Published in final edited form as:
Plant J. 2023 January 05; 113(5): 1021–1034. doi:10.1111/tpj.16097.

The *Saururus chinensis* genome provides insights into the evolution of pollination strategies and herbaceousness in magnoliids

Jia-Yu Xue[#1,2], **Zhen Li**[#3], **Shuai-Ya Hu**[#1], **Shu-Min Kao**[#3], **Tao Zhao**[#4], **Jie-Yu Wang**[#5], **Yue Wang**[2], **Min Chen**[2], **Yichun Qiu**[6], **Hai-Yun Fan**[1], **Yang Liu**[7,*], **Zhu-Qing Shao**[8,*], **Yves Van de Peer**[1,3,9,*]

[1]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

[2]Center for Plant Diversity and Systematics, Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing 210014, China

[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, VIB-UGent Center for Plant Systems Biology, B-9052 Ghent, Belgium

[4]State Key Laboratory of Crop Stress Biology for Arid Areas/Shaanxi Key Laboratory of Apple, College of Horticulture, Northwest A&F University, Yangling 712100, China

[5]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

[6]Max Planck Institute of Molecular Plant Physiology, Potsdam Science Park, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

[7]Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen 518004, Guangdong, China

[8]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, 210023, China

[9]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[*]To whom correspondence should be addressed: yang.liu0508@gmail.com (Y.L.), zhuqingshao@nju.edu.cn (Z.Q.S) or yves.vandepeer@psb.vib-ugent.be (Y.V.d.P.).

Author contributions

J.Y.X., Z.L., Y.L., Z.Q.S. and Y.V.d.P. conceived the study. J.Y.X. and M.C. collected samples. S.M.K. assembled and annotated the genome and conducted comparative transcriptome analyses. J.Y.X.,Z.L., J.Y.W., S.Y.H. and F.H.F. conducted phylogenetic analyses. Z.L. conducted WGD analyses. M.C. conducted anatomical experiments of the stem and Y.W. conducted transgenic experiments. J.Y.X. and Z.L. drafted the manuscript. Z.Q.S., Y.L., Y.Q., T.Z. and Y.V.d.P. participated in the revision of the manuscript. All authors read and approved the final manuscript.

Conflict of interest statement

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

Europe PMC Funders Author Manuscripts

# These authors contributed equally to this work.

Abstract

Saururus chinensis, an herbaceous magnoliid without perianth, represents a clade of early-diverging angiosperms that have gone through woodiness-herbaceousness transition and pollination obstacles: the characteristic white leaves underneath inflorescence during flowering time are considered to be a substitute for perianth to attract insect pollinators. Here, using the newly sequenced *S. chinensis* genome, we revisited the phylogenetic position of magnoliids within mesangiosperms, and recovered a sister relationship for magnoliids and Chloranthales. By considering differentially expressed genes, we identified candidate genes that are involved in the morphogenesis of the white leaves in *S. chinensis*. Among those genes, we verified - in a transgenic experiment with *Arabidopsis* - that increasing the expression of the 'pseudo-etiolation in light' gene (ScPEL) can inhibit the biosynthesis of chlorophyll. ScPEL is thus likely being responsible for the switches between green and white leaves, suggesting that changes in gene expression may underlie the evolution of pollination strategies. Despite being an herbaceous plant, *S. chinensis* still has vascular cambium and maintains the potential for secondary growth as a woody plant, because the necessary machinery, i.e., the entire gene set involved in lignin biosynthesis, is well preserved. However, similar expression levels of two key genes (CCR and CAD) between the stem and other tissues in the lignin biosynthesis pathway are possibly associated with the herbaceous nature of *S. chinensis*. In conclusion, the *S.chinensis* genome provides valuable insights into the adaptive evolution of pollination in Saururaceae and reveals a possible mechanism for the evolution of herbaceousness in magnoliids.

Keywords

herbaceous magnoliids; mesangiosperm radiation; leaf color; adaptation; pollination

Introduction

Saururus chinensis (Lour.) Baill., also known as Asian lizard's tail, is a typical perennial species in the family Saururaceae (order Piperales), containing many herbaceous species (Fig. 1a). *S. chinensis* is a diploid with 11 chromosomes (2n = 22) and widely distributed in East Asia. *S. chinensis* belongs to magnoliids, a clade of mesangiosperms characterized by three-merous flowers, one-pored pollen, and diverse secondary compounds of aromatic terpenoids (Palmer *et al.*, 2004; Endress & Doyle, 2009). The clade of magnoliids comprises ~9,000 species, accounting for ~3% of the species in angiosperms and includes many economically important species as sources of fruits, medicine, spices, and perfumes. The classification and phylogenetic position of magnoliids in mesangiosperms has been a long-lasting debate (Li *et al.*, 2019; One Thousand Plant Transcriptomes, 2019; Li & Van de Peer, 2021). Mesangiosperms include five lineages, namely monocots, eudicots, magnoliids, Chloranthales, and Ceratophyllales (APG IV, 2016). Morphologically, it is difficult to conclude on phylogenetic relationships between the five mesangiosperm lineages due to difficulties in distinguishing synapomorphic traits from plesiomorphic traits. Both eudicots and monocots, the two largest clades of mesangiosperms, share morphological traits with magnoliids. For instance, two cotyledons and reticular leaf venation in magnoliids resemble

Europe PMC Funders Author Manuscripts

those in eudicots, whereas pollen with one pore is characteristic for both magnoliids and monocots (Palmer *et al.*, 2004; Endress & Doyle, 2009). The rapid diversification of mesangiosperm lineages, plus phenomena such as incomplete lineage sorting, complicate resolving the exact evolutionary relationships of magnoliids and other mesangiosperm lineages, also using molecular data, despite many attempts employing a large amount of data from nuclear, plastid and mitochondrial genomes (Wickett *et al.*, 2014; Gitzendanner *et al.*, 2018; Li *et al.*, 2019; One Thousand Plant Transcriptomes, 2019; Xue *et al.*, 2021). The unresolved phylogeny of magnoliids hinders the evolutionary study of morphological traits of angiosperms.

As a flowering plant, *S. chinensis* presents white or light-yellow flowers on a spike inflorescence. Although perianth shows complete degeneration, during blooming time *S. chinensis* has two or three white leaves growing underneath the spike inflorescence, setting off the distinguishable inflorescence (Fig. 1a). After the seeds reach maturity, the white leaves gradually turn green again. Since the white leaves arise and fade in concert with the flowering time, they have been suggested to play an important role in attracting insect pollinators (Song *et al.*, 2018). Indeed, *S. chinensis* can be pollinated by insects, whereas its American relative *S. cernuus*, which has no such white leaves, only relies on wind pollination (Thien *et al.*, 1994), so the emergence of white leaves could be an adaptive evolutionary trait of *S. chinensis*.

Saururaceae are one of three families in Piperales, which is the only order in magnoliids with herbaceous plants, while the other three orders, Canellales, Laurales, and Magnoliales, are all comprised of woody trees. Woodiness is associated with secondary growth of the vascular cambium, a population of meristematic cells that are eventually divided into secondary xylem and phloem. In contrast, herbaceousness is a more complicated concept. For example, there are obvious differences between herbs in eudicots and monocots. All monocots lack vascular cambium, so they cannot be considered genuine woody plants with secondary growth (Spicer & Groover, 2010; Roodt *et al.*, 2019). Most herbaceous eudicots, however, still preserve vascular cambium, so genetically they still have the potential to resume secondary growth to become woody plants (Melzer *et al.*, 2008; Barra-Jimenez & Ragni, 2017). Because magnoliids have both woody and herbaceous plants and they occupy an interesting phylogenetic position with members that may still have retained some features from the common ancestor of eudicots and monocots, herbaceous magnoliids such as *S. chinensis* could shed light on origin of woodiness or herbaceousness in monocots and eudicots.

Here, we report a chromosome-level genome assembly of *S. chinensis* using a combination of three sequencing technologies. Together with other available angiosperm genomes, we study the phylogenetic relationships and evolution among magnoliids and other mesangiosperms. Combining plant organ anatomy, comparative genomics, gene expression analyses, and experimental validation, we reveal the molecular mechanism underlying herbaceousness of *S. chinensis*, as well as the evolution of white leaves underneath its spike inflorescence.

Europe PMC Funders Author Manuscripts

Results and discussion

Genome sequencing, assembly, and annotation

To estimate the genome size of *S. chinensis*, we employed flow cytometry analysis using rice (*Oryza sativa*) and *Nicotiana benthamiana* as references, respectively, and obtained two comparable estimates of 553 Mb and 523 Mb (see Methods and Supplementary Fig. S1, Table S1 and S2). Accordingly, we sequenced the *S. chinensis* genome by generating 244 Gb Illumina reads (~450×) and 50 Gb PacBio reads (93×) (Supplementary Table S3). The genome size of *S. chinensis* estimated by *k*-mer analysis of the Illumina reads is 502 Mb (Supplementary Fig. S3), similar to, but smaller than the flow cytometry estimates above. Also, *k*-mer analysis shows a high level of heterozygosity of 0.821% for the *S. chinensis* genome (Supplementary Fig. S3). Using both the PacBio and Illumina reads, we first obtained a preliminary assembled genome with a size of 698 Mb (see Methods), much larger than our estimates described above. Therefore, we mapped the Illumina reads to the preliminary assembly and observed two coverage peaks, with one at 44× and the other at 88× (Supplementary Fig. S4). Considering the high heterozygosity of the *S. chinensis* genome, the coverage peak at 44× is caused by assembled haplotypes from genomic regions with high heterozygosity, indicating that the preliminary assembly contains redundant allelic contigs. To refine the preliminary assembly and collapse haplotype regions, we used Purge Haplotigs to reassign allelic contigs and build a haplotype-fused assembly (Roach *et al.*, 2018). The genome size of the resulted assembly decreased from 698 Mb with 3,938 contigs (N50 = 1.0 Mb) to 537 Mb with 842 contigs (N50 = 1.4 Mb), falling in the range of our estimates of the genome size (Supplementary Fig. S4). We then made use of 60 Gb Hi-C data (111×) and assembled the *S. chinensis* genome into 38 scaffolds, with a scaffold N50 of 47.8 Mb (Supplementary Fig. S4 and S5). The 11 longest scaffolds, accounting for 533 Mb (99.2% of the total genome assembly), was selected to presumably correspond to the 11 chromosomes of the haplotype genome of *S. chinensis* (Fig. 1b). The genome-wide interaction heatmap shows a high-quality result of grouping and ordering contigs by the Hi-C data (Supplementary Fig. S6).

