LFS. **b** Specific mutations located in the binding pockets of alliinase and LFS. A comparison of the sequence logos of *Allium* species (633 sequences of alliinase and 515 of LFS) and non-*Allium* species (48 sequences of alliinase and 53 of LFS) is shown. **c** 3D structural models and LigPlot+ diagrams for alliinase/alliin. **d** 3D models and LigPlot+ diagrams for LFS/(E)-1-propene−1-sulfenic acid (1-PSA). The amino sites in the binding pockets are shown in light blue, substrates are shown in yellow, and PLP cofactors are shown in pink in the 3D models. Hydrogen bonds are displayed as green dashed lines, while hydrophobic (or non-bonded) interactions are displayed as red opposite arcs in LigPlot+ diagrams. Source data underlying panels (**a**) and (**b**) are provided as Source Data file.

noted). Next, orthologs were inferred using DISCO v.1.3.1[23]. The alignments for the orthologs were concatenated, and then an ML tree of Asparagales was built using the concatenated dataset. Additionally, individual gene trees were generated for each ortholog and a species tree was then inferred from these gene trees with ASTRAL v.5.7.3[24].

Moreover, a phylogenetic tree of Asparagales was constructed using 75 CDSs from chloroplast genomes to gain further insights into the evolutionary relationships within the group. Plastome sequences for 867 species were obtained from GenBank. As plastomes for the families Boryaceae, Ixioliriaceae, Lanariaceae, and Blandfordiaceae are inaccessible from GenBank, five CDSs for these families were accessed from GenBank. After sequence alignment, CDSs were concatenated, and an RAxML tree was constructed (Supplementary Fig. 2).

To investigate discordance, the nuclear ortholog trees were rooted with outgroups. Then, each rooted tree was compared against the ASTRAL tree using PhyParts v.0.0.1[28], with a bootstrap (BS) support cutoff of 50%, to obtain the proportion of concordant/conflicting bipartitions and ICA scores. An ICA close to 1 indicates strong concordance in the bipartition of interest, whereas an ICA closer to 0 indicates equal support for one or more conflicting bipartitions. A negative ICA indicates that the internode of interest conflicts with one or more bipartitions that are more frequent[27]. Additionally, an ICA value close to −1 indicates a lack of concordance for the bipartition of interest[27].

To distinguish nodes lacking support from those exhibiting signals of conflict, Quartet Sampling v.1.3.1[29] was performed with 100 replicates using the concatenated nuclear ortholog alignment and the ML tree inferred from the dataset. A QC value close to 1 suggests that all quartets are concordant, while a QC close to 0 suggests equivocal concordant/discordant quartets. A negative QC suggests that discordant quartets occur more frequently than concordant ones. No QD indicates no alternative topology (i.e., QC = 1). A QD close to 1 suggests that the two alternative topologies are present at equal frequencies, whereas a QD close to 0 suggests a preference for one of the two alternative topologies. Last, a QI value near 1 suggests that all replicates provide informative data, whereas a value close to 0 suggests uncertainty among the replicates.

Phylogenetic networks

To estimate phylogenetic networks that accommodate both reticulation (e.g., hybridization) and ILS, Bayesian inference was conducted with PhyloNet v.3.8.2[30]. Considering computational constraints and our specific interest in clades manifesting a distinct signal of conflict, we streamlined our sampling to 18 clades. The 18 clades included Asparagales (including 14 species), two clades in Asparagaceae (9 and 6 species separately), Amaryllidaceae (12 species), Asphodelaceae (9 species), two clades in Iridaceae (8 and 10 species separately), and 11 clades in Orchidaceae (eight to 14 species separately). We included only orthologs that are found in all species. In this way, we included 163 to 582 orthologs for these groups. Five independent runs were applied for each group. Searches were carried out, allowing up to three reticulation events. MCMC chains of 30 million with sample frequencies of 3000 were carried out for each group. Searches were performed using one cold chain with a temperature of 1.0 and two hot chains with temperatures of 2.0 and 3.0, respectively. Pseudo-likelihood was applied to speed up the searches. PhyloNet[30] calculated inheritance probabilities (γ) that represent the proportion of genes contributed by each parental population to a given hybrid node. The first 25% of the iterations were set as burn-in. Moreover, we used the function *quartetTreeTestInd* in the MSCquartets v.2.0[31] with the "T3 model" to evaluate the level of ILS within the 18 clades.

