

# Supplementary Material

## Continued Adaptation of C<sub>4</sub> Photosynthesis After an Initial Burst of Changes in the Andropogoneae Grasses

MATHEUS E. BIANCONI, JAN HACKEL, MARIA S. VORONTSOVA, ADRIANA ALBERTI, WATCHARA ARTHAN, SEAN V. BURKE, MELVIN R. DUVALL, ELIZABETH A. KELLOGG, SÉBASTIEN LAVERGNE, MICHAEL R. MCKAIN, ALEXANDRE MEUNIER, COLIN P. OSBORNE, PAWEENA TRAIPEM, PASCAL-ANTOINE CHRISTIN, GUILLAUME BESNARD

This Supplementary Material contains seven figures and six tables. Note that tables S1, S2, S3 and S4 are in separate .xlsx files. An additional 31 data files, including scripts and alignments, can be found in the compressed folder “Bianconi\_et\_al\_Syst\_Biol\_2019.zip” (see the accompanying README file for details).

**Figure S1.** Effect of different filtering stringencies and reference species on the set of nuclear genes retained for phylogenetic analyses. (a,f) Number of genes retained; (b,g) taxon occupancy (i.e. proportion of species represented); (c,h) proportion of the alignment that is complete (completeness); (d,i) number of parsimony informative sites; (e,j) proportion of the branches with bootstrap support values  $\geq 50\%$ . Analyses were performed on the reduced subset of 37 species (a-e) and the complete set of 66 species (f-j), and using either *Setaria italica* or *Sorghum bicolor* as the reference. Each analysis was repeated with a different level of trimming (no trim = no trimming; 50% trim = sites covered by less than 50% of sequences were trimmed; 70% trim = sites covered by less than 70% of sequences were trimmed). Finally, analyses considered either all gene trees ('all') or only gene trees with  $\geq 50\%$  of branches with bootstrap support  $\geq 50\%$ . Dashed rectangles highlight the datasets used to generate the multigene coalescent species tree in Fig. 1 (in black), and Fig. S3 (in grey).

**Figure S2.** Bayesian phylogram inferred from non-coding sequences of plastomes. Black circles on nodes indicate Bayesian posterior probability  $\geq 0.95$ .

**Figure S3.** Multigene coalescent species trees inferred from 3,127 nuclear genes using *Sorghum bicolor* as a reference. Pie charts on nodes indicate the proportion of quartet trees that support the main topology (in blue), the first alternative (in red), and the second alternative (in orange). Local

posterior probabilities are indicated near nodes. Branch lengths are given in coalescent units. The major groups of Panicoideae are indicated with shades, as in Fig. 1.

**Figure S4.** Bayesian phylograms inferred from individual nuclear markers. (a) *apo1*, (b) *arodeh*, (c) *dwarf8*, (d) *floricaula*, (e) *kn1*, (f) *phyB*, (g) *rep1* and (h) *waxy*. Bayesian posterior probability values  $\geq 70$  are indicated near nodes. The main groups of Panicoideae are delimited with shades.

**Figure S5.** Time-calibrated phylogenetic tree based on coding sequences of plastomes. Bars on nodes indicate the 95% HPD intervals for the ages. The main groups of Panicoideae are delimited with shades.

**Figure S6.** Bayesian phylograms inferred for C<sub>4</sub>-related genes. These phylograms were inferred based on 3<sup>rd</sup> positions of codons of several core C<sub>4</sub> genes: (a) *nadpmdh-1P1*, (b) *nadpme-1P4*, (c) *pck-1P1*, (d) *ppc-1P3*, (e) *ppdk-1P2*; paralogs of core C<sub>4</sub> genes: (f) *nadpme-1P1*, (g) *nadpme-1P2*, (h) *nadpme-1P3*, (i) *ppc-1P4*, (j) *ppc-1P5*, (k) *ppc-1P7*, (l) *ppdk-1P1*; and other nuclear phylogenetic markers: (m) *arodeh*, (n) *dwarf8*, (o) *kn1*, (p) *phyB* and (q) *waxy*. Branches in red lead to C<sub>4</sub> species. Bayesian posterior probability values  $\geq 70$  are indicated near nodes. The main groups of Panicoideae are delimited with shades.

**Figure S7.** Phylograms with branch lengths based on amino acid sequences. The tree topologies were fixed to those obtained using 3<sup>rd</sup> positions of codons used for branch model tests. Results are shown for core C<sub>4</sub> genes: (a) *nadpmdh-1P1*, (b) *nadpme-1P4*, (c) *pck-1P1*, (d) *ppc-1P3*, (e) *ppdk-1P2*; paralogs of core C<sub>4</sub> genes: (f) *nadpme-1P1*, (g) *nadpme-1P2*, (h) *nadpme-1P3*, (i) *ppc-1P4*, (j) *ppc-1P5*, (k) *ppc-1P7*, (l) *ppdk-1P1*; and other nuclear genes: (m) *arodeh*, (n) *dwarf8*, (o) *kn1*, (p) *phyB* and (q) *waxy*. Branches in red lead to C<sub>4</sub> species. Scale bars = 0.01 expected amino acid substitutions per site.

**Table S1.** Genomic data information.

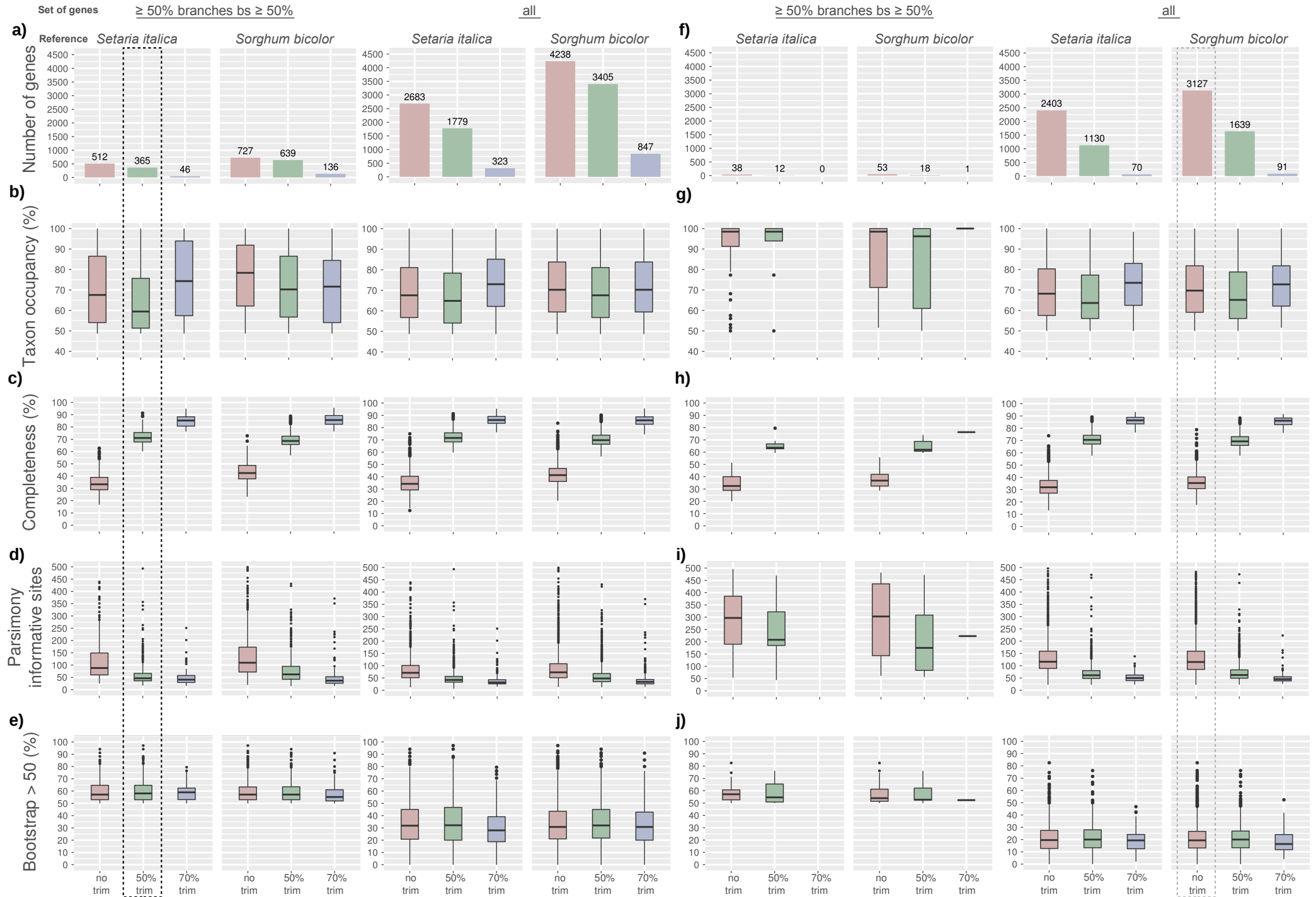
**Table S2.** Read mapping success and nuclear genome coverage in datasets assembled using *Setaria italica* or *Sorghum bicolor* sequences as references.

