

1 **ORPA: A Fast and Efficient Method for Constructing Genome-Wide**
2 **Alignments of Organelle Genomes for Phylogenetic Analysis**

3

4 **Guiqi Bi^{1,#*}, Xinxin Luan^{1,2#} Jianbin Yan^{1*}**

5 ¹*Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of*
6 *Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at*
7 *Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China*

8 ²*Zhengzhou Research Base, State Key Laboratory of Cotton Biology, School of Agricultural*
9 *Sciences, Zhengzhou University, Zhengzhou, China*

10

11 # These authors contributed equally to this work.

12 * Correspondence to: biguiqi@caas.cn (G.B.); jianbinlab@caas.cn (J.Y.)

13

14 **SUMMARY**

15 Creating a multi-gene alignment matrix for phylogenetic analysis using organelle genomes
16 involves aligning single-gene datasets manually, a process that can be time-consuming and prone
17 to errors. The HomBlocks pipeline has been created to eliminate the inaccuracies arising from
18 manual operations. The processing of a large number of sequences, however, remains a time-
19 consuming task. To conquer this challenge, we have developed a speedy and efficient method
20 called ORPA. ORPA quickly generates multiple sequence alignments for whole-genome
21 comparisons by parsing the result files of NCBI BLAST, completing the task in just one minute.
22 With increasing data volume, ORPA's efficiency is even more pronounced, over 300 times faster
23 than HomBlocks in aligning 60 high-plant chloroplast genomes. The tool's phylogenetic tree
24 outputs demonstrate equivalent results to HomBlocks, indicating its outstanding efficiency. Due
25 to its speed and accuracy, ORPA can identify species-level evolutionary conflicts, providing
26 valuable insights into evolutionary cognition.

27

28 **KEYWORDS**

29 Organelle phylogenomics, Phylogenomic conflict, Efficient pipeline

30

31 INTRODUCTION

32 Phylogenetic trees utilizing organelle genomes are becoming indispensable in
33 comparative genomics and systematics. They play a crucial role in elucidating the evolutionary
34 relationships among species, particularly when incomplete lineage sorting obscures these
35 relationships. This approach provides a broad perspective and facilitates a more accurate
36 assessment of the phylogenetic relationships among species. It is a valuable tool for studying the
37 complex evolutionary history of eukaryotic life (Li et al., 2019; Li et al., 2021).

38 Creating a precise multiple sequence alignment (MSA) is critical to constructing an
39 accurate phylogenetic tree. This typically involves aligning single-copy genes beforehand and
40 then concatenating them to form a super matrix. However, this can be laborious and error-prone,
41 often requiring manual adjustments. To tackle this challenge, we previously introduced
42 HomBlocks software (Bi et al., 2018), which enhances the accuracy and efficiency of MSA
43 construction for organelle genome sequences by eliminating the need for manual operations.
44 However, when dealing with a substantial number of sequences, HomBlocks may still struggle
45 with processing speed, which remains unsatisfactory. To address this issue more effectively, we
46 have developed an innovative approach called ORPA. This method ranks as the fastest software
47 for constructing multi-sequence alignments (MSA) of organelle genomes and delivers
48 exceptionally accurate results compared to HomBlocks. Our research findings offer compelling
49 evidence that ORPA is a highly viable option for HomBlocks, ensuring superior speed and
50 accuracy. Moreover, the ORPA can be employed in systematic investigations to promptly obtain
51 precise evolutionary relationships among species, resulting in significant research discoveries
52 such as species-level evolutionary conflicts.

53

54 **RESULTS**

55 **Comparison of tree topologies constructed from ORPA and HomBlocks sequence** 56 **alignment**

57 We evaluated the topological structures of phylogenetic trees constructed using ORPA
58 and HomBlocks by employing the benchmark data used in the HomBlocks publication (Bi et al.,
59 2018). As a direct comparison of the MSA results was impossible, we used this approach to
60 observe similarities and differences between the two methods. The construction of phylogenetic
61 trees was carried out using two approaches, the maximum likelihood method (RAxML
62 (Stamatakis, 2014)) and Bayesian method (MrBayes 3.2.5 (Ronquist et al., 2012)).

63 In the first test dataset, we evaluated the performance of two software programs using a
64 test dataset consisting of chloroplast genomes from 52 higher plants (Supplementary Table 1), as
65 obtained from Zhang et al. (Zhang et al., 2016). Our results showed that the alignment lengths
66 generated by ORPA and HomBlocks were 90,925 bp and 62,101 bp, respectively, with 8,270 bp
67 and 8,404 bp of parsimony-informative sites. The resulting phylogenetic trees constructed from
68 these two approaches exhibited identical topological structures with high support for all nodes
69 except five (Fig. 2). For our second test dataset, we employed 36 mitochondrial genomes from
70 Xenarthrans (Gibb et al., 2015) (Supplementary Table 2) to construct phylogenetic trees. Our
71 comparative analysis of the resulting tree structures indicates that at non-100% support nodes,
72 the support values from ORPA were slightly lower than those by HomBlocks. However, both
73 methods exhibited a high degree of consistency in the overall topology of the phylogenetic tree
74 (Fig. 3). These findings demonstrate the effectiveness of both approaches in generating reliable
75 phylogenetic trees. Notably, all Multiple Sequence Alignment (MSA) constructions were
76 accomplished within a time frame of less than 5 minutes using ORPA.