Divergence time estimation and biogeographic inference

Divergence time estimation was accomplished through BEAST v.2.6.3[36], utilizing seven calibration points. These calibration points included three within Asparagales, three within outgroups, and one at the root of all samples (Supplementary Data 2). Calibrations were only applied to the nodes with high support in phylogenetic analyses. Due to the vast size of the phylogenomic dataset, fifteen clock-like orthologs were selected using SortaDate[71] with the parameter "--order 1, 2, 3". In BEAST, the Gamma site model was applied with estimated Substitution Rate, estimated Proportion Invariant and Subst model, Relaxed Clock Log-Normal model, and Log-Normal priors. Six independent analyses were conducted, each running for one billion generations and sampled every 2000 generations. During the MCMC chain, the tree was fixed with the ML tree inferred from the 857 orthologs. The effective sample size scores for all relevant estimated parameters were checked to ensure values ≥200 using Tracer v.1.7.1[72]. The first 10% of trees were discarded as burn-in, and the remaining trees were used to generate a summary tree with TreeAnnotator v.2.6.3[36].

For biogeographic inference, a grafted phylogeny consisting of 310 Asparagales taxa was used. For the construction of the grafted phylogeny, we used the inter-family relationships depicted in the Asparagales species tree (Fig. 1a) as a backbone. Orchidaceae has more than 26,000 species and ca. 705 genera (the Plants of the World Online [http://www.plantsoftheworldonline.org/]). Sampling all these species or genera within Orchidaceae can be quite challenging due to their vast diversity. To represent Orchidaceae, 16 genera and 51 subtribes within Orchidaceae with relationships compiled from previous studies[2,38] were added to the backbone. Each of the family Amaryllidaceae, Asparagaceae, Asphodelaceae, and Iridaceae has more than 40 genera (World Flora Online: http://www.worldfloraonline.org). To represent the genera within these families, we first constructed ML trees using ITS sequences obtained from GenBank for the four families separately. Then, we added the genera that only existed in these ML trees to the reported phylogenies of the four families (Supplementary Data 4). The natural distribution of each taxon was accessed from the POWO (retrieved 12 October 2023). According to our phylogenetic analyses and the classification of POWO, all the 14 families and genera for which we proposed possible dispersal routes are monophyletic (Fig. 2). Eight geographical areas were designated: Europe (A), Africa (B), mainland Asia (C), South Asian islands (D), Australia and Papua New Guinea (E), North America (F), South America (G), and Pacific (H), similar to the study on Orchidaceae[14]. However, unlike the study[14], we have distinguished Europe and mainland Asia as separate regions, reflecting their significant geological separation during the period (100–40 Ma) when the Asparagales families originated. This approach aligns with methodologies adopted in other biogeographic studies, including those on *Dryopteris*[73]. Ancestral areas were reconstructed using BioGeoBEARS[39] with the best-fit model BAYAREALIKE + J. The maximum number of areas allowed at each node was set to three.

Compound measurement

The qualitative identification of eight compounds in the CSOs biosynthesis pathway was conducted across nine *Allium* species and seven other species within Asparagales. Bulbs or leaves were used. Each sample weighing 0.10–0.15 g was placed in a 2 mL centrifuge tube with two 3 mm steel balls and 1000 μL of a methanol/deionized water mixture (7:3, v/v). After being rubbed with a tissue grinder, the samples were homogenized for 30 seconds, sonicated for 30 minutes, and centrifuged at 13,000 rpm for 10 minutes at room temperature. The resulting extracts were filtered through a 0.22-μm filter and stored at −20 °C for later analysis.