**Table S3.** NCBI accession numbers of sequences of nuclear markers assembled in this study.

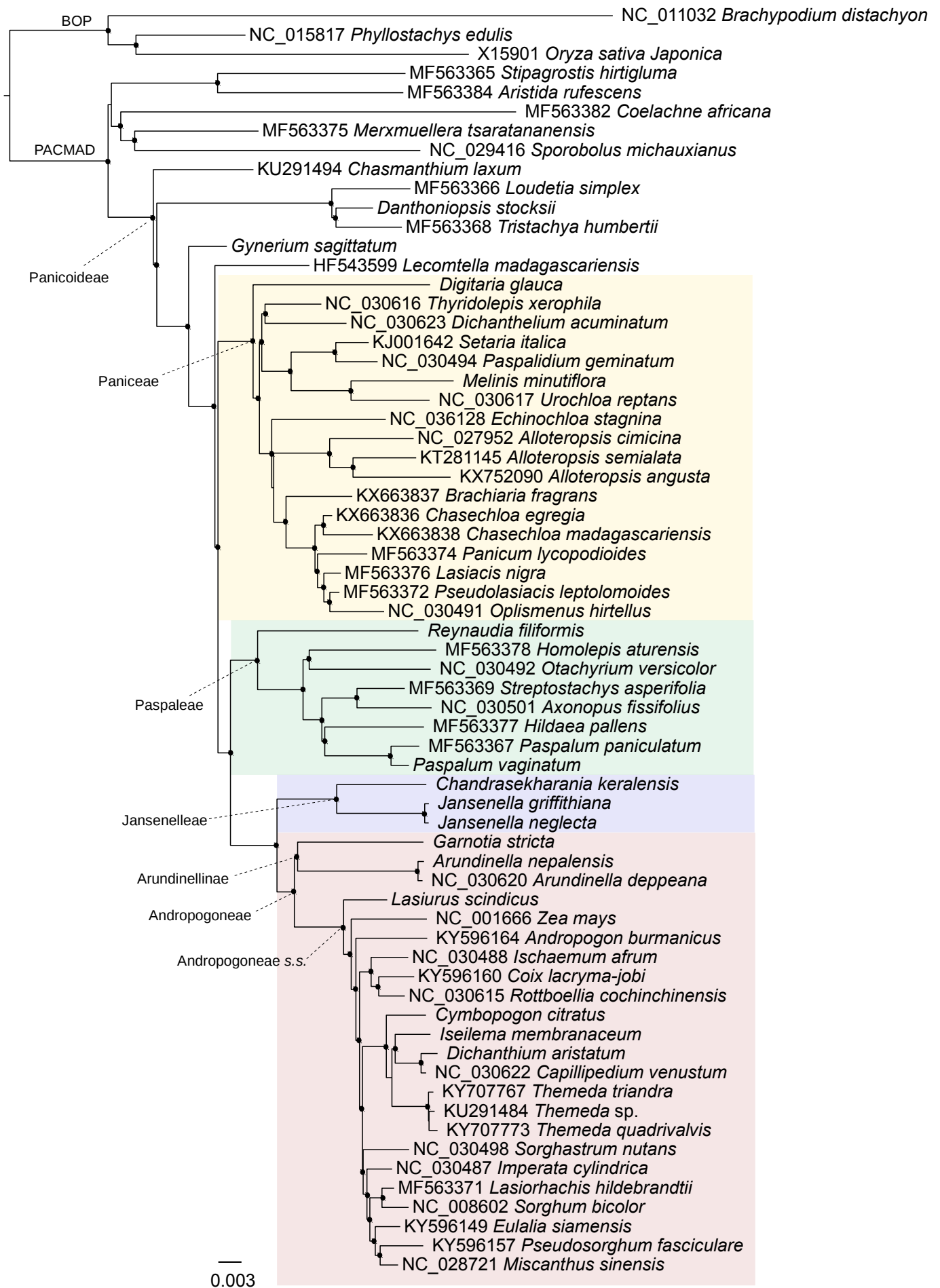
**Table S4.** Photosynthetic types and leaf anatomy data of the 66 grass species used in this study.

**Table S5.** Summary of multigene coalescent species trees inferred from nuclear data under different filtering stringencies.

**Table S6.** Summary of branch model comparisons performed using the species tree topology.

**Fig. S1**

**Fig. S2**



**Fig. S3**

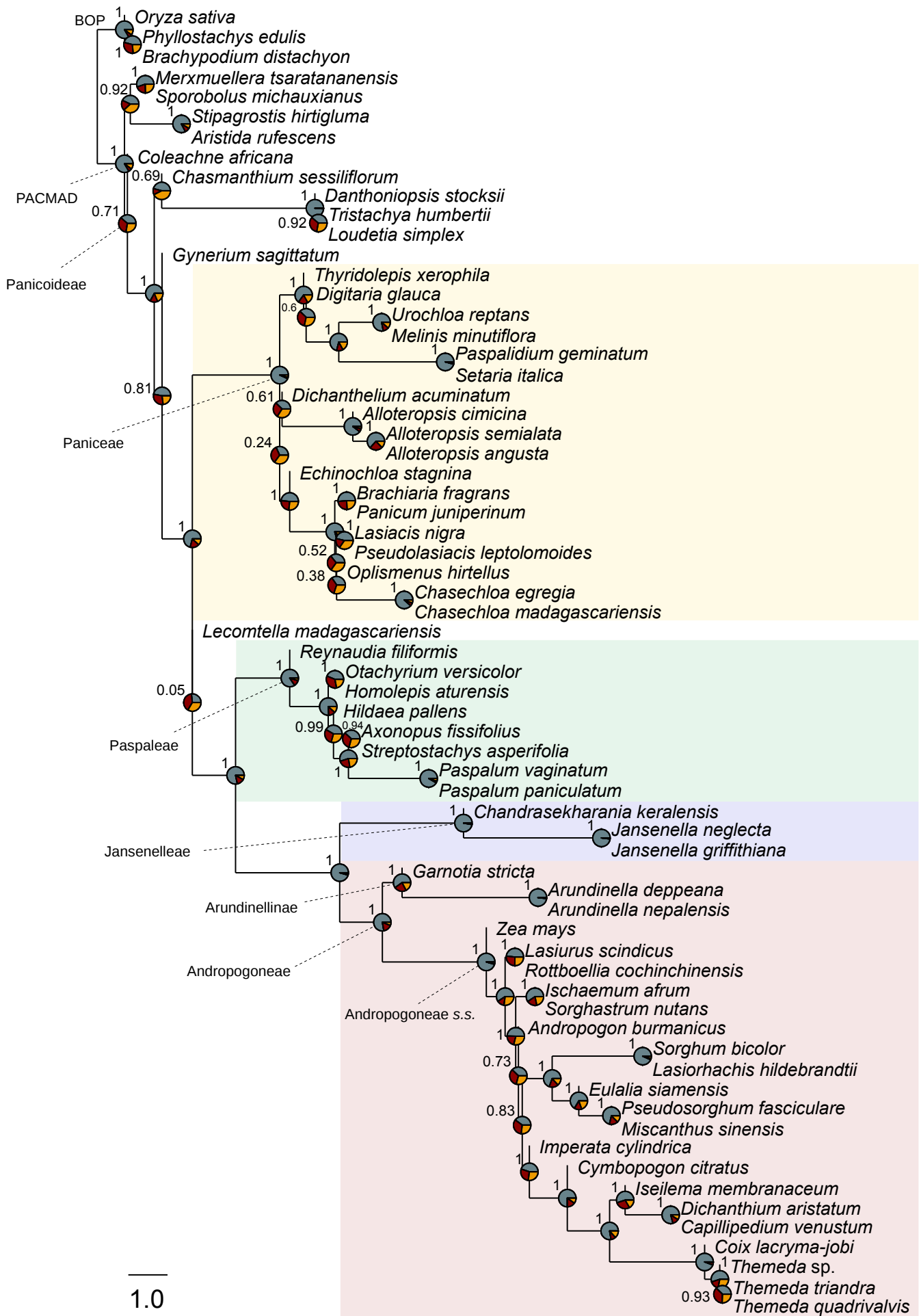
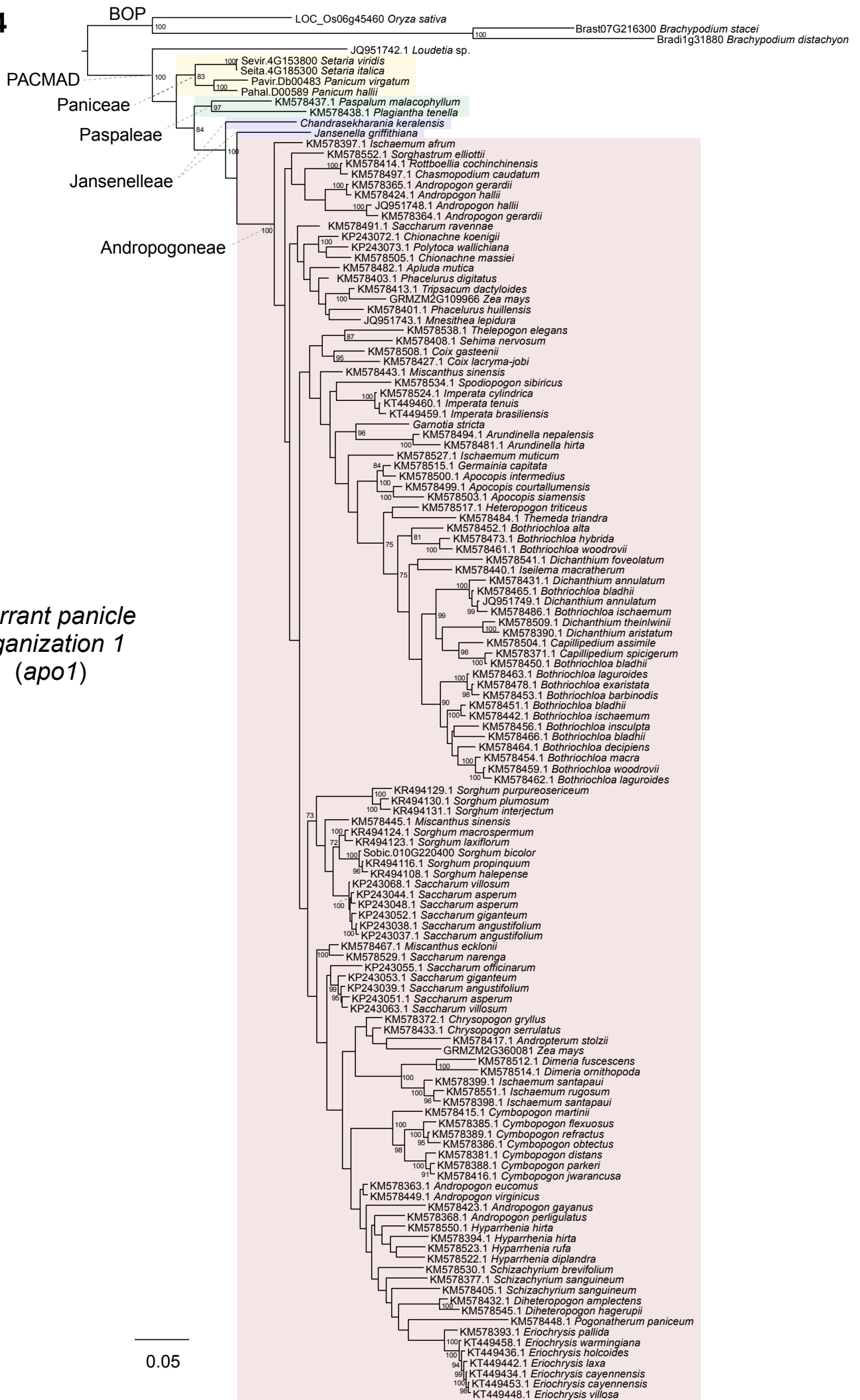