77 Moreover, we evaluated a set of higher plant mitochondrial datasets consisting of 18
78 publicly accessible mitochondrial sequences (Supplementary Table 3). Comparative assessments
79 of the resulting phylogenetic trees corroborated the dependability of the ORPA and HomBlocks
80 software suites (Fig. 3a). Additionally, we showcased the distribution of multiple sequence
81 alignment (MSA) derived from both tools aligning to a reference genome, and the findings
82 highlighted a remarkable concordance between the two methods, except for a few dissimilarities
83 at specific loci (Fig. 3b).

84 **Comparing the Runtime of ORPA and HomBlocks**

85 To directly compare the runtime differences of ORPA and HomBlocks on the same
86 dataset, we tested 60 higher plant chloroplast genomes (Supplementary Table 4). First, we
87 constructed genome-wide alignments using ORPA based on these 60 chloroplast sequences and
88 conducted Maximum likelihood tree reconstruction using IQ-TREE (Nguyen et al., 2014). Then,
89 we sampled based on the topology of the tree and compared the runtime of ORPA and
90 HomBlocks. The sampling range increased by 5 with each deepening of the evolutionary
91 relationship (Fig. 5a). We ran ORPA and HomBlocks for each dataset and calculated their
92 respective running times in minutes. Since ORPA typically runs quickly, we standardized the
93 comparison to 1 minute. Additionally, we used the alignment results from each dataset to
94 construct a fast ML tree using IQ-TREE and compared the resulting trees generated by the two
95 tools. To display the differences in running time, we presented a bar chart (Fig. 5b). For the
96 comparison of the resulting trees from each dataset, we expressed the results as a similarity
97 percentage (Fig. 5b).

98 Figure 5b shows a significant increase in HomBlocks' runtime beyond 30 sequences,
99 taking 313 minutes to complete the alignment process with 60 sequences. In contrast, based on

100 the same data source, ORPA can process faster than HomBlocks when the number of sequences
101 exceeds 10. Additionally, the runtime of HomBlocks increases exponentially with the number of
102 sequences, which could become more significant with datasets containing over 60 sequences.
103 Conversely, the script runtime for ORPA remains unaffected by the number of sequences, except
104 during the data preparation stage. This is mainly attributed to the utilization of BLAST as the
105 kernel for the alignment process, which avoids the need for single-threaded sequence comparison.

106 This highlights ORPA's advantage over HomBlocks when dealing with a large number of
107 sequences. Additionally, we conducted tree reconstructions for each pairwise test dataset and
108 compared the ML tree topologies generated by ORPA and HomBlocks using treedist
109 (<https://github.com/agormp/treedist>). Except for the comparison group with a sampling range of
110 25, the similarity between the tree topologies is 91%, indicating almost identical phylogenetic
111 tree topologies generated by ORPA and HomBlocks in the other 11 comparison groups. Our
112 findings suggest that ORPA outperforms HomBlocks in terms of speed and accuracy, offering
113 researchers a powerful tool for creating whole-genome alignments.

114 **Using ORPA for rapid detection of systematic evolutionary conflicts**

115 Advancements in sequencing technology have led to the accumulation of a vast amount
116 of organellar genome data. Effective utilization of this data has become a growing field of
117 interest. This is especially important for newly sequenced data, as rapid confirmation of species'
118 evolutionary relationships is crucial to verify sequencing accuracy. Additionally, constructing
119 phylogenetic trees with speed and accuracy to investigate evolutionary conflicts is a key area of
120 research in systematics biology. ORPA offers an elegant approach to achieving these goals. To
121 provide a more comprehensive demonstration, we utilized 52 Lamiales chloroplast datasets as an
122 example (Supplementary Table 5). We employed ORPA to build a multiple sequence alignment

123 of 101,454 characters, and constructed a corresponding phylogenetic tree to depict the
124 evolutionary relationships. Figure 5 demonstrates that there are two apparent conflicts between
125 the phylogenetic branches in regards to *Wightia speciosissima* and *Comoranthus minor*.

126 *Wightia speciosissima*, an angiosperm, has been assigned to a distinct family
127 (Wightiaceae) by the Angiosperm Phylogeny Group IV (The Angiosperm Phylogeny, 2016). Its
128 previous classification placed it within the Paulowniaceae family. However, analysis of its
129 evolutionary branching suggests that it shares closer evolutionary relationships with the
130 Phrymaceae family, thus representing a distinct lineage. This observation was also made by Xia
131 et al (Xia et al., 2019). Their study on plant phylogeny utilized data from nine chloroplast genes
132 and one mitochondrial *rps3* gene. Consequently, they advise against including *Wightia*
133 *speciosissima* in the Paulowniaceae family and suggest that it may instead be a hybrid origin
134 between early lineages of Phrymaceae and Paulowniaceae (Xia et al., 2019).

135 This phylogenetic tree also reveals the incongruity between the genus *Comoranthus* and
136 *Schrebera* in terms of their phylogenetic relationships. *Schrebera* is found in Africa and India,
137 while *Comoranthus* is only found in Madagascar and the Comoros Islands (Wallander et al.,
138 2000). Both species have similar fruit morphology: capsules with a woody ovary, ruffled
139 epidermis, and split in half when mature. They contain seeds and are suborbicular and pear-
140 shaped (Engel, 1968). Our analysis of the evolutionary relationships among these studied
141 species reveals a paraphyletic relationship, with *Comoranthus* found nested within *Schrebera*.
142 This outcome is consistent with recent findings by Hong-Wa et al., who suggest the genera
143 should be synonymized (Hong-Wa et al., 2023). Incorporating this finding into taxonomic
144 classification will aid in a more accurate understanding of the evolutionary history of these plant
145 groups.