Eight compounds were identified. Among them, serine, glutathione, valine, S-allylcysteine, alliin, and isoalliin were identified by comparing their retention time and *m/z* values of fragment ions to authentic standards. Isoalliin was identified at a retention time of approximately 5.1 min, while alliin was detected at approximately 5.3 min (Fig. 3b). Additionally, alliin presented a secondary ion fragment with an *m/z* value of approximately 137.0139 (Fig. 3c), a feature not observed for isoalliin. γ-glutamyl-S-allylcysteine and allicin were confirmed by comparing their *m/z* values of fragment ions to those in PubChem (https://pubchem.ncbi.nlm.nih.gov/) and MASSBANK. Glutathione and alliin were purchased from Shanghai Yuanye Biotechnology Company, Shanghai, China. S-allylcysteine and isoalliin were purchased from Caoyuankang Biotechnology Company, Chengdu, China. Serine and valine were purchased from Energy Chemical, Saen Chemical Technology Company, Shanghai, China. All reagents are analytical grade (≥98.0% pure). A concentration of 25–30 μg/mL for each standard was used.

Qualitative identification was performed on the SCIEX ZenoTOF™ 7600 (SCIEX, Foster City, CA, USA) with the ESI source coupled to a UPLC system (ExionLC AE system, Shimadzu, Japan). The UPLC conditions followed Liao et al.[17] with slight changes. The chromatographic separation was performed on a Waters BEH amide column (100 × 2.1 mm, 1.7 μm) maintained at 20 °C with a flow rate of 0.6 mL/min. The mobile phase A contained deionized water with 0.5% formic acid (v/v), and mobile phase B contained acetonitrile with 0.5% formic acid (v/v). The following gradient elution was used: 0–4 min, 10–15% A; 4–8 min, 15–60% A; 8–15 min, 10% A; and an injection volume of 1 μL. For QTOF, the mass spectrometer was operated in the positive ESI mode with a

SCIEX Turbo V™. The PeakView v.1.2 (AB SCIEX, Foster City, CA, USA) was used to analyze the data obtained from the information-dependent acquisition (IDA) method. The spray voltage and ion source temperature were set to 5.5 kV and 450 °C, respectively. The declustering potential (DP) was set to 60 V. The ion source gas 1, ion source gas 2, curtain gas, and CAD gas were set to 50, 50, 35, and 9 psi, respectively. The MS/MS spectra were generated in product ion scan mode at a collision energy (CE) of 15 V with a CE spread of 5 V. Parent ions scan ranged from *m/z* 50 to 800 Da with a 0.15 s accumulation time, and the product ions scan ranged from 30 to 800 Da with a 0.045 s accumulation time. Three biological replicates were used for all accessions.

Evolution of genes in the CSOs biosynthesis pathway

Thirteen genes involved in the pathway of CSOs biosynthesis were selected (Fig. 4a) according to previous studies[16,17,19,20]. To investigate gene copy numbers, we compiled a dataset comprising transcriptomes from 110 species. Each gene within the pathway was individually utilized as a "bait" to search against the dataset with SWIPE v.2.1.0[74]. A maximum of 100 hits was retained for each species. Candidate genes were filtered using Pfam domains with InterProScan[75] (Refer to Supplementary Data 7 for information about bait genes and Pfam domains). Orthologs for each gene family were classified based on their ML trees. Each gene family may include multiple ortholog groups. Orthologs that included the bait genes and had no gene duplication at the MRCA of Asparagales were treated as the orthologs most likely related to CSOs biosynthesis. Then, the number of gene copies in each species was determined from these ortholog groups (Supplementary Figs. 11–15). A two-sided Mann–Whitney U test was conducted using SPSS v.22 (IBM Corp. Released 2013). The mode of gene duplication for these genes was investigated using DupGen_finder-unique[44]. In addition, the WGD events within Amaryllidaceae were investigated. Whether the WGD events led to gene copy number increase was checked. WGD events within Amaryllidaceae were investigated using Tree2GD v.1.0.37[42] and the script *map_dups_mrca.py*[43]. These two methods calculate the number and proportion of duplicated gene clusters for each node within the Amaryllidaceae phylogeny. Nine species within Amaryllidaceae were used in Tree2GD. A duplicated gene cluster in a clade, which retains two subclades, indicates a signal of WGD event (AABB duplication)[76]. For *map_dups_mrca.py*, gene trees of homologs inferred from 501 samples were mapped to the Asparagales species tree, and the proportion of duplicated genes was counted. We employed a criterion to propose a WGD event, requiring ≥200 AABB gene clusters (inferred from Tree2GD) and ≥20% of duplicated gene clusters inferred from *map_dups_mrca.py* for a given clade. Under this criterion, we identified a WGD event at the MRCA of *Allium*, consistent with Hao et al.[20]. 1029 AABB gene clusters were found to support the WGD at the MRCA of *Allium*. Subsequently, we checked whether the 13 genes in the CSOs biosynthesis pathway were included in these 1029 gene clusters by comparing if there were identical sequence names between gene clusters and trees of the 13 genes. Divergence time for *alliinase* and *LFS* were separately estimated using TreePL v.1.0[77] with ML trees and branch length estimated using RAxML. Five calibration points were used (Supplementary Data 2).