We annotated repetitive sequences of the *S. chinensis* genome (see Methods) and found that they comprise 66.8% (360 Mb) of the assembled genome. LTR elements are the most abundant components (25.5%) among the repetitive sequences (Supplementary Table S5). Compared with other magnoliids genomes, *S. chinensis* has the highest proportion of TEs, despite that it has one of the smallest genomes (Supplementary Table S6).

To obtain gene models for the protein-coding genes in the *S. chinensis* genome, we used an approach integrating *de novo*, homology based, and RNA-Seq based gene predictions (see Methods). In total, 36,140 protein-coding genes were predicted with an average coding-sequence (CDS) length of 900 bp and an average of 3.96 exons per gene. Among all annotated genes, 29,555 (81.8%) were supported by transcriptome data, and 27,123 (74.1%) genes could be functionally annotated (Supplementary Table S7). Our annotation captures 93% of the embryophyta BUSCO (odb10) genes, either as single copy or in duplicate (Supplementary Fig. S7). Compared with the complete set of BUSCO genes of other published magnoliids genomes, such as 90% for *Cinnamomum kanehirae*, 85%

for *Piper nigrum* (85%), 81% for *Liriodendron chinense*, and 78% for *Persea americana* (Supplementary Fig. S7), the *S. chinensis* genome has more complete BUSCO genes, suggesting a high quality in terms of protein-coding gene predictions. In addition to the protein-coding genes, we identified 452 transfer RNAs, 344 ribosomal RNAs, 118 small nuclear RNAs (snRNAs), 685 small nucleolar RNAs (snoRNAs), and 4,637 other non-coding RNA including micro RNAs (miRNAs).

Phylogenetic position of Magnoliids among angiosperms and *Saururus* in magnoliids

We constructed a sequence dataset with 135 genomes representing the main angiosperm lineages (95 eudicots, 26 monocots, eight magnoliids, two Chloranthales, one Ceratophyllales and three species for the ANA grade) with three gymnosperms as outgroup (*Ginkgo biloba, Cycas panzhihuaensis*, and *Welwitschia mirabilis*). In total, 473 out of 1,614 BUSCO genes were kept in the sequence dataset because they existed in at least 75% of the taxa investigated (Manni *et al.*, 2021). For each BUSCO gene in the dataset, we prepared multiple sequence alignments of amino acids, CDSs, and the first and second codon positions thereof (Codon1+2) to infer phylogenetic relationships using both concatenated and multi-species coalescent (MSC) approaches (Supplementary Fig. S8-13). Among all results, Amborellales and Nymphaeales were successively resolved as the sister group of the other angiosperms, whereas diverse relationships were recovered for the five lineages of mesangiosperms. Ceratophyllales was recovered as the next diverging lineage based on the analyses of concatenated amino acid and Codon1+2 datasets, whereas the results from amino acid and CDS with MSC and concatenated CDS placed monocots as the sister group of all other mesangiosperms. Most results congruently resolved magnoliids as the sister group of Chloranthales (all datasets with MSC and the concatenated amino acid) with high support (≥95%), except that the concatenated CDS and Codon 1+2 placed magnoliids as the sister to eudicots-(Chloranthales-Ceratophyllales) and eudicots-Chloranthales, respectively, yet the two topologies received very low support (59% and 67%, respectively). Among all recovered topologies, the result derived from the concatenated amino acid alignment, grouping magnoliids with chloranthales (and both with eudicots) received the highest support for these early-diverging angiosperm lineages, above 95% for all deep nodes (Fig. 2a), while the results from other datasets always showed low support for some controversial relationships (Fig. 2b and Supplementary Fig. S8-13).

In previous studies (Chaw *et al.*, 2019; Chen *et al.*, 2019; Hu *et al.*, 2019; Rendon-Anaya *et al.*, 2019; Chen *et al.*, 2020; Lv *et al.*, 2020; Shang *et al.*, 2020; Yang *et al.*, 2020; Zhang *et al.*, 2020a; Guo *et al.*, 2021; Ma *et al.*, 2021), phylogenetic inconsistence of early-diverging angiosperm lineages could be observed between studies based on amino acid sequences and the corresponding CDS, and between studies based on different methods like concatenation of multiple genes, or MSC. Some studies also indicated that the discordance in phylogenetic reconstruction was due to potential incomplete lineage sorting (ILS) during the early radiation of mesangiosperms (Soltis & Soltis, 2019), where polymorphic allele states in ancestral populations do not have enough time to fix because of rapid species divergence among ancestors of extant lineages. In this study, despite of a larger taxon and gene sampling, discordance among different methods and datasets remained. Among single gene trees with regards to the relationships of mesangiosperms, 36%, 33.4% and 33.4% of

all 473 gene trees generated by the CDS, Codon1+2, and amino acid datasets respectively supported Ceratophyllales as the sister lineage to all other mesangiosperms, and 37%, 33.5% and 33.2% supported monocots as first diverging lineages within mesangiosperms. The sister relationship of magnoliids and Chloranthales received relatively higher 40%, 39% and 42.7% support, and the sister group of eudicots and the magnoliids-Chloranthales lineage received 36%, 32.8% and 35.5% support from the gene trees (Fig. 2c and Supplementary Fig. S14). However, none of the relationships were predominant among all single gene trees, and alternative relationships still received considerable support (Supplementary Fig. S14), suggesting substantial ILS during the evolution of meagsangiosperm lineages. As ILS and other factors likely affect phylogenetic inference, recently, different approaches have been used to study the phylogenetic position of magnolids, for instance based on genomic features such as synteny, and these have suggested that magnoliids are more closely related to monocots than to eudicots (Zhao *et al.*, 2021).

The phylogenetic relationship among magnoliids, monocots, and eudicots has been long debated (Soltis & Soltis, 2019), and three different relationships are usually favored: 1) (monocots, (magnoliids, eudicots)), 2) (magnoliid, (monocots, eudicots)), and 3) (eudicots, (magnoliids, monocots)). Although magnoliids still share morphological features with both monocots and eudicots, they seem to share more common features with eudicots. That is why the lineage has been considered, together with eudicots, as the so-called dicots (Cronquist, 1981). The closer relationship of magnoliids and eudicots, at least morphologically, might be related to the more similar gene sets in the two lineages. We identified gene families shared by either two lineages among monocots, eudicots and magnoliids, while being absent in the third and indeed found that magnoliids and eudicots share more gene families than either of the two with monocots (Fig. 2d), which are functionally enriched in RNA polymerase II transcription regulatory region sequence-specific DNA binding (Supplementary Fig. S15). Such gene families may be responsible for the fact that magnoliids and eudicots are morphologically more alike, although our analyses of gene families here lend little support to any phylogenetic relationships among the mesangiosperms, as gene families existing in the most recent common ancestor of mesangiosperms might have been differentially lost in different lineages.

Whole-genome duplications in Piperales

So-called age distributions of synonymous substitutions per synonymous sites (K_S) for all paralogs (paranome) and for paralogs retained in collinear regions (anchor pairs) both show a signature peak at $K_S \approx 1.1$, suggestive of a whole-genome duplication (WGD) event in the *Saururus* genome (Fig. 3a and Supplementary Fig. S16). Intragenomic analysis of paralogous gene orders also reveals collinear regions with two paralogous segments (Supplementary Fig. S17), further supporting a WGD event identified in the *Saururus* genome. GO enrichment analysis reveals that retained paralogous genes are mainly involved in “regulation of transcription, DNA-templated”, “regulation of flower development”, “phospholipid transport”, and “seed development” with respect to ‘biological progress’, and “DNA-binding transcription factor activity”, “sequence-specific DNA binding”, “DNA binding”, and “calmodulin binding” with respect to ‘molecular function’ (Supplementary Fig. S18).

To place the WGD in the magnoliids clade, we compared one-to-one orthologous K_S distributions between *Saururus* and *P. nigrum* and *Liriodendron chinense* (Fig. 3a). The K_S peak value for the WGD in the *Saururus* genome is smaller than the K_S value representing the divergence between *Saururus* and *Liriodendron*, but larger than the K_S value representing the divergence between *Saururus* and *Piper*. Interestingly, when comparing the one-to-one orthologous K_S distributions between *Vitis* and *Saururus*, *Piper*, and *Liriodendron* (Supplementary Fig. S19), we observed different K_S peaks for the same speciation event between eudicots and magnoliids, suggesting different synonymous substitution rates among the three species, with *Piper* as the fastest and *Liriodendron* as the slowest, resulting in an overestimate for the divergence between *Piper* and *Saururus* and an underestimate for the divergence between *Liriodendron* and *Saururus*, respectively. We hence employed a relative rate test to adjust the synonymous substitution rates in *Piper* and *Liriodendron* compared to the rate of *Saururus,* giving an even stronger support for the WGD in the *Saururus* genome to be shared with *Piper,* but having occurred after the divergence between *Saururus* and *Liriodendron* (see arrows and dotted orthologous K_S distributions and the denoted WGD K_S peak in Fig. 3a).