146 In summary, ORPA has demonstrated considerable promise in the realm of systematic
147 taxonomy, as demonstrated by the aforementioned use cases.

148

149 **EXPERIMENTAL PROCEDURES**

150 **Methods**

151 The framework of ORPA is written in Perl. As tools for aligning genomes, HomBlocks uses
152 a method of identifying locally collinear blocks (LCBs), while the main difference with ORPA is
153 its strategy of directly parsing the NCBI BLAST online tool results. By avoiding the need for
154 software installation and various dependencies, this approach simplifies genome alignment for
155 novices in the field of bioinformatics (Fig. 1). The core of ORPA is based on the widely-used
156 BLAST tool (McGinnis et al., 2004), which offers significant improvements in the efficiency and
157 speed of sequence alignments. Compared to HomBlocks, ORPA is able to construct alignment
158 files within 5 minutes on average. In contrast, HomBlocks requires an increasing amount of
159 processing time as the number of sequences being aligned grows due to the single-threaded
160 operation of its core software, Mavue (Darling et al., 2004). Therefore, ORPA offers a more
161 efficient and versatile alternative to HomBlocks.

162 ORPA also provides users with four trimming methods, namely Gblocks (Castresana, 2000),
163 trimAl (Capella-Gutiérrez et al., 2009), Noisy (Dress et al., 2008), and BMGE (Crisuolo et al.,
164 2010), which are same to those offered by HomBlocks. Importantly, users can directly use the
165 output results from ORPA to facilitate the construction of a phylogenetic tree, thus streamlining
166 the sequence alignment process.

167 **Implementation**

168 ORPA is a rapid tool for constructing multiple sequence alignments of organelles. It is a
169 command-line tool and functional under any version of Linux without the need for external
170 installation. The Perl source code of ORPA is freely available for download at
171 <https://github.com/BGQ/ORPA.git>, and comprehensive documentation and tutorials can be
172 found at <https://github.com/BGQ/ORPA.git>.

173

174 **ACKNOWLEDGMENTS**

175 We sincerely thank the editors and reviewers for their valuable suggestions and comments on
176 this study. This work was supported by the National Key R&D Program of China
177 (2018YFA0903200), Science Technology and Innovation Commission of Shenzhen
178 Municipality of China (ZDSYS 20200811142605017). It was also supported by the Elite Young
179 Scientists Program of CAAS and the Agricultural Science and Technology Innovation Program.

180

181 **AUTHOR CONTRIBUTIONS**

182 GQB and JBY conceived and designed the study. GQB and XXL collected data. GQB and XXL
183 performed data analysis. GQB and JBY wrote the manuscript with other authors providing
184 recommendations for modifications.

185

186 **CONFLICT OF INTEREST**

187 The authors declare that they have no conflicts of interest associated with this work.

188

189 **DATA AVAILABILITY STATEMENT**

190 The attached file contains a list of example data and organelle genome sequences that are
191 referred to in the main body of this study.

192

193 **SUPPORTING INFORMATION**

194 Additional Supporting Information may be found in the online version of this article.

195 **Table S1.** Accession numbers of 52 higher plant chloroplast genomes from GenBank used in the
196 first example dataset.

197 **Table S2.** Accession numbers of 36 xenarthrans mitochondrial genomes from GenBank used in
198 the second example dataset.

199 **Table S3.** Accession numbers of 18 higher plant mitochondrial genomes from GenBank used in
200 the third example dataset.

201 **Table S4.** Accession numbers of 60 higher plant chloroplast genomes from GenBank used in the
202 fourth example dataset.

203 **Table S5.** Accession numbers of Lamiales chloroplast genomes from GenBank used in the fifth
204 example dataset.