To investigate the gene expression level, RNA sequencing was conducted for 16 species with three replicates. Among the 16 species, nine are from *Allium*, and seven are from other lineages within Asparagales. In total, 48 transcriptomes were generated. De novo transcriptomes for each species were assembled using Trinity v.2.3.2. The expression level of each gene was measured with Transcript per million (TPM) by aligning RNA-seq reads to the transcriptome of each species using Salmon v.0.9.1[78]. Then, the TPM for genes in those ortholog groups was extracted.

We examined the motifs of 13 genes within the CSOs biosynthesis pathway. We compiled a dataset consisting of 42 species (Supplementary Data 11), for which whole-genome sequencing data were available. Each gene in the pathway was individually utilized as a "bait" to search the dataset using SWIPE v.2.1.0 to identify homologs. Please refer to Supplementary Data 7 for details regarding the bait genes. Motifs were predicted using MEME v.5.5.5[79], and 15 motifs were allowed for each gene.

Furthermore, the active residues for the five genes in the downstream sub-pathway were examined. Specifically, the protein structures of PCS and GGT from *Allium* were predicted using AlphaFold2 online (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). The protein structures of FMO, alliinase, and LFS from *Allium* were accessed from the RCSB PDB (https://www.rcsb.org/) with entry IDs 6WPU, 1LK9, and 5GTF, respectively. Then, the active amino acids for PCS, GGT, and FMO were predicted using AutoDockTools v.1.5.6[80] and gathered from previous research (Supplementary Data 13). The active amino acids of alliinase[60] and LFS[61] were also gathered. The protein structures of genes from non-*Allium* species were predicted using AlphaFold2 online. The 3D conformers of the substrates were obtained from PubChem and converted to PDB format. Subsequently, we compared the variation in active sites between *Allium* and non-*Allium* species. 3D molecular structures were visualized using PyMOL in AMDock[81] and ligand-protein interactions were visualized using LigPlot+ v.2.2.8[82].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequence reads of the 244 samples generated in this study have been deposited in the NCBI Sequence Read Archive under BioProject nos. PRJNA1107703 and PRJNA1107706. Raw reads or annotations for other samples were accessed from the internet with accession nos. provided in Supplementary Data 1. The assembled transcriptomes for the samples sequenced in this study, CDSs and PEPs for all the 501 samples used in phylogenetic analyses, as well as those for the 16 samples used in gene expression level analyses, are available at Figshare: [https://doi.org/10.6084/m9.figshare.25516204]. Additionally, the sequences of orthologs, data matrices for phylogenetic analyses, divergence times, ancestral area reconstructions, and Source Data are also available at Figshare: [https://doi.org/10.6084/m9.figshare.25516204]. Specimens have been deposited at the Herbarium of Guangxi Botanical Garden of Medicinal Plants.

Code availability

Codes used in this study have been deposited at Figshare: [https://doi.org/10.6084/m9.figshare.25516204].

References

1. Byng, J. W. et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
2. Zhang, G. et al. Comprehensive phylogenetic analyses of Orchidaceae using nuclear genes and evolutionary insights into epiphytism. *J. Integr. Plant Biol.* **65**, 1204–1225 (2023).
3. Seberg, O. et al. Phylogeny of the Asparagales based on three plastid and two mitochondrial genes. *Am. J. Bot.* **99**, 875–889 (2012).
4. Bremer, K. & Janssen, T. Gondwanan origin of major monocot groups inferred from dispersal-vicariance analysis. *Aliso* **22**, 22–27 (2006).
5. Givnish, T. J. et al. Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. *Am. J. Bot.* **105**, 1888–1910 (2018).

