Comparative genomics in Orobanchaceae provides insight into the origin and evolution of plant parasitism

Yuxing Xu[a], Jingxiong Zhang[a], Canrong Ma[a, c], Yunting Lei[a], Guojing Shen[a], Jianjun Jin[d], Deren A. R. Eaton[d], Jianqiang Wu[a, b]

[a] Department of Economic Plants and Biotechnology, Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China; [b] CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China; [c] University of Chinese Academy of Sciences, Beijing 100049, China; [d]Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY 10027, USA

*Jianqiang Wu

Email: wujianqiang@mail.kib.ac.cn

Author Contributions: Y.X. and J.W. designed research; J.Z., C.M., and G.S. prepared plant tissues, DNA samples, and RNA samples; Y.X., J.W., Y.L., G.S., and J.J. analyzed data; Y.X., J.W., J.J., and D.E. wrote the paper.

Competing Interest Statement: The authors declare no competing interest.

Classification: Biological Sciences: Plant Biology and Evolution.

Keywords: Genomics, Orobanchaceae, Parasitism, Whole-Genome Duplication, Gene Loss, Gene family expansion.

Abstract

Orobanchaceae is the largest family of plant parasites comprising autotrophic and parasitic plants with all degrees of parasitism, making it by far the best family for studying the origin and evolution of plant parasitism. Here we provide three high-quality genomes of orobanchaceous plants, the autotrophic *Lindenbergia luchunensis* and holoparasitic *Phelipanche aegyptiaca* and *Orobanche cumana*. Phylogenomic analysis on these three genomes and the previously published genomes and transcriptomes of other orobanchaceous species created a robust phylogenetic framework of Orobanchaceae. An ancient whole-genome duplication (WGD; about 79.29 Mya) was detected in the common ancestor of Orobanchaceae and *Mimulus guttatus*, and this event might have contributed to the origin of parasitism; however, except for *Striga*, no WGD events occurred in any lineages of orobanchaceous parasites after divergence from their autotrophic common ancestor, and the expansion of the genomes of orobanchaceous parasitic plants was mainly due to retrotransposons. We detected evident convergent gene loss in all parasites within Orobanchaceae and between Orobanchaceae and dodder *Cuscuta australis*. The gene families in the orobanchaceous parasites showed a clear pattern of recent gains and expansions, and the expanded gene families are enriched in functions related to development of haustorium. Thus, recent gene family expansions may have facilitated the adaptation of orobanchaceous parasites to different hosts. These data well support the previously proposed model of parasitic plant genome evolution (1). This work illustrates the trajectory of the evolution of orobanchaceous parasites, and will facilitate future studies on parasitism and the control of parasitic plants in agriculture.

Introduction

Most plants are autotrophs. However, about 4750 angiosperms are parasitic plants (2), which completely or partly depend on hosts to survive by extracting water and nutrients from the host plants. Some parasitic plants still have chlorophyll and photosynthesize, and they are named hemiparasitic plants, while some no longer photosynthesize and these are called holoparasites. According to whether hosts are needed for the completion of life cycles, parasitic plants can also be classified as the facultative ones and the obligate ones (3). In angiosperms, parasitic plants independently evolved 12 or 13 times from autotrophic plants (3). In eight parasitic plant lineages, hemiparasitic species became extinct, leaving only holoparasites remaining until today (3), whereas Orobanchaceae (Lamiales), is the largest family of parasitic plants, flourished with all types of parasitic lifestyles: facultative hemiparasites and obligate hemi- and holoparasites (Fig. 1) (3, 4). Orobanchaceae even has three genera of autotrophs: *Rehmannia*, *Triaenophora*, and *Lindenbergia*, and *Lindenbergia* is the closest relative to the parasitic lineages (5). Furthermore, some of the orobanchaceous parasites are economically important for being devastating pests in agriculture. For example, the annual economic loss caused by the witchweed (*Striga*) and broomrapes (*Orobanche* and *Phelipanche*) is over 10 billion US dollars in Asia and Africa (6). Orobanchaceae is thus the best extant system in which to study the evolutionary transition to plant parasitism, as well as to develop genomic resources for future agricultural applications (3).

Phylogenetic relationships among the major lineages of Orobanchaceae remain unclear. Previous research by McNeal et al. (7) split the family into six major clades based on three nuclear and two plastid markers, while a more recent study by Li et al. (5) analyzed five additional markers and recognized eight major clades, with incongruent relationships compared to the previous treatment. Following the Angiosperm Phylogeny Group nomenclature the Li et al. treatment recognized Rehmannieae, Lindenbergieae, Cymbarieae, Orobancheae, *Brandisia*, Rhinantheae, Buchnereae, and Pedicularideae. A surprising result of this study was the placement of the holoparasitic clade Orobancheae sister to all other parasitic Orobanchaceae, which suggests holoparasitism may have evolved earlier than other forms of parasitism. This conflicts with the proposed that the common ancestor of orobanchaceous parasitic plants, which was a facultative hemiparasite, evolved from an autotrophic plant, and later obligate hemiparasites appeared, and the latest ones were the obligate holoparasites, which no longer photosynthesize (3).

The evolution of parasitism requires major morphological and physiological innovations to perceive and recognize potential host plants and to form parasitic connections. In this respect, parasitic plants evolved a unique organ, haustorium, using which these parasites can attach and

60 penetrate host tissues (8). The haustoria of hemiparasites, such as *Triphysaria* and *Striga*, fuse
61 with the xylem of host plants, enabling these hemiparasites to obtain water and nutrients in the
62 host xylem sap. Holoparasites respectively fuse their haustorial xylem and phloem with the xylem
63 and phloem of hosts, and the fusion between parasite and host phloem tissues form symplastic
64 continuity by plasmodesmatal connections (8). In Orobanchaceae, the obligate hemiparasites and
65 holoparasites can perceive strigolactones released from host roots and activate the germination
66 process (9, 10). Recent studies revealed that in the orobanchaceous obligate parasites, the
67 karrikin receptor gene *KAI2* evolved *KAI2d* by duplication and neofunctionalization, and KAI2d
68 perceive the host root-released strigolactones, enabling these parasites to detect the existence of
69 host plants in the vicinity (11–13). After germination, some parasites also sense certain host
70 signals, named haustorium-inducing factors, such as DMBQ (2,6-dimethoxy-p-benzoquinone),
71 which is a cell wall-derived simple quinone (8), and thus develop haustoria.
72
73 Previous studies on *Cuscuta australis* (Convolvulaceae) (14), a shoot parasitic plant, *Striga*
74 *asiatica* (Orobanchaceae) (15), and an endoparasite *Sapria himalayana* (Rafflesiaceae) (16)
75 indicated that gene loss is common in parasitic plants. Wicke et al. (17) studied the plastomes of
76 various species of Orobanchaceae and found that the plastomes of orobanchaceous parasites
77 obviously experienced large-scale gene loss, especially the plastomes of holoparasites, in which
78 most of the photosynthesis-related genes are missing. Gene loss was detected even in the
79 orobanchaceous hemiparasites, such as the *ndh* genes (17). The degree of gene loss is
80 generally considered to be positively associated with the degree of parasitism (16), as a greater
81 degree of parasitism requires higher dependency on the host and fewer physiological processes
82 are needed by the parasite itself, and gene loss may enable the parasite to synchronize its own
83 physiology with that of the host, probably resulting in better fitness (18).
84
85 Thus far, four orobanchaceous species' genomes have been released, the autotrophic
86 *Rehmannia glutinosa* (19) and the obligate hemiparasite *Striga asiatica* (15) and facultative
87 hemiparasite *Phtheirospermum japonicum* (20) and *Pedicularis cranolopha* (21), but the genomes
88 of orobanchaceous holoparasites are still lacking. In this study, we sequenced and assembled
89 high-quality genomes of the autotroph *Lindenbergia luchunensis* and the holoparasites
90 *Orobanche cumana* and *Phelipanche aegyptiaca*, among which the latter two are notorious
91 parasitic pests in agriculture. Detailed comparative genomic analyses were performed using
92 these new genome data and the previously published genomes and transcriptomes. We
93 reconstructed a robust phylogeny and analyzed the trajectory of the genome evolution in
94 Orobanchaceae. Furthermore, our analysis on orobanchaceous genomes revealed the genomic
95 basis for parasitism evolution. This study fills the gap in our understanding of parasitism evolution
96 and provides new insight into the physiology and ecology of these parasites.

Results

Sequencing, assembly, and annotation of the genomes of *Lindenbergia luchunensis*, *Orobanche cumana*, and *Phelipanche aegyptiaca*

On a Nanopore platform, 36.47Gb data were generated for *L. luchunensis* (Llu), and on a PacBio Sequel II platform, 158.30 Gb and 396.00 Gb data were generated respectively for *O. cumana* (Ocu) and *P. aegyptiaca* (Pae). Using an MGI-SEQ 2000 platform, short reads (38.19 Gb, 141.20 Gb, and 329.33 Gb, respectively) were produced for Llu, Ocu, and Pae. Genome survey analysis using the short reads indicated low heterozygosities (SI Appendix, Table S1). Long reads were thus directly used for *de novo* assembling the genomes, and thereafter the short reads were used to correct the genome assemblies, forming contigs. The total length of the genome assembly of Pae is 3876.90 Mb (N50 9.97 Mb); using the Hi-C data, we obtained chromosome-level Llu and Ocu genome (Llu: 20 pseudochromosomes, 212.21 Mb; Ocu: 19 pseudochromosomes, 1417.94 Mb) (SI Appendix, Table S1). The accuracy of these three genomes all exceeded 99.9% (SI Appendix, Table S1). Thereafter, we first annotated repetitive sequences, including simple repeats, transposable elements, and RNA genes. In the genomes of Llu, Ocu, and Pae, the repetitive sequences respectively occupied 30.72%, 79.82%, and 84.40% (SI Appendix, Table S2). Mapping the RNA-seq data (SI Appendix, Table S3) from different organs to the respective genomes resulted in 29,669, 42,525, and 50,484 protein-coding genes in Llu, Ocu, and Pae. Benchmarking universal single-copy orthologs (BUSCO) analysis indicated that the completeness of these three genomes is 97.1% (Llu), 80.5% (Ocu), and 76.3% (Pae) (SI Appendix, Table S4). The low completeness values of the genes in Ocu and Pae genome is due to the large-scale gene loss in Ocu and Pae (see below for analysis), which is associated with their holoparasitic lifestyle.