205

206 **REFERENCES**

- 207 **Bi, G., Mao, Y., Xing, Q. & Cao, M.** (2018) HomBlocks: a multiple-alignment construction pipeline for
208 organelle phylogenomics based on locally collinear block searching. *Genomics*, **110**, 18-22.
- 209 **Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T.** (2009) trimAl: a tool for automated
210 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.
- 211 **Castresana, J.** (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in
212 Phylogenetic Analysis. *Molecular Biology and Evolution*, **17**, 540-552.
- 213 **Criscuolo, A. & Gribaldo, S.** (2010) BMGE (Block Mapping and Gathering with Entropy): a new
214 software for selection of phylogenetic informative regions from multiple sequence alignments.
215 *BMC Evolutionary Biology*, **10**, 210.
- 216 **Darling, A.C.E., Mau, B., Blattner, F.R. & Perna, N.T.** (2004) Mauve: Multiple Alignment of
217 Conserved Genomic Sequence With Rearrangements. *Genome Research*, **14**, 1394-1403.
- 218 **Dress, A.W., Flamm, C., Fritsch, G., Grunewald, S., Kruspe, M., Prohaska, S.J. et al.** (2008) Noisy:
219 identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular*
220 *Biology*, **3**, 7.
- 221 **Engel, P.P.** (1968) The induction of biochemical and morphological mutants in the moss *Physcomitrella*
222 *patens*. *American Journal of Botany*, **55**, 438-446.
- 223 **Gibb, G.C., Condamine, F.L., Kuch, M., Enk, J., Moraes-Barros, N., Superina, M. et al.** (2015)
224 Shotgun Mitogenomics Provides a Reference Phylogenetic Framework and Timescale for Living
225 Xenarthrans. *Molecular Biology and Evolution*, **33**, 621-642.
- 226 **Hong-Wa, C., Dupin, J., Frasier, C., Schatz, G. & Besnard, G.** (2023) Systematics and biogeography
227 of Oleaceae subtribe Schreberinae, with recircumscription and revision of its Malagasy members.
228 *Botanical Journal of the Linnean Society*.
- 229 **Li, H.T., Luo, Y., Gan, L., Ma, P.F., Gao, L.M., Yang, J.B. et al.** (2021) Plastid phylogenomic insights
230 into relationships of all flowering plant families. *BMC Biology*, **19**, 1-13.
- 231 **Li, H.T., Yi, T.S., Gao, L.M., Ma, P.F., Zhang, T., Yang, J.B. et al.** (2019) Origin of angiosperms and
232 the puzzle of the Jurassic gap. *Nature Plants*, **5**, 461-470.
- 233 **Mcginnis, S. & Madden, T.L.** (2004) BLAST: at the core of a powerful and diverse set of sequence
234 analysis tools. *Nucleic Acids Research*, **32**, W20-W25.
- 235 **Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q.** (2014) IQ-TREE: A Fast and Effective
236 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and*
237 *Evolution*, **32**, 268-274.
- 238 **Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S. et al.** (2012)
239 MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large
240 Model Space. *Systematic Biology*, **61**, 539-542.
- 241 **Stamatakis, A.** (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
242 phylogenies. *Bioinformatics*, **30**, 1312-1313.
- 243 **The Angiosperm Phylogeny, G.** (2016) An update of the Angiosperm Phylogeny Group classification
244 for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*,
245 **181**, 1-20.
- 246 **Wallander, E. & Albert, V.A.** (2000) Phylogeny and classification of Oleaceae based on rps16 and trnL-
247 F sequence data. *American Journal of Botany*, **87**, 1827-1841.
- 248 **Xia, Z., Wen, J. & Gao, Z.** (2019) Does the Enigmatic *Wightia* Belong to Paulowniaceae (Lamiales)?
249 *Frontiers in Plant Science*, **10**, 528.
- 250 **Zhang, D., Li, K., Gao, J., Liu, Y. & Gao, L.-Z.** (2016) The Complete Plastid Genome Sequence of the
251 Wild Rice *Zizania latifolia* and Comparative Chloroplast Genomics of the Rice Tribe Oryzaceae,
252 Poaceae. *Frontiers in Ecology and Evolution*, **4**, 88.

254 **FIGURE LEGENDS**

255 **Fig. 1 Comparison of ORPA and HomBlocks efficiency in the conventional workflow for**
256 **the phylogenetic tree construction of organelle genomes.**

257
258 **Fig. 2 Comparison of topology between the HomBlocks tree (left) and the ORPA tree (right)**
259 **of 52 higher plant chloroplast genomes.** The phylogenetic trees were constructed using
260 maximum likelihood (ML) and Bayesian inference (BI) methods with the HomBlocks alignment
261 (62,101 characters) and the ORPA alignment (90,925 characters), respectively. The support
262 values inferred from RAxML (left) and Bayesian posterior probability (right) are indicated by
263 the numbers on the nodes. Fully resolved nodes are not labeled with numbers. These results
264 provide insights into the comparative performance of the two alignment methods for
265 phylogenetic analysis of chloroplast genomes in higher plants.

266
267 **Fig. 3 Topology comparison of two phylogenetic trees of 36 xenarthran mitochondrial**
268 **genomes.** The HomBlocks tree (left) and the ORPA tree (right) were constructed using different
269 alignment methods, one with 15,170 characters and the other with 8,696 characters. Maximum
270 likelihood and Bayesian inference methods were used to construct the trees, and the support
271 values derived from RAxML (left) and Bayesian posterior probability (right) are indicated by the
272 numerical values on the nodes. Fully resolved nodes are unlabeled.

