

14 **Abstract**

15 Few viruses have been documented in plants outside angiosperms. Endogenous viral
16 elements (paleoviruses) provide ‘molecular fossils’ for studying the deep history and
17 macroevolution of viruses. Endogenous plant pararetroviruses (EPRVs) are
18 widespread across angiosperms, but little is known about EPRVs in earlier branching
19 plants. Here we use a large-scale phylogenomic approach to investigate the diversity
20 and macroevolution of plant pararetroviruses (formally known as *Caulimoviridae*).
21 We uncover an unprecedented and unappreciated diversity of EPRVs in the genomes
22 of gymnosperms and ferns. The known angiosperm viruses only constitute a minor
23 part of the *Caulimoviridae* diversity. By characterizing the distribution of EPRVs, we
24 show that no major euphyllophyte lineages escape the activity of *Caulimoviridae*,
25 raising the possibility that many exogenous *Caulimoviridae* remain to be discovered
26 in euphyllophytes. We find that the copy numbers of EPRVs are generally high,
27 suggesting that EPRVs define a unique group of repetitive elements and represent
28 major components of euphyllophyte genomes. Phylogenetic analyses reveal an
29 ancient monilophyte origin of *Caulimoviridae* and at least three independent origins
30 of *Caulimoviridae* in angiosperms by cross-division transmissions. Our findings
31 uncover the remarkable diversity of *Caulimoviridae* and have important implications
32 in understanding the origin and macroevolution of plant pararetroviruses.

33 Endogenous viral elements (EVEs), the viral sequences integrated into their hosts'
34 genomes, document past virus (paleovirus) infections and provide 'molecular fossils'
35 for studying the deep history of viruses¹. EVEs lay the foundation of an emerging
36 field, Paleovirology^{1,2}. The best characterized EVEs are endogenous retroviruses
37 (ERVs)³. The replication of retroviruses requires integration into their hosts' genomes.
38 On occasion, retroviruses infect germ lines of their hosts, and the integrated
39 retroviruses, namely ERVs, become vertically inherited. ERVs are widespread and
40 highly abundant in the genomes of vertebrates³; for example, ERVs make up 5%-8%
41 of the human genome⁴. Recently, endogenous non-retroviral elements have been
42 increasingly identified by comparative genomic analyses, which reveal the remarkable
43 diversity, deep history, and macroevolution of related viruses^{5,6}. EVEs (especially
44 ERVs) were pervasively co-opted for the hosts' biology, ranging from placentation, to
45 inhibition of exogenous viral infection, to regulation of innate immunity⁷⁻⁹.

46

47 Like retroviruses, two families of viruses with double-stranded DNA genomes
48 replicate through RNA intermediates, known as pararetroviruses or DNA
49 reverse-transcribing viruses¹⁰. Unlike retroviruses, these pararetroviruses lack
50 integrase, and thus integration into host genomes is not essential for their replication.
51 Pararetroviruses infect vertebrates (*Hepadnaviridae*) and plants (*Caulimoviridae*).
52 Evolutionary analyses suggest that *Hepadnaviridae* and *Caulimoviridae* originated
53 independently from retrotransposons with long terminal repeats (LTRs)¹¹.
54 Endogenous hepadnaviruses have been increasingly identified in the genomes of
55 many species of Aves and reptiles¹²⁻¹⁴. The copy number of endogenous
56 hepadnaviruses in the host genomes is very low (usually around ten copies)¹²⁻¹⁴. The
57 identification of endogenous hepadnaviruses reveals the prevalent nature and deep
58 history (more than 207 million years) of *Hepadnaviridae* in vertebrates¹²⁻¹⁵.

59

60 The *Caulimoviridae* family, pararetroviruses infecting plants, is classified into
61 eight genera, namely *Caulimovirus*, *Soymovirus*, *Tungrovirus*, *Badnavirus*,
62 *Solendovirus*, *Cavemovirus*, *Rosadnavirus*, and *Petuvirus*¹⁶. The genome size of

63 *Caulimoviridae* is usually between 6,000 and 8,000 base pairs (bps), encoding one to
64 eight open reading frames (ORFs). The proteins (or domains) common to
65 *Caulimoviridae* include movement protein (MP), coat protein (CP), aspartic protease
66 (AP or PR), reverse transcriptase (RT), and RNase H1 (RH)¹⁶. While the replication of
67 *Caulimoviridae* does not require integration into host genomes, endogenous plant
68 pararetroviruses (EPRVs) were identified in many angiosperms at pre-genomic era¹⁷,
69 for example banana¹⁸ and tobacco¹⁹. Genome-scale data provide important resource to
70 explore the distribution and diversity of EPRVs within plant genomes, which would
71 improve our understanding of the macroevolution of *Caulimoviridae* and the
72 relationship between viruses and their hosts. By mining a variety of plant genomes,
73 Geering et al.²⁰ identified a novel lineage of EPRVs in flowering plants (angiosperms),
74 which was designated ‘Florendovirus’ and was thought to constitute a new genus
75 within *Caulimoviridae*. However, EPRVs have not been detected in the genomes of
76 plants outside angiosperms²⁰.

77

78 In this study, we use a large-scale phylogenomic approach to investigate whether
79 EPRVs are present in the genomes of plants outside angiosperms. By mining ten
80 gymnosperm and six fern genomes, we identified EPRVs in the genomes of nearly all
81 these gymnosperms and ferns. Phylogenetic analyses using the newly identified
82 EPRVs together with other angiosperm viruses reveal an unappreciated diversity of
83 *Caulimoviridae* and show that the known angiosperm viruses only constitute a minor
84 part of the *Caulimoviridae* diversity. The newly identified EPRVs in gymnosperms
85 and ferns provide important and novel insights into the diversity, distribution, and
86 macroevolution of *Caulimoviridae*.

87

88 **Results**

89 **Identification of EPRVs in gymnosperms and ferns**

90 We used a combined similarity search and phylogenetic analysis approach to
91 screen the genomes of ten gymnosperms, six ferns, and four other earlier branching
92 plant species (*Selaginella moellendorffii*, *Physcomitrella patens*, *Marchantia*

93 *polymorpha*, and *Klebsormidium flaccidum*) for the presence of EPRVs (Fig. 1 and
94 Supplemental Table 1). Briefly, similarity search with the protein sequences of
95 representative *Caulimoviridae* was performed against these plant genomes (Fig. 1 and
96 Supplemental Table 1). Given RT and RH of *Caulimoviridae* share significant
97 similarity with retrotransposons and other reverse-transcribing viruses, EPRVs were
98 further identified and confirmed by phylogenetic analyses (see Methods). We found
99 that EPRVs are present in the genomes of nearly all the gymnosperms and ferns
100 investigated in this study (Fig. 1), suggesting that EPRVs are prevalent and
101 widespread in gymnosperms and ferns. EPRVs were not identified in the genome of
102 the fern *Ceratopteris richardii*, which does not necessarily indicate the absence of
103 EPRVs but is more likely due to the low-density coverage (1.082×) of its genome
104 sequencing (only 1% of its genome is covered by the genome assembly)²¹. No EPRV
105 was detected in the genomes of the lycophyte *S. moellendorffii*, the moss *P. patens*,
106 the liverwort *M. polymorpha*, and the charophyte *K. flaccidum* (Fig. 1). Together with
107 previous reports of EPRVs in angiosperms¹⁷⁻²⁰, we conclude that EPRVs are
108 widespread in the genomes of euphyllophytes (ferns and seed plants).

109
110 The copy numbers of EPRVs within the genomes of gymnosperms and ferns
111 appear to vary widely across plant species. We estimated that the genomes of the
112 gymnosperms and ferns contain 112 – 20,579 copies of EPRVs (Table 1). However,
113 the genomes of some gymnosperms and ferns are of low coverage, these estimated
114 numbers should be taken with cautions. On the other hand, for the three genomes
115 that >90% are covered by genome assembly (*Pinus taeda*, *Picea glauca*, and *Ginkgo*
116 *biloba*), the EPRV copy numbers were estimated to be 6,733, 2,520, and 481,
117 respectively. Our results suggest that EPRVs might represent major components of
118 plant genomes.

119

120 **Diversity and classification of *Caulimoviridae***

121 To explore the relationship between the newly identified gymnosperm and fern
122 EPRVs and the known angiosperm *Caulimoviridae*, we inferred a phylogeny of

123 representative exogenous and endogenous viruses of *Caulimoviridae* using the highly
124 conserved RT-RH proteins with the retrotransposon *Ty3* as the outgroup. Our
125 phylogenetic analysis reveals an extraordinarily large diversity of the *Caulimoviridae*
126 family, which has never been appreciated previously (Fig. 2). The known eight viral
127 genera and florendoviruses fell well within the diversity of EPRVs of gymnosperms
128 and ferns. It follows that the previously known angiosperm *Caulimoviridae* constitute
129 only a minor part of its diversity.