Fig. S4

a)

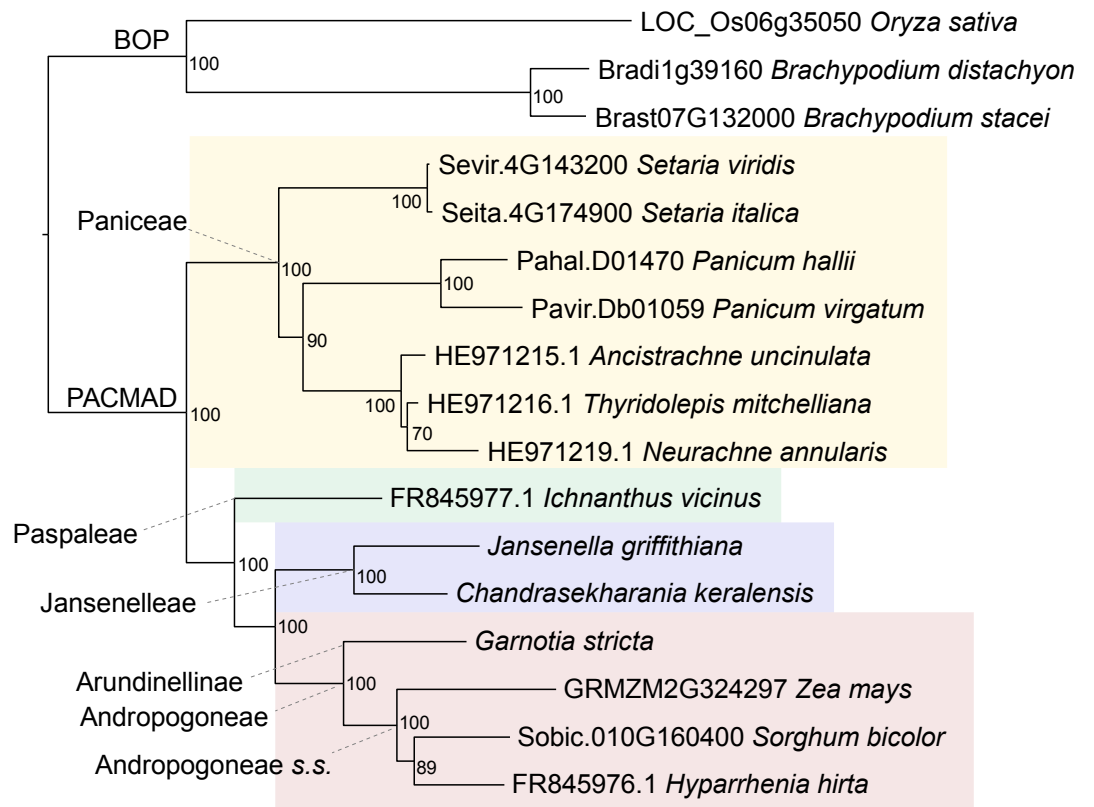


aberrant panicle organization 1 (apo1)

0.05

b)

Arogenate dehydrogenase  
(*arodeh*)



c)

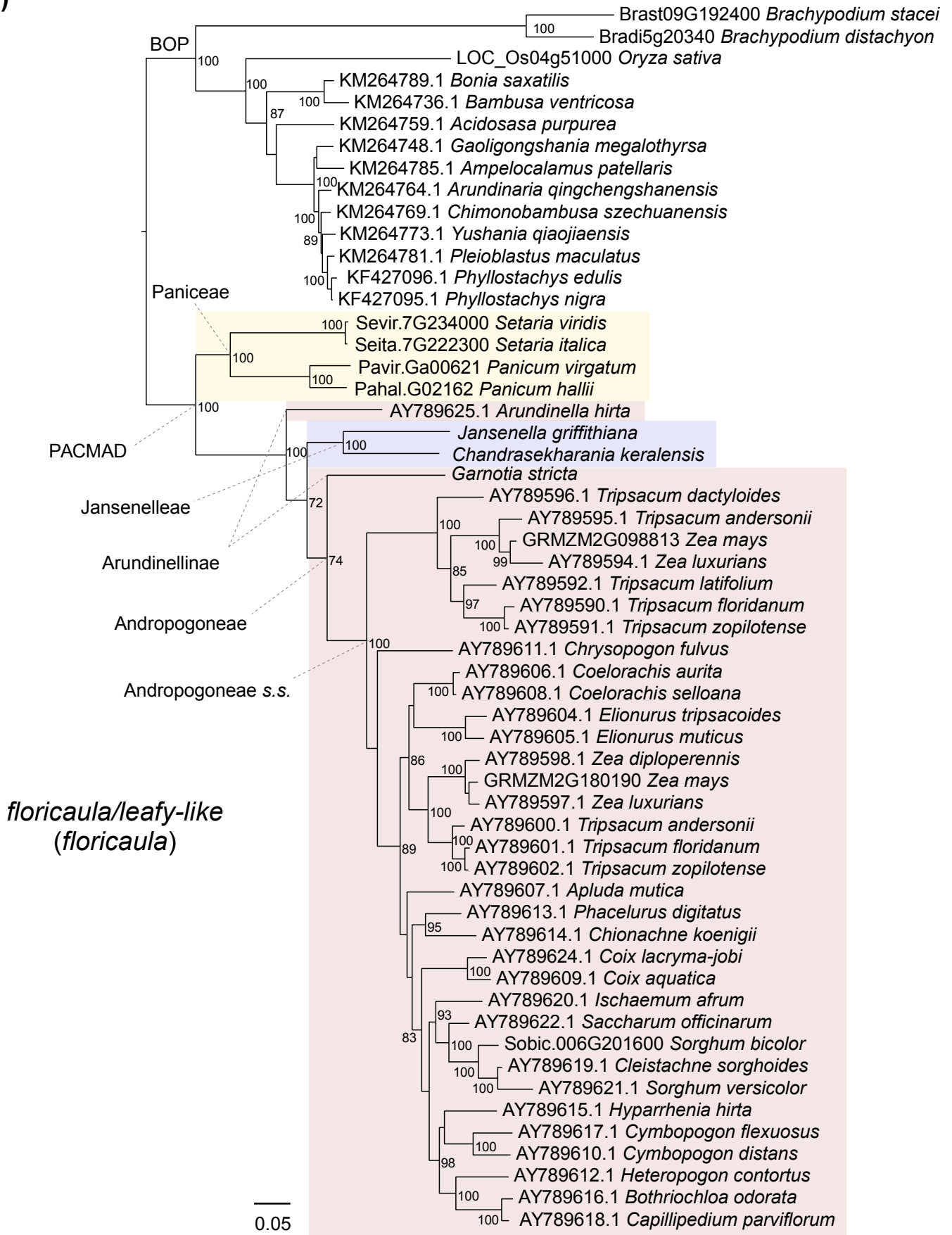


DELLA protein-encoding gene *dwarf 8* (*dwarf8*)

0.05

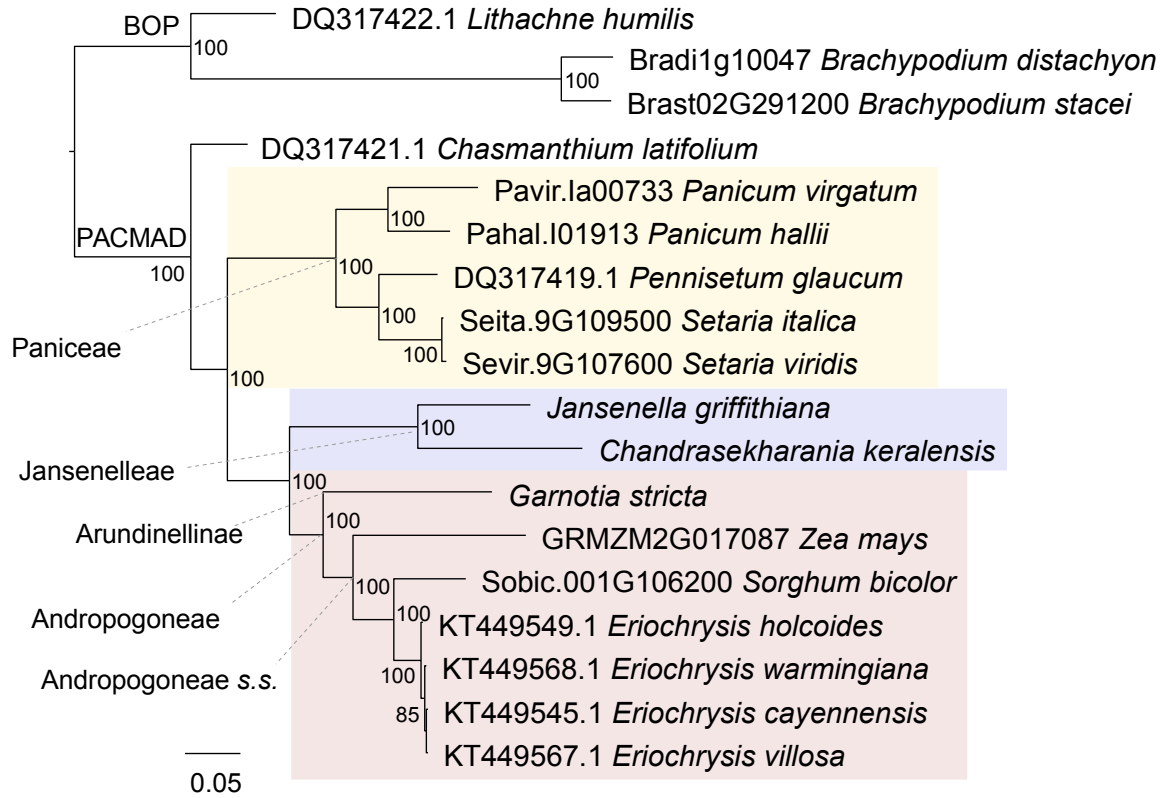


d)



e)