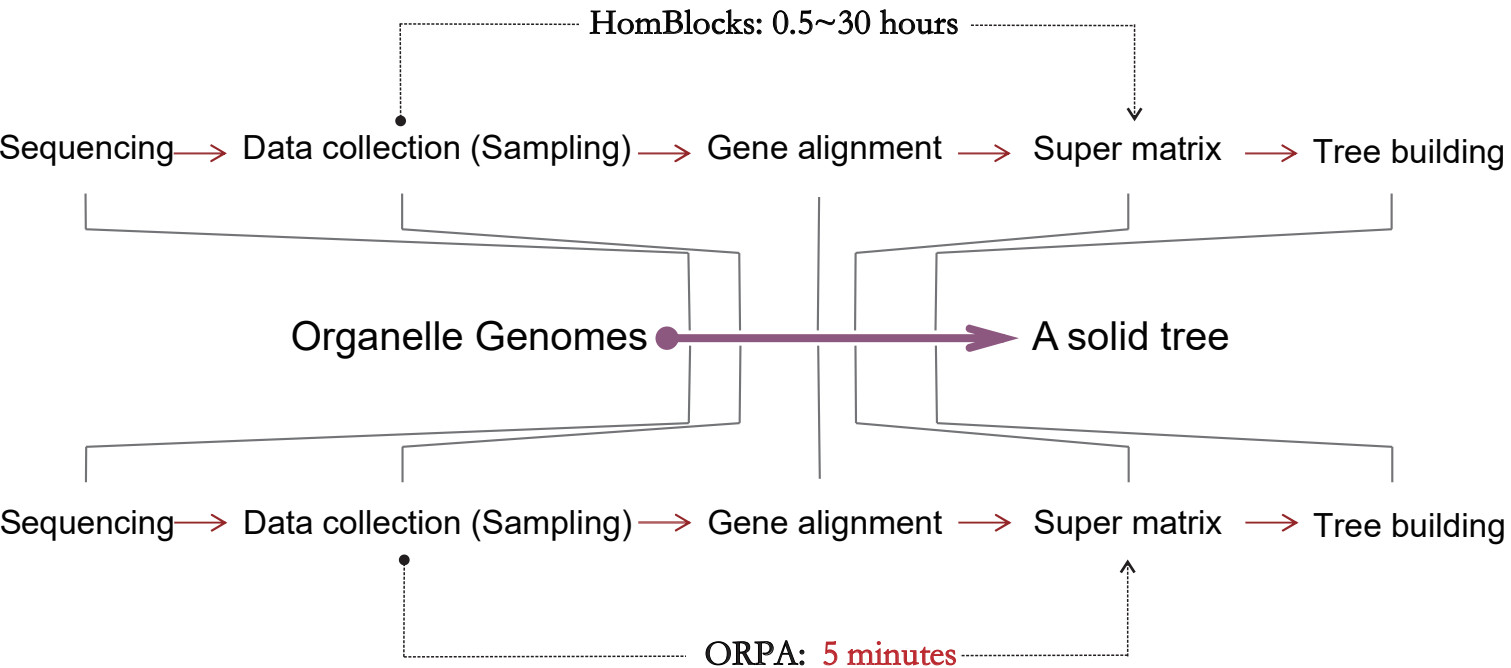
273
274 **Fig. 4 Comparison of phylogenetic trees and alignment methods for 18 higher plant**
275 **mitochondrial genomes. a,** Two phylogenetic trees of 18 higher plant mitochondrial genomes
276 were constructed using ORPA and HomBlocks alignment methods, respectively. The trees were

277 constructed with maximum likelihood and Bayesian inference methods, and the support values
278 derived from RAxML and Bayesian posterior probability are indicated on each node. Fully
279 resolved nodes are unlabeled. **b**, Distributional differences of phylogenetic alignments obtained
280 from ORPA and HomBlocks methods using *Ajuga reptans* as the reference sequence. The circos
281 plot illustrates the differing sequence composition sites between the two methods, with green and
282 gray dots indicating the variation between the alignments.

283
284 **Fig. 5 Comparison of ORPA and HomBlocks runtime efficiency.** **a**, Comparison of runtime
285 for 60 higher plant chloroplast datasets. A maximum likelihood tree shows the evolutionary
286 relationship among 60 samples. Nodes with 100% support are unspecified, and other partially
287 supported nodes are labeled with bootstrap and aLTR values. Sampling begins at the base of the
288 tree and proceeds with increasing sample sizes of 5 until all data are used, resulting in a total of
289 12 comparison groups. **b**, Comparison of ORPA and HomBlocks runtime. The sample size
290 corresponds to the sampling range in Figure 5a. The percentage on the bar chart represents the
291 similarity in systemic tree topology generated by the two software programs.

292
293 **Fig. 6 Identification of species-level evolutionary conflicts using ORPA.** A total of 52
294 Lamiales chloroplast trees were constructed using 101,544 characters from the ORPA alignment.
295 Maximum likelihood and Bayesian inference methods were used to construct the trees, and the
296 support values derived from RAxML (left) and Bayesian posterior probability (right) are
297 indicated by numerical values on the nodes. Fully resolved nodes are indicated by red dots.
298 *Wightia speciosissima*, which has a controversial position in Lamiales, is labelled in red. The
299 morphology of four species from *Schrebera* and *Comoranthus* genera is shown on the right side

300 of the figure. Additionally, the results reveal a paraphyletic relationship, with *Comoranthus*
301 *minor* nested within *Schrebera*, leading to the synonymization of these genera.