130

131 Our phylogenetic analysis identified at least seven monophyletic groups of
132 EPRVs with high supports (Bayesian posterior probability > 0.95) in gymnosperms
133 and ferns. These clades were designated gymnosperm endogenous florendovirus-like
134 virus 1-3 (GEFLV 1-3), fern endogenous florendovirus-like virus (FEFLV),
135 gymnosperm endogenous caulimovirus-like virus 1-2 (GECLV 1-2), and fern
136 endogenous caulimovirus-like virus (FECLV), respectively (Fig. 2). The host of each
137 clade is restricted to one plant division (except GECLV 1) (Fig. 2). The divergence
138 within one of these clades is comparable to and even greater than that of one known
139 *Caulimoviridae* genus. Some gymnosperm and fern EPRVs are not readily classified,
140 either because one virus formed one branch or because the branches are not strongly
141 supported.

142

143 **Macroevolutionary mode of *Caulimoviridae***

144 To estimate the relative importance of co-speciation and host switching in the
145 macroevolution of *Caulimoviridae*, we performed a global assessment of the
146 correspondence between *Caulimoviridae* and host phylogenetic trees using the
147 event-based approach. The analyses found no significant signal for co-speciation (p
148 values > 0.05; Table 2), suggesting co-speciation might not play a predominant role in
149 the diversification of *Caulimoviridae*.

150

151 Our phylogenetic analysis shows that some newly described fern EPRVs occupy
152 important phylogenetic positions -- basal to all the other known *Caulimoviridae*.

153 Ancestral state reconstruction reveals that the *Caulimoviridae* family originated in
154 ferns (Supplemental Fig. 1). Our phylogenetic analysis also shows that the
155 angiosperm viruses form three independent monophyletic groups: two consist of
156 known eight genera of exogenous viruses and one consists of florendoviruses (Fig. 2).
157 The three angiosperm viral groups are only distantly related to each other. The
158 phylogenetic relationship among euphyllophyte viruses indicate that the angiosperm
159 viruses originated multiple times probably through cross-division transmission from
160 gymnosperms.

161

162 **Genome structure evolution of *Caulimoviridae***

163 To explore the genome structure evolution within the *Caulimoviridae* family, we
164 reconstructed the consensus genome sequences of EPRVs (Supplemental Data 1-5).
165 Given the fern genomes are of low coverage, we only reconstructed the genomes of
166 gymnosperm EPRVs, and one representative for each of the five gymnosperm EPRV
167 clades were inferred. These gymnosperm EPRV genomes vary wildly in size (from
168 6,061 to 8,109 bps) and ORF organization (Fig. 3 and Supplementary Data).

169 Conserved Domain (CD) searches show that protein domains common to all the
170 EPRVs include MP, AP, RT, and RH, suggesting that the gymnosperm EPRVs exhibit
171 a similar protein architecture as the angiosperm *Caulimoviridae* (Fig. 3). No CP
172 homologs were identified in the consensus genome sequences, possibly due to the
173 rapid nature of its evolution. However, we identified the zinc-finger CCHC motif, a
174 hallmark of the CP protein, in PtaeV_2 (GECLV1) and GbilV (GECLV2) (Fig. 3). CD
175 searches did not find any integrase-like domain, a pattern similar to the angiosperm
176 *Caulimoviridae* and indicates integration might not be necessary for the replication of
177 gymnosperm EPRVs either.

178

179 **Age estimate of EPRV bursts**

180 Because the genomes of loblolly pine (*P. taeda*) and ginkgo (*G. biloba*) were of
181 relatively high quality and contain rather distinct numbers of EPRVs (Table 1), they
182 were used to infer the evolutionary dynamics of EPRVs within the host genomes.

183 Mixture model analyses of the genetic divergence between EPRV copies and their
184 consensus nucleotide sequence show that there are four and three peaks in *P. taeda*
185 and *G. biloba* (Supplemental Table 2), suggesting at least four and three independent
186 EPRV integration events occurring along the lineages leading to *P. taeda* and *G.*
187 *biloba*, respectively. Based on the mixture analyses and phylogenetic analyses
188 (Supplemental Fig. 2), the ERPVs within the genomes of *P. taeda* and *G. biloba* were
189 classified into four and three families.

190

191 We failed to find any orthologous integration of EPRVs in different species and
192 cannot directly estimate the time of viral integration. The median pairwise genetic
193 distance within each family was calculated to examine the age of burst for each EPRV
194 family^{22, 23} (Supplemental Table 3). Our results show that the EPRV proliferation
195 dynamics were of difference between *P. taeda* and *G. biloba* (Supplemental Table 3).
196 But we found all of the EPRV families investigated here experienced proliferation
197 peaks tens or hundreds of million years ago. Consistently, the EPRV copies contain
198 many frame-shift mutations and premature stop codons (Supplemental Fig. 4).
199 However, those analyses come with two caveats: i) it is uncertain whether the EPRV
200 proliferation activity within the host genome follows the Gaussian distribution; ii) the
201 evolutionary rate of EPRVs remains unclear. Nevertheless, our results suggest that
202 EPRVs evolved within their host genomes for hundreds of millions of years and
203 indicate an ancient origin of *Caulimoviridae*.

204

205 Discussion

206 In this study, we report the identification of EPRVs within the genomes of
207 gymnosperms and ferns. Together with the previous reports of exogenous and
208 endogenous *Caulimoviridae* in angiosperms, our results demonstrate that all the major
209 lineages of euphyllophytes (ferns and seed plants) are/were infected by the
210 *Caulimoviridae* family. No EPRVs detected in the genome of *C. richardii* does not
211 necessarily mean the absence of EPRVs, which is probably due to the low coverage

212 nature of its genome assembly. Few viruses have been documented in plant species
213 outside angiosperm²⁴. The identification of EPRVs in gymnosperms and ferns makes
214 *Caulimoviridae* the only known viral family that infects all major lineages of
215 euphyllophytes.

216

217 Our findings show that the newly identified EPRVs exhibit an unprecedented
218 diversity, and the known angiosperm viral diversity only accounts for a minority of
219 the *Caulimoviridae* diversity. The current *Caulimoviridae* classification system¹⁶
220 cannot readily account for the diversity of EPRVs in gymnosperms and ferns. Indeed,
221 the divergence within one clade of gymnosperm or fern EPRVs is comparable to the
222 divergence of one exogenous viral genus or florendoviruses. Therefore, an updated
223 classification incorporating gymnosperm and fern EPRVs should be developed. Most
224 of the EPRV clades lack the exogenous counterparts, either because the ancient viral
225 lineages completely died out, or because many exogenous viruses remain to be
226 discovered.

227

228 Two possible macroevolutionary modes of *Caulimoviridae* could be conceived: i)
229 Co-speciation model: the viruses have co-evolved with their euphyllophyte hosts
230 for >400 million years and undergone sporadic cross-species transmission (Fig. 5a); ii)
231 Cross-species transmission model: frequent cross-species transmissions predominated
232 in the evolution of *Caulimoviridae* (Fig. 5b). In this study, we failed to find
233 co-speciation signal between *Caulimoviridae* and its hosts, suggesting co-speciation
234 might not be predominant in the macroevolution of *Caulimoviridae* (Table 2 and
235 Supplemental Fig. 3). Indeed, it appears that the angiosperm viruses originated
236 multiple times via independent cross-division transmission events (gymnosperms to
237 angiosperms). For plant viruses, cross-division transmission events were rarely
238 documented, partially because much remains unknown about the virosphere in plants
239 outside angiosperms. Given multiple cross-division transmission events took place in
240 *Caulimoviridae*, we think that cross-division transmission might not be a rare event
241 for other plant viruses.

242

243 Phylogenetic analysis and ancestral state reconstruction show that the
244 *Caulimoviridae* family might have a monilophyte origin (Supplemental Fig. 1), which
245 is compatible with the fact that we did not find any EPRV in earlier branching plants
246 (lycophytes and nonvascular plants). One might argue that the absence of EPRVs in
247 earlier branching plants is due to no viral integration occurring. However, this
248 possibility seems to be unlikely, given viral integration is so widespread in
249 euphyllophytes. The paleoviruses provide ‘molecular fossils’ for estimating the age of
250 related viruses. Previously, the integration of banana streak virus into the *Musa*
251 *balbisiana* genome was estimated to occur 0.63 million years ago^{25, 26}. The
252 endogenization of florendoviruses in *Oryza* species was estimated to take place at
253 least 1.8 million years ago²⁰. Although we cannot directly date when the EPRV
254 endogenization events occurred by finding orthologous integration of EPRVs in
255 closely related species, the proliferation dynamics analyses of EPRVs within two
256 representative genomes indicate they might have activated within the host genomes
257 for hundreds of millions of years. Taken together, our findings pinpoint a possible
258 ancient monilophyte origin of *Caulimoviridae*.