*knotted 1*  
(*kn1*)



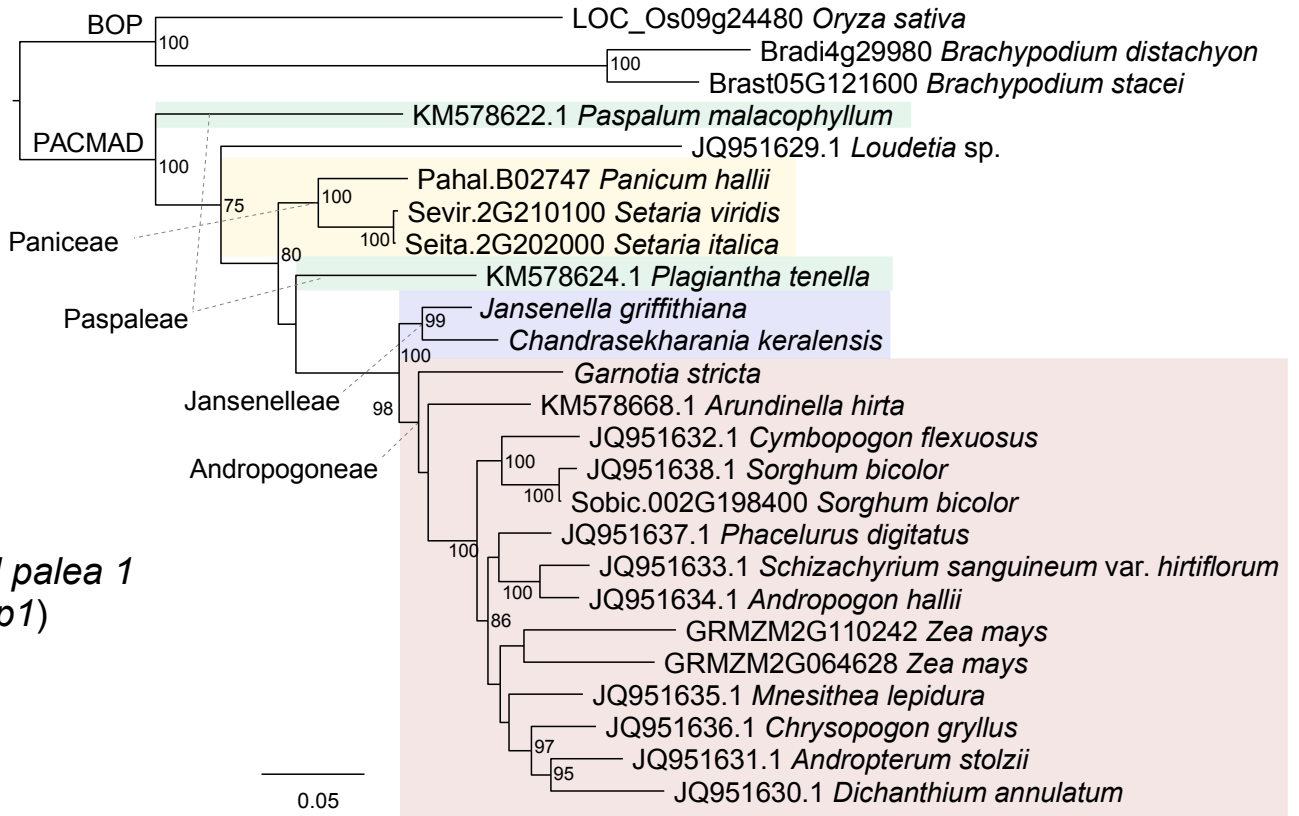
f)



Phytochrome B  
(*phyB*)

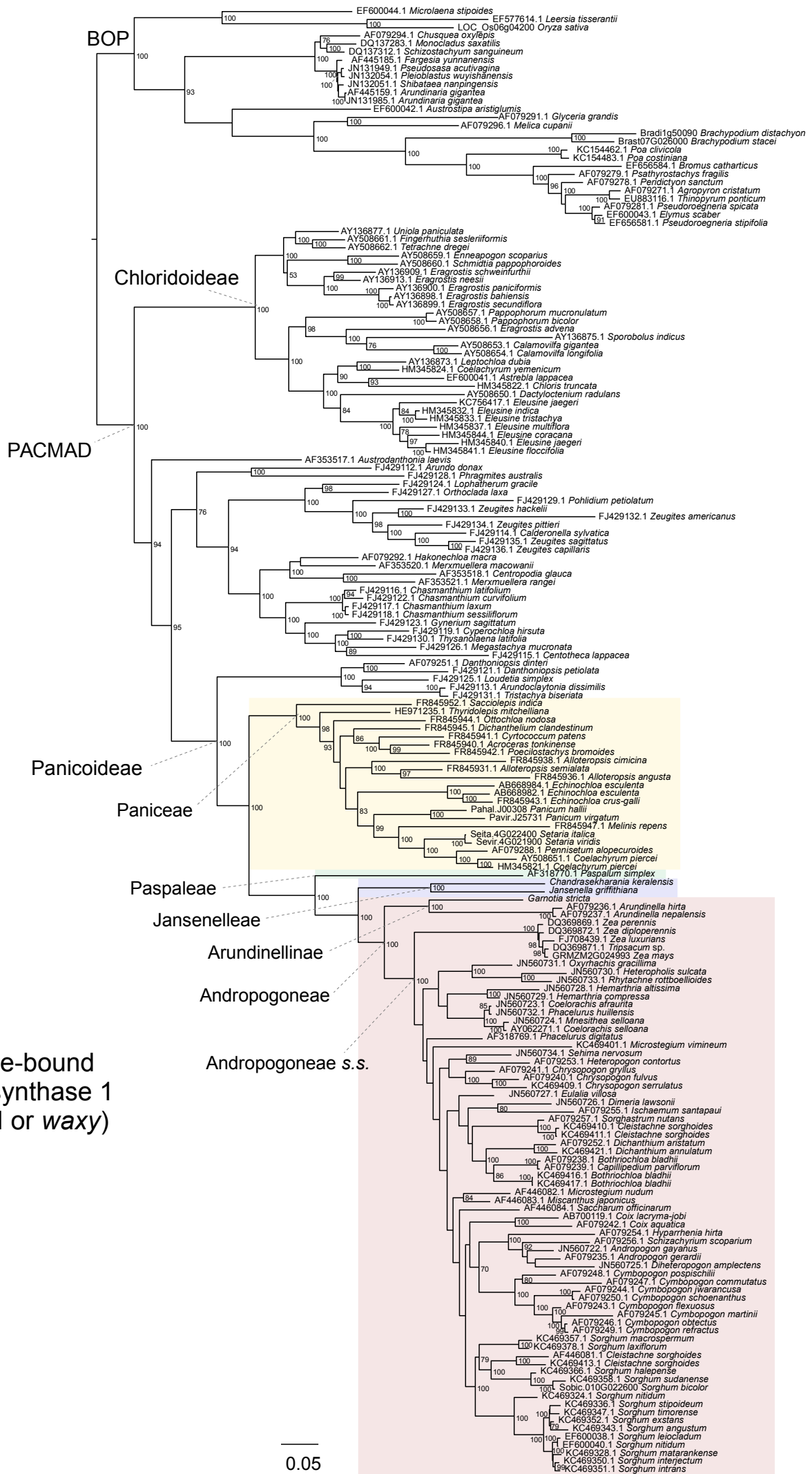
0.05

g)



retarded palea 1  
(rep1)

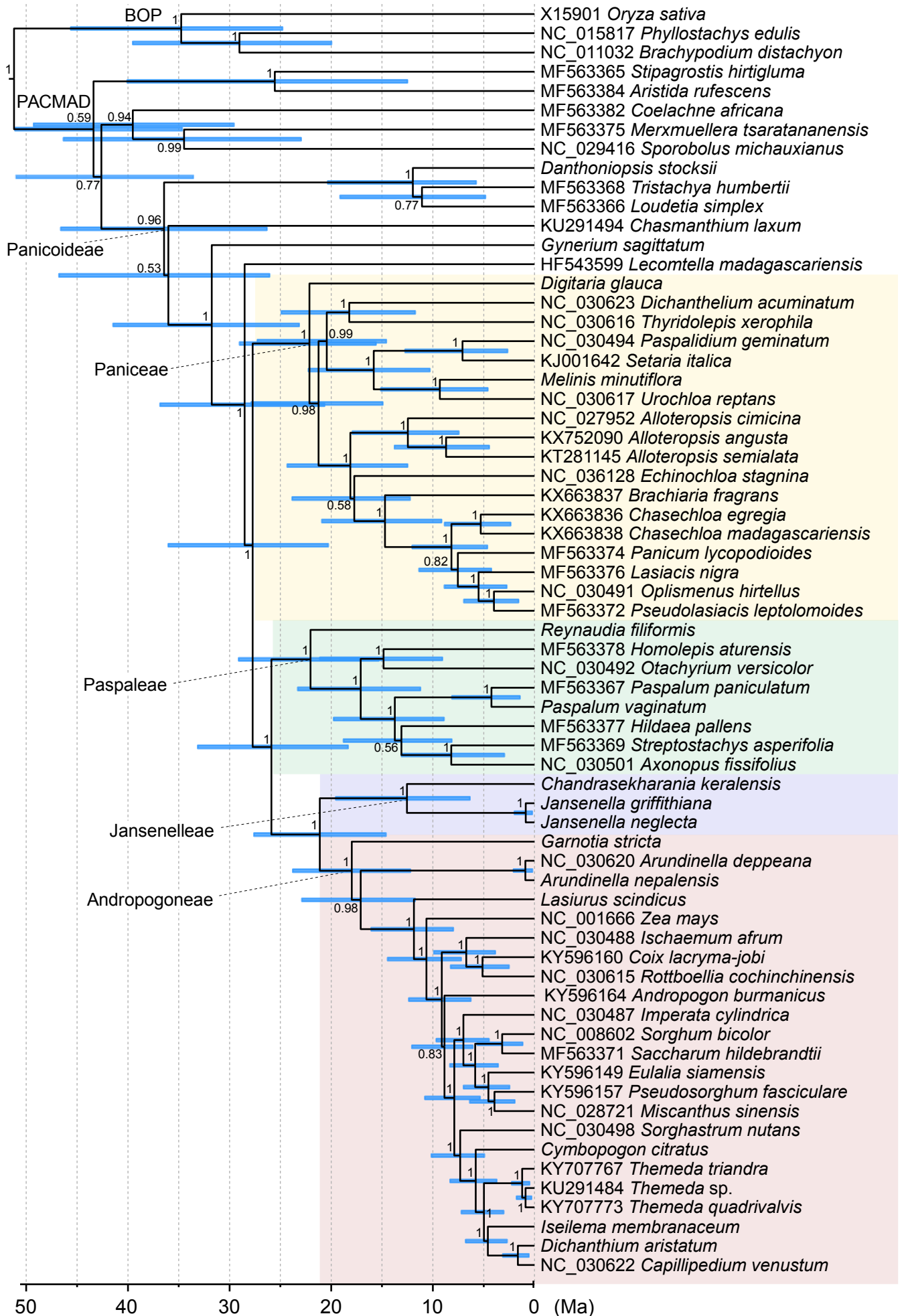
h)