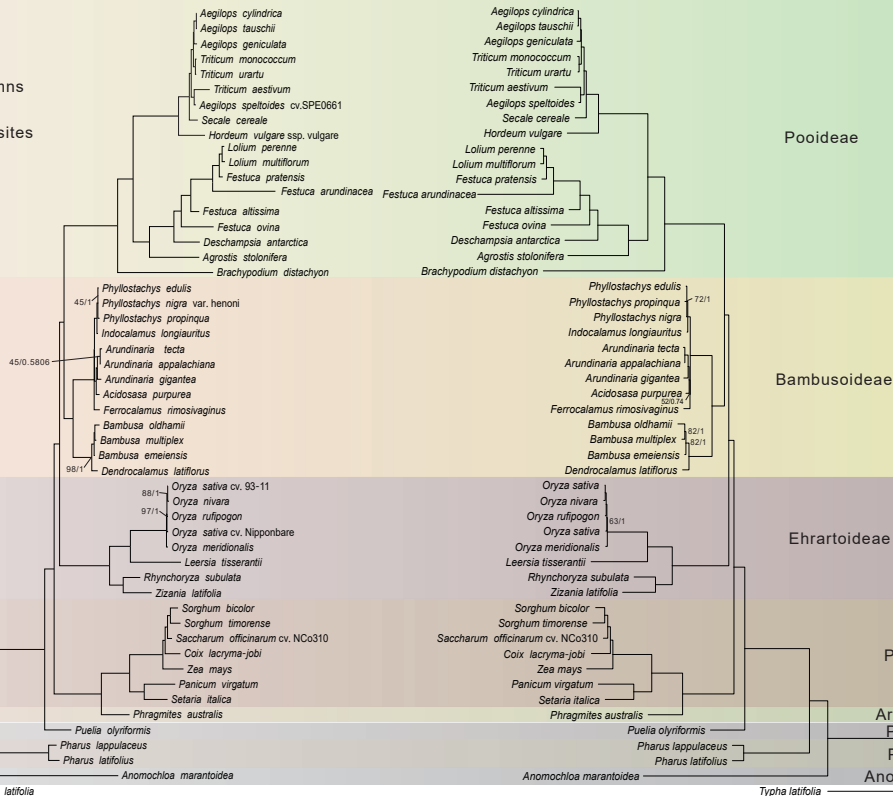


HomBlocks

Alignment with 62,101 columns
 5,824 distinct patterns
 8,404 parsimony-informative sites
 7,378 singleton sites
 46,319 constant sites

ORPA

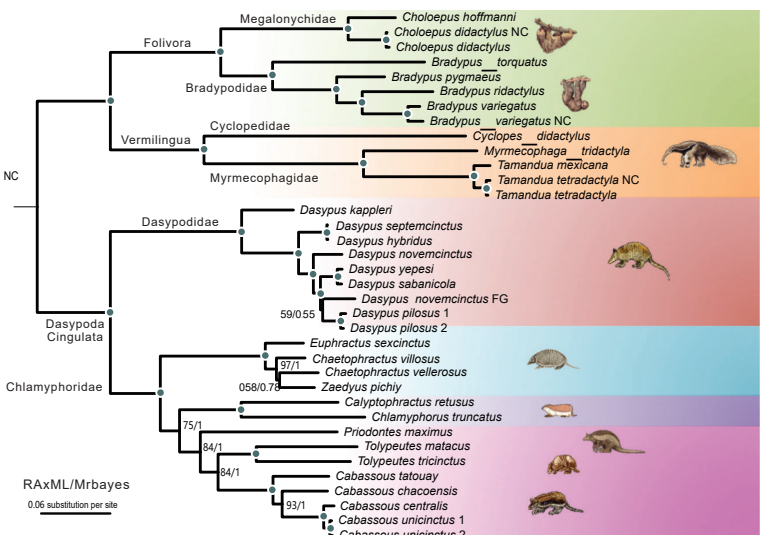
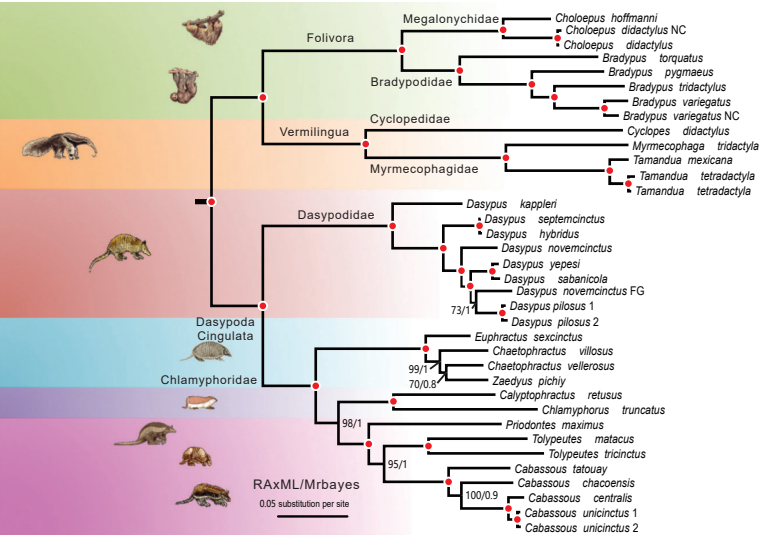
Alignment with 90,925 columns
 5,459 distinct patterns
 8,270 parsimony-informative sites
 8,679 singleton sites
 73,976 constant sites



Node supports
 RAXML/Mrbayes

HomBlocks

ORPA



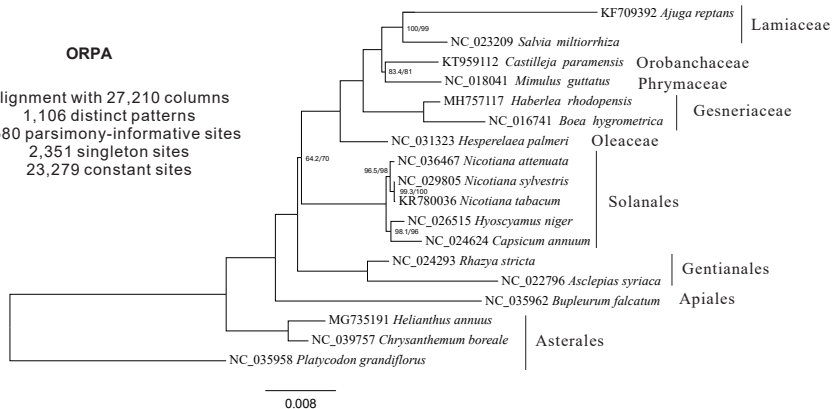
Alignment has 15,170 columns, 6,201 distinct patterns
6,722 parsimony-informative sites, 919 singleton sites, 7,529 constant sites