259

260 Unlike EVEs of non-retroviral source, the copy numbers of EPRVs are generally
261 high (>1,000 copies for 10 out of 16 gymnosperm and fern species), suggesting that
262 EPRVs contribute significantly to the complexity of host genomes. *Caulimoviridae*
263 are closely related to LTR retrotransposons¹¹. On the other hand, EPRVs lack LTRs,
264 which makes it inappropriate to be classified as an LTR retrotransposon. It seems to
265 be more appropriate to define EPRVs as a unique group of transposable elements¹⁹.

266

267 Our findings suggest that *Caulimoviridae* integrated into and amplified within
268 host genomes multiple times. However, the integration and amplification mechanisms
269 of EPRVs remain unclear, as the *Caulimoviridae* genomes lack integrase-like proteins
270 and integration is not essential for their replication. Several potential mechanisms
271 might be involved: i) Unlike the highly sequestered nature of animal germlines, plant

272 germline tissues might allow more frequent infections of viruses. However, the copy
273 numbers of EVEs within animals seem to be generally similar to EVEs within plants⁵.
274 ⁶. It appears that the high copy numbers of EPRVs are due to its own biology. ii)
275 EPRVs encode a ‘cryptic’ integrase without significant similarity with the known
276 proteins that function in integration. No integrase domain was found in the *Petunia*
277 *vein clearing virus* (PVCV) genome. But one of its proteins encodes two distinctive
278 motifs [HHCC and DD(35)E] that are shared by the integrase domain of retroviruses
279 and LTR retrotransposons²⁷. However, it remains unknown whether that protein
280 performs function similar to integrase. iii) Microhomology-mediated recombination
281 between EPRV sequences and host sequences during host gap repair process could
282 result in the integration of viral sequences into the host genomes¹⁷. This mechanism
283 requires the free ends of open circular viral sequences produced during virus
284 replication^{19,28}. iv) Like short interspersed elements (SINE), EPRVs might integrate
285 and amplify themselves within the host genomes via hijacking the integrase of other
286 retrotransposons^{29,30}.

287

288 **Method**

289 **Identification of EPRVs in plant genomes**

290 The genome sequences of twenty plant species were used to screen the presence
291 of EPRVs, including ten gymnosperms (*P. taeda*, *Pinus lambertiana*, *Pinus sylvestris*,
292 *Picea abies*, *Picea glauca*, *G. biloba*, *Gnetum gnemon*, *Juniperus communis*, *Taxus*
293 *baccata*, and *Abies sibirica*), six ferns (*C. richardii*, *Dipteris conjugata*, *Plagiogyria*
294 *formosana*, *Pteridium aquilinum*, *Polypodium glycyrrhiza*, and *Cystopteris protrusa*),
295 one moss (*P. patens*), one liverwort (*M. polymorpha*), one lycophyte (*S.*
296 *moellendorffii*), and one charophyte (*K. flaccidum*)^{21,31,32} (Table S1). To identify
297 putative EPRVs within these genomes, we employed a two-step phylogenomic
298 approach. First, the tBLASTn algorithm was employed to search against the plant
299 genomes using the RT-RH domain sequences of *Rice tungro bacilliform virus* (RTBV)
300 and PVCV as queries with an *e* cutoff value of 10⁻¹⁰. Next, all the significant hits

301 obtained were aligned with RT-RH sequences of representative LTR retrotransposons,
302 retroviruses, *Hepadnaviridae*, and *Caulimoviridae*³³ using MAFFT with default
303 parameters³⁴. Putative EPRVs, which form a monophyletic group with other
304 *Caulimoviridae* with high support values, were identified based on phylogenetic
305 analyses. EPRVs were confirmed by further rounds of phylogenetic analyses with
306 putative EPRVs and representative LTR retrotransposons, retroviruses, *Hepadnaviridae*,
307 and *Caulimoviridae*. Phylogenetic analyses were performed using an approximate
308 maximum likelihood method implemented in FastTree 2.1.9 with default parameters³⁵.
309 The copy number of EPRVs within each species was then counted. If length between
310 hits was less than 5,000 bps and the hits were in the same order as the query, the hits
311 were treated as a single copy.

312

313 **Phylogenetic analysis**

314 To further analyze the relationship among *Caulimoviridae*, phylogenetic analysis
315 was performed using the RT-RH protein sequences from representative EPRV
316 sequences of each gymnosperm and fern species, exogenous viruses, and
317 florendoviruses. The *Ty3* retrotransposon sequence was used as the outgroup. These
318 protein sequences were aligned using MAFFT algorithm with an accurate method
319 with the L-INS-i strategy³⁴. Ambiguous regions within the alignment were removed
320 using Gblocks 0.91b³⁶. The phylogenetic analysis was performed using a Bayesian
321 method implemented in MrBayes 3.2.6 (ref. 37). The RtRev substitution model was
322 used. A total of 17,000,000 generations in four chains were run, sampling posterior
323 trees every 100 generations. The first 25% of the posterior trees were discarded for
324 further analysis.

325

326 **Reconstruction of consensus genome sequences**

327 Relatively complete EPRV sequences with *MP-AP-RT-RH* domains with
328 extended flanking regions (~5,000 bps for each end) were extracted. These sequences
329 were then used as queries to search sequences with high similarity within their own
330 host genome using the BLASTn algorithm with an *e* cutoff value of 10^{-25} . The

331 significant hits were aligned using MAFFT³⁴ and consensus sequences were generated
332 using Geneious 10 (ref. 38) and manually edited. ORFs with the minimum size of 500
333 were found using Geneious 10 (ref. 38). Protein domains within these reconstructed
334 genomes were detected using the CD search³⁹.

335

336 **Analysis of EPRV activity within host genomes**

337 Because the genomes of *P. taeda* and *G. biloba* were of relatively high quality, we
338 used their genomes to infer the evolutionary dynamics of EPRVs within the host
339 genomes. The *AP-RT-RH* nucleotide sequences with length >500 bps of all the EPRVs
340 within *P. taeda* and *G. biloba* genomes were extracted independently. These
341 sequences were aligned using MAFFT and consensus sequences were inferred using
342 Geneious 10 (ref. 38). The genetic distance between the consensus sequences and
343 EPRVs was calculated based on the Kimura two-parameter model. To identify
344 significant peaks in the genetic distance data sets, Gaussian mixture models were
345 fitted using the R package mclust. The number of components (each component is
346 modeled by the Gaussian distribution) was estimated by fitting models. The Bayesian
347 Information Criterion (BIC) was used as the model selection criterion.

348

349 The phylogenetic trees of the *AP-RT-RH* nucleotide sequences extracted above
350 within each host genome were reconstructed using FastTree 2.1.9 with a GTR+CAT
351 model. The different EPRV families within the phylogenetic trees were allocated
352 based on the results by mixture model analyses. The nucleotide sequences of each
353 EPRV family were extracted and aligned using MAFFT³⁴. Pairwise genetic distances
354 were calculated based on the Kimura two-parameter model. The age of burst (T) for
355 each EPRV family was estimated through $T = D/2\mu$, where D represents the median
356 pairwise distance and μ represents the evolutionary rate of host ($\sim 1.43 \times 10^{-9}$ - 2.2×10^{-9}
357 substitutions per site per year³¹).

358

359 **Co-speciation analysis**

360 We explore the host-virus co-speciation signal at the level of class, for the

361 complex evolutionary history of EPRVs after integration might complicate
362 co-speciation analysis. The relationships between virus and host phylogenetic trees
363 were assessed using an event-based method implemented in Jane 4 (ref. 40). Briefly,
364 five events (cospeciation, duplication, duplication & host switch, loss and failure to
365 diverge) were assigned to a cost. The number of each event was estimated by finding
366 the solution with the minimum total cost. The event-cost schemes
367 (cospeciation-duplication-duplication & host switch-loss-failure to diverge) were set
368 as follows, -1-0-0-0-0 (refs. 41, 42) and 0-1-1-2-0 (ref. 40). Host-virus phylogeny
369 congruence was assessed by statistical tests with both two randomization methods,
370 random tip mapping and random parasite tree, with the sample size of 500.

371

372 **Reconstruction of ancestral states**

373 To detect the macroevolutionary pattern among *Caulimoviridae*, we performed
374 ancestral state reconstruction with Mesquite 3.10 (ref. 43). We assigned the 96 viral
375 taxa (Fig. 2) using their hosts (gymnosperm, angiosperm, and fern) as characters. For
376 the outgroup Ty3, a '?' character was assigned. The parsimony model was used to
377 trace character evolution over the posterior trees sampled in the aforementioned
378 Bayesian analysis.

379

380 **Acknowledgements**

381 This work was supported by the Natural Science Foundation of Jiangsu Province
382 (BK20161016) and the Priority Academic Program Development (PAPD) of Jiangsu
383 Higher Education Institutions.