Although only one WGD at $K_S \approx 0.1$ has been identified in the published *Piper* genome (Hu *et al.*, 2019), our results suggest that at least two WGDs have occurred in *Piper*, one recent WGD specific to *Piper* and one WGD shared between *Piper* and *Saururus*. Surprisingly, intergenomic collinear analysis between *Saururus* and *Piper* shows evidence of even three WGD events in the *Piper* genome with one genomic segment in the *Saururus* genome corresponding to up to eight genomic segments in the *Piper* genome (Fig. 3b). Based on the K_S values of orthologs in collinear segments between *Saururus* and *Piper* (as reflected as different colored dots in Fig. 3b), four of the eight genomic segments in the *Piper* genome are more similar to the segments in the *Saururus* genome as the orthologs in the four genomic segments have smaller K_S values (in orange) than do the orthologs from the other four segments (in yellow and green).

Also, compared with the collinear segments that have orthologs with larger K_S values, the collinear segments having orthologs with smaller K_S values are more continuous (Fig. 3b). Together, these patterns can be explained by an ancient WGD that has occurred before the divergence of *Saururus* and *Piper* and two WGDs that have occurred later, independently in the lineage leading to *Piper*. Phylogenomic analyses of anchor pairs in the *Saururus* and *Piper* genomes are also in support of a shared WGD in both species (Fig. 3c). In conclusion, we identified one WGD shared by *Saururus* and *Piper,* in the Piperales order. The two independent WGDs unique to the *Piper* genome are visible in K_S age distributions for the whole paranome and anchor pairs with K_S peak values at 0.1 and 0.8. The signal K_S peak for the shared WGD between *Piper* and *Saururus* is, however, concealed by the two younger WGDs (Supplementary Fig. S20) and the faster synonymous substitution rate. Considering the shared WGD has a $K_S \approx 1.1$ in the *Saururus* genome, the expected K_S peak for the shared WGD in a genome with higher synonymous substitution rates than *Saururus* would be larger than K_S 1.1, hence falling in a K_S range where K_S values reach saturation (Vanneste *et al.*, 2013; Zhang *et al.*, 2020b).

Morphogenesis of white leaves and adaptive evolution of pollination strategy of *S. chinensis*

The perianth is an important part of angiosperm flowers for attracting pollinating insects, and loss of the perianth often suggests the lack of attraction to pollinators. Although Saururaceae are a family in which no species has perianth, *S. chinensis* has the most noticeable morphological character, namely two or three white leaves growing underneath the inflorescence during flowering time. The white leaves in *S. chinensis* have proven to play an important role in attracting insect pollinators (Song *et al.*, 2018). To identify the regulatory genes underlying the development of white leaves, we collected tissues of the white top leaves during flowering time, and after flowering when the top leaves have turned as green as other regular leaves and sequenced their transcriptomes for expression analyses. Additionally, regular, ever green leaves beneath the white leaves were also collected during flowering time and sequenced as a reference or control. In total, 146 differentially expressed genes (DEGs) (Supplementary Table S10) were identified by DESeq2 (Love *et al.*, 2014). Functional annotation of these genes suggested that 27 genes were enriched for 27 genes were enriched for chlorophyll biogenesis and 23 of them showed significantly lower expression levels in white leaves (Supplementary Fig. S21), thus should theoretically block chlorophyll accumulation. In addition to the genes involved in the chlorophyll biogenesis pathway, another DEG Sc004_1478, seems also responsible for the color change of top leaves,because it is a homolog to the pseudo-etiolation in light (PEL) gene in *Arabidopsis*. It has been well acknowledged that the expression of the PEL gene has a negative correlation with chlorophyll content and bolting time in Arabidopsis (Ichikawa *et al.*, 2006).

To further verify the functional role of Sc004_1478, or ScPEL, in the formation of green and white leaves, a transgenic and hetero-expression experiment was conducted in Arabidopsis. Concisely, the ORF of the ScPEL gene, combined with Spe I and Asc I restriction endonucleases sites, was inserted into the pMDC83-GFP expression vector driven by the 35S promoter. We then transformed the vector of pMDC83-Sc004_1478-GFP into an Arabidopsis Col line and successfully observed pseudo-etiolation in the transgenic individuals: the transgenic lines showed obvious light-yellow leaves and even stems, while the wild type individuals had normal green leaves and stems (Fig. 4a). We further tested the chlorophyll content in leaves and found a significant decrease of chlorophyll in the transgenic lines compared with the wild types (Fig. 4b and Supplementary Table S12), suggesting a negative association between the expression of ScPEL and the chlorophyll content. Therefore, high expression of ScPEL is likely to lead to,or one possible reason for the transition of top leaves from green to white in *S. chinensis* (Supplementary Fig. S23).

The above expression analyses and transgenic experiments revealed the molecular mechanisms underlying the white leaves of *S. chinensis* during flowering time. In the context of macroevolution, the rise of white leaves in *S. chinensis* could be associated with the loss of perianth and considered as a success of adaptive evolution of pollination strategy (Fig. 4c). Because no species in Saururaceae has perianth (APG IV, 2016), the absence of the perianth is likely a synapomorphy for the family, resulting in decreased attraction to pollinating insects. So developing white-color organs during the blooming stage seems to be an alternative strategy to attract pollinators under certain environments, which have

successfully compensated for decreased attraction due to the loss of perianth, and help multiple Saururaceae taxa to survive. For example, *Houttuynia cordata* and *Gymnotheca involucrata* have either a white involucre or white bracts to attract pollinators, and *S. chinensis* adopted to develop white top leaves by increasing the expression of the ScPEL gene,as shown above. However, since involucres, bracts, and leaves are different organs, the 'white-color' organs in these Saururaceae species must have been independently evolved for entomophily, suggestinga convergent evolution in Saururaceae lineages.Convergent evolution is further supported by the fact that a close relative to *S. chinensis* in America, *S. cernuus*, has no white leaves during flowering time and largely relies on wind pollination (Thien *et al.*, 1994) (Fig. 4c). Apparently, the two closely related species show different adaptation strategies by adopting two kinds of pollinating strategies,correlated with green or white colors of top leaves during flowering time (Liang, 1995).

Anatomical and molecular evidence indicates that 'magnoliid herbs' resemble eudicot herbs

Whether a plant can develop genuine wood depends on its ability of secondary growth through vascular cambium. Herbaceous plants either do not initiate secondary growth even if they have vascular cambium, or have no ability of secondary growth due to the lack of vascular cambium. Most herbaceous eudicots, such as *Arabidopsis* and *Medicago*, belong to the former, whereas all monocots belong to the latter. However, whether magnoliid herbs like *S. chinensis* have vascular cambium or not remains unclear.

Using tissue section dye technique, we studied the cross-section of the stems and roots of *S. chinensis*. Our anatomical results indicate that the stems of *S. chinensis* are homologous to those in eudicots, since all vascular bundles of *S. chinensis* are arranged in circles (Fig. 5a). More importantly, when looking at the organization of a single vascular bundle in *S. chinensis*, vascular cambium can be observed (Fig. 5b), suggesting a morphological similarity between herbaceous plants in magnoliids and eudicots.

Since herbaceous plants from both magnoliids and eudicots have vascular cambium, the reason why they do not grow into woody shrubs or trees can be attributed to lacking initiation of secondary growth. This is supported by some evidence from herbaceous eudicots. Morphological transition between woodiness and herbaceousness can often be observed among close-related species (e.g., legumes, Boraginaceae, Lamiaceae, Lythraceae), implying frequent and recurrent transitions in eudicot lineages; and *Arabidopsis* can grow into woody plants when two regulatory genes (SOC1 and FUL) are simultaneously knocked out (Melzer *et al.*, 2008). As *S. chinensis* has similar vascular cambium to eudicots, we speculate that the woody-herbaceous transformation in magnoliids is more likely a matter of gene expressional regulation, rather than loss of relevant genes involved in the development of vascular cambium, as observed in monocots (Roodt *et al.*, 2019).

As expected, we could identify all genes involved in the lignin biosynthesis pathway (LBP) in the *S. chinensis* genome. As a matter of fact, all LBP genes were preserved in all surveyed 135 angiosperm genomes (Supplementary Tables S13), and we did not detect significant correlation between copy number of the LBP genes and woody/herbaceous phenotypes. As an herbaceous plant, the stems of *S. chinensis* must have lower levels of lignin, the

major component of wood. Therefore, we considered the expression of LBP genes in *S. chinensis*, and found that two genes catalyzing the critical final two steps in LBP - **C**innamoyl **C**oA **r**eductase (CCR), the rate limiting enzyme catalyzing hydroxycinnamoyl COA thioesters to corresponding aldehydes (Chabannes *et al.*, 2001),and **c**innamyl **a**lcohol **d**ehydrogenase (CAD) (Yan *et al.*, 2019), responsible for the final step catalyzing the reduction of cinnamaldehyde to cinnamyl alcohol - are not differentially expressed in the stems compared to other organs, such as roots, green leaves, white leaves, flowers, and fruits of *S. chinensis*. The expression pattern of CCR and CAD in different organs of *S. chinensis* is in line with those in other investigated herbaceous plants across angiosperms, while CCR and CAD in woody plants are usually more highly expressed in the stems (Supplementary Tables S14 and S15). We believe that the restricted expression of both CCR and CAD in stems of *S. chinensis* may therefore explain its herbaceous appearance (Fig. 5c).