Granule-bound starch synthase 1 (GBSSI or waxy)

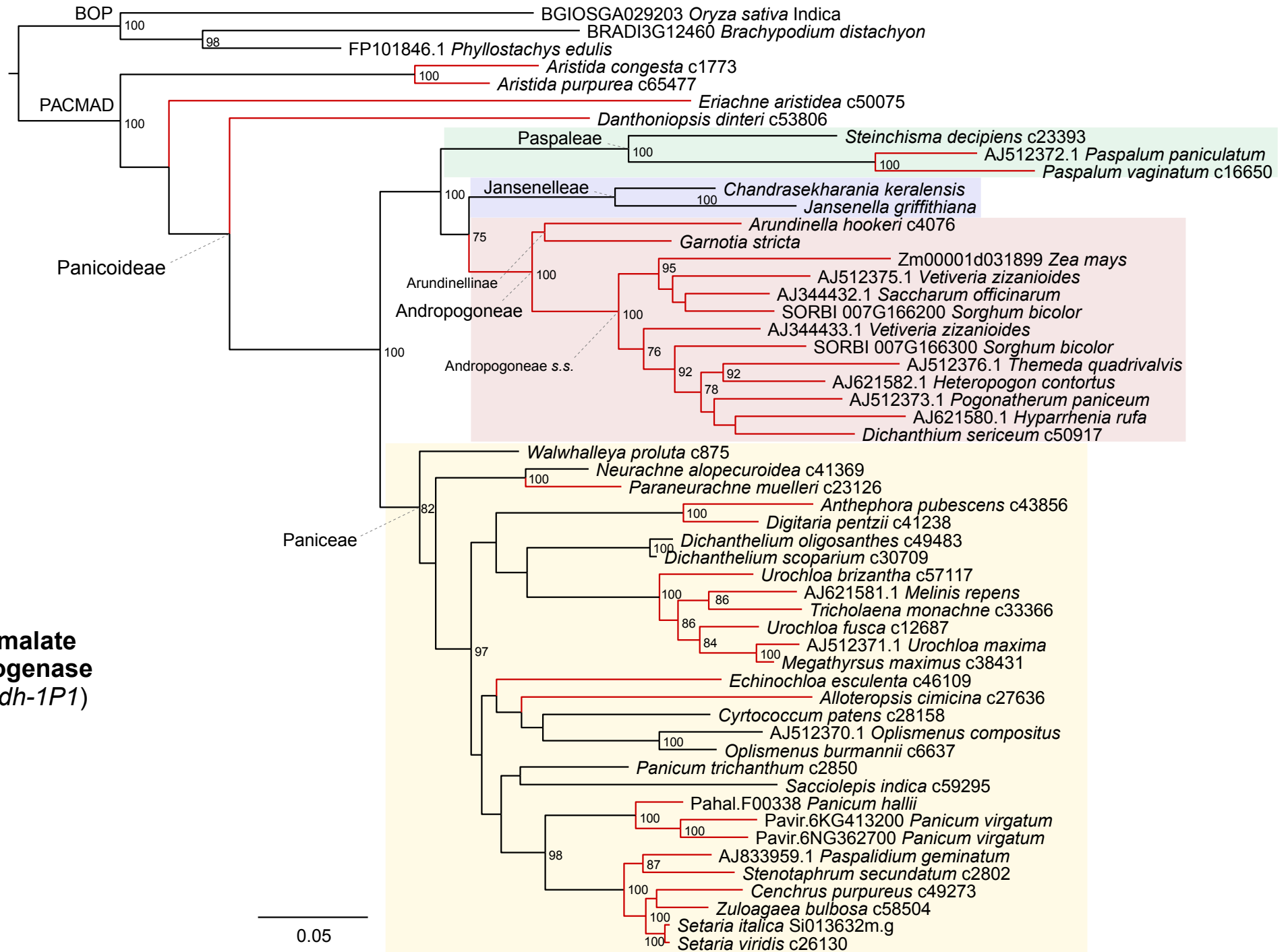
0.05

**Fig. S5**



**Fig. S6**

a)



b)





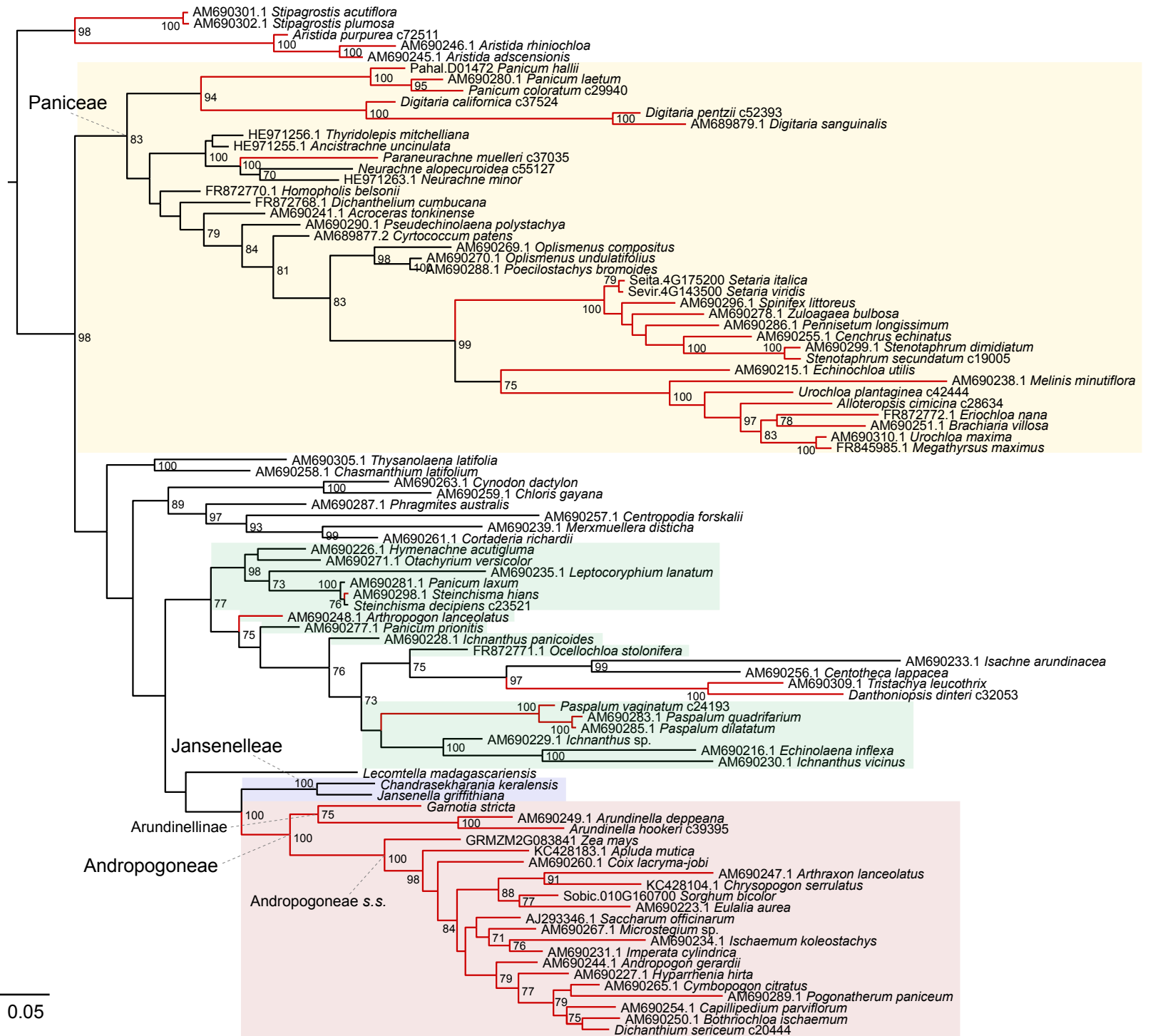
c)



Phosphoenolpyruvate  
carboxykinase  
(pck-1P1)