Phylogenetic relationships in Orobanchaceae based on nuclear genes

Previously, it was established that there are eight major clades in Orobanchaceae; however, the phylogenetic relationships among the clades were still uncertain (5, 7, 22). The nuclear genes from all available genomes/transcriptomes of all the 19 Orobanchaceae species (SI Appendix, Table S5), covering seven out of the eight clades of Orobanchaceae, and three outgroup species (SI Appendix, Table S5) were used for the phylogenetic analysis pipeline (SI Appendix, Supplementary Methodology). Finally, after filtering out the orthogroups (OGs) which had low quality and/or high copy numbers, 907 OGs were selected for phylogenetic reconstruction. Coalescent and concatenation method yielded identical topologies with 100% support for all the nodes (Fig. 2). In this phylogenetic tree, consistent with previous findings (5, 7, 22, 23), the autotrophic lineages Rehmannieae and Lindenbergieae are the basal clades, and Lindenbergieae is the sister clade of all the parasitic plant lineages (Fig. 2). Within the group of parasitic plants,

134 Cymbarieae is the sister of all the rest lineages, a result different from Li et al. (5), which was
135 based on a few nuclear genes. In the rest four clades, Buchnereae and Orobancheae are sisters,
136 and so are Pedicularideae and Rhinantheae (Fig. 2); the topology of these four clades is distinct
137 from all previous studies (5, 7, 22, 23), except that it is similar to the topologies based on two
138 *matK* and *rps2* (7). Previously, several studies indicated that parasitic plants have higher
139 nucleotide substitution rates than typical autotrophic plants (24, 25); however, in our phylogenetic
140 tree, only the branch length of the holoparasitic plant *Aeginetia indica* in the clade Buchnereae is
141 exceptionally long, the branches of all holoparasitic plants in the clade *Orobancheae* were not
142 much longer than those of the autotrophic plants (SI Appendix, Fig. S1). We calculated the K_s
143 values of all the lineages in Orobanchaceae, but indeed no branches showed high substitution
144 rates (SI Appendix, Fig. S2).
145
146 Based on this tree, we estimated the divergence times for all the nodes using a Bayesian
147 molecular dating analysis. The estimated origin time of Orobanchaceae is about 46.51 million
148 years ago (Mya) (95% highest posterior density (HPD), 37.91-55.39 Mya), and the time
149 parasitism appeared was around 38.58 Mya (95% HPD, 31.27-45.81 Mya). Holoparasites
150 evolved independently at least three times in Orobanchaceae (Fig. 2 and (26)), among which,
151 Orobancheae appeared the earliest at ca. 33.95 Mya (95% HPD, 27.41-40.43 Mya).
152
153 **Genome evolution in Orobanchaceae**
154 Genome sizes and chromosome numbers vary in Orobanchaceae, and they seem to be positively
155 associated with the degrees of parasitism (4, 27). It is seemingly plausible that multiple rounds of
156 WGD events occurred during the evolution of the Orobanchaceae species (4, 27). To reveal the
157 WGD events in Orobanchaceae comprehensively, six species with reference genomes (Pae,
158 Ocu, and Llu from this study; *Striga asiatica* (Sas), *Phtheirospermum japonicum* (Pja), and
159 *Pedicularis cranolopha* (Pcr) from published/released data) and four outgroup species (*Mimulus*
160 *guttatus* (Mgu; Lamiales), *Solanum lycopersicum* (Solanales, lamiids), *Coffea canephora*
161 (Gentianales, lamiids), and *Arabidopsis thaliana* (Brassicales, rosids) were chosen for colinear
162 fragment comparisons. Instead of the commonly used methods based on colinear fragments
163 obtained by homology search, an analysis based on "orthologous colinear fragments (OCFs)"
164 was specifically designed for this study (see SI Appendix, Supplementary Methodology), which
165 enabled us to investigate the most recent WGD events in Orobanchaceae.
166
167 Our analysis based on OCFs from all species pairs (Fig. 3A) indicated that the Sas lineage
168 experienced an independent WGD event (named α^B event) (Fig. 3A), and this was consistent with
169 a previous study (15, 28). However, no WGD events were detected in any other orobanchaceous
170 lineages, including those containing Pae and Ocu (Fig. 3A). The ancestor of Orobanchaceae did

171 not experience any independent WGD events, even though there was a WGD event during the
172 evolution of the common ancestor of Orobanchaceae and *Mimulus*, which is named the β^L event
173 (Fig. 3A). Notably, the detected WGD events in the outgroup species are consistent with the
174 previous reports (29–31), supporting the reliability of our novel WGD analysis pipeline. Next, we
175 selected four nodes (S1-S4 in Fig. 3A) within the Lamiales clade, and all the orthologous genes
176 originated from these speciation events were selected for calculating the K_s values (SI Appendix,
177 Fig. S3). We found that the peak values of the K_s distributions had a linear relationship with the
178 divergence times of these nodes (R^2 = 0.9926, Fig. 3A). Therefore, the K_s values from the
179 paralogous genes resulted from the α^B and β^L events were used to infer the times of these two
180 events: α^B and β^L occurred at 31.00 ± 5.39 and 79.29 ± 5.39 Mya, respectively (Fig. 3A). The α^B
181 event likely occurred in Buchnereae soon after it diverged from Orobancheae at 35.78 Mya (95%
182 HPD, 28.97-42.52 Mya) (Fig. 2, 3A). The occurrence of β^L event happened to be close to the
183 Cretaceous-Paleocene boundary, when the well-known mass extinction event occurred (32) and
184 many plant lineages experienced WGD events (33, 34). We speculated that the β^L event may
185 have played an important role in enabling the common ancestor of Mgu and Orobanchaceae to
186 survive the extreme environmental changes.
187
188 Chromosome numbers is important for understanding the genome evolution of eukaryotes. The
189 chromosome numbers vary in Orobanchaceae (4, 27). Based on the chromosome-level genome
190 data of Ocu, Pcr, Llu and the outgroup species Mgu, we obtained 18 protochromosomes by
191 reconstructing the ancestral karyotype of Orobanchaceae. Analysis of the syntenic chromosomal
192 fragment indicated that Llu and Ocu exhibited good collinearity with the ancestral
193 protochromosomes (Fig. 3B, SI Appendix, Fig. S4). For example, the Chromosome 9 of Ocu and
194 the Chromosome 5 of Llu appeared to be very similar to the ancestral Chromosome 8. However,
195 Pcr has only eight chromosomes and its chromosomes experienced at least 11 fusion events.
196
197 Orobanchaceous species have variable genome sizes, ranging from 220 Mb to 8.7 Gb, and
198 parasites with greater degrees of parasitism generally have larger genome sizes (5). However,
199 our analysis suggests that most lineages in Orobanchaceae did not experience WGD events,
200 ruling out the possibility that the large genomes were resulted from ploidization. To gain insight
201 into the driving force of genome expansions, we compared the genome sizes of six
202 orobanchaceous species (Llu, Pja, Pcr, Sas, Ocu, and Pae) and the respective repetitive
203 elements (REs) in their genomes (Fig. 4A). Indeed, the large genome sizes were mainly resulted
204 from increased numbers of REs: in Ocu, Pja, and Pcr, which are about 1 Gb, 64-79% of their
205 genomes are REs; Pae has the largest genome (3.877 Gb), and 84% of the Pae gnome are REs,
206 among which the Long Terminal Repeats (LTRs) family is the most abundant. Notably, analysis
207 of the sequences of the LTRs in Llu, Pcr, Ocu, and Pae suggested that different subfamilies of

208 LTRs caused genome expansions in different species, especially specific LTR subfamilies
209 expanded to extreme extents very recently in Pae and Ocu (SI Appendix, Fig. S5). To determine
210 whether the massively increased numbers of LTRs could affect the coding regions in these
211 species, we compared the intron sizes in Pae and Llu, which have the largest and smallest
212 genomes in the orobanchaceous species of this study. We found that most introns retained their
213 lengths, regardless of genome sizes (Fig. 4B). Moreover, the contents of REs in the flanking
214 regions of genes became larger as the genome size increased, suggesting a possibility that the
215 REs may influence the expression of neighboring genes (Fig. 4C). We also noticed that even
216 though no WGD events were involved, Pja, Pcr, Pae, and Ocu with large genomes have many
217 more genes than does Llu, and this was mainly due to tandem or segmental duplication (30.96%-
218 56.33%), but not LTR-assisted "copy-paste" duplication (6.55%-17.40%) (SI Appendix, Table S6).
219 Given that the Chromosome 9 of Llu and Chromosome 5 of Ocu are homologous, we specifically
220 inspected the differences between these two chromosomes. Ocu's Chromosome 5 is eight times
221 larger than is Llu's Chromosome 9 (82 Mb vs 10 Mb). We found that many Class I
222 retrotransposons, including LTRs, in the proximal region of the Chromosome 5 of Ocu, where
223 gene densities are relatively low, supporting that expansion of LTRs is the main driving force of
224 increased genome sizes of orobanchaceous parasites (Fig. 4D). This finding is consistent with a
225 previous conclusion obtained by genome skimming (35).
226
227 **Evolution of genes and gene families in Orobanchaceae**
228 We analyzed all the gene families in these orobanchaceous species (Pae, Ocu, Sas, Pja, Pcr,
229 and Llu), *Cuscuta australis* (Cau), and their relatives (for orobanchaceous species: Mgu and
230 *Sesamum indicum*; for Cau: *Ipomoea nil* and *Solanum lycopersicum*), and two outgroup species,
231 *Coffea canephora* and *Arabidopsis thaliana* (Dataset S1). We found that, like Cau, all parasites of
232 Orobanchaceae experienced frequently gene family contractions; however, many gene families in
233 orobanchaceous parasites exhibited even more severe expansions, which is rare in Cau (Fig. 5).
234 Based on the phylogenetic tree (Fig. 2), we calculated the numbers of gene families that
235 underwent significant expansions and contractions across all branches (Fig. 5). Contractions
236 were frequent when parasitism and obligate parasitism first appeared in Orobanchaceae, and
237 thereafter, most lineages appeared to have independent gene family contractions. Most of the
238 gene family expansions happened independently in different species. Among the expanded gene
239 families, genes related to transcriptional regulation, transport, cell wall formation, phloem/xylem
240 histogenesis, and lateral root formation were enriched (Dataset S2). These gene may be involved
241 in the evolution of haustorium. Furthermore, we screened the transcriptomic data from Pja, Sas,
242 Ocu, and Pae for genes with high expression levels in haustoria (expression level of a given gene
243 is at least 1-fold greater in haustorium tissues than in other tissues), and found that at least
244 69.06% of these highly expressed genes in haustoria evolved through duplication (Fig. 6, SI