Alignment 8,696 columns, 2,891 distinct patterns
3,015 parsimony-informative sites, 494 singleton sites, 5,187 constant sites

a

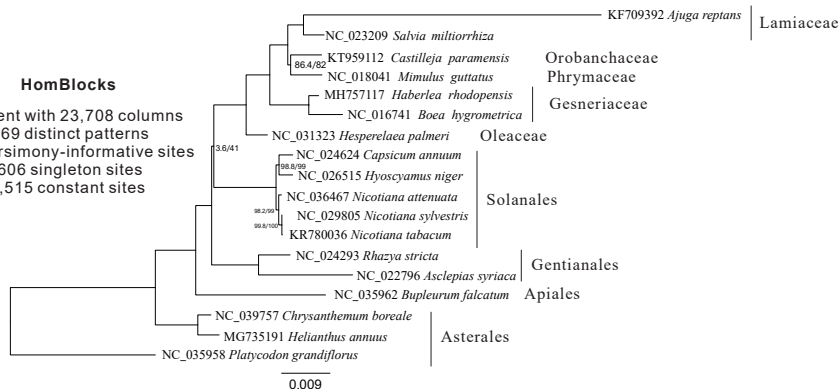
ORPA

Alignment with 27,210 columns
1,106 distinct patterns
1,580 parsimony-informative sites
2,351 singleton sites
23,279 constant sites

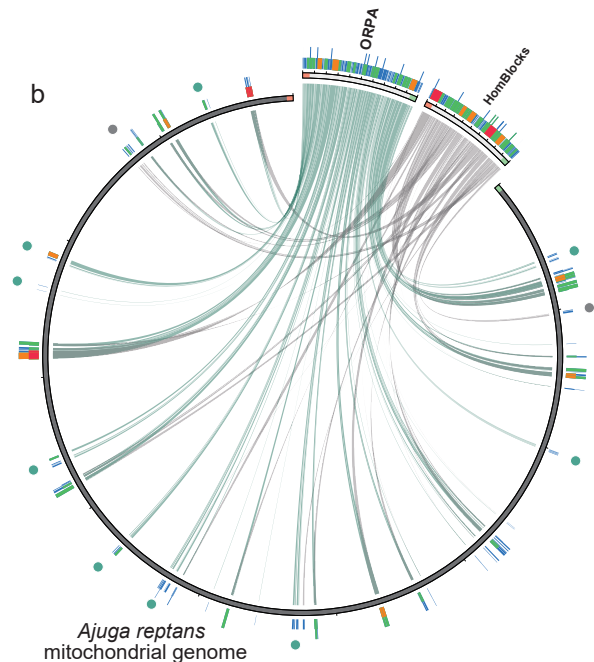


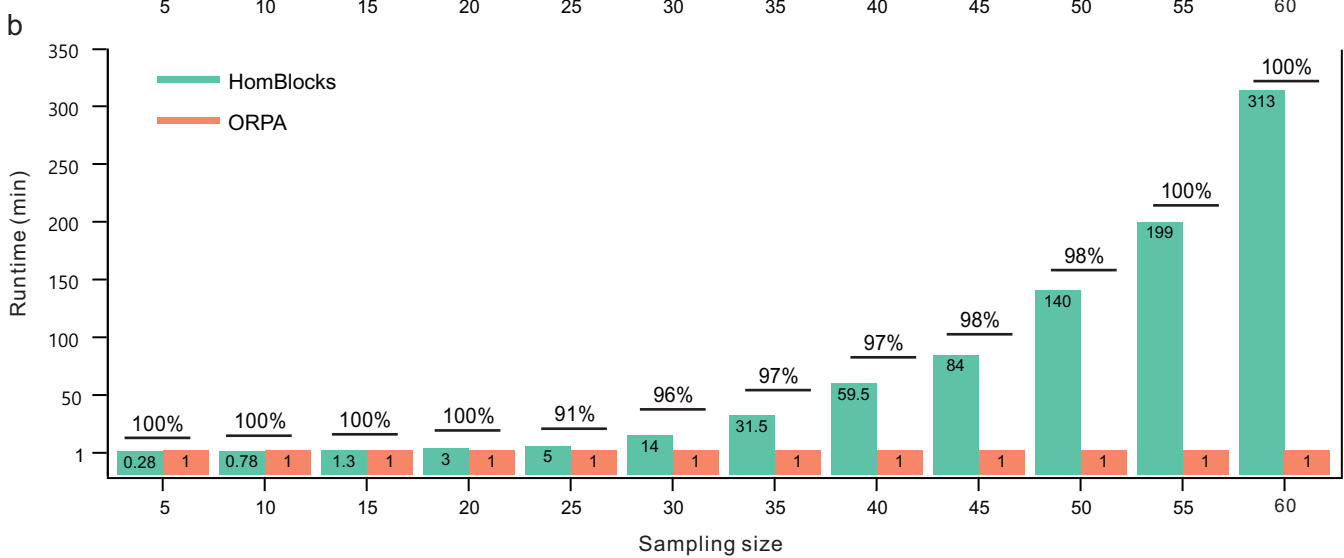
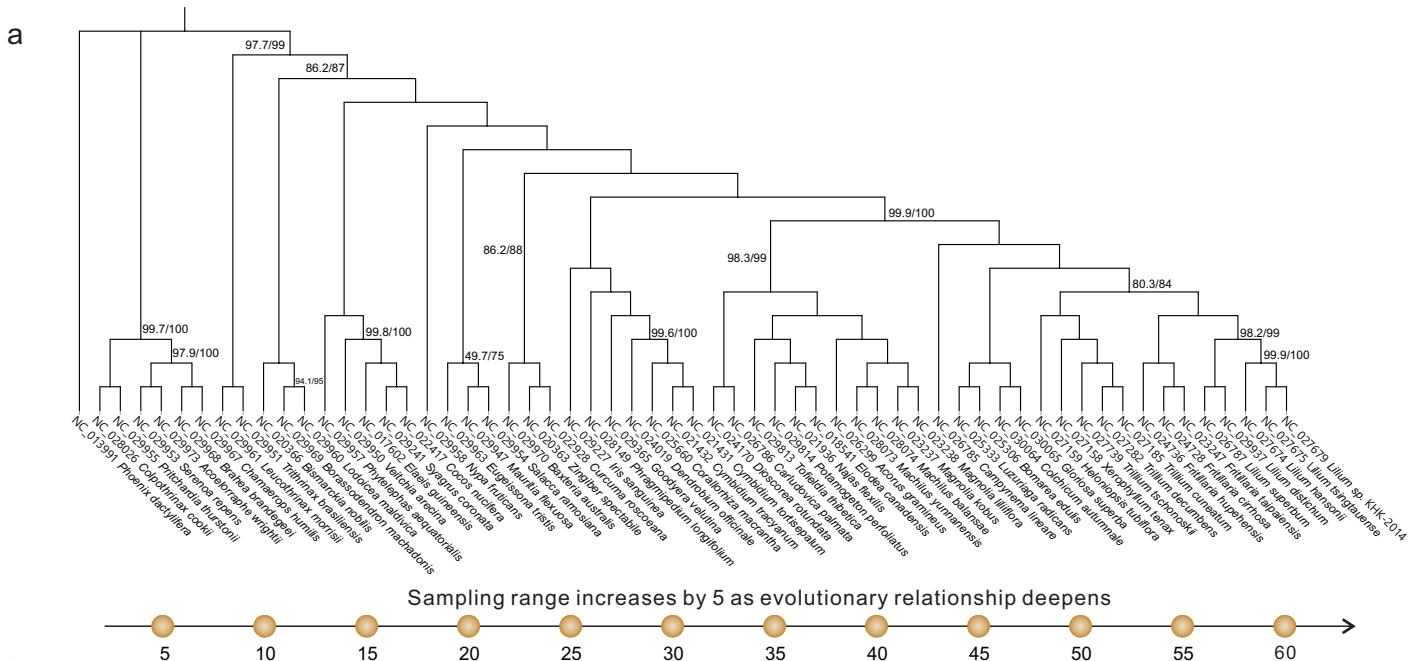
HomBlocks