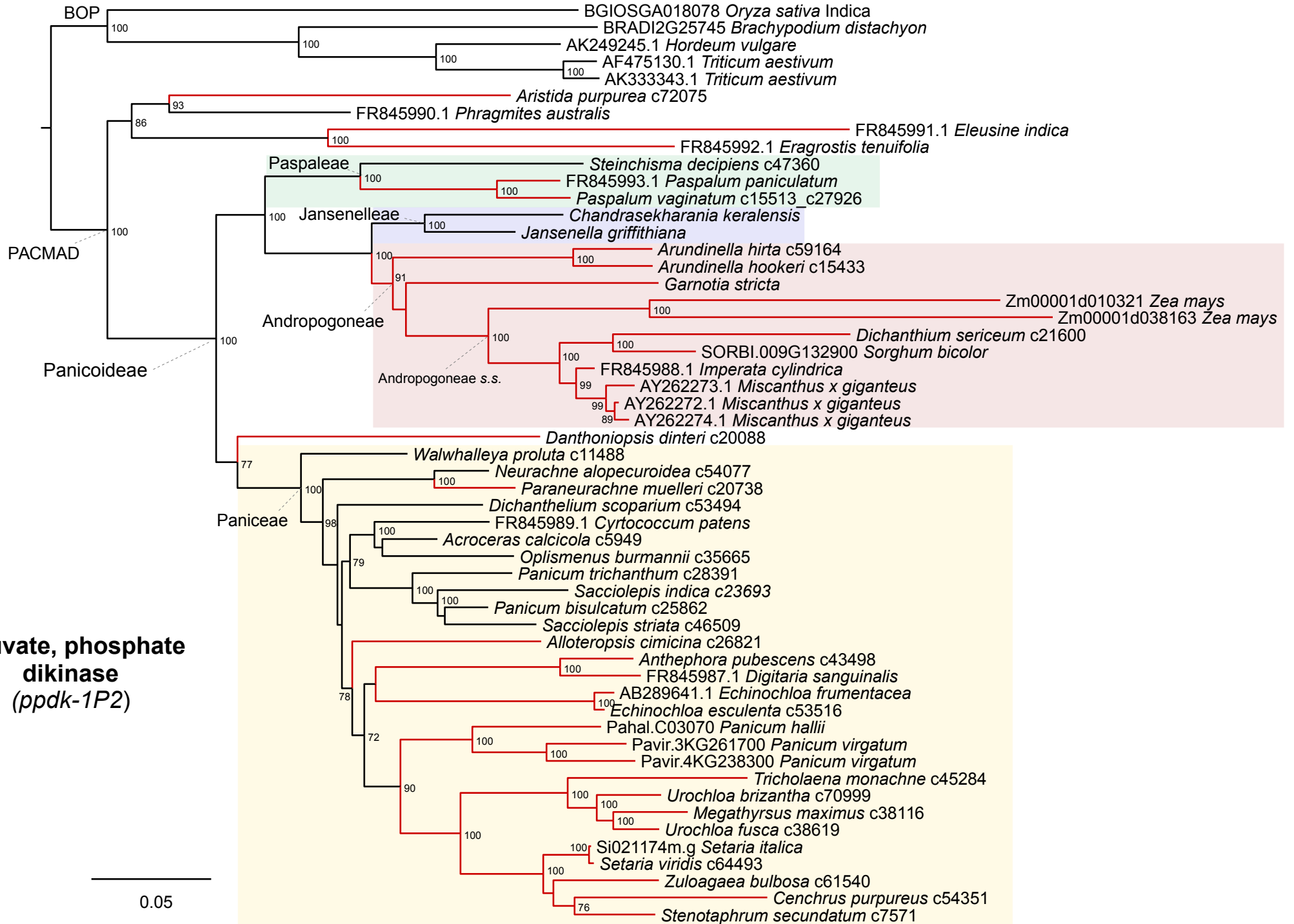
0.05

d)

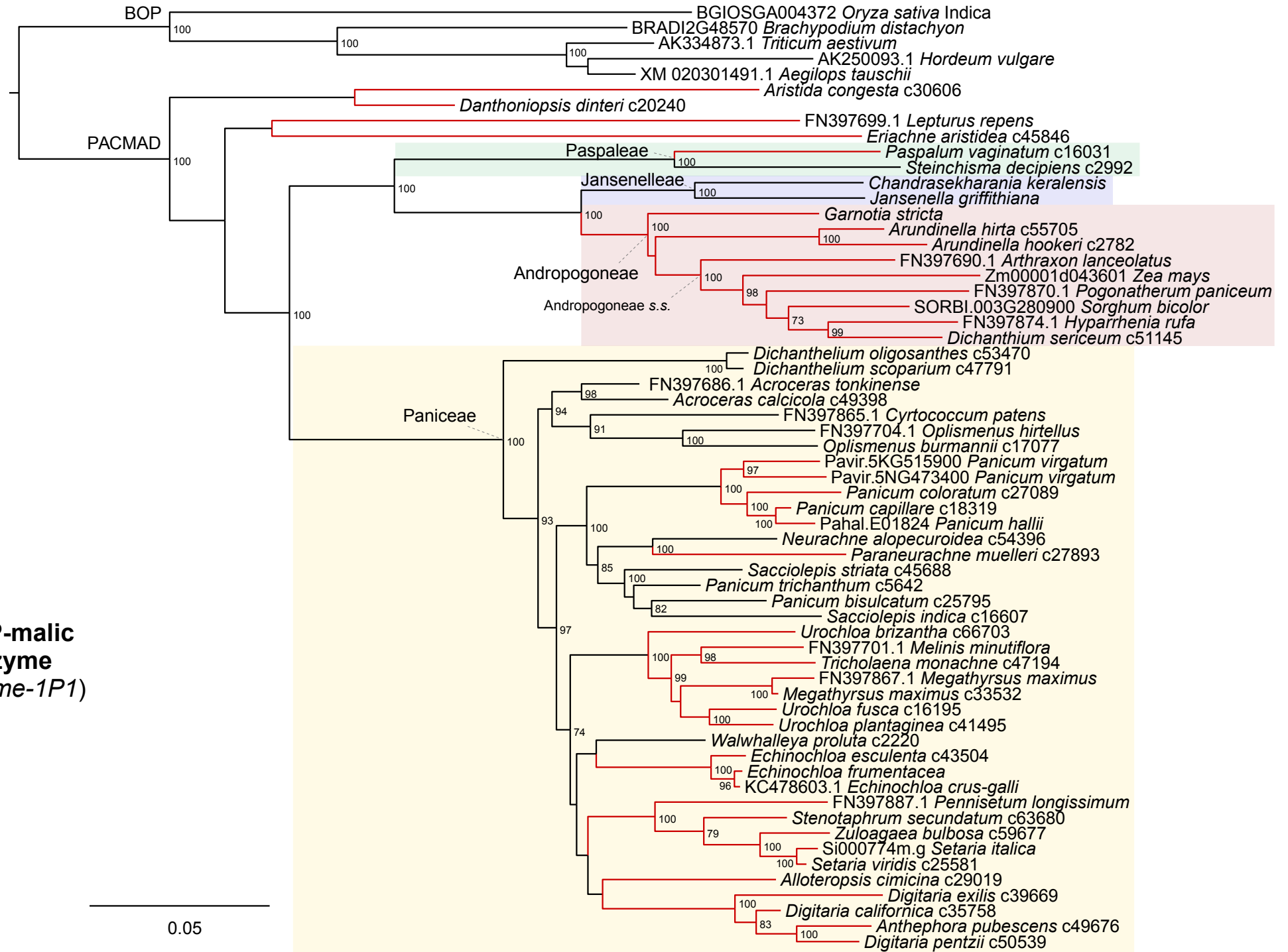


Phosphoenolpyruvate  
carboxylase  
(ppc-1P3)

e)

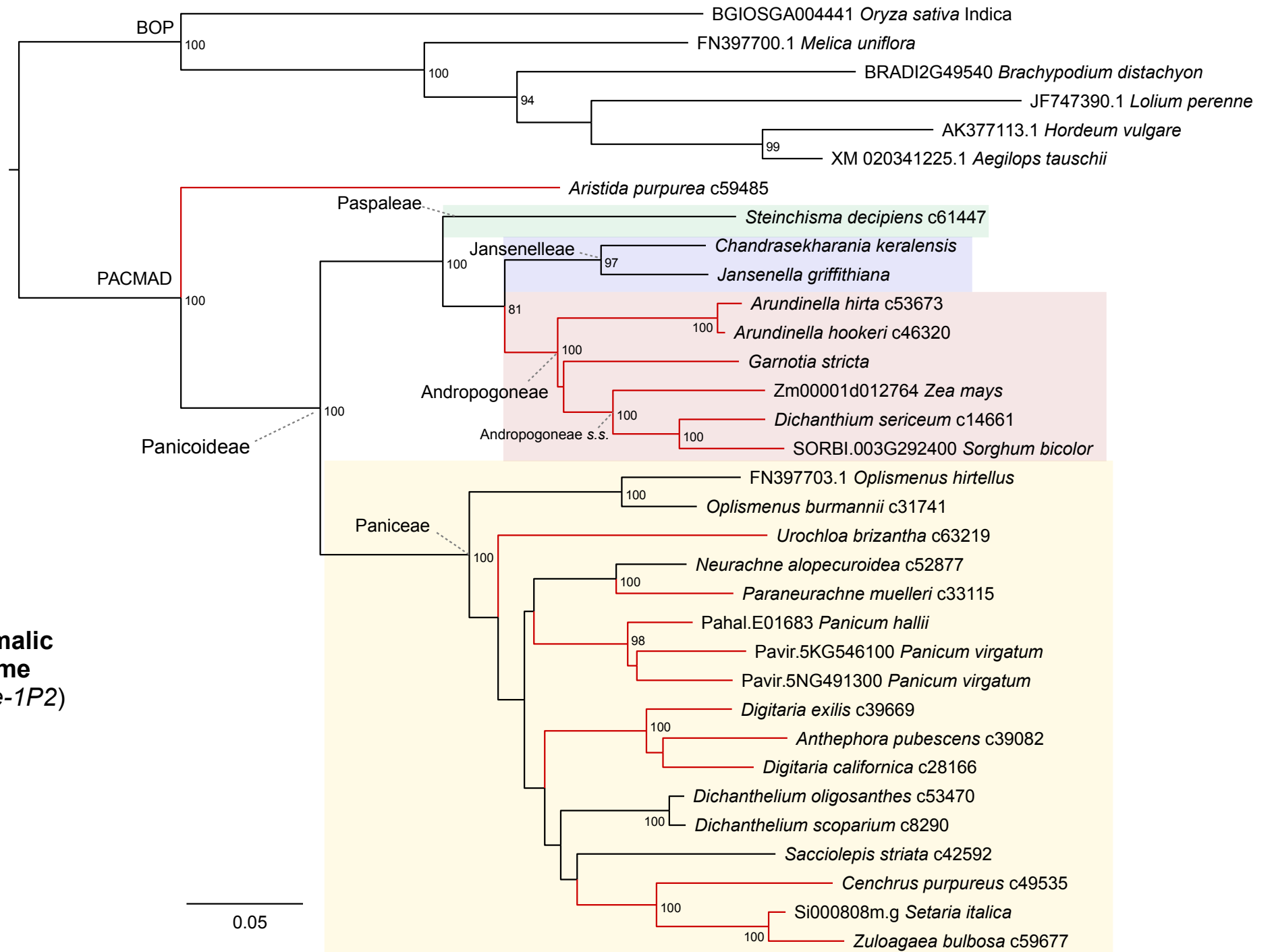


f)