384

385 **Author contributions**

386 G.-Z.H. designed the study. G.Z. and G.-Z.H performed the experiments. G.-Z.H
387 and G.Z. analyzed the data and wrote the manuscript.

388

389 **Competing Financial Interests**

390 The authors declare no competing financial interests.

391

392 **References**

393 1. Katzourakis, A. & Gifford, R. J. Endogenous Viral Elements in Animal Genomes.
394 *PLOS Genet.* **6**, e1001191 (2010).

395 2. Emerman, M. & Malik, H. S. Paleovirology--modern consequences of ancient
396 viruses. *PLOS Biol.* **8**, e1000301 (2010).

397 3. Hayward, A., Grabherr, M. & Jern, P. Broad-scale phylogenomics provides
398 insights into retrovirus–host evolution. *Proc. Natl Acad. Sci. USA* **110**,
399 20146-20151 (2013).

400 4. Belshaw, R. *et al.* Long-term reinfection of the human genome by endogenous
401 retroviruses. *Proc. Natl Acad. Sci. USA* **101**, 4894–4899 (2004).

402 5. Feschotte, C. & Gilbert, C. Endogenous viruses: insight into viral evolution and
403 impact on host biology. *Nat. Rev. Genet.* **13**, 283-296 (2012).

404 6. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and
405 ancient viral evolution. *Virology* **479-480**, 26-37 (2015).

406 7. Aswad, A. & Katzourakis, A. Paleovirology and virally derived immunity. *Trends*
407 *Ecol. Evol.* **27**, 627–636 (2012).

408 8. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate
409 immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087
410 (2016).

411 9. Esnault, C., Cornelis, G., Heidmann, O. & Heidmann, T. Differential
412 Evolutionary Fate of an Ancestral Primate Endogenous Retrovirus Envelope
413 Gene, the EnvV *Syncytin*, Captured for a Function in Placentation. *PLOS Genet.*
414 **9**, e1003400 (2013).

415 10. Temin, H. M. Reverse transcription in the eukaryotic genome: retroviruses,
416 pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* **2**,
417 455-468 (1985).

- 418 11. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon
419 their reverse transcriptase sequences. *EMBO J.* **9**, 3353-62 (1990).
- 420 12. Gilbert, C. & Feschotte, C. Genomic Fossils Calibrate the Long-Term Evolution
421 of Hepadnaviruses. *PLOS Biol.* **8**, e1000495 (2010).
- 422 13. Gilbert, C. *et al.* Endogenous hepadnaviruses, bornaviruses and circoviruses in
423 snakes. *Proc. Biol. Sci.* **281**, 20141122 (2014).
- 424 14. Suh, A. *et al.* Early mesozoic coexistence of amniotes and hepadnaviridae. *PLOS*
425 *Genet.* **10**, e1004559 (2014).
- 426 15. Dill, J. A. *et al.* Distinct viral lineages from fish and amphibians reveal the
427 complex evolutionary history of hepadnaviruses. *J. Virol.* **90**, 7920-7933 (2016).
- 428 16. King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. *Virus Taxonomy:*
429 *Classification and Nomenclature of Viruses: Ninth Report of the International*
430 *Committee on Taxonomy of Viruses.* (Elsevier Academic Press, 2012).
- 431 17. Staginnus, C. & Richert-Pöggeler, K. R. Endogenous pararetroviruses: two-faced
432 travelers in the plant genome. *Trends Plant Sci.* **11**, 485-91 (2006).
- 433 18. Harper, G., Osuji, J. O., Heslop-Harrison, J. S. & Hull, R. Integration of banana
434 streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence.
435 *Virology* **255**, 207-213 (1999).
- 436 19. Jakowitsch, J., Mette, M. F., van der Winden, J., Matzke, M. A. & Matzke, A. J.
437 M. Integrated pararetroviral sequences define a unique class of dispersed
438 repetitive DNA in plants. *Proc. Natl Acad. Sci. USA* **96**, 13241-13246 (1999).
- 439 20. Geering, A. D. W. *et al.* Endogenous florendoviruses are major components of
440 plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5**, 5269 (2014).
- 441 21. Wolf, P. G. *et al.* An Exploration into Fern Genome Space. *Genome Biol. Evol.* **7**,
442 2533-2544 (2015).
- 443 22. Vandewege, M. W., Platt, R. N., Ray, D. A. & Hoffmann, F. G. Transposable
444 Element Targeting by piRNAs in Laurasiatherians with Distinct Transposable
445 Element Histories. *Genome Biol. Evol.* **8**, 1327-1337 (2016).
- 446 23. Ray, D. A. *et al.* Multiple waves of recent DNA transposon activity in the bat,
447 *Myotis lucifugus*. *Genome Res.* **18**, 717-728 (2008).

- 448 24. Hull, R. *Plant Virology (Fifth edition)*, p47, (Elsevier Academic Press, 2014).
- 449 25. Gayral, P. *et al.* A single Banana streak virus integration event in the banana
450 genome as the origin of infectious endogenous pararetrovirus. *J. Virol.* **82**,
451 6697-710 (2008).
- 452 26. Gayral, P. *et al.* Evolution of endogenous sequences of banana streak virus: what
453 can we learn from banana (*Musa sp.*) evolution? *J. Virol.* **84**, 7346-59 (2010).
- 454 27. Richert-Pöggeler, K. R. & Shepherd, R. J. Petunia vein-clearing virus: a plant
455 pararetrovirus with the core sequences for an integrase function. *Virology* **236**,
456 137-146 (1997).
- 457 28. Kunii, M. *et al.* Reconstruction of putative DNA virus from endogenous rice
458 tungro bacilliform virus-like sequences in the rice genome: implications for
459 integration and evolution. *BMC Genomics* **5**, 80 (2004).
- 460 29. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in
461 structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187-215
462 (2011).
- 463 30. Richardson, S. R. *et al.* The Influence of LINE-1 and SINE Retrotransposons on
464 Mammalian Genomes. *Microbiol Spectr.* **3**, MDNA3-0061-2014 (2015).
- 465 31. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome
466 evolution. *Nature* **497**, 579-584 (2013).
- 467 32. Guan, R. *et al.* Draft genome of the living fossil Ginkgo biloba. *GigaScience* **5**,
468 49 (2016).
- 469 33. Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H. & Moya, A. Network
470 dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol.*
471 *Direct.* **4**, 41 (2009).
- 472 34. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software
473 Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**,
474 772-780 (2013).
- 475 35. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately
476 Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
- 477 36. Talavera, G. & Castresana, J. Improvement of phylogenies after removing

- 478 divergent and ambiguously aligned blocks from protein sequence alignments.
479 *Syst. Biol.* **56**, 564-577 (2007).
- 480 37. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and
481 Model Choice Across a Large Model Space. *Syst. Biol.* **61**, 539-542 (2012).
- 482 38. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software
483 platform for the organization and analysis of sequence data. *Bioinformatics* **28**,
484 1647-1649 (2012).
- 485 39. Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic*
486 *Acids Res.* **43**, D222-D226 (2015).
- 487 40. Conow, C., Fielder, D., Ovadia, Y. & Libeskind-Hadas, R. Jane: a new tool for the
488 cophylogeny reconstruction problem. *Algorithms Mol. Biol.* **5**, 16 (2010).
- 489 41. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early
490 Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).
- 491 42. Ronquist, F. Phylogenetic approaches in coevolution and biogeography. *Zool. Scr.*
492 **26**, 313-322 (1997).
- 493 43. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary
494 analysis. Version 3.10. (2016) Available at: <http://mesquiteproject.org>. Accessed
495 on 28 Jun 2016.
- 496 44. Chang, C., Bowman, J. L. & Meyerowitz, E. M. Field Guide to Plant Model
497 Systems. *Cell* **167**, 325-339 (2016).
- 498 45. Chaw, S. M., Zharkikh, A., Sung, H. M., Lau, T. C. & Li, W. H. Molecular
499 phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear
500 18S rRNA sequences. *Mol. Biol. Evol.* **14**, 56-68 (1997).
- 501 46. Bowe, L. M., Coat, G. & dePamphilis, C. W. Phylogeny of seed plants based on
502 all three genomic compartments: extant gymnosperms are monophyletic and
503 Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA* **97**, 4092-4097
504 (2000).