6. Timilsena, P. R. et al. Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. *Front. Plant Sci.* **13**, 876779 (2022).
7. Baker, W. J. et al. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Syst. Biol.* **71**, 301–319 (2022).
8. Zuntini, A. R. et al. Phylogenomics and the rise of the angiosperms. *Nature* **629**, 843–850 (2024).
9. Stull, G. W. et al. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants* **7**, 1015–1025 (2021).
10. Yu, J. et al. Integrated phylogenomic analyses unveil reticulate evolution in *Parthenocissus* (Vitaceae), highlighting speciation dynamics in the Himalayan–Hengduan Mountains. *New Phytol.* **238**, 888–903 (2023).
11. Birch, J. L. & Keeley, S. C. Dispersal pathways across the Pacific: the historical biogeography of *Astelia* s.l. (Asteliaceae, Asparagales). *J. Biogeogr.* **40**, 1914–1927 (2013).
12. Kocyan, A. et al. Molecular phylogenetics of Hypoxidaceae-evidence from plastid DNA data and inferences on morphology and biogeography. *Mol. Phylogenet. Evol.* **60**, 122–136 (2011).
13. Goldblatt, P. et al. Iridaceae 'Out of Australasia'? phylogeny, biogeography, and divergence time based on plastid DNA sequences. *Syst. Bot.* **33**, 495–508 (2008).
14. Givnish, T. J. et al. Orchid historical biogeography, diversification, Antarctica and the paradox of orchid dispersal. *J. Biogeogr.* **43**, 1905–1916 (2016).
15. Guo, C. et al. Phylogenomics and the flowering plant tree of life. *J. Integr. Plant Biol.* **65**, 299–323 (2023).
16. Yoshimoto, N. & Saito, K. S-Alk(en)ylcysteine sulfoxides in the genus *Allium*: proposed biosynthesis, chemical conversion, and bioactivities. *J. Exp. Bot.* **70**, 4123–4137 (2019).
17. Liao, N. et al. Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in *Allium* crops. *Nat. Commun.* **13**, 6690 (2022).
18. Venâncio, P. C. et al. Antimicrobial activity of two garlic species (*Allium sativum* and *A. tuberosum*) against staphylococci infection. In vivo study in rats. *Adv. Pharm. Bull.* **7**, 115–121 (2017).
19. Sun, X. et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Mol. Plant* **13**, 1328–1339 (2020).
20. Hao, F. et al. Chromosome-level genomes of three key *Allium* crops and their trait evolution. *Nat. Genet.* **55**, 1976–1986 (2023).
21. Liu, N. et al. Transcriptome landscapes of multiple tissues highlight the genes involved in the flavor metabolic pathway in Chinese chive (*Allium tuberosum*). *Genomics* **113**, 2145–2157 (2021).
22. Jones, M. G. et al. Biosynthesis of the flavour precursors of onion and garlic. *J. Exp. Bot.* **55**, 1903–1918 (2004).
23. Willson, J., Roddur, M. S., Liu, B., Zaharias, P. & Warnow, T. DISCO: species tree inference using multicopy gene family tree decomposition. *Syst. Biol.* **71**, 610–629 (2022).
24. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* **19**, 153 (2018).
25. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
26. Stevens, P. F. *Angiosperm Phylogeny Website*. Version 14. http://www.mobot.org/MOBOT/research/APweb/ (2001).
27. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).
28. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).
29. Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E. & Smith, S. A. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* **105**, 385–403 (2018).
30. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67**, 735–740 (2018).
31. Rhodes, J. A., Baños, H., Mitchell, J. D. & Allman, E. S. MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. *Bioinformatics* **37**, 1766–1768 (2021).
32. Solís-Lemus, C., Bastide, P. & Ané, C. PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* **34**, 3292–3298 (2017).
33. Blair, C. & Ané, C. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Syst. Biol.* **69**, 593–601 (2020).
34. Morales-Briones, D. F. et al. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. *Syst. Biol.* **70**, 219–235 (2021).
35. Sousa, V. & Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* **14**, 404–414 (2013).
36. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
37. Li, H. T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461–470 (2019).
38. Pérez-Escobar, O. A. et al. The origin and speciation of orchids. *New Phytol.* **242**, 700–716 (2024).
39. Matzke, N. J. BioGeoBEARS: bioGeography with Bayesian (and likelihood) evolutionary analysis with R scripts, version 1.1.1. Zenodo https://doi.org/10.5281/zenodo.1478250 (2018).
40. Wang, Y. et al. Progress in systematics and biogeography of Orchidaceae. *Plant Divers.* **46**, 425–434 (2024).
41. Markwick, P. J. *Paul's Palaeo Pages*. http://www.palaeogeography.net (2011).
42. Chen, D. Y., Zhang, T. K., Chen, Y. M., Ma, H. & Qi, J. Tree2GD: a phylogenomic method to detect large-scale gene duplication events. *Bioinformatics* **38**, 5317–5321 (2022).
43. Yang, Y. et al. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytol.* **217**, 855–870 (2018).
44. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
45. Condamine, F. L., Clapham, M. E. & Kergoat, G. J. Global patterns of insect diversification: towards a reconciliation of fossil and molecular evidence? *Sci. Rep.* **6**, 19208 (2016).
46. Ellis, S. A. & Scatcherd, J. E. Bean seed fly (*Delia platura*, *Delia florilega*) and onion fly (*Delia antiqua*) incidence in England and an evaluation of chemical and biological control options. *Ann. Appl. Biol.* **151**, 259–267 (2007).
47. Ding, S. et al. The phylogeny and evolutionary timescale of Muscoidea (Diptera: Brachycera: Calyptratae) inferred from mitochondrial genomes. *PLoS ONE* **10**, e0134170 (2015).
48. Xuan, J. L. et al. The phylogeny and divergence times of leaf-mining flies (Diptera: Agromyzidae) from anchored phylogenomics. *Mol. Phylogenet. Evol.* **184**, 107778 (2023).
49. Lancaster, J. E. & Boland, M. J. *Flavor Biochemistry: Onions and Allied Crops* (CRC Press, 1990).
50. Auger, J. et al. Insecticidal and fungicidal potential of *Allium* substances as biofumigants. *Agroindustria* **3**, 367–370 (2004).
51. Christ, B. et al. Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. *Nat. Commun.* **10**, 3206 (2019).