245 Appendix, Table S7). Next, these duplicated genes were used to determine the history of the
246 duplication events. Many of the duplication events happened in Mgu and Orobanchaceae's
247 MRCA, and according to our colinear fragment comparisons, 98.73% (545/552) of the duplication
248 events were caused by the β^{L} WGD event; the rest of the duplication events, most occurred
249 during the specific evolution of Pja, Sas, and Pae, but not in the common ancestor of
250 orobanchaceous parasites (Fig. 6, SI Appendix, Table S7).
251
252 Gene loss is common in parasitic plants (14, 16). We specifically split all the gene families into
253 OGs, and these OGs were thereafter used for gene loss analysis. Gene family contractions were
254 not additionally analyzed in other ways, as comparison of gene loss provides more detailed
255 information on the biological significance of the missing genes. A bioinformatic pipeline was also
256 employed to screen for false positives generated by errors from gene annotation and
257 orthogrouping (SI Appendix, Supplementary Methodology). As a comparison, Cau was included
258 in the gene loss analysis with five parasitic Orobanchaceae species, as *Cuscuta* and
259 Orobanchaceae are both Lamiids. *Sapria himalayana* (Rafflesiaceae) (16) was not included in our
260 analysis as it is phylogenetically very distant from Lamiids and is an endoparasite. We found that
261 the severity of gene loss is positively associated with the degree of parasitism: the facultative
262 hemiparasites Pja and Pcr respectively lost 2.41% and 2.95% of the conserved genes, the
263 obligate hemiparasite Sas lost 5.71%, and the obligate holoparasites Ocu, Pae, and Cau
264 respectively lost 13.11%, 13.85%, and 14.88% (Fig. 7A and B). The history of gene loss events in
265 orobanchaceous parasites was estimated: the common ancestor of all orobanchaceous parasites
266 lost only 74 genes, and 191 and 777 genes were lost respectively during the evolution of the
267 ancestors of obligate parasites and holoparasites (Fig. 7A). Notably, even though hundreds of
268 genes were lost during the species-specific evolution of these parasites (Fig. 7A), many genes
269 were convergently lost (Fig. 7A): by comparing the lost genes in Pedicularideae (Pja and Pcr) and
270 Orobancheae (Ocu and Pae), we found that 74-122 genes were convergently lost, which was well
271 supported by the hypergeometric test (Fig. 7A). Furthermore, in the closely related holoparasite
272 Ocu and Pae, 33.6% and 29.9% of the lost genes are species-specific, and selection analysis
273 indicated that significant portions of these Ocu- and Pae-retained genes, which had been
274 specifically lost in the other species, are under relaxed selection (SI Appendix, Fig. S6). It is
275 possible that Ocu and Pae are still experiencing genes loss, and these genes will be convergently
276 lost in the future due to relaxed selection.
277
278 We next compared the lost genes in the orobanchaceous parasites and Cau and it was found that
279 among the 1687 lost genes in Cau, more than half (889 genes, 57.14%) were also lost in Pae and
280 Ocu, and 75 genes were also lost in the hemiparasites Sas, Pja, and Pcr (Fig. 7B). It is evident
281 that all orobanchaceous parasites and Cau convergently lost various genes, suggesting that

282 some of these gene loss events could contribute to their adaptation to the parasitic lifestyle. To
283 gain further insight into the functions of the lost genes in Cau, Pae, and Ocu, which are all
284 holoparasites, we first performed gene ontology (GO) enrichment analysis on all their lost genes,
285 and the GO terms were next used for interspecies pairwise comparisons. The resulting common
286 and unique GO terms were further categorized by a network analysis (Fig. 7C). Pae and Ocu-lost
287 genes exhibited very similar GO terms, while the enriched GO terms from the Cau lost genes
288 appeared to be quite different from those of Pae or Ocu. In Cau, most of the lost genes' GO terms
289 are related to ion transport in root, ethylene biosynthesis, and response to nutrient and oxygen
290 levels, and this is consistent with the rootless morphology of Cau (Fig. 7C). The majority of the
291 lost genes in Pae and Ocu are involved in photosynthesis, chloroplast organization, and
292 chlorophyll biosynthesis; although Cau also lost genes related to these processes, the numbers
293 of GO terms under these processes are remarkably fewer than those from Pae or Ocu (Fig. 7C).
294
295 The regulation of flowering in some parasitic plants, especially in holoparasites, is considered to
296 be unique (18). We specifically inspected the flowering regulatory genes in Cau and all the
297 parasites in Orobanchaceae (Fig. 7D). Many important flowering regulatory genes are missing in
298 the holoparasites Pae, Ocu, and Cau, including *CO*, *SVP*, *ELF4*, *AGL19* and *AGL16*, and *CDF1*
299 and *CDF3* (Fig. 7D). In terms of the regulatory pathways of flowering, Cau lost more genes
300 related to the autonomous and hormone pathway than did Pae and Ocu, while Pae and Ocu lost
301 more genes involved in the circadian and photoperiod pathway than did Cau. We suppose that
302 this was due to the fact that Cau uses host plant's flowering signal FT to activate its own flowering
303 (18) and Pae and Ocu likely do not rely on light signaling to flower, as they spend a significant
304 amount of time living underground.
305
306 **The evolution of KAI2 gene family**
307 After perceiving the strigolactones (SLs) released from adjacent host roots, seeds of obligate
308 parasites in Orobanchaceae germinate (4, 9, 36). In normal autotrophic plants, D14, an
309 alpha/beta hydrolase protein, perceives SLs and regulates physiological processes of
310 development and stress resistance (37). In *Striga*, however, the receptor of host-released SLs
311 has been identified to be encoded by *KAI2d*, and importantly, *KAI2d* did not evolve from *D14*, but
312 from the *KAI2* gene, which encodes the receptor of karrikin, through tandem duplication, large
313 gene family expansion, and neofunctionalization (11, 15).
314
315 We analyzed the *KAI2* gene families in Orobanchaceae and a few autotrophic species from
316 Lamiids, and Arabidopsis was used as the outgroup species (Fig. 8A). Only a single *KAI2* was
317 identified in Arabidopsis, whereas in *Coffea canephora*, *Solanum lycopersicum*, and
318 Orobanchaceae at least two copies exist, suggesting that *KAI2* duplicated in the ancestor of

319 Lamiids resulting in the two paralogs, *KAI2c* and *KAI2i* (Fig. 8A). We specifically inspected the
320 *KAI2* gene-containing chromosomal fragments in Llu and *Solanum lycopersicum*, and found that
321 in both species, *KAI2c* and *KAI2i* are either on different chromosomes (Llu) or distantly located on
322 the same chromosome (~ 20 Mb in *Solanum lycopersicum*). Collinearity analysis indicated that
323 the two *KAI2* gene-containing chromosomal fragments are obviously homologous with the
324 fragment containing *KAI2* gene in Arabidopsis (Fig. 8B). Given that no WGD events were
325 detected in the ancestor of Lamiids ((30) and this study, Fig. 3), the duplication of *KAI2* in Lamiids
326 is a result of segmental duplication. Thereafter, *KAI2d* is derived from duplication of *KAI2i* (Fig.
327 8A); given that in Pja, Pcr, and Sas, *KAI2d* and *KAI2i* are adjacently located (Fig. 8B), *KAI2i* gave
328 rise to *KAI2d* by tandem duplication. In Pja and Pcr, the chromosome fragments containing *KAI2i*
329 and *KAI2d* show good collinearity with the chromosome fragment containing *KAI2i* in Llu, while in
330 Sas, Ocu, and Pae, the collinearity was not detected (Fig. 8B). Thus, we inferred that this tandem
331 duplication event, in which *KAI2d* was born, happened in the common ancestor of
332 orobanchaceous parasitic plants, and in Pja and Pcr the locations of these duplicated genes
333 remained, whereas in Sas, Pae, and Ocu both genes translocated to other loci (Fig. 8B).
334
335 **Discussion**
336
337 Orobanchaceae is the only family of plants with all types of parasitic lifestyles, and it is preferred
338 for studying the origin and evolution of parasitism. The phylogeny of Orobanchaceae has been
339 studied using plastome fragments (23, 38), ITS (7, 23), and individual nuclear genes (5, 7). Even
340 though most of these studies identified eight monophyly clades, the phylogenetic relationship
341 among these clades was not completely resolved. In this study, we assembled three genomes of
342 orobanchaceous plants, Llu, Ocu, and Pae. Together with the published/released transcriptome
343 and genome data of other orobanchaceous plants, this study covered seven out of eight major
344 clades of Orobanchaceae. Using 907 nuclear OGs, we reconstructed a highly resolved and well-
345 supported phylogeny of Orobanchaceae, and the divergence times were estimated using a
346 Bayesian molecular dating approach (Fig. 2). Based on the phylogeny, we propose that 1) the
347 common ancestor of Orobanchaceae, an autotrophic plant, evolved parasitism (haustorium) at
348 38.58 Mya (95% HPD, 31.27-45.81 Mya); 2) the Cymbarieae, which is composed of only
349 facultative hemiparasites, was the earliest diverged parasitic lineage with low biodiversity; 3) by
350 Oligocene (33.90-23.03 Mya), a facultative hemiparasitic ancestor rapidly evolved into two
351 lineages; one diverged into Buchnereae (facultative hemiparasitic, obligate hemiparasitic, and
352 holoparasitic species) and Orobancheae (holoparasites), while the other evolved to be
353 Pedicularideae (facultative hemiparasitic species) and Rhinantheae (facultative hemiparasitic and
354 holoparasitic); these four clades have diverse species. Grasslands rapidly expanded during
355 Oligocene (39), and this event may have provided orobanchaceous parasites with opportunities