505 **Table 1. Copy number of EPRVs identified and estimated in plant genomes**

506

Species	Division	Proportion of genome covered by assembly (%)	No. of EPRVs identified	No. of EPRVs estimated ²
<i>Pinus lambertiana</i>	Pinophyta	89.04	4556	5117
<i>Pinus taeda</i>	Pinophyta	92.00	6194	6733
<i>Picea glauca</i>	Pinophyta	100	2520	2520
<i>Juniperus communis</i>	Pinophyta	16.04	18	112
<i>Taxus baccata</i>	Pinophyta	27.79	112	403
<i>Abies sibirica</i>	Pinophyta	14.08	130	923
<i>Pinus sylvestris</i>	Pinophyta	30.20	314	1040
<i>Picea abies</i>	Pinophyta	61.22	1238	2022
<i>Ginkgo biloba</i>	Ginkgophyta	100	481	481
<i>Gnetum gnemon</i>	Gnetophyta	55.66	1328	2386
<i>Dipteris conjugata</i>	Pteridophyta	9.51	144	1515
<i>Plagiogyria formosana</i>	Pteridophyta	0.31	64	20579
<i>Ceratopteris richardii</i>	Pteridophyta	3.11	0	0
<i>Pteridium aquilinum</i>	Pteridophyta	6.34	57	898
<i>Polypodium glycyrrhiza</i>	Pteridophyta	0.53	21	3947
<i>Cystopteris protrusa</i>	Pteridophyta	1.00	12	1199
<i>Physcomitrella patens</i>	Bryophyta	ND ¹	0	0
<i>Selaginella moellendorffii</i>	Lycopodiophyta	ND ¹	0	0
<i>Marchantia polymorpha</i>	Marchantiophyta	ND ¹	0	0
<i>Klebsormidium flaccidum</i>	Charophyta	ND ¹	0	0

507

508 ¹ Not determined.

509 ² No. of EPRVs estimated = No. of EPRVs identified/Proportion of genome covered by assembly.

510

Table 2. Number of events experienced by virus lineages

511

Event costs ¹	Total cost	Cospeciation ²	Duplication ²	Duplication & host switching ²	Loss ²	Failure to diverge ²	P-value ³	P-value ⁴
-1, 0, 0, 0, 0	-8	8-8	2-3	13-14	7-20	0-0	>0.05	>0.05
0, 1, 1, 2, 0	20	4-4	3-5	15-17	0-0	0-0	>0.05	>0.05

512

513 ¹Event costs are for cospeciation, duplication, duplication & host switching, loss, and failure to
514 diverge, respectively.

515 ²Number of events is expressed as ranges that result in the same cost.

516 ³Random tip mapping method with sample size of 500.

517 ⁴Random parasite tree method with sample size of 500.

518 **Figure Legends**

519 **Figure 1. Distribution of EPRVs within plant genomes.** The phylogenetic
520 relationship of plant species is based on Refs. 21, 31, 32, 44-46. Different plant
521 divisions were labelled in different colors, and angiosperms, gymnosperms, and ferns
522 were labelled in pink, green, and blue, respectively. The presence and absence of
523 EPRVs were marked with solid and open circles around the related species. The
524 half-filled circle indicates that EPRVs have been identified in some but not all the
525 angiosperms.

526

527 **Figure 2. Phylogenetic relationship of representative exogenous and endogenous**
528 ***Caulimoviridae*.** The phylogenetic tree was inferred based on the RT-RH protein
529 sequences using a Bayesian method. The tree was rooted using the *Ty3* LTR
530 retrotransposon as the outgroup. Bayesian posterior probabilities were shown on the
531 selected nodes. *Caulimoviridae* of angiosperms, gymnosperms, and ferns were
532 highlighted in orange, green, and blue, respectively. Virus abbreviations: CVMV,
533 Cassava vein mosaic virus; SPCV, Sweet potato caulimo-like virus; SPVCV, Sweet
534 potato vein clearing virus; TVCV, Tobacco vein clearing virus; BSOLV, Banana streak
535 OL virus; CoYMV, Commelina yellow mottle virus; RTBV, Rice tungro bacilliform
536 virus; CaMV, Cauliflower mosaic virus; FMV, Figwort mosaic virus; RYV, Rose
537 yellow vein virus; PCSV, Peanut chlorotic streak virus; SoyCMV, Soybean chlorotic
538 mottle virus; CitPRV, Citrus endogenous pararetrovirus; PVCV, Petunia vein clearing
539 virus; VvinCV_sc1, *Vitis vinifera* C virus sequence cluster 1; CsatAV, *Cucumis*
540 *sativus* A virus; MescV, *Manihot esculenta* virus; FvesV_sc1, *Fragaria vesca* virus
541 sequence cluster 1; AtrichAV, *Amborella trichopoda* A virus; AtrichCV, *Amborella*
542 *trichopoda* C virus; EgranV_sc1, *Eucalyptus grandis* virus sequence cluster 1;
543 OsatBV, *Oryza sativa* B virus; SbicV, *Sorghum bicolor* virus. Species abbreviations:
544 Pabi, *Picea abies*; Pgla, *Picea glauca*; Ptae, *Pinus taeda*; Plam, *Pinus lambertiana*;
545 Psyl, *Pinus sylvestris*; Asib, *Abies sibirica*; Gbil, *Ginkgo biloba*; Jcom, *Juniperus*
546 *communis*; Ggne, *Gnetum gnemon*; Tbac, *Taxus baccata*; Pfor, *Plagiogyria formosana*;

547 Pgly, *Polypodium glycyrrhiza*; Cpro, *Cystopteris protrusa*; Paqu, *Pteridium aquilinum*;
548 Pgly, *Polypodium glycyrrhiza*.

549

550 **Figure 3. Genome structures of representative gymnosperm EPRVs and**
551 **exogenous *Caulimoviridae*.** Virus name abbreviation: RTBV, Rice tungro bacilliform
552 virus; PVCV, Petunia vein clearing virus; Pglav_1, *Picea glauca* virus 1; Pglav_2,
553 *Picea glauca* virus 2; Ptaev_1, *Pinus taeda* virus 1; Ptaev_2, *Pinus taeda* virus 2;
554 GbilV, *Ginkgo biloba* virus. The tRNA^{Met} which represents the beginning of the viral
555 replication, was indicated by an asterisk. The grey lines represent the genomes and the
556 rectangles represent the putative open reading frames (ORFs). The conserved protein
557 domains were labelled in different colors. Abbreviations: MP, movement protein;
558 PR/AP: protease/pepsin-like aspartate protease; RT, reverse transcriptase; RH,
559 ribonuclease H1; zf_CCHC, zinc-finger CCHC motif.

560

561 **Figure 4. Proliferation dynamics of EPRVs within the *P. taeda* (a and c) and *G.***
562 ***biloba* (b and d) genomes.** Yellow rectangles in **a** and **c** represent the distribution of
563 the genetic distances between EPRV copies and their consensus sequences. **b** and **d**
564 indicate Gaussian mixture models fitted.

565

566 **Figure 5. The models of the *Caulimoviridae* macroevolution.** The evolution of plant
567 hosts and viruses were indicated by blue lines and yellow dash lines, respectively. **(a)**
568 Co-speciation model: the viruses have co-evolved with their euphyllophyte hosts and
569 undergone sporadic cross-species transmission. **(b)** Cross-species transmission model:
570 frequent cross-species transmission predominated in the evolution of *Caulimoviridae*.

EPRVs

bioRxiv preprint doi: <https://doi.org/10.1101/170415>; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Angiosperms