52. Xu, Z. et al. The genome of *Corydalis* reveals the evolution of benzylisoquinoline alkaloid biosynthesis in Ranunculales. *Plant J.* **111**, 217–230 (2022).
53. Chen, L. Y. et al. Phylogenomic analyses of Alismatales shed light into adaptations to aquatic environments. *Mol. Biol. Evol.* **39**, msac079 (2022).
54. Beran, F., Kollner, T. G., Gershenzon, J. & Tholl, D. Chemical convergence between plants and insects: biosynthetic origins and functions of common secondary metabolites. *New Phytol.* **223**, 52–67 (2019).
55. Hurni, S. et al. Rye *Pm8* and wheat *Pm3* are orthologous genes and show evolutionary conservation of resistance function against powdery mildew. *Plant J.* **76**, 957–969 (2013).
56. Consortium, M. E. G. et al. Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. *Mol. Plant* **11**, 1084–1096 (2018).
57. Yang, X. et al. Parallel analysis of global garlic gene expression and alliin content following leaf wounding. *BMC Plant Biol.* **21**, 174 (2021).
58. Yoshimoto, N. et al. Identification of a flavin-containing S-oxygenating monooxygenase involved in alliin biosynthesis in garlic. *Plant J.* **83**, 941–951 (2015).
59. Qin, L. et al. Metabolomics and transcriptomics analyses provides insights into S-alk(en)yl cysteine sulfoxides (CSOs) accumulation in onion (*Allium cepa*). *Sci. Hortic.* **310**, 111727 (2023).
60. Kuettner, E. B., Hilgenfeld, R. & Weiss, M. S. The active principle of garlic at atomic resolution. *J. Biol. Chem.* **277**, 46402–46407 (2002).
61. Arakawa, T. et al. Dissecting the stereocontrolled conversion of short-lived sulfenic acid by lachrymatory factor synthase. *ACS Catal.* **10**, 9–19 (2020).
62. Song, L. & Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* **4**, 48 (2015).
63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
65. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
66. Haas, B. J. TransDecoder. GitHub https://github.com/TransDecoder/TransDecoder (2016).
67. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421 (2009).
68. van Dongen, S. M. *Graph Clustering by Flow Simulation* (University of Utrecht, 2000).
69. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
70. Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715–716 (2008).
71. Smith, S. A., Brown, J. W. & Walker, J. F. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS ONE* **13**, e0197433 (2018).
72. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
73. Sessa, E. B., Zimmer, E. A. & Givnish, T. J. Phylogeny, divergence times, and historical biogeography of New World *Dryopteris* (Dryopteridaceae). *Am. J. Bot.* **99**, 730–750 (2012).
74. Rognes, T. Faster Smith-Waterman database searches with intersequence SIMD parallelisation. *BMC Bioinf.* **12**, 221 (2011).
75. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
76. Huang, C. H. et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* **33**, 394–412 (2016).
77. Smith, S. A. & O'Meara, B. C. TreePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
78. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
79. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
80. Morris, G. M. et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
81. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A. & Moreno, E. AMDock: a versatile graphical tool for assisting molecular docking with Autodock Vina and Autodock4. *Biol. Direct* **15**, 1–12 (2020).
82. Laskowski, R. A. & Swindells, M. B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **51**, 2778–2786 (2011).