Conclusions

Our sequenced *S. chinensis* genome provides a new resource for the resolution of the phylogenetic ambiguity, and recovered a sister group relationship for magnollids and Chloranthales. The characteristic white leaves of *S. chinensis* functionally serve as an alternative of the lost perianth in attracting insect pollinators, and we explained the molecular mechanism by the combination of bioinformatics analysis and experimental verification, that the increased expression of ScPEL gene can inhibit the chlorophyll biosynthesis and may lead to switches between green and white leaves. Such adaptive evolution of pollination strategies could be observed in multiple Sauraceae species, but by changing the color of different organs, thus should be the results of convergent evolution. Magnoliids comprise mainly woody trees and shrubs with few herbaceous lineages. So we explored the mechanism for the formation of magnoliid herbs. Anatomical and molecular evidence both suggest *S. chinensis* maintain the potential for secondary growth like a woody plant, yet the low expression of two key genes in the lignin biosynthesis pathway is possibly the reason that makes *S. chinensis* remain an herb. Therefore, the expressional regulation maybe the key to the woodiness-herbaceousness transition in magnliids and eudicots, and could explain the frequent woodiness-herbaceousness transition along the eudicot evolution.

Materials and methods

Sample preparation and sequencing

S. chinensis cultivated at Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, were chosen to provide experimental tissues. Fresh leave tissues were collected to extract genomic DNA. For genome survey analysis, a short paired-end Illumina DNA library with a 350 bp insert size was sequenced on the Illumina HiSeq 2500 sequencer. For PacBio Sequel sequencing, 50 μg of high molecular-weight (HMW) genomic DNA were prepared to generate two standard SMRTbell libraries with 20 Kb insertions. PacBio long reads were sequenced on the PacBio Sequel System (Pacific Biosciences) with SMRTbell Template Prep Kit 1.0-SPv3 (Pacific Biosciences), yielding 50 Gb PacBio data (read length N50 = 11.3 Kb).

Europe PMC Funders Author Manuscripts

Estimation of the Genome Size

Two approaches were used to estimate the size of *S. chinensis* genome: flow cytometry and k-mer spectra from 89 Gb of Illumina short reads. The flow cytometry technique was conducted with rice and *Nicotiana benthamiana* as the references. Additionally, we used the k-mer analysis to assess the genome size from short Illumina reads. K-mer frequencies were calculated using Jellyfish version 2.2.6 with k-mer size 21 (Marcais & Kingsford, 2011), and the resulting data was sent to GenomeScope version 1.0 for the estimation (Vurture *et al.*, 2017). The genome size was calculated with the following formula: genome size = k-mer number/peak depth.

Genome assembly and assessment

We used Canu version 1.8 to correct and assemble the raw PacBio reads with default parameters except for parameters: “genomeSize=550m correctedErrorRate=0.040” (Koren *et al.*, 2017). Following Canu, we polished the long-read assembled contigs using Pilon version 1.23 for eight iterative rounds (Walker *et al.*, 2014). In total, 244 Gbp of Illumina short reads were mapped to the contigs using Bowtie version 2.3.4.3 with parameter: “--sensitive-local” (Langmead & Salzberg, 2012), followed by Pilon polishing with parameter: “--fix bases” for four iterative rounds (Walker *et al.*, 2014). Another four iterative rounds of polishing with parameters: “--fix all” (bases, gaps, and local) were further performed to fix all the variants. Purge haplotigs version 1.1.0 was used to refine the assembly and collapse the homologous regions (Roach *et al.*, 2018), with parameters: “-l 10 -m 63 -h 195” for “purge_haplotigs contigcov” (“purge_haplotigs hist” and “purge_haplotigs purge” were performed with default parameters). The HI-C data were processed by the AllHiC pipeline (Zhang *et al.*, 2019) in the scaffolding process.

Identification of transposable elements and repetitive DNA

We used both *de novo* and homology-based predictions to identify the repetitive elements in the *Saururus chinensis* genome. We built a *de novo* repeat library by RepeatModeler 2.0 with the support of LTRStruct by adding parameter: “-LTRStruct” (McCarthy & McDonald, 2003). Additionally, we used TransposonPSI to detect (retro-) transposon ORF. We merged the libraries from RepeatModeler and TransposonPSI by USEARCH with 80% identity as the minimum threshold for combining similar sequences in the *de novo* libraries (Edgar, 2010). The resulting non-redundant *de novo* repeat library was further scanned to remove potential plant protein-coding sequences.

Finally, we used RepeatMasker version 4-1-0 with parameter: “-e rmblast -a -s -norna -no_is -xsmall -gff -html -lib” to identify and classify repeats in the genome assembly. The soft-masked genome sequence was used for downstream gene prediction and RNA-Seq mapping analysis.

Predictions of genes and noncoding RNAs

Transcriptome-based, homology-based, and *ab initio* prediction methods were applied to predict the gene models of *S. chinensis*. For transcriptome-based prediction, we mapped the RNA-seq reads to the repeat-masked genome using HISAT2 v2.0.5 from 12 libraries of 6 tissues including inflorescence, fruit, root, stem, leaf, and bract (Kim *et al.*, 2015). The

mapping results were used by Class (Song & Florea, 2013), Scallop (Shao & Kingsford, 2017), and StringTie (Pertea *et al.*, 2015) to create gene models, followed by reconciliation using Mikado (Venturini *et al.*, 2018). The transcripts from the genome-referenced assembly were further applied to obtain reliable transcripts with the longest open reading frames using TransDecoder (https://github.com/TransDecoder/TransDecoder). For the homology-based prediction, the protein sequences of uniprot (taxonomy viridiplantae with TE removed) were aligned to the genome using Exonerate ("--percent 50 --maxintron 50000") and genomeThreader ("-species arabidopsis -gff3out -skipalignmentout"). For the *ab initio* gene prediction, we used BRAKER2 in the "--etpmode" with the plant protein sequences (OrthoDB release 10) mapping results from the suggested pipeline (GaTech) (Hoff *et al.*, 2016). Finally, we used EVidenceModeler to reconcile the gene models obtained from the abovementioned approaches to generate consensus gene structures (Haas *et al.*, 2008).

Noncoding RNAs were predicted using tRNAscan-SE version 1.31 (Schattner *et al.*, 2005) and Infernal version 1.1.2 (Nawrocki *et al.*, 2009) (with cmsearch -E 0.001 against Rfam12 database).

Functional annotation

We used InterProScan version 5.39-77 (Jones *et al.*, 2014) to annotate the functions of detected motifs and domains by searching public databases (GO, PFAM, and KEGG). We used Mercator 4 from the MapMan website to annotate for the MapMan functional category (Lohse *et al.*, 2014) and used KofamKOALA to annotate the gene models with KEGG orthologues ID based on profile HMM (Aramaki *et al.*, 2020).

Phylogenetic reconstruction

To identify probable orthologous genes for phylogenetic analysis, we firstly downloaded the sequenced genomes from public resources (see Supplementary Table S8). Since the gene duplication and loss events occurred frequently during angiosperm history, to minimize the influence of paralogues and identify the most probable orthologues, 1614 genes from embryophyte_odb10 of BUSCO were used to identify the most probable orthologues within single-copy gene families using BUSCO v5.2.2 (Manni *et al.*, 2021). Then, the orthogroups were filtered under following criterions: 1) orthologous genes present in ≥75% genomes; 2) copy number ≤ 5 in each orthologous gene group in each species; 3) only the longest copy was chosen for phylogenetic analyses. This procedure developed the final refined single-copy genes (RSCGs).

Multiple sequence alignment of proteins was conducted using MAFFT 7.187 (Katoh & Standley, 2013), and the trimming of alignment and the coding sequence alignment reconstruction were conducted by trimAl (Capella-Gutierrez *et al.*, 2009). For the concatenated nucleotide sequence partitioning strategy, both first two codon positions were derived from the CDS alignments using a customized python script. IQTREE (Nguyen *et al.*, 2015)was used for model selection and subsequent reconstruction of the Maximum-likelihood phylogenetic tree for both CDS and protein alignments, consensus tree was obtained after 1000 times ultra-fast bootstrap. ML tree of each RSCG was reconstructed by IQTREE (Nguyen *et al.*, 2015)as mentioned beyond. The coalescence tree was reconstructed

by ASTRAL 5.7.3 (Mirarab & Warnow, 2015) based on gene trees from RSCGs, and the quartet supports were also obtained from ASTRAL using a parameter of -t=32. DISCOVISTA (Sayyari, et al. 2018) was used to summarize and visualize species trees.

Identification of whole-genome duplications

K_S-based age distributions for the whole paranome – all paralogs – in the *S. chinensis* genome was constructed as previously described (Vanneste *et al.*, 2013). Briefly, the paranome was constructed by performing an all-against-all protein sequence similarity search using BLASTP with an E-value cutoff of 1×10^{-10}, after which gene families were built with the mclblastline pipeline (v10-201) (micans.org/mcl) (Enright *et al.*, 2002). Each gene family was aligned using MUSCLE (v3.8.31) (Edgar, 2004), and K_S estimates for all pairwise comparisons within a gene family were obtained using maximum likelihood in the CODEML program (Goldman & Yang, 1994) of the PAML package (v4.4c) (Yang, 2007). We then subdivided gene families into subfamilies for which K_S estimates between members did not exceed 5. To correct for the redundancy of K_S values (a gene family of n members produces $n(n–1)/2$ pairwise K_S estimates for $n–1$ retained duplication events), we inferred a phylogenetic tree for each subfamily using PhyML (Guindon *et al.*, 2010) with the default settings. For each duplication node in the resulting phylogenetic tree, all m K_S estimates between the two child clades were added to the K_S distribution with a weight of $1/m$ (where m is the number of K_S estimates for a duplication event), so that the sum of the weights of all K_S estimates for a single duplication event was one. To identify synteny or collinear segments in the genome of *S. chinensis*, i-ADHoRe (v3.0) was used with the parameters level_2_only=FALSE (Proost *et al.*, 2012). The K_S distribution of paralogs located on collinear segments (anchor pairs) was calculated using maximum likelihood in the CODEML program of the PAML package (v4.4c) (Yang, 2007).