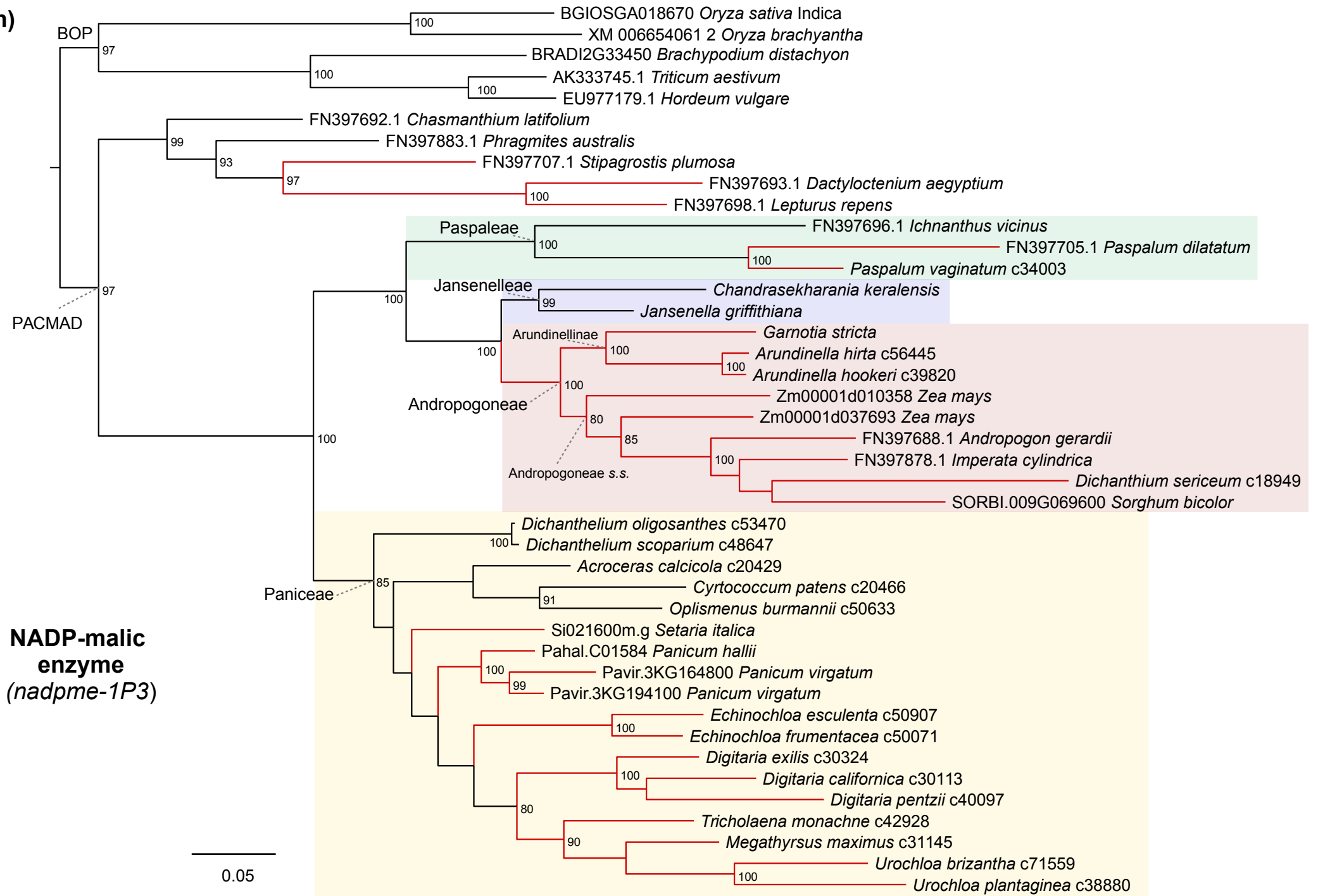


g)

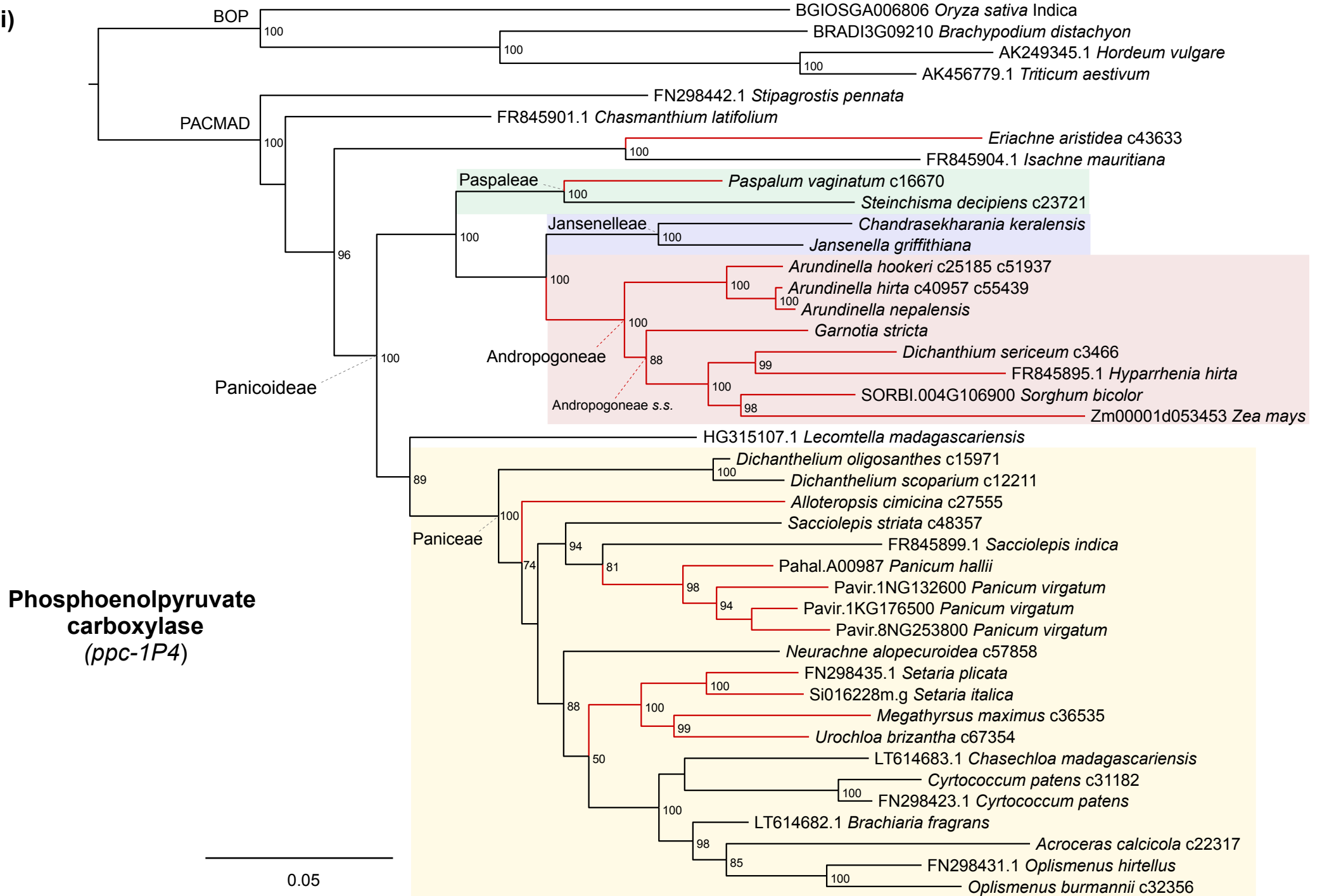
**NADP-malic  
enzyme  
(*nadpme-1P2*)**