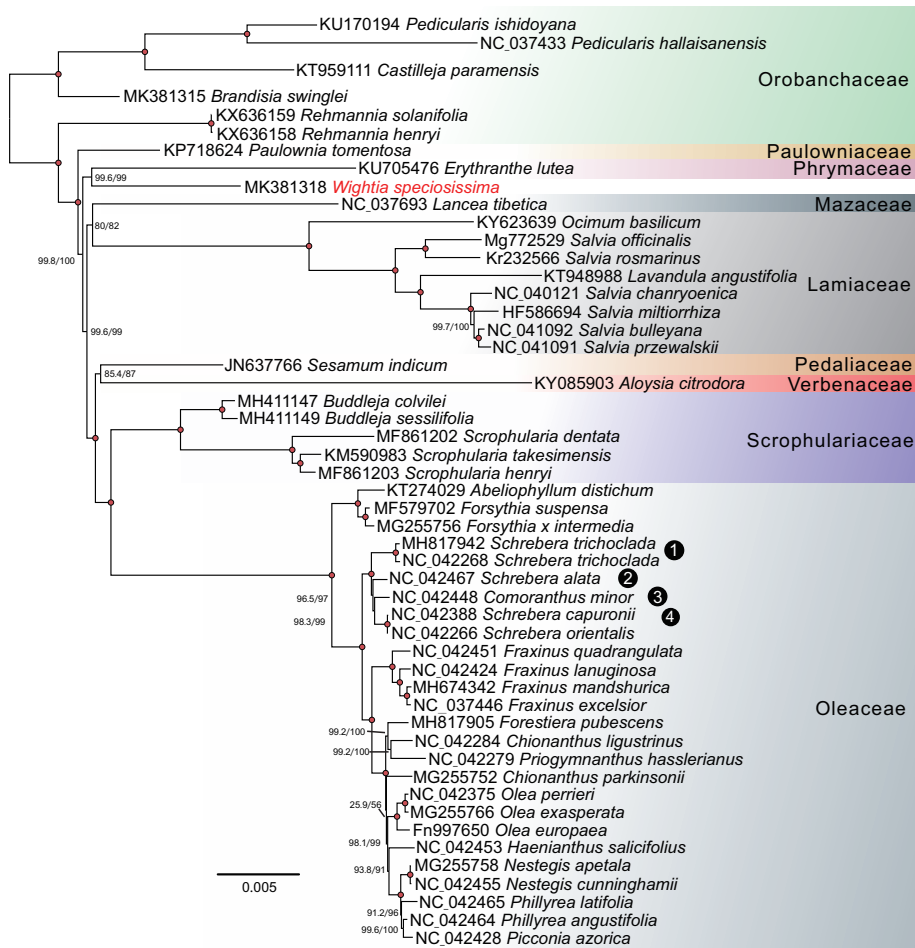
Alignment with 23,708 columns
1,169 distinct patterns
1,587 parsimony-informative sites
2,606 singleton sites
19,515 constant sites



b







Schrebera trichoclada



Schrebera alata



Comoranthus minor



Schrebera capuronii