The K_S-based orthologue age distributions were constructed by identifying one-to-one orthologues between species by selecting reciprocal best hits (Moreno-Hagelsieb & Latimer, 2008), followed by K_S estimation using the CODEML program, as above. K_S distributions for one-to-one orthologs between the outgroup species *V. vinifera* and *S. chinensis*, *P. nigrum*, and *L. chinense* were used to compare the relative timing of the WGD in *S. chinensis* with speciation events within magnoliids. To quantify the differences in substitution rates, we used the orthologous K_S distributions obtained above to perform relative rate tests. First, the K_S distance between two species was estimated by the mode of their orthologous K_S distribution. Then, using the K_S distance between *S. chinensis* and *P. nigrum*, the K_S distance between *V. vinifera* and *S. chinensis*, and the K_S distance between *V. vinifera* and *P. nigrum*, we calculated distances to *S. chinensis* and *P. nigrum* after their divergence, respectively. Then, orthologous K_S between *S. chinensis* and *P. nigrum* was corrected by the double of the K_S distance to *S. chinensis*, assuming that *P. nigrum* has the same substitution rate as *S. chinensis*. Similarly, we corrected the orthologous K_S between *S. chinensis* and *L. chinense*.

To identify the duplication events that resulted in the 1,354 anchor pairs in *S. chinensis* and the 30,406 anchor pairs in *P. nigrum*, we performed phylogenomic analyses employing protein-coding genes from eight species, including six species from magnoliids plus *V.*

vinifera and *Amborella trichopoda*. OrthoFinder (v2.3.11) (Emms & Kelly, 2019) was used with default parameters to identify gene families based on sequence similarities. Then, 3,884 of the 31,760 anchor pairs with K_S values greater than five were removed. If the remaining anchors fell into different gene families, indicating incorrect assignment of gene families by OrthoMCL (Fischer *et al.*, 2011), we merged the corresponding gene families. In this way, we obtained 58,275 multi-gene gene families. Next, phylogenetic trees were constructed for the subset of 5,916 gene families with no more than 200 genes that had at least one pair of anchors and one gene from *A. trichopoda*. Multiple sequence alignments were produced by MUSCLE (v3.8.31) (Edgar, 2004) using default parameters. These were trimmed by trimAl (v1.4) (Capella-Gutierrez *et al.*, 2009) to remove low-quality regions based on a heuristic approach (-automated1) that depends on a distribution of residue similarities inferred from the alignments for each gene family. RAxML (v8.2.0) (Stamatakis, 2006) was then used with the GTR+Γ model to estimate a maximum likelihood tree starting with 200 rapid bootstraps followed by maximum likelihood optimizations on every fifth bootstrap tree. Gene trees were rooted based on genes from *A. trichopoda* if these formed a monophyletic group in the tree; otherwise, mid-point rooting was applied. The timing of the duplication event for each anchor pair relative to the lineage divergence events with bootstrap values greater than or equal to 80% was then inferred using the approach described in (Zhang *et al.*, 2017).

RNA-Seq and Transcriptomic data analysis

To discover the critical genes involved in the reversible changes of bract color, we sequenced the tissues of white and green leaves with three replicates, respectively (in addition to other tissues including inflorescence, fruit, root, and stem). <a table listing all RNA-Seq data?> The expression profile (read counts) of each library was calculated using the LSTrAP pipeline (Proost *et al.*, 2017), which utilizes Trimmomatic (Bolger *et al.*, 2014), Bowtie 2 (Langmead & Salzberg, 2012) and HTSeq (Anders *et al.*, 2015) for the read QC, read mapping, and read counting processes, respectively. After getting the count table, we used DESeq2 to identify the differentially expressed genes between bracts and leaves (Love *et al.*, 2014). Significant expressed genes were defined as genes with p-value < 0.05.

Functional characterization of ScPEL

Cloning of ScPEL—Primers of the ScPEL gene were designed based on the 5’ and 3’ UTR region sequences (Supplementary Table S11) to clone the complete coding sequence of ScPEL (start codon to stop codon) from the cDNA of white leaves. The PCR program for amplification of ScPEL is as follows: pre-denaturation: 95°C, 3 min, 1 cycle; denaturation:95°C, 30 s, annealing 56°C-72°C, 15 s, extension 72°C, 35 cycles; final extension: 72°C, 5 min, 1 cycle. The amplification product (blunt ends) was detected and the gel DNA was extracted for cloning and transformation.

Genetic transformation in *Arabidopsis thaliana*—Homologous recombinant primers were designed and the plant expression vector pMDC83, which is under the control of CaMV 35S promoter, was used to construct homologous recombinant vectors. *Agrobacterium tumefaciens* strain GV3101, carrying the recombinant vector, was used to infect *Arabidopsis* (Eco-type Col-0) plants by floral dip method (Clough and Bent 1998).

Putative transformants were identified as hygromycin-resistant seedlings that produced green leaves and well-established roots within the selective medium, and further examined by positive identification by PCR experiments (Supplementary Fig. S20). After the T2 generation seeds were harvested and dried, sterilized seeds were suspended in sterile water and plated on 1/2 MS culture medium (without antibiotic). After 14 days, the colors of the seedling leaves on the same plate were different, some of them showed dark green, and the others showed light green. The qRT-PCR results verified that the ScPEL gene was not expressed in the wild type lines, whereas highly expressed in the transgenic lines (Supplementary Fig. S21).

Chlorophyll Content Assays—A chlorophyll bleaching assay was used to measure the chlorophyll content of *Arabidopsis thaliana* leaves. (1) Cut 0.1 g (Fresh weight) leaves and immersed in 20 mL 80% ethanol at room temperature for 24 h (gently agitating in the dark); (2) The chlorophyll concentration was quantified using a spectrophotometer at wavelengths of 664 nm and 647 nm; (3) The formula of calculate the amount of chlorophyll extraction was as follow: Chlorophyll extraction amount (mmol/g) =7.93*A664nm+19.53*A647nm.

Identification of lignin biosynthesis pathway genes

All genes from different plant genomes were classified into orthologous gene families by OrthoFinder (Emms & Kelly, 2019), and the functionally characterized genes of *Arabidopsis thaliana* were used as markers to identify corresponding gene families, and gene numbers of different species were thus determined according to the genes assigned to each family.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J.Y.X. acknowledges the grant from the Fundamental Research Funds for the Central Universities (No. KYCXJC2022003), Y.V.d.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01), Z.Q.S. acknowledges funding from the Outstanding Young Teacher of "QingLan Project"of Jiangsu Province. The computational resources and services were provided by the Bioinformatics Center of Nanjing Agricultural University.

This work was funded by European Research Council (DOUBLE-TROUBLE 833522).

Data availability

The processed Illumina and PacBio reads have been deposited at NCBI with BioProject ID: PRJCA008755.The genome assembly, along with the gene models and the functional annotation, can be accessed at ORCAE: https://bioinformatics.psb.ugent.be/orcae/overview/Sauch.

References

!!! INVALID CITATION !!! (Chabannes et al., 2001).
!!! INVALID CITATION !!! (Yan et al., 2019).

Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31 (2) 166–169. [PubMed: 25260700]

APG IV A-p-g. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society. 2016; 181 (1) 1–20.

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020; 36 (7) 2251–2252. [PubMed: 31742321]

Barra-Jimenez A, Ragni L. Secondary development in the stem: when Arabidopsis and trees are closer than it seems. Current Opinion in Plant Biology. 2017; 35: 145–151. [PubMed: 28013083]

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30 (15) 2114–2120. [PubMed: 24695404]

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25 (15) 1972–1973. [PubMed: 19505945]

Chabannes M, Barakate A, Lapierre C, Marita JM, Ralph J, Pean M, Danoun S, Halpin C, Grima-Pettenati J, Boudet AM. Strong decrease in lignin content without significant alteration of plant development is induced by simultaneous down-regulation of cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD) in tobacco plants. Plant Journal. 2001; 28 (3) 257–270.

Chaw SM, Liu YC, Wu YW, Wang HY, Lin CYI, Wu CS, Ke HM, Chang LY, Hsu CY, Yang HT, et al. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nature Plants. 2019; 5 (1) 63–73. [PubMed: 30626928]

Chen JH, Hao ZD, Guang XM, Zhao CX, Wang PK, Xue LJ, Zhu QH, Yang LF, Sheng Y, Zhou YW, et al. Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation. Nature Plants. 2019; 5 (3) 328. [PubMed: 30675017]

Chen SP, Sun WH, Xiong YF, Jiang YT, Liu XD, Liao XY, Zhang DY, Jiang SZ, Li Y, Liu B, et al. The Phoebe genome sheds light on the evolution of magnoliids. Horticulture Research. 2020; 7 (1)

Cronquist, A. An Integrated System of Classification of Flowering Plants. Columbia University Press; 1981.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32 (5) 1792–1797. [PubMed: 15034147]

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26 (19) 2460–2461. [PubMed: 20709691]

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019; 20 (1) 238. [PubMed: 31727128]

Endress PK, Doyle JA. Reconstructing the ancestral angiosperm flower and its initial specializations. Am J Bot. 2009; 96 (1) 22–66. [PubMed: 21628175]

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30 (7) 1575–1584. [PubMed: 11917018]

Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ Jr. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics. 2011; Chapter 6

Gitzendanner MA, Soltis PS, Wong GKS, Ruhfel BR, Soltis DE. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. American Journal of Botany. 2018; 105 (3) 291–301. [PubMed: 29603143]

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution. 1994; 11 (5) 725–736. [PubMed: 7968486]

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59 (3) 307–321. [PubMed: 20525638]

Guo X, Fang D, Sahu SK, Yang S, Guang X, Folk R, Smith SA, Chanderbali AS, Chen S, Liu M, et al. Chloranthus genome provides insights into the early diversification of angiosperms. Nat Commun. 2021; 12 (1) 6930. [PubMed: 34836973]

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008; 9 (1) R7 [PubMed: 18190707]

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016; 32 (5) 767–769. [PubMed: 26559507]

Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, et al. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nat Commun. 2019; 10 (1) 4702. [PubMed: 31619678]

Ichikawa T, Nakazawa M, Kawashima M, Iizumi H, Kuroda H, Kondou Y, Tsuhara Y, Suzuki K, Ishikawa A, Seki M, et al. The FOX hunting system: an alternative gain-of-function gene hunting technique. Plant Journal. 2006; 48 (6) 974–985.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30 (9) 1236–1240. [PubMed: 24451626]

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30 (4) 772–780. [PubMed: 23329690]

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12 (4) 357–360. [PubMed: 25751142]

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27 (5) 722–736. [PubMed: 28298431]

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9 (4) 357–359. [PubMed: 22388286]

Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, et al. Origin of angiosperms and the puzzle of the Jurassic gap. Nat Plants. 2019; 5 (5) 461–470. [PubMed: 31061536]

Li Z, Van de Peer Y. A non-duplicated magnoliid genome. Nat Plants. 2021; 7 (9) 1162–1163. [PubMed: 34475527]

Liang HX. On the evolution and distribution of Saururaceae. Acta Botanica Yunnanica. 1995; 17 (3) 255–267.

Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ. 2014; 37 (5) 1250–1258. [PubMed: 24237261]

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15 (12) 550. [PubMed: 25516281]

Lv QD, Qiu J, Liu J, Li Z, Zhang WT, Wang Q, Fang J, Pan JJ, Chen ZD, Cheng WL, et al. The Chimonanthus salicifolius genome provides insight into magnoliid evolution and flavonoid biosynthesis. Plant Journal. 2020; 103 (5) 1910–1923.

Ma JX, Sun PC, Wang DD, Wang ZY, Yang J, Li Y, Mu WJ, Xu RP, Wu Y, Dong CC, et al. The Chloranthus sessilifolius genome provides insight into early diversification of angiosperms. Nature Communications. 2021; 12 (1)

Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 2021; 38 (10) 4647–4654. [PubMed: 34320186]

Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27 (6) 764–770. [PubMed: 21217122]

McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003; 19 (3) 362–367. [PubMed: 12584121]

Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T. Flowering-time genes modulate meristem determinacy and growth form in Arabidopsis thaliana. Nature Genetics. 2008; 40 (12) 1489–1492. [PubMed: 18997783]

Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 2015; 31 (12) 144–52.

Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008; 24 (3) 319–324. [PubMed: 18042555]

Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25 (10) 1335–1337. [PubMed: 19307242]

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32 (1) 268–274. [PubMed: 25371430]

One Thousand Plant Transcriptomes I. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019; 574 (7780) 679–685. [PubMed: 31645766]

Palmer JD, Soltis DE, Chase MW. The plant tree of life: an overview and some points of view. Am J Bot. 2004; 91 (10) 1437–1445. [PubMed: 21652302]

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015; 33 (3) 290–295. [PubMed: 25690850]

Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. 2012; 40 (2) e11 [PubMed: 22102584]

Proost S, Krawczyk A, Mutwil M. LSTrAP: efficiently combining RNA sequencing data into co-expression networks. BMC Bioinformatics. 2017; 18 (1) 444. [PubMed: 29017446]

Rendon-Anaya M, Ibarra-Laclette E, Mendez-Bravo A, Lan TY, Zheng CF, Carretero-Paulet L, Perez-Torres CA, Chacon-Lopez A, Hernandez-Guzman G, Chang TH, et al. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences of the United States of America. 2019; 116 (34) 17081–17089. [PubMed: 31387975]

Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018; 19 (1) 460. [PubMed: 30497373]

Roodt D, Li Z, Van de Peer Y, Mizrachi E. Loss of Wood Formation Genes in Monocot Genomes. Genome Biol Evol. 2019; 11 (7) 1986–1996. [PubMed: 31173081]

Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33 (Web Server issue) W686–689. [PubMed: 15980563]

Shang JZ, Tian JP, Cheng HH, Yan QM, Li L, Jamal A, Xu ZP, Xiang L, Saski CA, Jin SX, et al. The chromosome-level wintersweet (Chimonanthus praecox) genome provides insights into floral scent biosynthesis and flowering in winter. Genome Biology. 2020; 21 (1)

Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol. 2017; 35 (12) 1167–1169. [PubMed: 29131147]

Soltis DE, Soltis PS. Nuclear genomes of two magnoliids. Nature Plants. 2019; 5 (1) 6–7. [PubMed: 30626927]

Song B, Stocklin J, Armbruster WS, Gao Y, Peng D, Sun H. Reversible colour change in leaves enhances pollinator attraction and reproductive success in Saururus chinensis (Saururaceae). Ann Bot. 2018; 121 (4) 641–650. [PubMed: 29325003]

Song L, Florea L. CLASS: constrained transcript assembly of RNA-seq reads. BMC Bioinformatics. 2013; 14 (Suppl 5) S14

Spicer R, Groover A. Evolution of development of vascular cambia and secondary growth. New Phytol. 2010; 186 (3) 577–592. [PubMed: 20522166]

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22 (21) 2688–2690. [PubMed: 16928733]

Thien LB, Ellgaard EG, Devall MS, Ellgaard SE, Ramp PF. Population structure and reprodctive biology of Saururus cenuus L. (Saururaceae). Plant Species Biology. 1994; 9: 47–55.

Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013; 30 (1) 177–190. [PubMed: 22936721]

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience. 2018; 7 (8)

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017; 33 (14) 2202–2204. [PubMed: 28369201]

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9 (11) e112963 [PubMed: 25409509]

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111 (45) E4859–E4868. [PubMed: 25355905]

Xue JY, Dong SS, Wang MQ, Song TQ, Zhou GC, Li Z, Van de Peer Y, Shao ZQ, Wang W, Chen M, et al. Mitochondrial genes from 18 angiosperms fill sampling gaps for phylogenomic inferences of the early diversification of flowering plants. Journal of Systematics and Evolution. 2021.

Yan XJ, Liu J, Kim H, Liu BG, Huang X, Yang ZC, Lin YCJ, Chen H, Yang CM, Wang JP, et al. CAD1 and CCR2 protein complex formation in monolignol biosynthesis in Populus trichocarpa. New Phytologist. 2019; 222 (1) 244–260. [PubMed: 30276825]

Yang L, Su D, Chang X, Foster CSP, Sun L, Huang CH, Zhou X, Zeng L, Ma H, Zhong B. Phylogenomic Insights into Deep Phylogeny of Angiosperms Based on Broad Nuclear Gene Sampling. Plant Commun. 2020; 1 (2) 100027 [PubMed: 33367231]

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24 (8) 1586–1591. [PubMed: 17483113]

Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, Wang JY, Lin YC, Xu Q, Chen LJ, et al. The Apostasia genome and the evolution of orchids. Nature. 2017; 549 (7672) 379–383. [PubMed: 28902843]

Zhang LS, Chen F, Zhang XT, Li Z, Zhao YY, Lohaus R, Chang XJ, Dong W, Ho SYW, Liu X, et al. The water lily genome and the early evolution of flowering plants. Nature. 2020a; 577 (7788) 79. [PubMed: 31853069]

Zhang LS, Wu SD, Chang XJ, Wang XY, Zhao YP, Xia YP, Trigiano RN, Jiao YN, Chen F. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. Plant Cell and Environment. 2020b; 43 (12) 2847–2856.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019; 5 (8) 833–845. [PubMed: 31383970]

Zhao T, Zwaenepoel A, Xue JY, Kao SM, Li Z, Schranz ME, Van de Peer Y. Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun. 2021; 12 (1) 3498. [PubMed: 34108452]

Europe PMC Funders Author Manuscripts

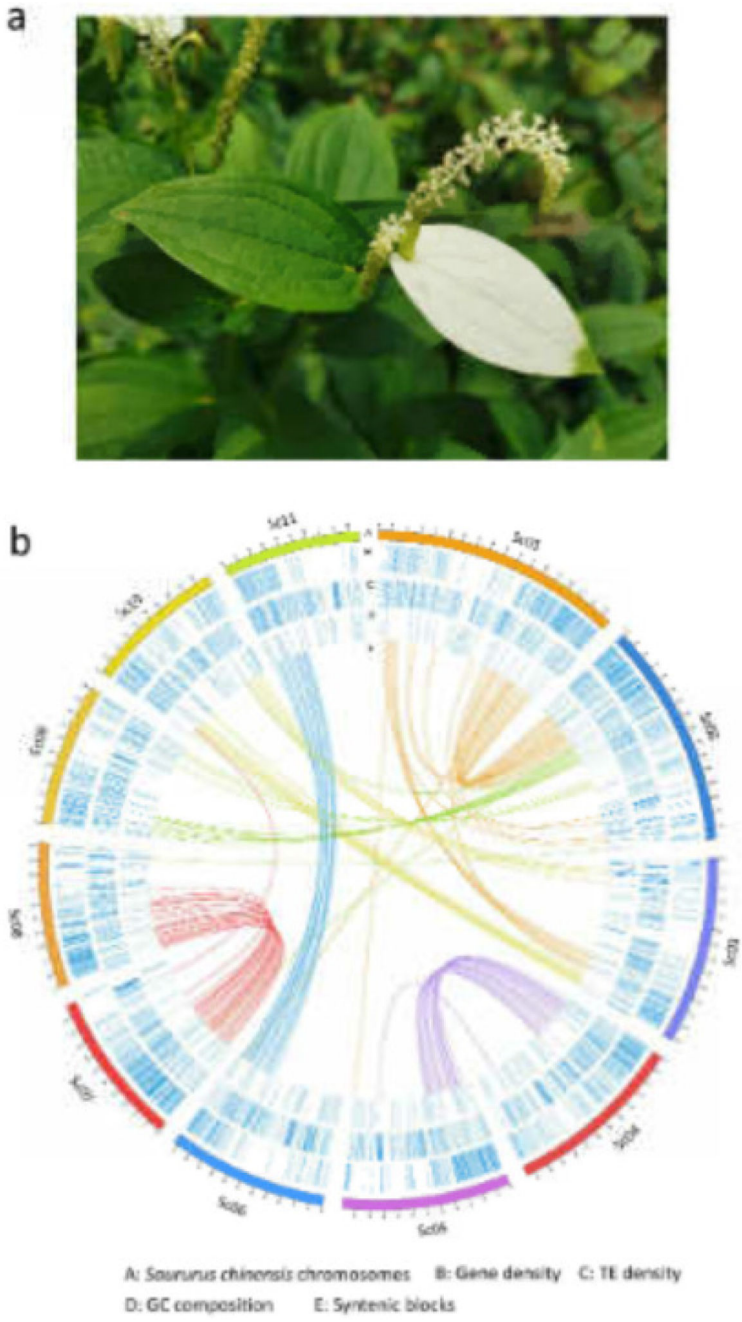

Fig. 1. *Saururus chinensis* photo and genome.

a) Photo of flowering *S. chinensis*.

b) Landscape of the *S. chinensis* genome. Concentric circles, from outermost to innermost, showing (A) Eleven longest pseudo-molecules corresponding to 11 chromosomes; (B) gene density; (C) TE density; (D) GC composition; (E) syntenic blocks.