356 to exploit grasses and herbs, enabling the expansion of these parasites (7, 40). Notably, different
357 from a recent study (5), in our scenario, hemiparasites evolved earlier than did holoparasites, and
358 this is consistent with the progressive evolution of increasing dependency of hosts.
359
360 Most species in Orobanchaceae have 8 to 20 pairs of chromosomes, but the chromosome
361 number can reach 72 in some species (4, 27). Based on the chromosome number variation, Lyko
362 and Wicke (27) proposed at least three ancient WGD events during the evolution of
363 Orobanchaceae, including one WGD event in *Striga* and another WGD event in the Orobancheae
364 clade. The WGD event in *Striga* was confirmed by the recent genome analysis (15). However, our
365 collinearity analysis of the genome data of Llu, Pga, Pcr, Sas, Ocu, and Pae did not find other
366 WGD events in Orobanchaceae, except a single WGD event (α^B) that occurred during the
367 evolution of the ancestor of Buchnereae (31.00 ± 5.39 Mya) (Fig. 3A), a clade where Sas resides.
368 Moreover, the 1:1 collinearity of the Llu and Ocu's chromosomes (SI Appendix, Fig. S4) also
369 strongly supports that no ploidization happened after the divergence between Llu and Ocu. Even
370 though no WGD events were detected at the times of the emergence of parasitism and
371 holoparasitism, there was a very ancient WGD event (β^L, 79.29 ± 5.39 Mya) in the common
372 ancestor of Orobanchaceae and *Mimulus* (Fig. 3A). In this study, we reconstructed the 18
373 protochromosomes of Orobanchaceae ancestor using the genomes of species from *Orobanche*,
374 *Pedicularis*, and *Lindenbergia* and the outgroup *Mimulus*. The chromosome numbers and
375 collinearity are well retained in *Lindenbergia* and *Orobanche*; however, the chromosome-level
376 genome of Pcr indicates that Pcr has only eight chromosomes (Fig. 3B). Thus, Pcr must have
377 experienced large-scale chromosome recombination and fusion events. *Striga* has only 12
378 chromosomes (4), even though WGD happened during the evolution of *Striga*. We inferred that
379 *Striga* also experienced many chromosome recombinations and fusions.
380
381 The dodder Cau and the endoparasite *Sapria himalayana* are both non-orobanchaceous species.
382 Cau has only 19671 genes (14), and it is among the higher plants with the least number of
383 protein-coding genes. *Sapria himalayana* genome harbors 55179 genes; however, most of them
384 are transposon-like sequences, leaving only about 10000 low-copy genes (16). By contrast,
385 orobanchaceous parasites have many more genes than do Cau and *Sapria himalayana* (Fig. 5),
386 and this was due to species-specific gene family expansions in different orobanchaceous
387 parasites (Fig. 5). Furthermore, among these expanded gene families, genes likely related to
388 haustorium development and function were identified and many of the genes highly expressed in
389 haustoria were evolved from lineage-specific duplications (Fig. 6). We propose that species-
390 specifically expanded gene families may be the driving force underlying the adaptation of the
391 parasites to different hosts.
392

393 Similar to Cau and *Sapria himalayana*, large-scale gene loss was also detected in
394 orobanchaceous parasites, especially in the holoparasites Ocu and Pae. The reconstructed gene-
395 loss history indicated that only 74 genes were missing in the common ancestor of
396 orobanchaceous parasites, when parasitism evolved; by contrast, 777 genes were deleted from
397 the genome of the MRCA of Ocu and Pae. Thus, gene loss became more severe as the degree
398 of parasitism increased. Considering our data of gene loss in Orobanchaceae and the previous
399 published data from Orchidaceae (41) and *Sapria himalayana* (16), in parasitic plants and
400 mycoheterotrophic plants, which are all heterotrophs, the degree of gene loss and parasitism
401 seems to be well positively correlated as well. Accordingly, heterotrophic plants could be divided
402 into four classes with increasing levels of heterotrophy and degrees of gene loss: in the first class,
403 heterotrophs with 2-3% gene loss are facultative hemiparasites; in the second class, heterotrophs
404 with 6-7% gene loss require hosts to complete certain developmental stages, and these are
405 obligate hemiparasites and initial mycoheterotrophic plants, such as most species of
406 Orchidaceae; in the third class, heterotrophs lost 13-15% genes, and these are holoparasites or
407 full mycoheterotrophs with vegetative organs, such as *Cuscuta*, *Orobanche*, and *Gastrodia*; in the
408 fourth class, gene loss is over 30% and these are endoparasites with only reproductive organs,
409 such as *Sapria himalayana*. The lost genes in the heterotrophs of the first and second class seem
410 to be rather random, while the genes lost in the heterotrophs of the third class are related to
411 photosynthesis, transcriptional regulation, stress adaptation, and organogenesis; the lost genes in
412 endoparasite *Sapria himalayana*, from the fourth class, includes almost all those lost in Pae, Ocu,
413 Cau, and *Gastrodia*, representing the extreme form of parasitism.
414
415 Importantly, comparisons among orobanchaceous species and between Orobanchaceae and
416 Cau indicated convergent evolution of gene loss, suggesting that abandonment of certain
417 physiological functions is positive selected in parasitic plants. In *Striga*, the *protein phosphatase*
418 *2C* gene is mutated, leading to abscisic acid (ABA) insensitivity, and the impaired ABA signaling
419 pathway led to very high transpiration in *Striga*, enabling this parasite to absorb nutrients from
420 host xylem sap, even under drought conditions (15, 42). The leafless *Cuscuta australis* lost
421 almost all the flowering regulatory pathways, and its *FT* gene, which encodes an important
422 flowering signal, had pseudogenized; however, Cau uses the FT proteins from host plants to
423 activate its own flowering, enabling it to synchronize its flowering times with those of the hosts
424 (18). In this manner, Cau can parasitize a large number of host species with varying flowering
425 times, and thus, losing the flowering genes is an elegant strategy of adaptation in Cau (18). It
426 would be interesting to study what pathways are affected by gene loss in orobanchaceous
427 parasites and importantly the physiological and ecological impact incurred by gene loss.
428

429 Searcy and MacInnis (1) proposed a model of parasitic plant evolution: first, development of
430 haustorium which enables physical connections with host plants; after haustorium evolved, the
431 parasites start to lose genes that are no longer needed by the parasites and parasites evolve to
432 adapt to their specific hosts. The evolution of orobanchaceous parasites well supports this model.
433 The β^L WGD event in the common ancestor of Orobanchaceae and *Mimulus* generated many
434 duplicated genes which likely contributed to the evolutionary gain of a haustorium; thereafter,
435 gene loss events occurred in orobanchaceous parasites, especially in holoparasites, and in
436 parallel, expansions in specific gene families may have contributed to specialization and
437 adaptation of the parasites to the hosts.
438
439 In this study, our phylogenomic analysis constructed a robust phylogenetic framework of the
440 major lineages of Orobanchaceae. Furthermore, genome evolution and comparative genomics
441 analysis shed new light on the origin and evolution of parasitism in Orobanchaceae. These
442 genome resources and further molecular, genetic, and genomic studies on Orobanchaceae
443 species will provide many interesting insights into the evolution underlying the drastic
444 physiological and morphological changes in these fascinating plants.
445
446 **Materials and Methods**
447
448 **Genome sequencing**
449 Leaves of Llu and aboveground parts of stems of Ocu and Pae were harvested and cleaned with
450 tap water and used for DNA extraction. For short-read sequencing: DNA libraries were
451 constructed using a VAHTS Universal Plus DNA Library Prep Kit for MGI (Vazyme, Nanjing,
452 China) following the manufacturer's instructions, and the libraries were sequenced on an MGI-
453 SEQ 2000 (MGI Tech, Shenzhen, China) to generate 150-bp paired-end data. For long-read
454 sequencing: for Llu, we used the PromethION platform from Oxford Nanopore Technologies; for
455 Ocu and Pae, we used the PacBio Sequel II platform.
456
457 Tissues of Llu and Ocu were used for the Hi-C pipeline following a previously described method
458 (43), including crosslinking, chromatin digestion with MboI (New England BioLabs, MA), labeling
459 of DNA ends, ligation, purification, shearing, and biotin pull-down. The Hi-C libraries were
460 sequenced on an MGI-SEQ 2000 (MGI Tech, Shenzhen, China) platform in the 150-bp paired-
461 end mode.
462
463 To aid genome annotation, we obtained RNA-seq data from different tissues of orobanchaceous
464 plants (SI Appendix, Table S3). The total RNA was extracted from these tissues using the TRI
465 Reagent (Sigma). The MGIEasy RNA library preparation kit was used for library construction.

466 From a DNBSEQ-T7 platform (MGI Tech, Shenzhen, China), 150bp pair-end sequencing reads
467 were obtained.
468
469 **Genome assembly**
470 We first surveyed the genome of Llu, Ocu, and Pae based on *k*-mer (*k*=17) depth frequency
471 distribution analysis using the GCE software (v1.0.0) (44) with short-read data, then obtained the
472 genome size, repeat content, and heterozygosity for these species (SI Appendix, Table S1).
473
474 For Ocu and Pae, the subreads from the Pacbio sequencing were corrected, trimmed, and
475 assembled by CANU (v2.2) (45). Contigs were polished with the PacBio data using GCpp (v1.9.0)
476 (https://github.com/PacificBiosciences/gcpp) and then corrected with short-reads data using Pilon
477 (v1.23) (46). For Llu, the long reads of ONT were de novo assembled using NextDenovo (v2.3.1)
478 (https://github.com/Nextomics/NextDenovo), and the obtained contigs were corrected with
479 NextPolish (v1.1.0) (47) using both short reads and long reads.
480
481 To obtain chromosome-scale assemblies for Llu and Ocu, we mapped the Hi-C data to contigs
482 and constructed the Hi-C contact matrices, which were used to correct the assembly of contigs
483 and build scaffolds using the 3D-DNA pipeline (v180922) (48).
484
485 In order to evaluate the quality of the assembly, short reads were mapped to the assemblies
486 using HISAT2 (49), and the results from mapping were used for calling SNPs and indels using
487 SAMtools (v1.9) (50) and BCFtools (v1.9) (51). The homozygous SNPs and indels were
488 considered as assembly error, and thus the correct rates of assembly were calculated.
489
490 **Genome annotation**
491
492 We first used the EDTA pipeline (v1.8.3) (52) to annotate repetitive sequences in the genomes,
493 which integrates with the LTR_FINDER (53), LTRharvest (54), LTR_retriever (55), Generic
494 Repeat Finder (56), TIR-learner (57), HelitronScanner (58), and TEsorter (59). EDTA first used
495 these software to de novo annotate genome-wide transposons, and then the resulting
496 transposons were integrated with the Dfam database (https://dfam.org/) into RepeatMasker
497 (http://repeatmasker.org) to annotate repetitive elements.
498
499 To annotate the protein-coding genes, four steps were used to guide the annotation. 1) Augustus
500 (v3.4.0) (60) for gene model prediction was trained with the MAKER-P pipeline (61) using the
501 RNA-seq data and the protein-coding genes in the close species (*Striga asiatica*, *Mimulus*
502 *guttata*, *Arabidopsis thaliana*, *Solanum lycopersicum*, *Coffea canephora*, and *Cuscuta australis*).