Picea glauca



Picea abies



Pinus sylvestris



Pinus taeda



Pinus lambertiana



Abies sibirica



Taxus baccata



Juniperus communis



Gnetum gnemon



Ginkgo biloba



Cystopteris protrusa



Polypodium glycyrrhiza



Pteridium aquilinum



Ceratopteris richardii



Plagiogyria formosana



Dipteris conjugata



Selaginella moellendorffii



Physcomitrella patens



Marchantia polymorpha



Klebsormidium flaccidum

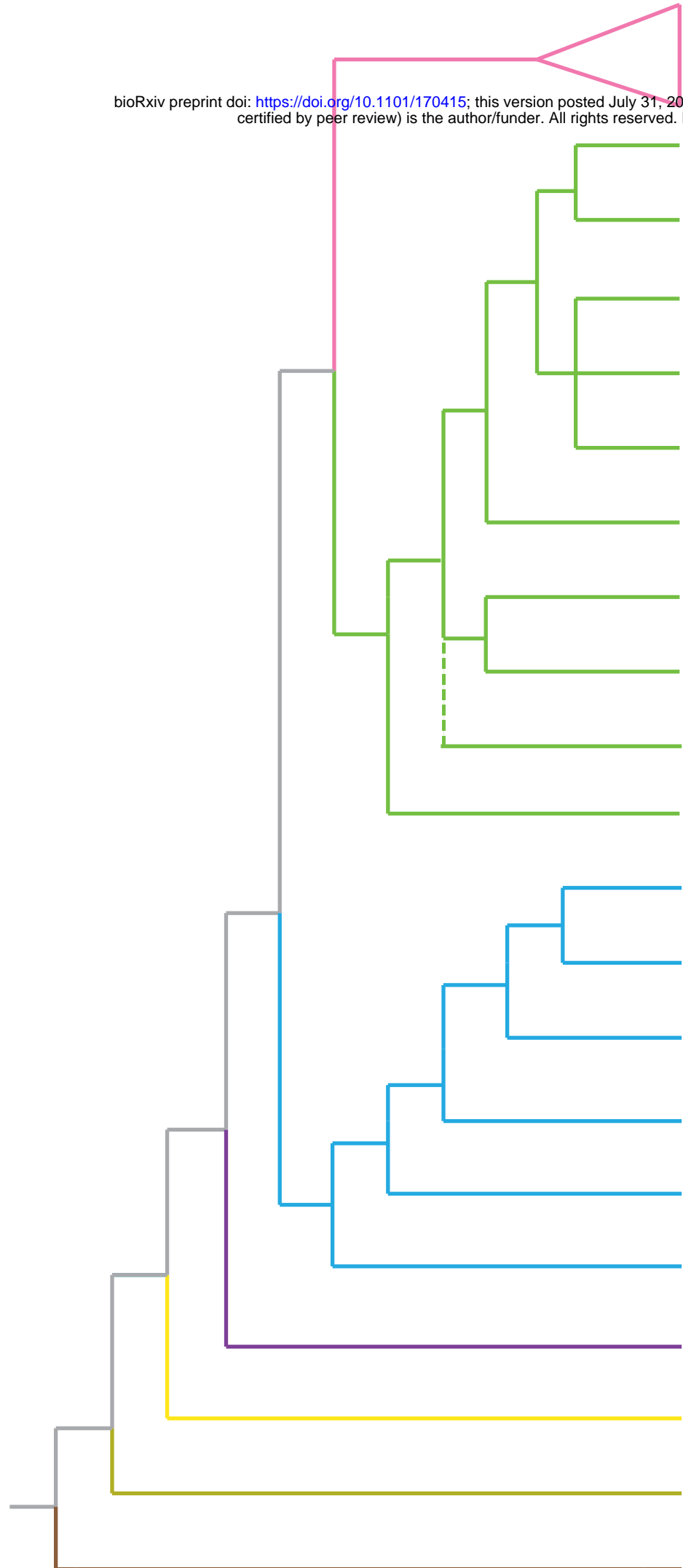


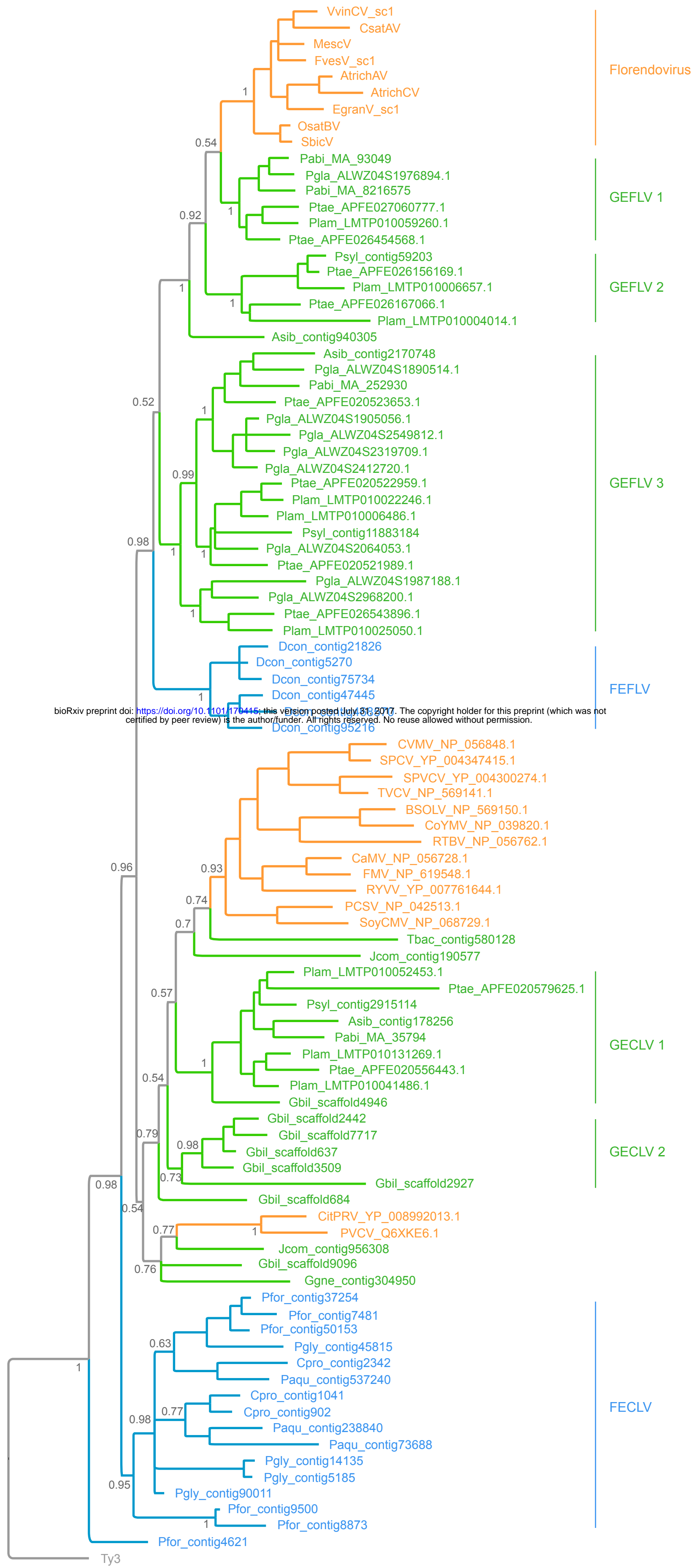
Gymnosperms

Euphyllophytes

Ferns

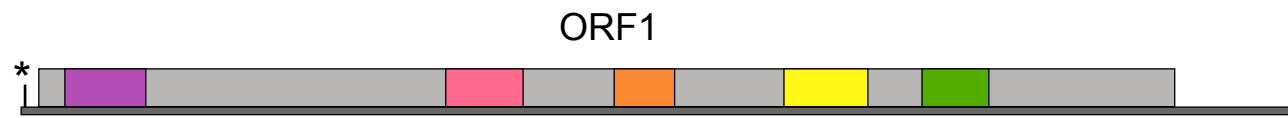
Vascular plants



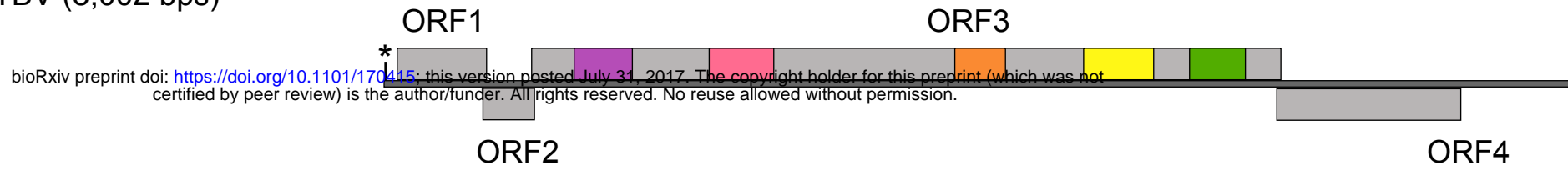


bioRxiv preprint doi: <https://doi.org/10.1101/170445>; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

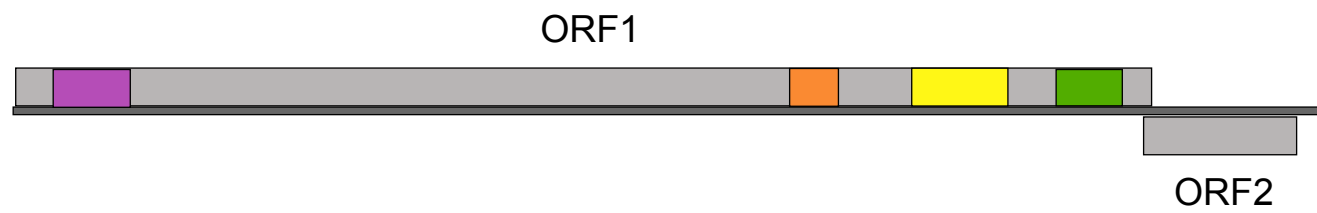
PVCV (7,206 bps)