Acknowledgements

We thank the Kunming Institute of Botany, Chinese Academy of Sciences (CAS), Xishuangbanna Tropical Botanical Garden (CAS), Guiyang Medicinal Botanical Garden, and Shenyang Botanical Garden for assistance in plant collection; Can Dai (Hubei University) for discussion on this study; Michael L. Moody (University of Texas at El Paso) for discussion and providing plant photo; Xu Lu, Hui-Ying Wang (China Pharmaceutical University) and Xiang-Yang Leng (SCIEX, China) for technical assistance; Xiao-Wei Xin (Shandong Drug and Food Vocational College) for providing plant materials; Bing Liu (Institute of Botany, CAS), Zhong Zhang (Jinggangshan National Nature Reserve) and Ye-Chun Xu (Guangdong Academy of Agricultural Sciences) for providing plant photos; Ya-Dong Zhou (Nanchang University) for assisting in data analyses and providing plant photos; Guo-Yong Xie (China Pharmaceutical University) for assisting in plant cultivation; Li Feng, Xing-Ze Li, Wen-Da Zhang, Wen-Fang Zheng, and Wen-Yu Du (China Pharmaceutical University) for assisting in data analyses; John A. Rhodes (University of Alaska Fairbanks) for explaining the results of MSCquartets; Ji Yang (Fudan University) and Tao Wan (Wuhan Botanical Garden) for suggestions on this manuscript. This work was supported by the Guangxi Innovation-Driven Development Project (no. GuiKe AA18242040 to K.-H.W.), the Construction of Southern Medicine Germplasm Resource Base for Guangdong Northern (20231206 to K.-H. W.), the National Natural Science Foundation of China (32370242 to L.-Y.C.), the Fundamental Research Funds for the Central Universities (2632024TD04 to L.-Y.C.), and the Sino-Africa Joint Research Center (SAJC202101 to Q.-F.W.).

Author contributions

Q.-F.W., L.-Y.C., K.-H.W., and N.-H.T. designed the research. Q.-F.W., K.-H.W., Y.H., X.-M.W., G.-W.H., X.-X.W., and L.-Y.C. contributed to the taxon sampling and sequencing. X.-X.W., C.-H.H., D.F.M.-B., N.Z., P.-G.Z., X.-Y.H., and L.-Y.C. performed data analyses. X.-X.W. performed wet lab experiments. X.-X.W., X.-Y.W., D.F.M.-B., and L.-Y.C. prepared the figures and tables. L.-Y.C. drafted the manuscript. Q.-F.W., C.-H.H., D.F.M.-B., X.-X.W., and L.-Y.C. revised this manuscript. All the authors read this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53943-6.

Correspondence and requests for materials should be addressed to Kun-Hua Wei, Ning-Hua Tan, Qing-Feng Wang or Ling-Yun Chen.

Peer review information *Nature Communications* thanks Deng-Feng Xie and Mingfang Zhang for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.