Europe PMC Funders Author Manuscripts

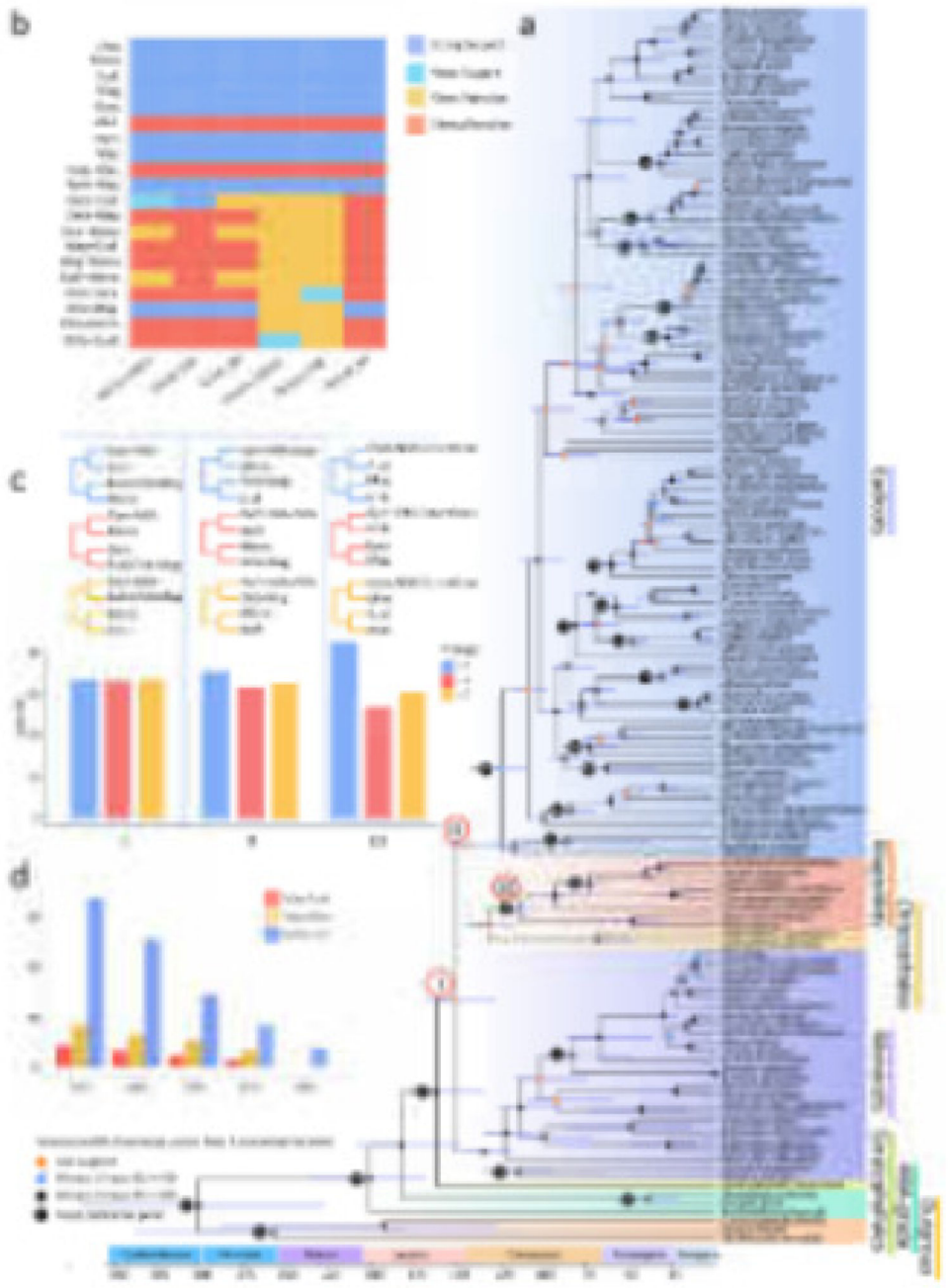

Fig. 2. Phylogenomic relationships and evolutionary timescale of 135 angiosperms and three gymnosperms.

a) Concatenation-based ML tree and divergence times based on 473 genes. Divergence times were estimated using 25 calibration points in MCMCTree. Fossil constraints are shown by red dots, and the details of fossils are available in Supplementary Table S9. Horizontal bars represent 95% credibility intervals.

b) Species tree analysis via DiscoVista. Rows correspond to focal splits, and columns correspond to the results of different methods reported in 138-taxa dataset. Yellow indicates

rejection of a clade; and weakly rejected clades correspond to clades that are not present in the tree, but are compatible if low support branches (below 90%) are contracted. AA: amino acid sequences; NT: nucleotide coding sequences; NT12: the combined first and second codon positions. ASTRAL: coalescent tree inference method using Astral software; CONCAT: maximum likelihood tree inferred with RaxML based on concatenated datasets. The chronogram of 138 seed plant species was inferred with MCMCTree based on 473 nuclear genes with concordant evolutionary histories.

c) Estimated proportions of the 473 gene trees with different topologies based on CDS, amino acid and Codon1+2 alignments. The x-axis labels q1, q2, and q3 refer to the quartet support for the main topology (blue), the first alternative (red), and the second alternative (yellow), respectively.

d) Numbers of shared gene families among either two lineages of eudicots, monocots and magnoliids, meanwhile absent in the third. Percentage indicates the minimum proportion of species in each lineage that have a certain gene family.

Europe PMC Funders Author Manuscripts

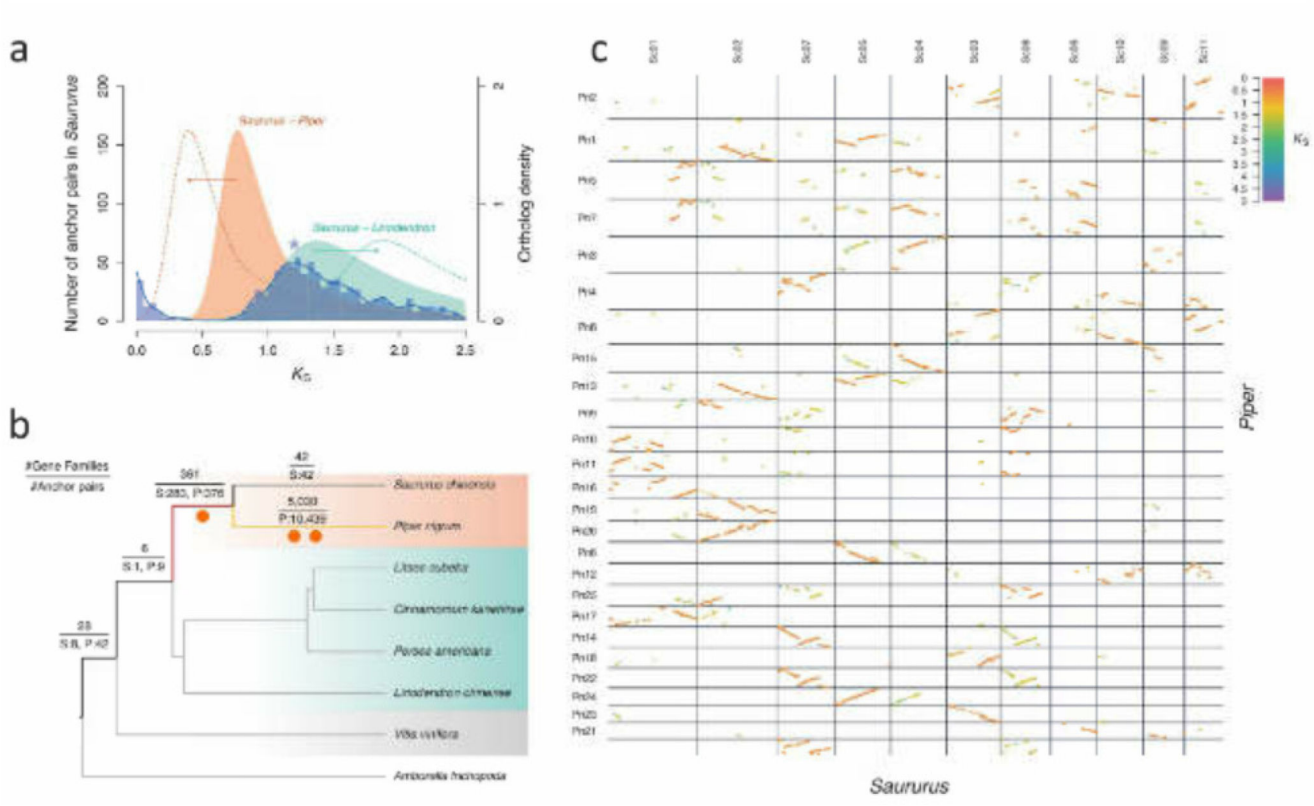

Fig. 3. Whole-genome duplication (WGD) in the *Saururus* genome.