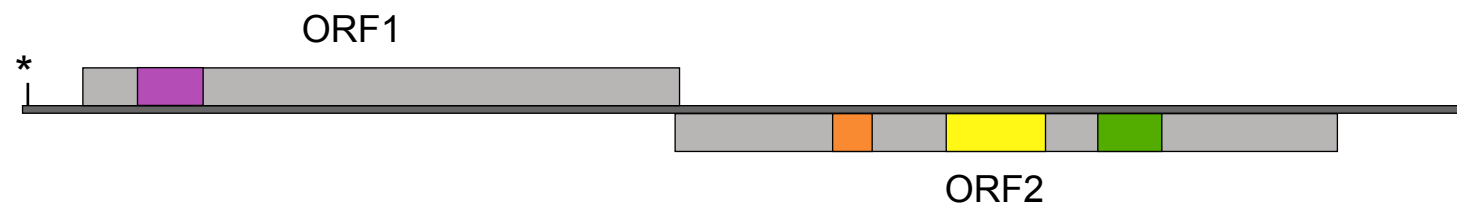
RTBV (8,002 bps)



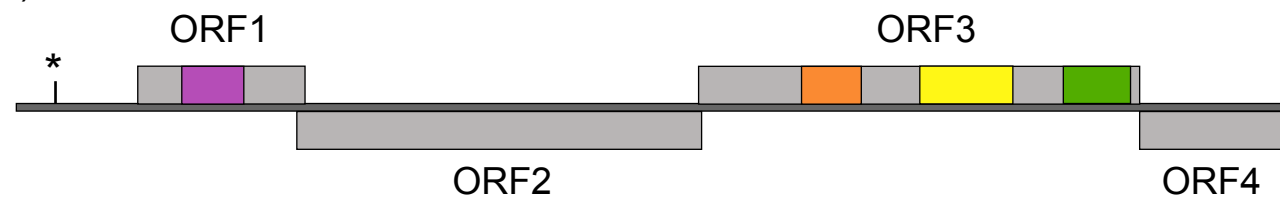
GEFLV 1: Pglav_1 (7,421 bps)



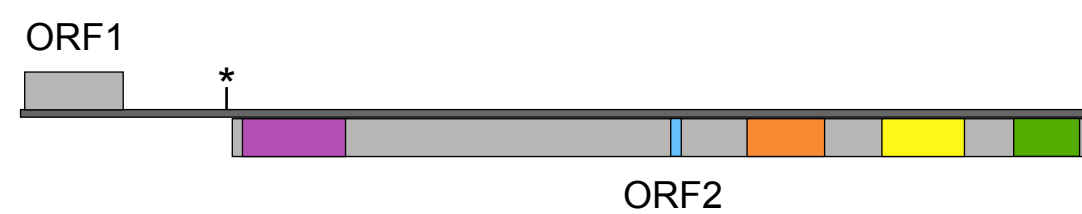
GEFLV 2: Ptaev_1 (8,109 bps)



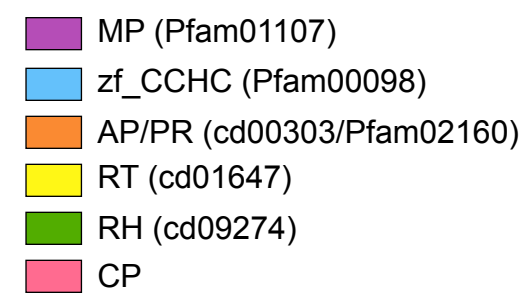
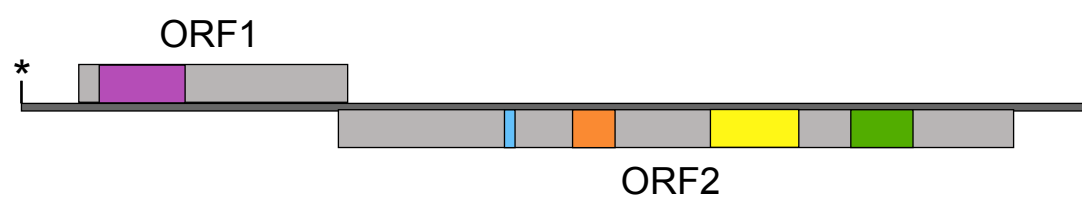
GEFLV 3: Pglav_2 (7,219 bps)