503 After three rounds of training of Augustus, de novo annotation of protein-coding genes was done.
504 2) The RNA-seq data were mapped to unmasked genomes using TopHat2 (v2.1.1) (62), and the
505 mapped reads were assembled to gene structures by Cufflinks (v2.2.1) (63); TRINITY (v2.8.5)
506 (64) was used to de novo assemble the RNA-seq reads and thereafter PASA2 (v2.4.1) (65) was
507 used to map the assembled contigs to the genomes and annotate the gene structures. These
508 annotation results were merged. 3) The protein-coding genes in the closely related species were
509 aligned to the unmasked genomes using GenBlastA (v1.0.1) (66), and GeneWise2 (v2.4.1) (67)
510 was employed to generate gene structures. 4) The gene structures obtained from the above three
511 steps were combined by EVidenceModeler (v1.1.1) (68). Then, the gene annotation results were
512 compared with the repetitive element annotation results to remove the genes that overlap with
513 repetitive elements. The final annotations were achieved.
514
515 See SI Appendix, Supplementary Methodology for phylogenetic analysis and comparative
516 genomics analysis.
517
518 **Data availability**
519 The genome assembly, gene models, and sequence reads of Llu, Ocu, and Pae are available at
520 the National Genomics Data Center (https://ngdc.cncb.ac.cn/) under BioProject number
521 PRJCA008847. All data are also available from the corresponding author upon request.
522
523 **Acknowledgments**
524
525 We thank Dr. Xiaopeng Yun (Inner Mongolia Academy of Agricultural and Animal Husbandry
526 Sciences) for providing Ocu and Pae seeds, and the Germplasm Bank of Wild Species in
527 Southwest China for providing Llu seeds. Drs. Jun Zhao and Jian Zhang (Inner Mongolia
528 Agricultural University) are thanked for identification of Ocu. We are grateful to the High-
529 Performance Computing Facility of Department of Economic Plants and Biotechnology at
530 Kunming Institute of Botany for providing computing services. This study was supported the
531 National Natural Science Foundation of China (No. 32000179, Y.X.), Strategic Priority Research
532 Program of Chinese Academy of Sciences (No. XDPB16, J.W.), CAS "Light of West China"
533 Program (G.S.), the Special Research Assistant of Chinese Academy of Sciences (Y.X. and J.Z.),
534 and the Postdoctoral Directional Training Foundation of Yunnan Province (Y.X.).
535
536 **References**
537
538 1. D. G. Searcy, A. J. MacInnis, Measurements by DNA renaturation of the genetic basis of
539 parasitic reduction. *Evolution* **24**, 796–806 (1970).

2. D. L. Nickrent, Parasitic angiosperms: How often and how many? *TAXON* **69**, 5–27 (2020).
3. J. H. Westwood, J. I. Yoder, M. P. Timko, C. W. dePamphilis, The evolution of parasitism in plants. *Trends in Plant Science* **15**, 227–235 (2010).
4. D. M. Joel, J. Gressel, L. J. Musselman, Eds., *Parasitic Orobanchaceae* (Springer Berlin Heidelberg, 2013) https:/doi.org/10.1007/978-3-642-38146-1 (October 17, 2021).
5. X. Li, T. Feng, C. Randle, G. M. Schneeweiss, Phylogenetic relationships in Orobanchaceae inferred from low-copy cuclear genes: Consolidation of major clades and identification of a novel position of the non-photosynthetic orobanche clade sister to all other parasitic Orobanchaceae. *Front. Plant Sci.* **10**, 902 (2019).
6. A. D. Twyford, Parasitic plants. *Current Biology* **28**, R857–R859 (2018).
7. J. R. McNeal, J. R. Bennett, A. D. Wolfe, S. Mathews, Phylogeny and origins of holoparasitism in Orobanchaceae. *American Journal of Botany* **100**, 971–983 (2013).
8. S. Yoshida, S. Cui, Y. Ichihashi, K. Shirasu, The haustorium, a specialized invasive organ in parasitic plants. *Annu. Rev. Plant Biol.* **67**, 643–667 (2016).
9. C. E. Cook, L. P. Whichard, B. Turner, M. E. Wall, G. H. Egley, Germination of witchweed (Striga lutea Lour.): Isolation and properties of a potent stimulant. *Science* **154**, 1189–1190 (1966).
10. C. E. Cook, *et al.*, Germination stimulants. II. Structure of strigol, a potent seed germination stimulant for witchweed (Striga lutea). *J. Am. Chem. Soc.* **94**, 6198–6199 (1972).
11. C. E. Conn, *et al.*, Convergent evolution of strigolactone perception enabled host detection in parasitic plants. *Science* **349**, 540–543 (2015).
12. S. Toh, *et al.*, Structure-function analysis identifies highly sensitive strigolactone receptors in Striga. *Science* **350**, 203–207 (2015).
13. Y. Tsuchiya, *et al.*, PARASITIC PLANTS. Probing strigolactone receptors in Striga hermonthica with fluorescence. *Science* **349**, 864–868 (2015).
14. G. Sun, *et al.*, Large-scale gene losses underlie the genome evolution of parasitic plant Cuscuta australis. *Nat Commun* **9**, 2683 (2018).
15. S. Yoshida, *et al.*, Genome sequence of Striga asiatica provides insight into the evolution of plant parasitism. *Curr Biol* **29**, 3041-3052.e4 (2019).
16. L. Cai, *et al.*, Deeply altered genome architecture in the endoparasitic flowering plant Sapria himalayana Griff. (Rafflesiaceae). *Current Biology* **31**, 1002-1011.e9 (2021).
17. S. Wicke, *et al.*, Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. *PNAS* **113**, 9045–9050 (2016).
18. G. Shen, *et al.*, *Cuscuta australis* (dodder) parasite eavesdrops on the host plants' FT signals to flower. *Proc Natl Acad Sci USA* **117**, 23125–23130 (2020).

575 19. L. Ma, *et al.*, De novo genome assembly of the potent medicinal plant Rehmannia glutinosa
576 using nanopore technology. *Computational and Structural Biotechnology Journal* **19**, 3954–
577 3963 (2021).
578 20. S. Cui, *et al.*, Ethylene signaling mediates host invasion by parasitic plants. *Science*
579 *Advances* (2020) https:/doi.org/10.1126/sciadv.abc2385 (January 4, 2022).
580 21. J. Jin, Pedicularis cranolopha genome reference project.
581 https://www.ncbi.nlm.nih.gov/bioproject/805569
582 22. G. M. Schneeweiss, "Phylogenetic relationships and evolutionary trends in Orobanchaceae"
583 in *Parasitic Orobanchaceae*, D. M. Joel, J. Gressel, L. J. Musselman, Eds. (Springer Berlin
584 Heidelberg, 2013), pp. 243–265.
585 23. A. D. Wolfe, C. P. Randle, L. Liu, K. E. Steiner, Phylogeny and biogeography of
586 Orobanchaceae. *Folia Geobot* **40**, 115–134 (2005).
587 24. B. Lemaire, S. Huysmans, E. Smets, V. Merckx, Rate accelerations in nuclear 18S rDNA of
588 mycoheterotrophic and parasitic angiosperms. *J Plant Res* **124**, 561–576 (2011).
589 25. L. Bromham, P. F. Cowman, R. Lanfear, Parasitic plants have increased rates of molecular
590 evolution across all three genomes. *BMC Evolutionary Biology* **13**, 126 (2013).
591 26. W. Fu, *et al.*, Testing the hypothesis of multiple origins of holoparasitism in Orobanchaceae:
592 Phylogenetic evidence from the last two unplaced holoparasitic genera, Gleadovia and
593 Phacellanthus. *Frontiers in Plant Science* **8**, 1380 (2017).
594 27. P. Lyko, S. Wicke, Genomic reconfiguration in parasitic plants involves considerable gene
595 losses alongside global genome size inflation and gene births. *Plant Physiology* **186**, 1412–
596 1423 (2021).
597 28. N. J. Wickett, *et al.*, Transcriptomes of the parasitic plant family Orobanchaceae reveal
598 surprising conservation of chlorophyll synthesis. *Curr Biol* **21**, 2098–2104 (2011).
599 29. S. Sato, *et al.*, The tomato genome sequence provides insights into fleshy fruit evolution.
600 *Nature* **485**, 635–641 (2012).
601 30. F. Denoeud, *et al.*, The coffee genome provides insight into the convergent evolution of
602 caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).
603 31. C. Simillion, K. Vandepoele, M. C. E. V. Montagu, M. Zabeau, Y. V. de Peer, The hidden
604 duplication past of Arabidopsis thaliana. *PNAS* **99**, 13627–13632 (2002).
605 32. P. Schulte, *et al.*, The Chicxulub asteroid impact and mass extinction at the Cretaceous-
606 Paleogene boundary. *Science* **327**, 1214–1218 (2010).
607 33. J. A. Fawcett, S. Maere, Y. V. de Peer, Plants with double genomes might have had a better
608 chance to survive the Cretaceous–Tertiary extinction event. *PNAS* **106**, 5737–5742 (2009).
609 34. S. Wu, B. Han, Y. Jiao, Genetic contribution of paleopolyploidy to adaptive evolution in
610 angiosperms. *Molecular Plant* **13**, 59–71 (2020).

611 35. M. Piednoël, *et al.*, Next-generation sequencing reveals the impact of repetitive DNA across
612 phylogenetically closely related genomes of Orobanchaceae. *Molecular Biology and*
613 *Evolution* **29**, 3601–3611 (2012).
614 36. T. Yokota, H. Sakai, K. Okuno, K. Yoneyama, Y. Takeuchi, Alectrol and orobanchol,
615 germination stimulants for Orobanche minor, from its host red clover. *Phytochemistry* **49**,
616 1967–1973 (1998).
617 37. M. T. Waters, C. Gutjahr, T. Bennett, D. C. Nelson, Strigolactone signaling and evolution.
618 *Annu. Rev. Plant Biol.* **68**, 291–322 (2017).
619 38. N. D. Young, K. E. Steiner, C. W. dePamphilis, The evolution of parasitism in
620 Scrophulariaceae/Orobanchaceae: Plastid gene sequences refute an evolutionary transition
621 series. *Annals of the Missouri Botanical Garden* **86**, 876 (1999).
622 39. T. H. Torsvik, L. R. M. Cocks, "Paleogene" in *Earth History and Palaeogeography*,
623 (Cambridge University Press, 2016), pp. 240–255.
624 40. O. Eriksson, K. Kainulainen, The evolutionary ecology of dust seeds. *Perspectives in Plant*
625 *Ecology, Evolution and Systematics* **13**, 73–87 (2011).
626 41. Y. Xu, *et al.*, A chromosome‐scale *Gastrodia elata* genome and large‐scale comparative
627 genomic analysis indicate convergent evolution by gene loss in mycoheterotrophic and
628 parasitic plants. *The Plant Journal* **108**, 1609–1623 (2021).
629 42. H. Fujioka, *et al.*, Aberrant protein phosphatase 2C leads to abscisic acid insensitivity and
630 high transpiration in parasitic Striga. *Nat. Plants* **5**, 258–262 (2019).
631 43. S. S. P. Rao, *et al.*, A 3D map of the human genome at kilobase resolution reveals principles
632 of chromatin looping. *Cell* **159**, 1665–1680 (2014).
633 44. B. Liu, *et al.*, Estimation of genomic characteristics by analyzing k-mer frequency in de novo
634 genome projects. *arXiv* (2020) (March 14, 2022).
635 45. S. Koren, *et al.*, Canu: scalable and accurate long-read assembly via adaptive *k* -mer
636 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
637 46. B. J. Walker, *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection
638 and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
639 47. J. Hu, J. Fan, Z. Sun, S. Liu, NextPolish: a fast and efficient genome polishing tool for long-
640 read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
641 48. O. Dudchenko, *et al.*, De novo assembly of the Aedes aegypti genome using Hi-C yields
642 chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
643 49. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment
644 and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
645 50. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
646 2079 (2009).