a) K_S age distributions for anchor pairs of *S. chinensis* (left-hand y-axis; dark blue histogram and line; peaks under the star represent a WGD event) and for one-to-one orthologs between *S. chinensis* and *P. nigrum* and *L. chinense* (right-hand y-axis; red and green filled curves of kernel-density estimates, respectively; a peak represents a species divergence event). The arrows in different colors indicate under- (to the left) and overestimations (to the right) of the divergence events and point to the K_S values after corrections of different substitution rates based on that in *S. chinensis* (see Methods). The dotted curves also show the orthologous distributions after substitution rate corrections.

b) Phylogenomic analysis of the WGDs in *S. chinensis* and *P.nigrum*. The numbers on the branches of the species tree indicate the number of gene families with one or more anchor pairs from at least one of the *S. chinensis* and *P. nigrum* genomes that coalesced on the respective branch (top), as well as the individual contributions of anchor pairs (bottom) from the *S. chinensis* genome (S) and *P. nigrum* genome (P). The branches with WGD events are highlighted by two dots, and one dot under the yellow and red branches, respectively. All the duplication events have bootstrap values over 80% (see Methods).

c) Syntenic dot plot of the intergenomic comparison between *S. chinensis* and *P.nigrum*.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

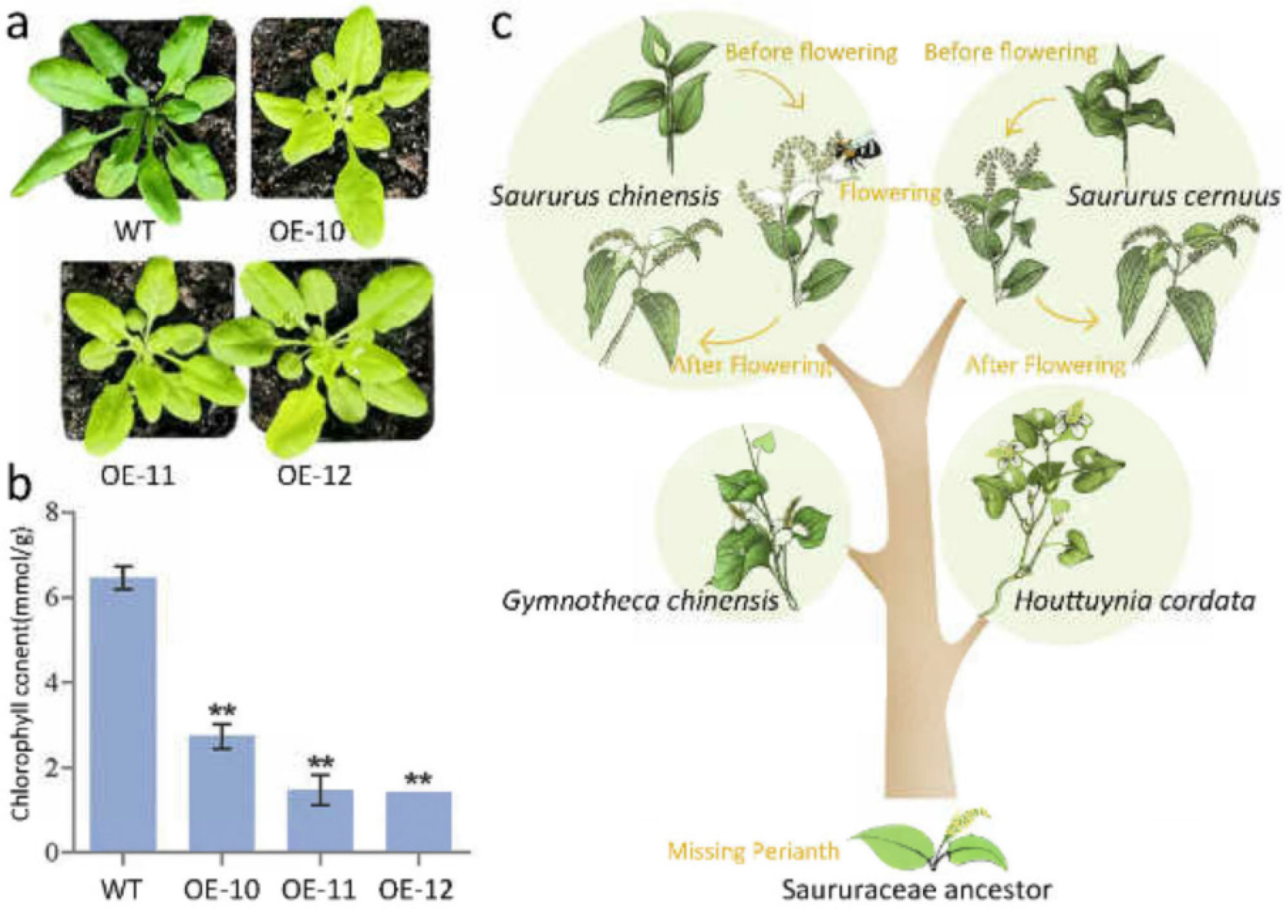

Fig. 4. Functional validation of *ScPEL* in *Arabidopsis thaliana* and inferred evolution for white organs and pollination in Saururaceae.

a) Rosette etiolation phynotype of *ScPEL* overexpressing plants and the wide types (WTs).

b) Chlorophyll content of *ScPEL* overexpressing plants and the WTs.

c) Hypothesized evolution of white organs and pollination in Saururaceae.

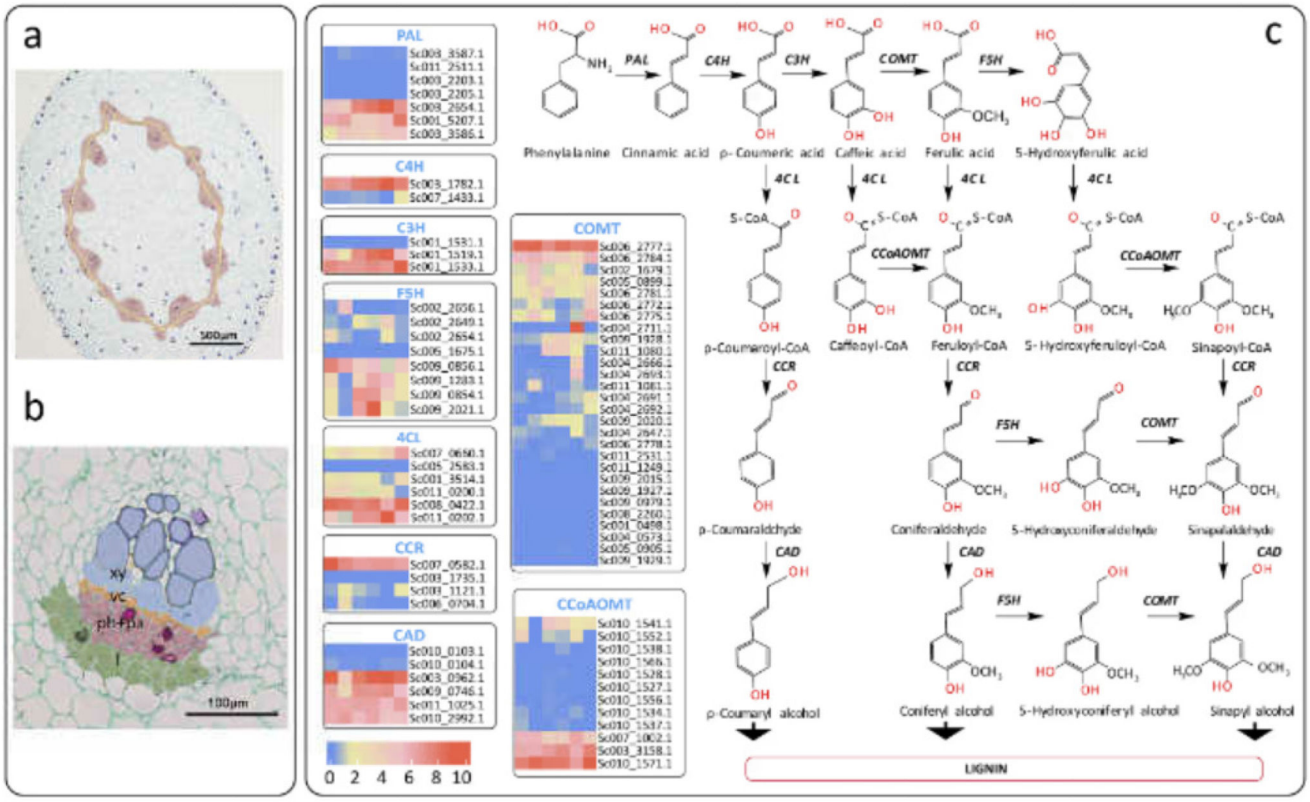

Fig. 5. Morphological and genetic characteristics of herbaceous *Saururus chinensis*.

a) Cross-section anatomy of *S. chinensis* stem; red, vascular bundles and tissue in cross-section; yellow, vascular cambium.

b) Labeled vascular components in one vascular bundle; f (green), fibers; ph+pa (red), phloem and vascular parenchyma; vc (yellow), vascular cambium; xy (purple), xylem.

c) Expressional profile of genes in the lignin biosynthesis pathway in *S. chinensis*. The right part shows the lignin biosynthesis pathway and enzymes (bold and italic) responsible for each catalytic step, and the left heatmaps display the expression of genes in (left to right) stems, roots, green leaves, white leaves, inflorescence and fruits, respectively.

Europe PMC Funders Author Manuscripts