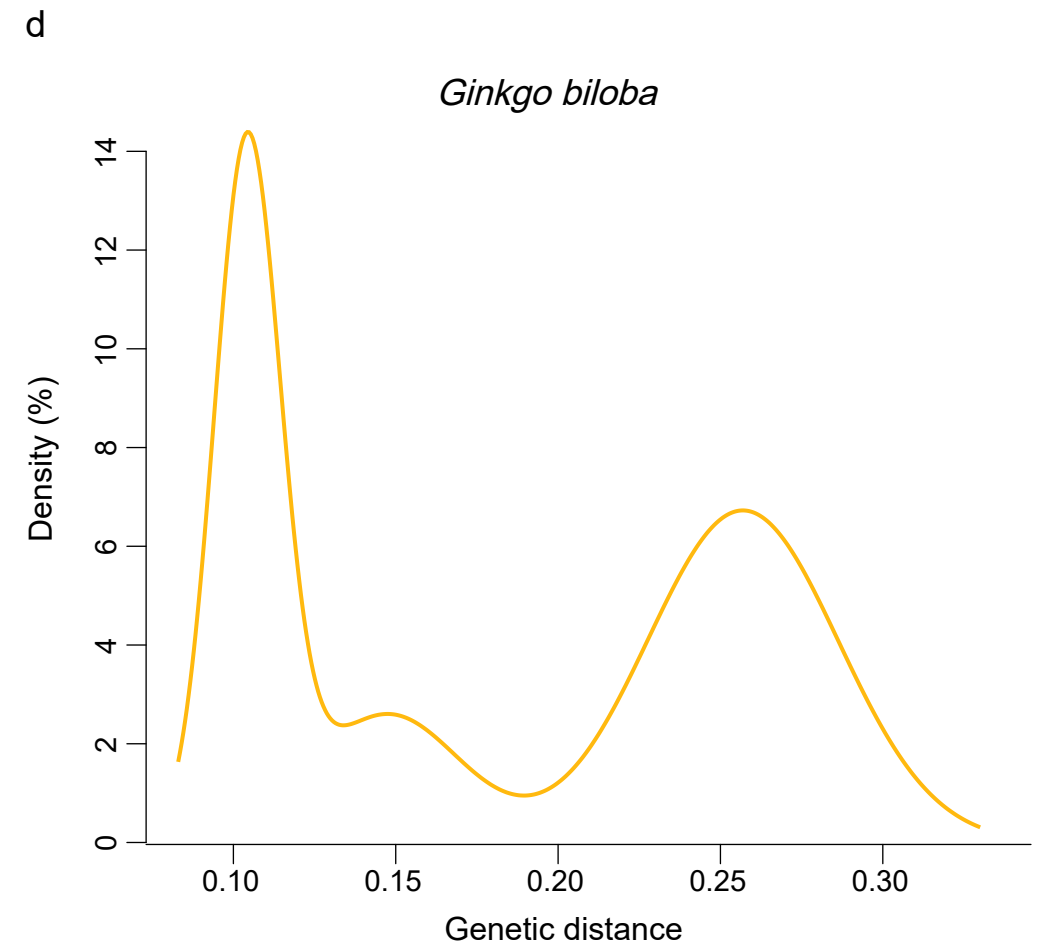
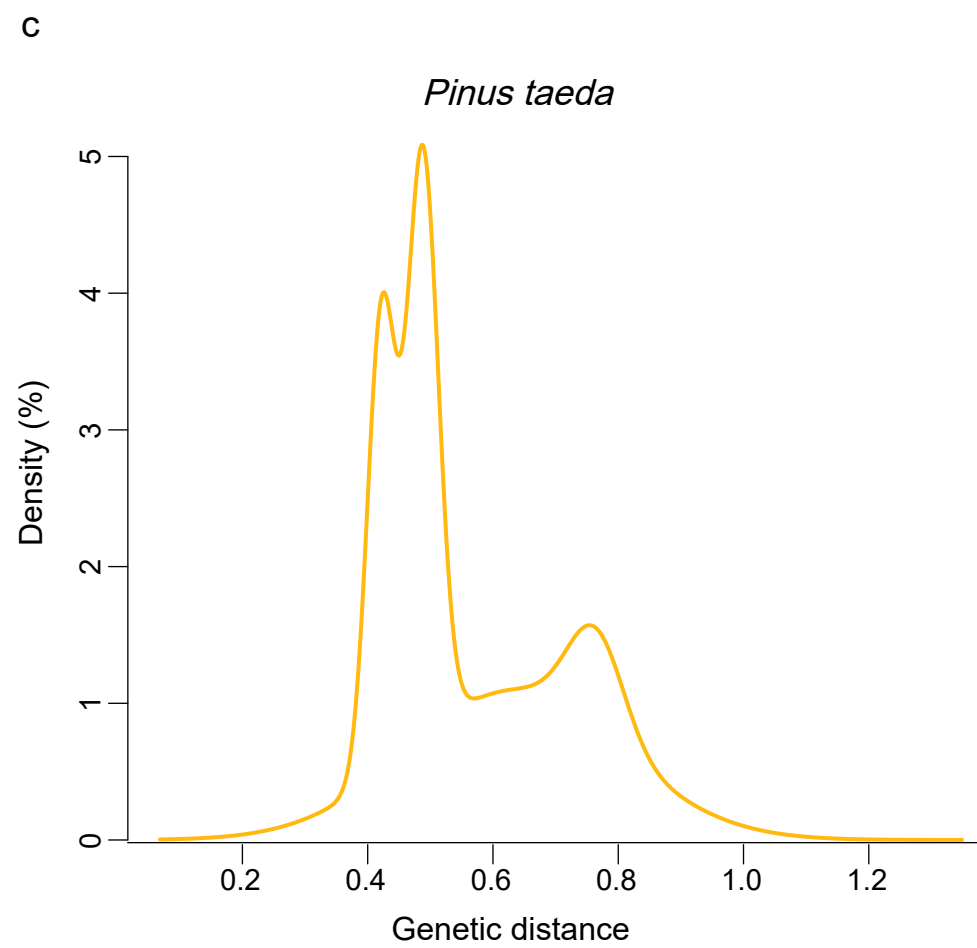
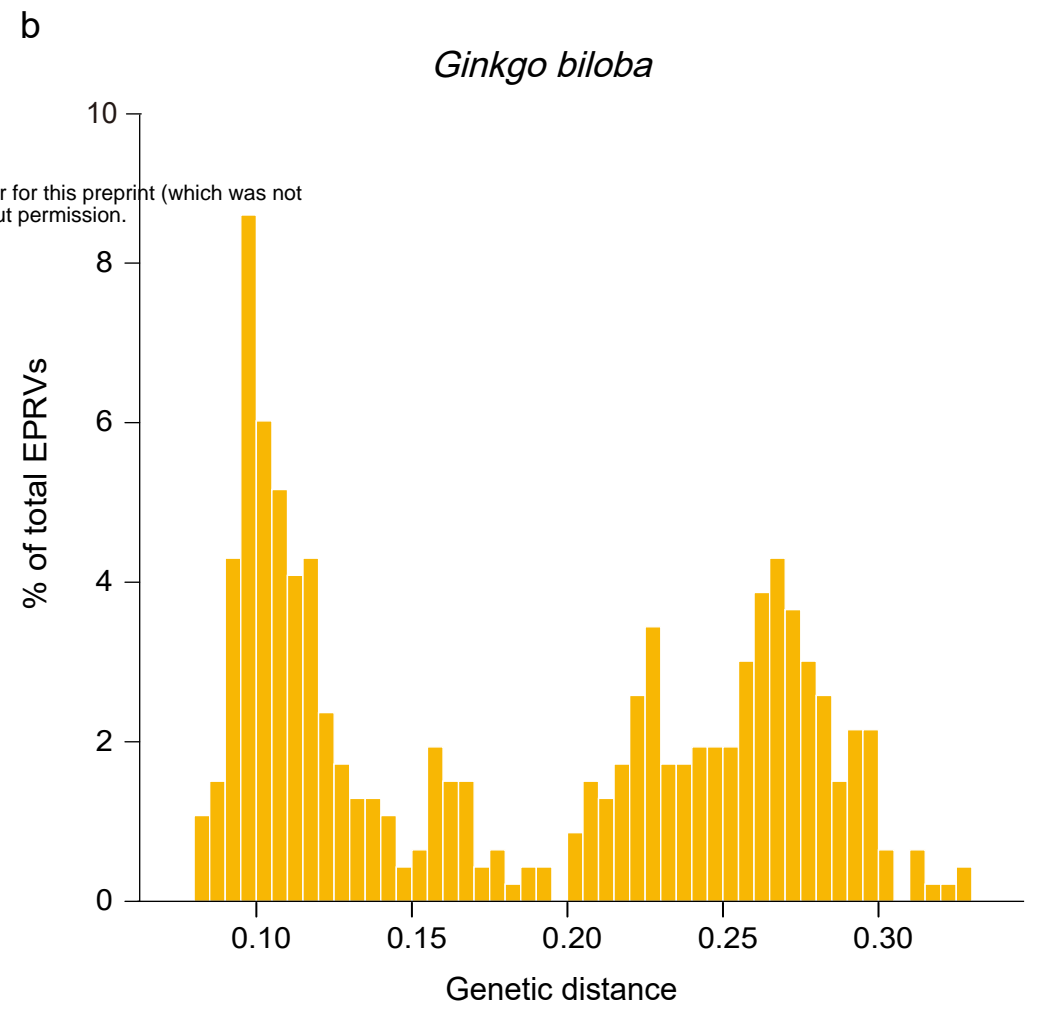
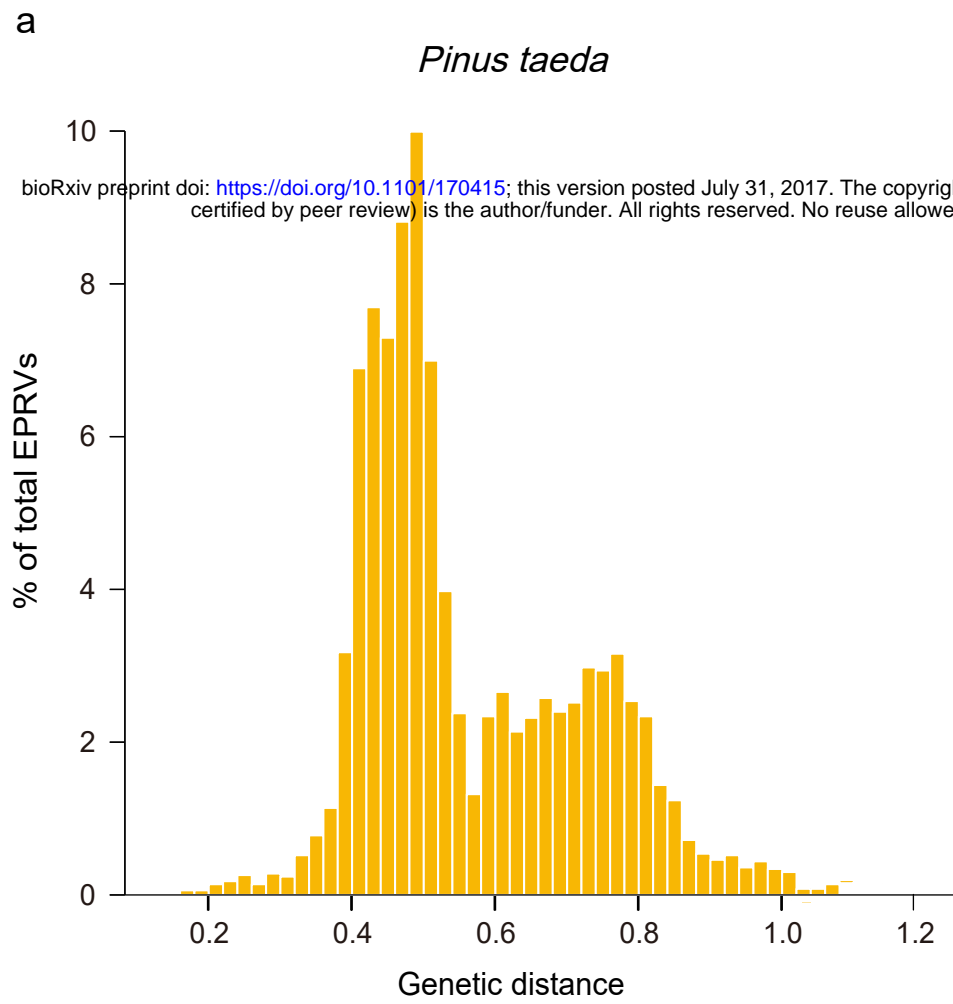


GECLV 1: Ptaev_2 (6,061 bps)

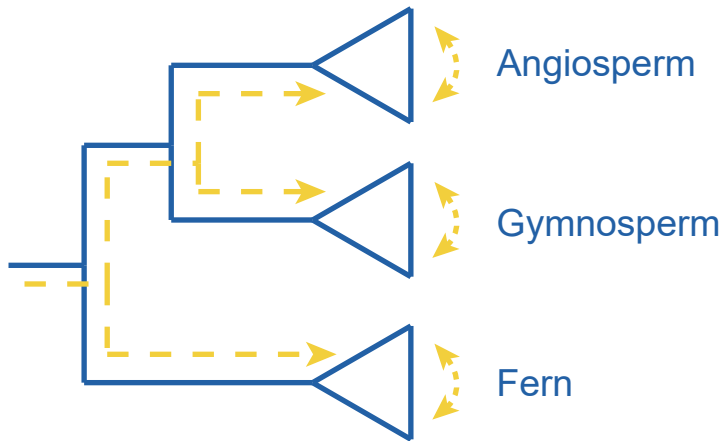


GECLV 2: Gbilv (6,081 bps)





a



b