647 51. P. Danecek, *et al.*, Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008
648 (2021).
649 52. S. Ou, *et al.*, Benchmarking transposable element annotation methods for creation of a
650 streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
651 53. Z. Xu, H. Wang, LTR_FINDER: an efficient tool for the prediction of full-length LTR
652 retrotransposons. *Nucleic Acids Res* **35**, W265-268 (2007).
653 54. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de
654 novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
655 55. S. Ou, N. Jiang, LTR_retriever: a highly accurate and sensitive program for identification of
656 Long Terminal Repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
657 56. J. Shi, C. Liang, Generic Repeat Finder: a high-sensitivity tool for genome-wide de novo
658 repeat detection. *Plant Physiol* **180**, 1803–1815 (2019).
659 57. W. Su, X. Gu, T. Peterson, TIR-Learner, a new ensemble method for TIR transposable
660 element annotation, provides evidence for abundant new transposable elements in the
661 maize genome. *Mol Plant* **12**, 447–460 (2019).
662 58. W. Xiong, L. He, J. Lai, H. K. Dooner, C. Du, HelitronScanner uncovers a large overlooked
663 cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci U S A* **111**,
664 10263–10268 (2014).
665 59. R.-G. Zhang, Z.-X. Wang, S. Ou, G.-Y. Li, TEsorter: lineage-level classification of
666 transposable elements using conserved protein domains. *bioRxiv*, 800177 (2019).
667 60. O. Keller, M. Kollmar, M. Stanke, S. Waack, A novel hybrid gene prediction method
668 employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
669 61. M. S. Campbell, *et al.*, MAKER-P: a tool kit for the rapid creation, management, and quality
670 control of plant genome annotations. *Plant Physiology* **164**, 513–524 (2014).
671 62. D. Kim, *et al.*, TopHat2: accurate alignment of transcriptomes in the presence of insertions,
672 deletions and gene fusions. *Genome Biology* **14**, R36 (2013).
673 63. C. Trapnell, *et al.*, Differential gene and transcript expression analysis of RNA-seq
674 experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
675 64. B. J. Haas, *et al.*, De novo transcript sequence reconstruction from RNA-seq using the
676 Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).
677 65. B. J. Haas, *et al.*, Improving the Arabidopsis genome annotation using maximal transcript
678 alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
679 66. R. She, J. S.-C. Chu, K. Wang, J. Pei, N. Chen, GenBlastA: enabling BLAST to identify
680 homologous gene sequences. *Genome Res* **19**, 143–149 (2009).
681 67. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res* **14**, 988–995
682 (2004).

683 68. B. J. Haas, *et al.*, Automated eukaryotic gene structure annotation using EVidenceModeler
684 and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
685

Figures

Lindenbergia	***Pedicularis***	***Striga***	***Orobanche***
Autotrophic	Facultative Hemiparasitic	Obligate Hemiparasitic	Obligate Holoparasitic

Figure 1. Plant species in the family Orobanchaceae spanning from non-parasitic to increasing levels of parasitism.

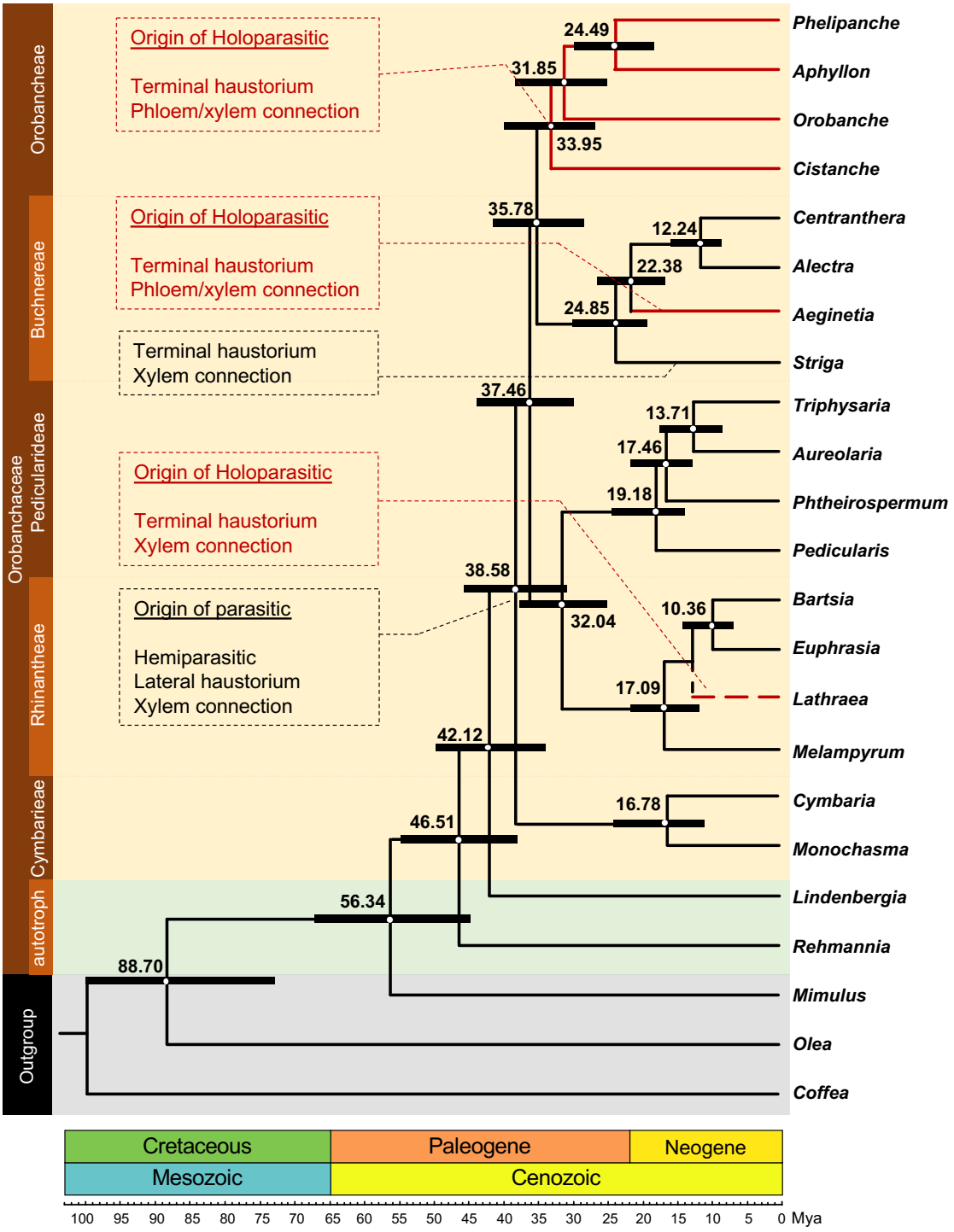

Figure 2. Phylogenetic and molecular clock dating analysis of the major clades in Orobanchaceae.

The phylogenetic tree was constructed using low-copy OGs from 19 orobanchaceous species and three outgroup plants. The gray background indicates autotrophic outgroups; the light grayish-green background represents orobanchaceous autotrophic plant clades; the pale-yellow background represents orobanchaceous parasitic plants; the red branches indicate holoparasites. The red dashed line for *Lathraea* highlights that its position is from the work of Li et al. (5). The numbers near the internodes represent divergence times (Mya) based on molecular clocking dating, and the black bars at the internodes represent 95% HPD for divergence times. All internodes are 100% supported. Important evolutionary events are shown in the dashed boxes. The vertical bar on the left indicates five monophyletic clades (Cymbarieae, Orobancheae, Rhinantheae, Buchnereae, and Pedicularideae) and two autotrophic clades (Rehmannieae and Lindenbergieae, which are shown as “autotroph”). The bar on the bottom indicates the geologic time scale.

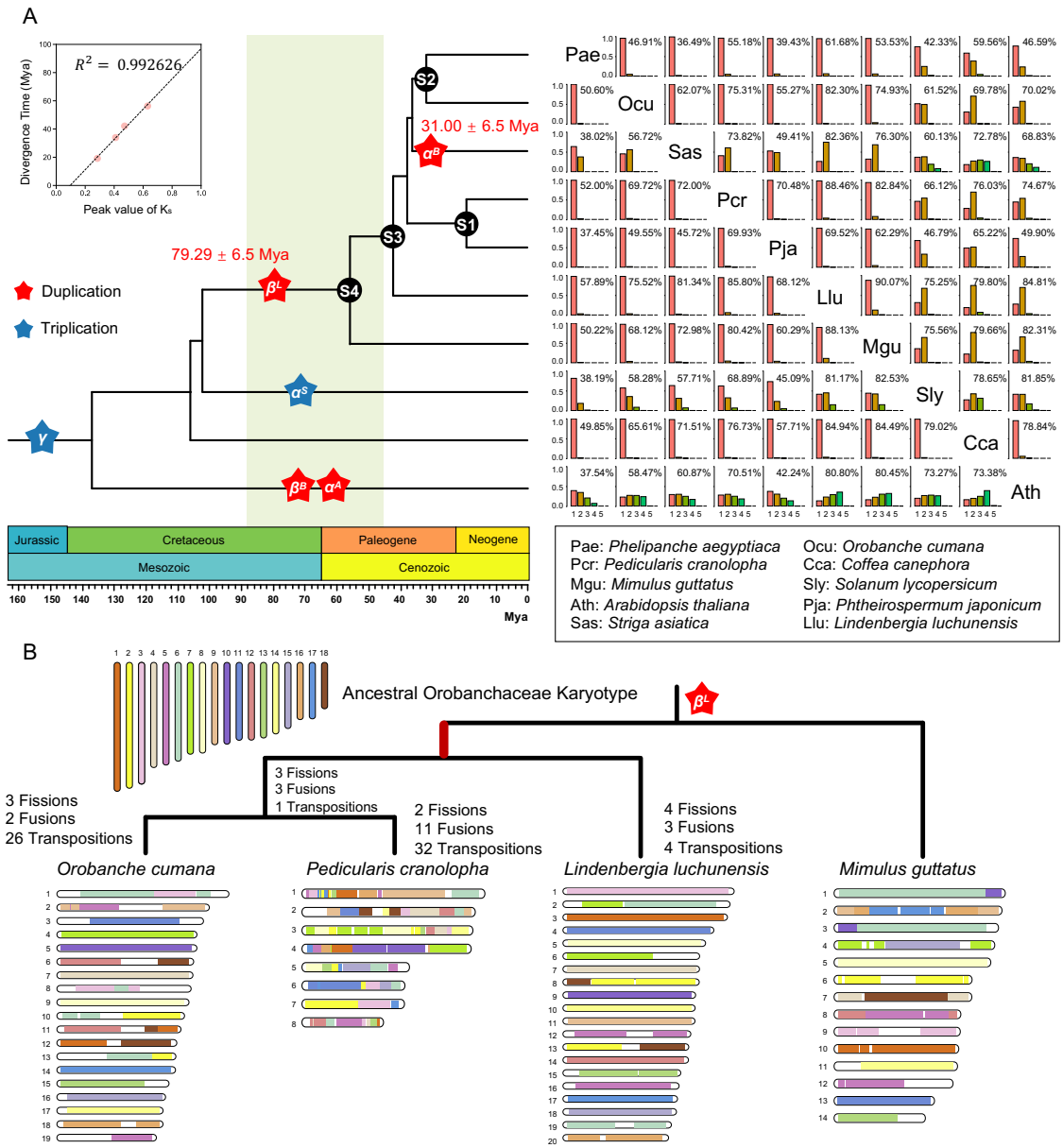

Figure 3. Genome evolution in Orobanchaceae.

(A) The palaeopolyploidization in Orobanchaceae and outgroup lineages. The right histogram matrix indicates the coverage depths of orthologous colinear fragments (OCFs) from pairwise species comparisons. The OCFs from a given species pair were respectively mapped to each species to create two histograms indicating the coverage depths of OCFs, and the max depths of OCFs mapped to species B can be used to infer the WGD events in species A after speciation between species A and B, and vice versa (see methods). In the matrix, each histogram's row and column indicate the two species used for creating the OCFs, and the species of the column is the

one that was the mapping-target species. The percentage in the upper right corner of each histogram represents the proportion of genes covered by OCF in the target species to the total genes. The WGD events inferred from the histogram matrix are marked with stars in the species tree on the left (blue for triplication, red for duplication), and the positions of the stars represent the times of WGD events, which were inferred from the peak value of K_s. The black circles on the species tree represent the four speciation events which were used to calculate the correlation between K_s values and divergence times (see SI Appendix, Supplementary Methodology). The correlation between the peak value of K_s and divergence times is shown in the upper left corner. (B) Eighteen ancestral Orobanchaceae protochromosomes were constructed using *Mimulus guttatus* as an outgroup.

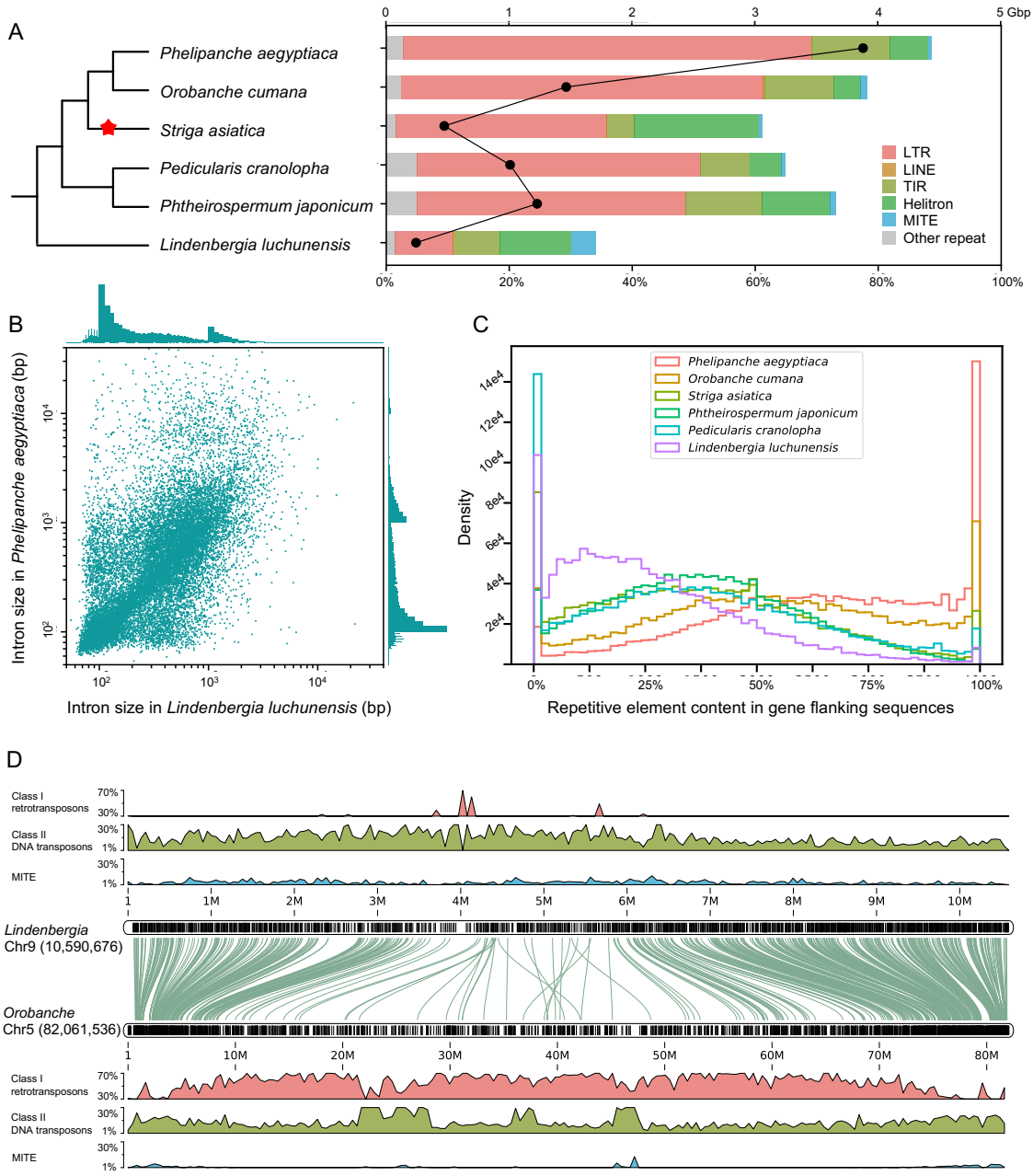

Figure 4. Expansion of repetitive elements in the genomes of Orobanchaceae.

(A) Genome sizes and proportions of repetitive elements in six species of Orobanchaceae. The black dots and X-axis on top represent the genome sizes, and the bars and X-axis on bottom represent the percentages of different types of repetitive elements in the genomes. (B) Size distribution of introns that are conserved in Llu and Pae. (C) Repetitive element content in gene flanking regions in six species of Orobanchaceae. (D) Comparison of chromosomal repetitive elements and gene positions in chromosome 9 of Llu and chromosome 5 of Ocu. Black lines in

the boxes indicate the positions of coding genes on the chromosomes; area graphs indicate the proportion of different types of repeat sequences in the corresponding chromosomal regions; curved lines connect orthologous gene pairs. The length of the bin used for plotting is 1/1000 of the total chromosome length.

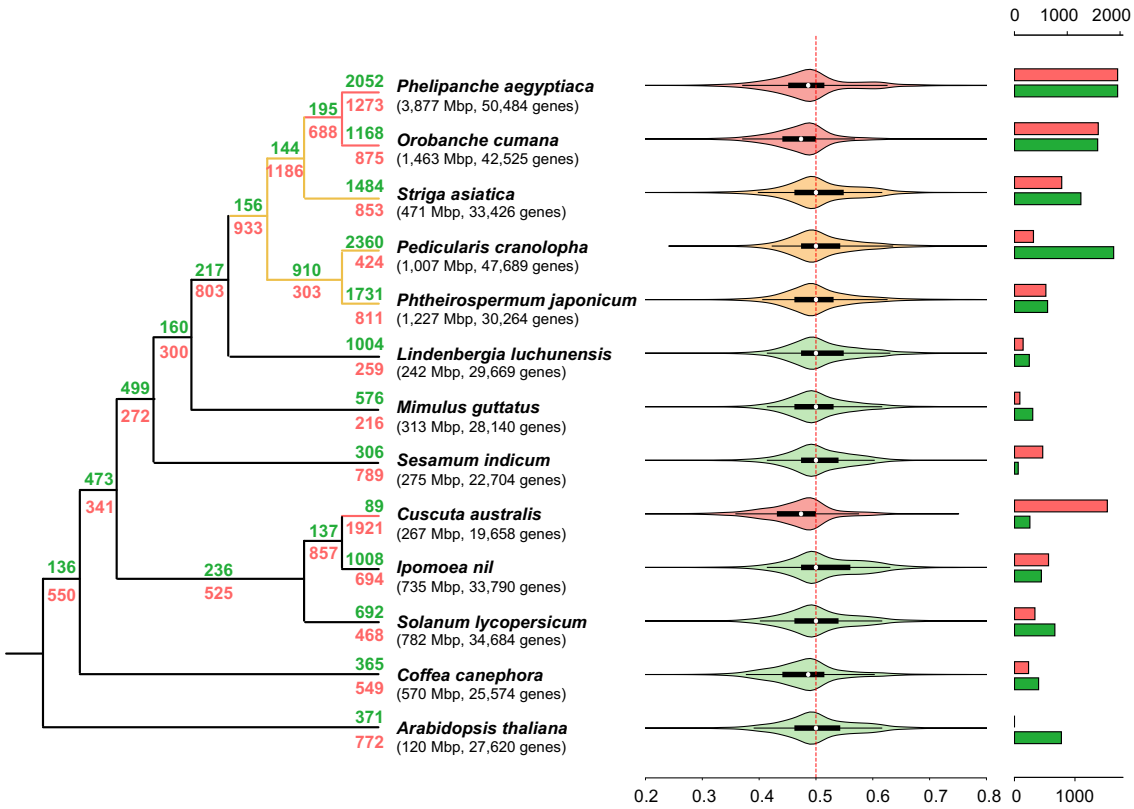

Figure 5. Gene family expansion and contraction in Orobanchaceae.

On each branch of the species tree, the green number indicates the numbers of expanded gene families, and the red number indicates the number of contracted gene families. Each density curve in the violin plot shows the F-index distribution of conserved gene families in each species indicated on the left. Species whose density curves filled with soft red, soft orange, and grayish green are holoparasitic, hemiparasitic, and autotrophic plants, respectively. The bar plot on the right shows the numbers of conserved gene families which are completely lost in each species (red bars with X-axis on bottom) and the numbers of species-specific gene families (green bars with X-axis on top).

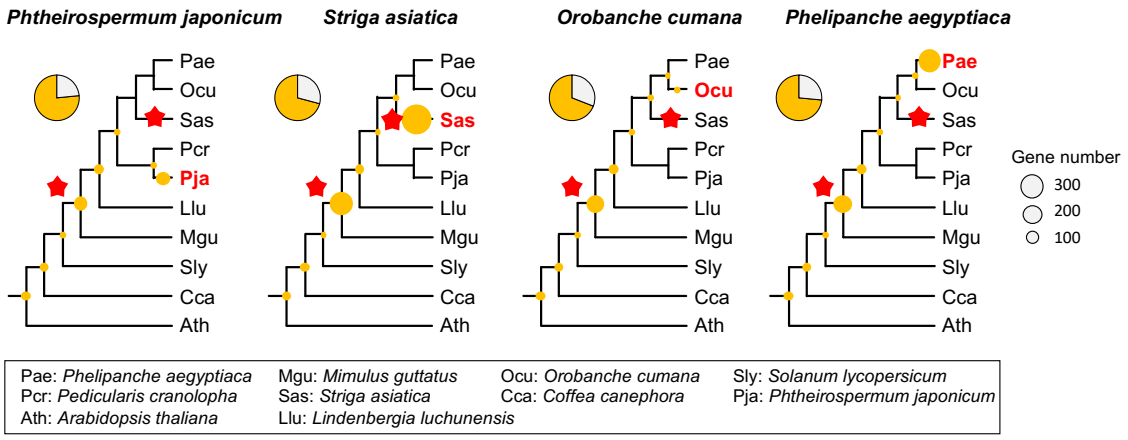

Figure 6. Phylogenetic placement of gene duplications detected in the highly expressed genes in the haustorium.

The highly expressed genes in each species' haustorium were identified. Gene trees were constructed for all the gene families comprising these highly expressed genes. The duplication events were inferred from the topologies of the gene trees. The yellow areas of the pie charts respectively represent the proportions of genes that experienced duplication(s). The size of the yellow circle on each tree node represents the number of gene duplication events occurred during that time.

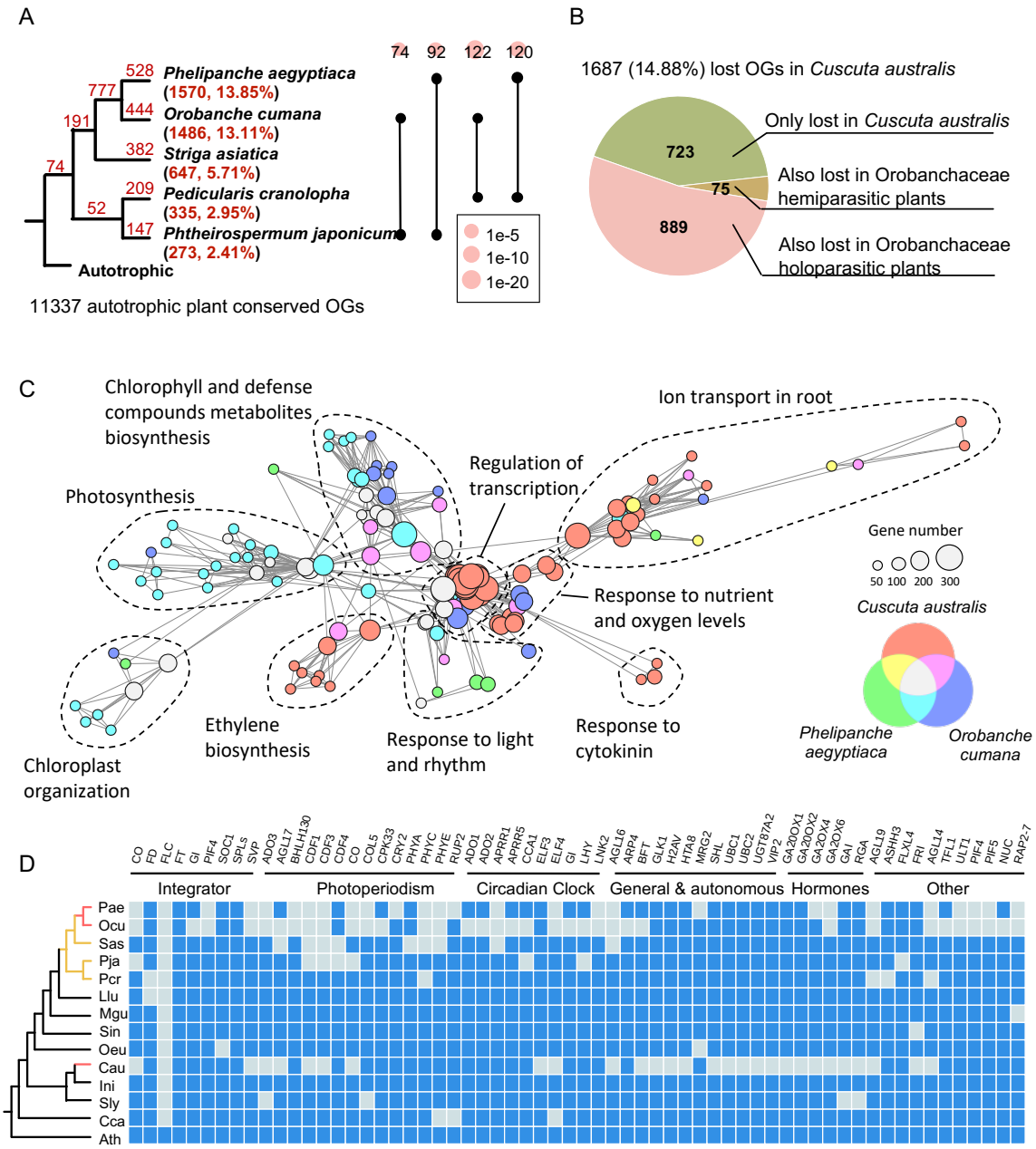

Figure 7. Gene loss in the orobanchaceous parasitic plants and Cau.

(A) History and convergent evolution of gene loss in orobanchaceous parasitic plants. Left: the red numbers on the species tree branches represent the numbers of lost OGs; the number and percentage in brackets under the species name respectively indicates the total number of lost OGs and the proportion of total number of lost OGs to the number of conserved OGs in autotrophic plants. Right: the numbers on top indicate the numbers of the same lost OGs between the two species indicated by the dumbbells after their divergence from the MRCA; the

size of soft red circle on each number indicates the *p*-value of hypergeometric test. (B) The numbers of commonly and specifically lost OGs in Cau and orobanchaceous hemiparasitic and holoparasitic plants. (C) Gene ontology (GO) enrichment analysis of lost OGs in Cau, Pae and Ocu. Each node in the networks plot represents a GO term of a biological process that was enriched (p-value < 0.02), and its size indicates the number of OGs associated with the GO term; the colors of nodes are assigned according to the species in which the gene loss occurred specifically or in common with other species; each edge represents the similarity between two GO terms based on the number of common OGs. (D) Loss of genes related to flowering time regulation in orobanchaceous parasitic plants and Cau. Grey and blue boxes respectively represent absence and presence of the specific gene. Pathway information was retrieved from FIOR-ID (http://www.phytosystems.ulg.ac.be/florid/).

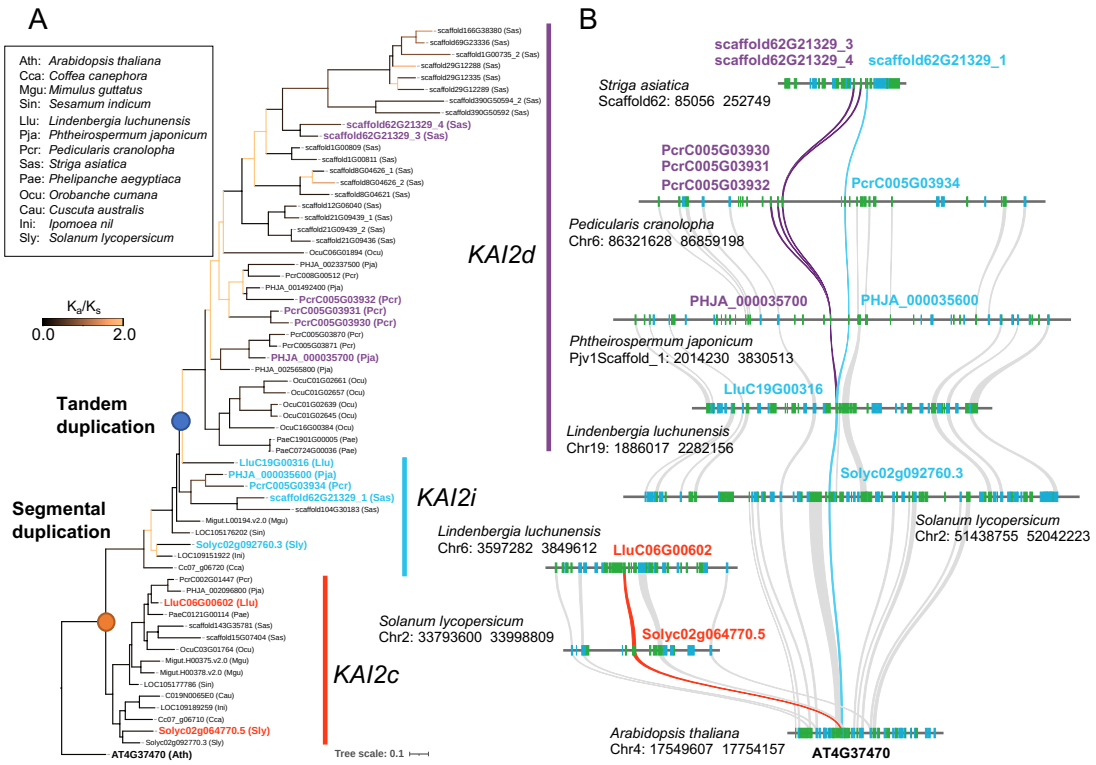

Figure 8. Molecular evolution of *KAI2* gene family.

(A) The phylogenetic tree of *KAI2* family in 13 species. Blue and orange dots indicate duplication events, and the colors of branches represent the values of K_a/K_s. (B) The collinearity relationships of chromosomal fragments containing KAI2 gene in Orobanchaceae parasitic plants and closely related autotrophic species. Syntenic gene pairs are connected by curves. Curves connecting orthologs of *KAI2c*, *KAI2i*, and *KAI2d* genes are highlighted by different colors.