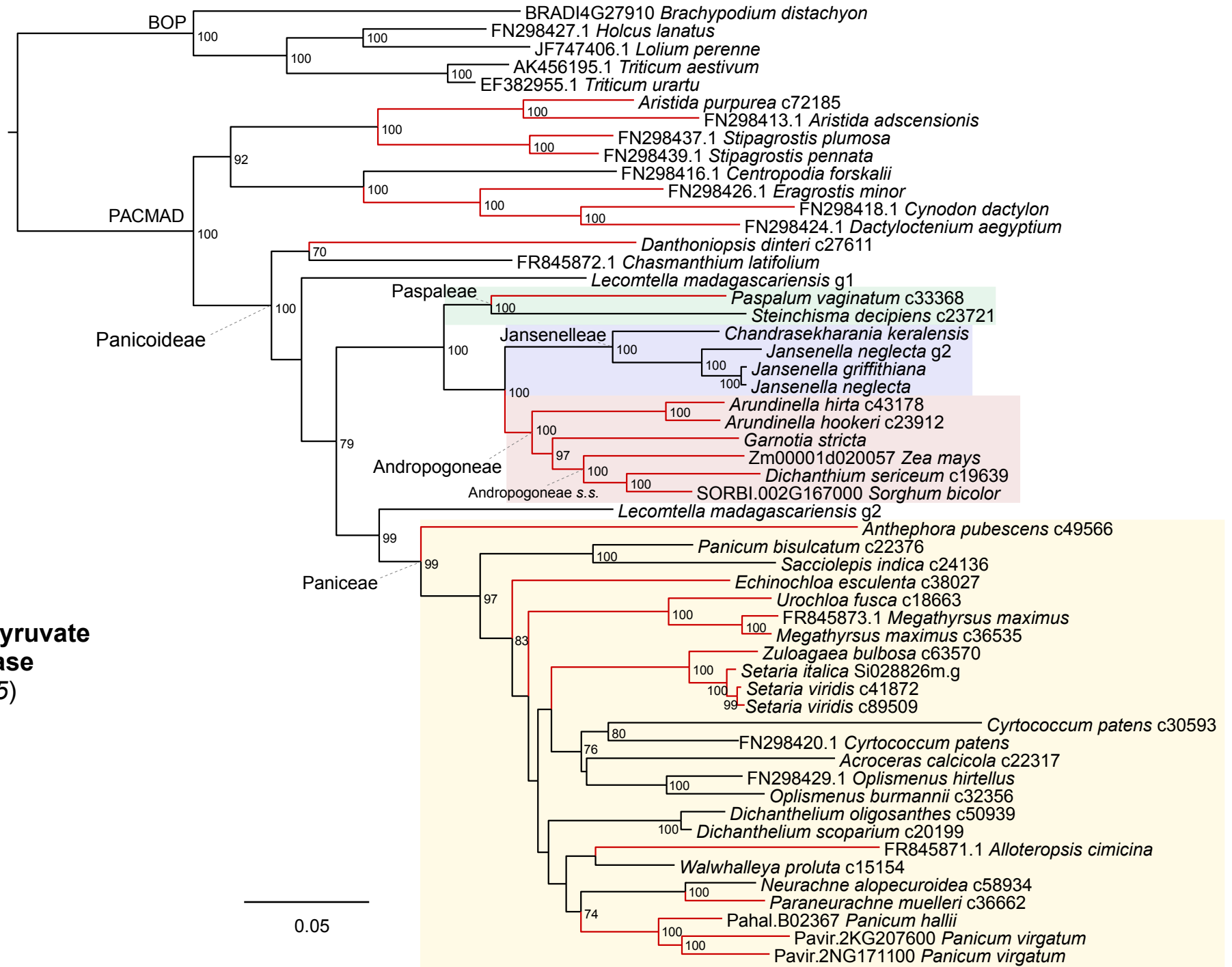
h)



i)

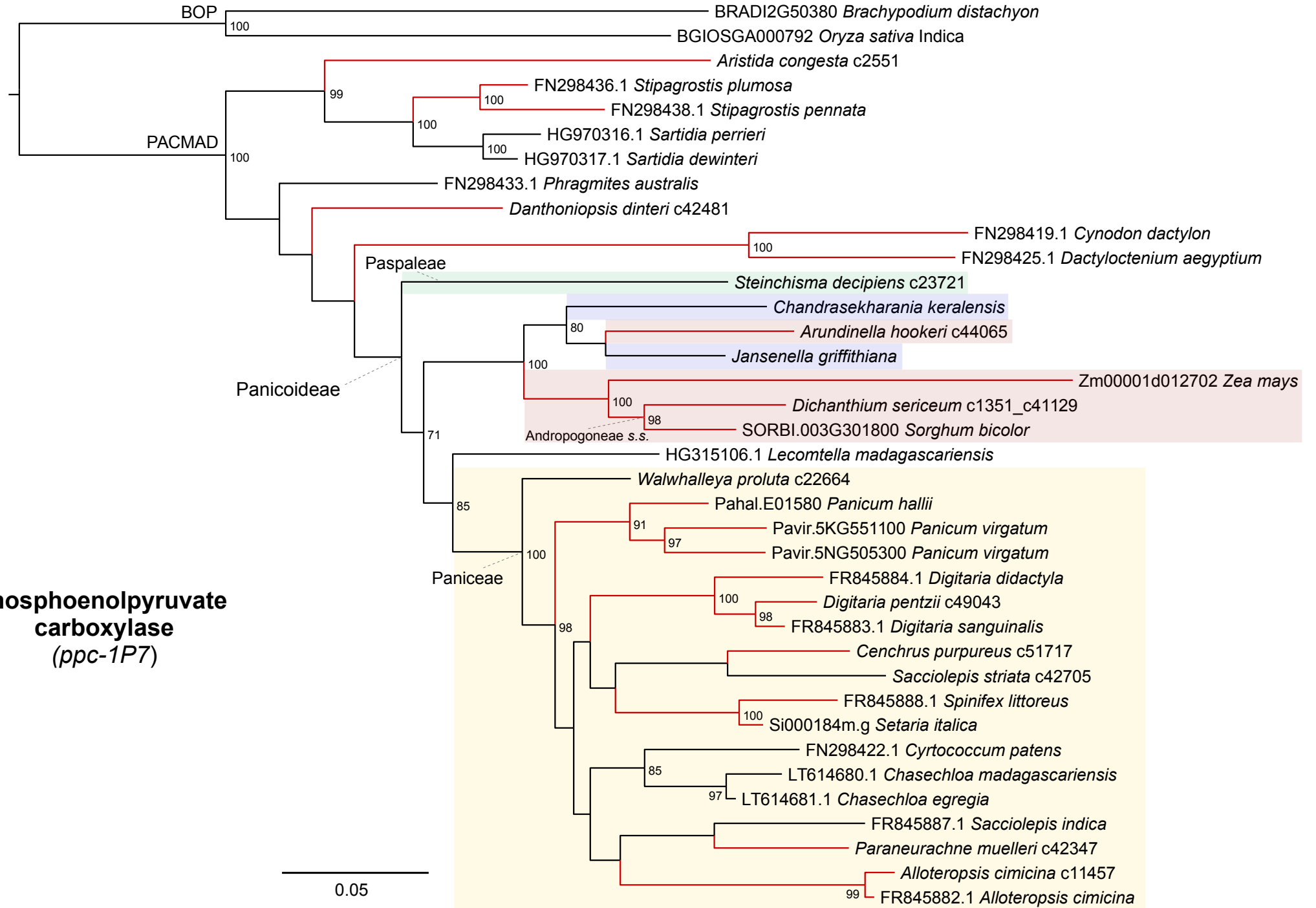


j)

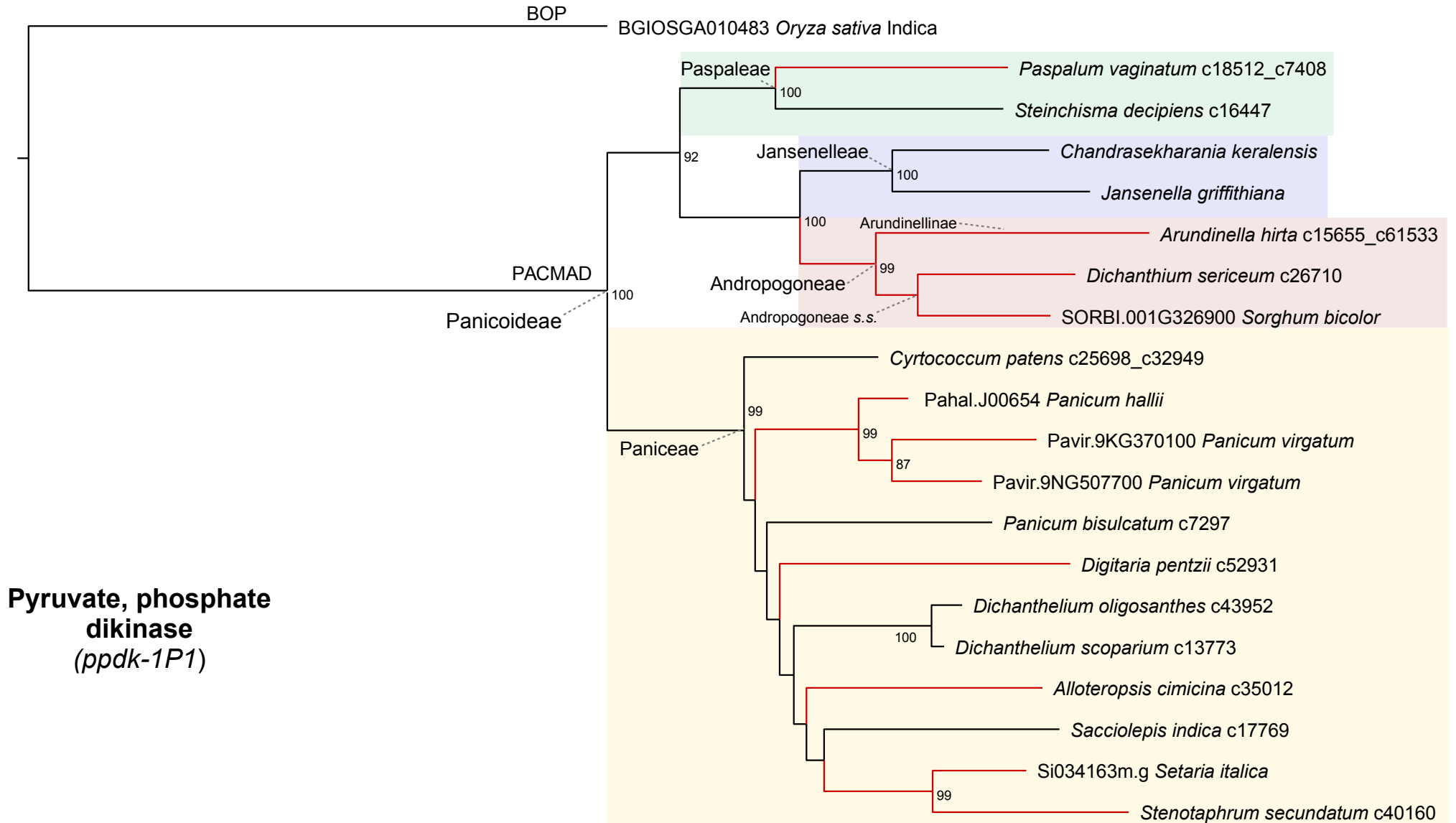




k)



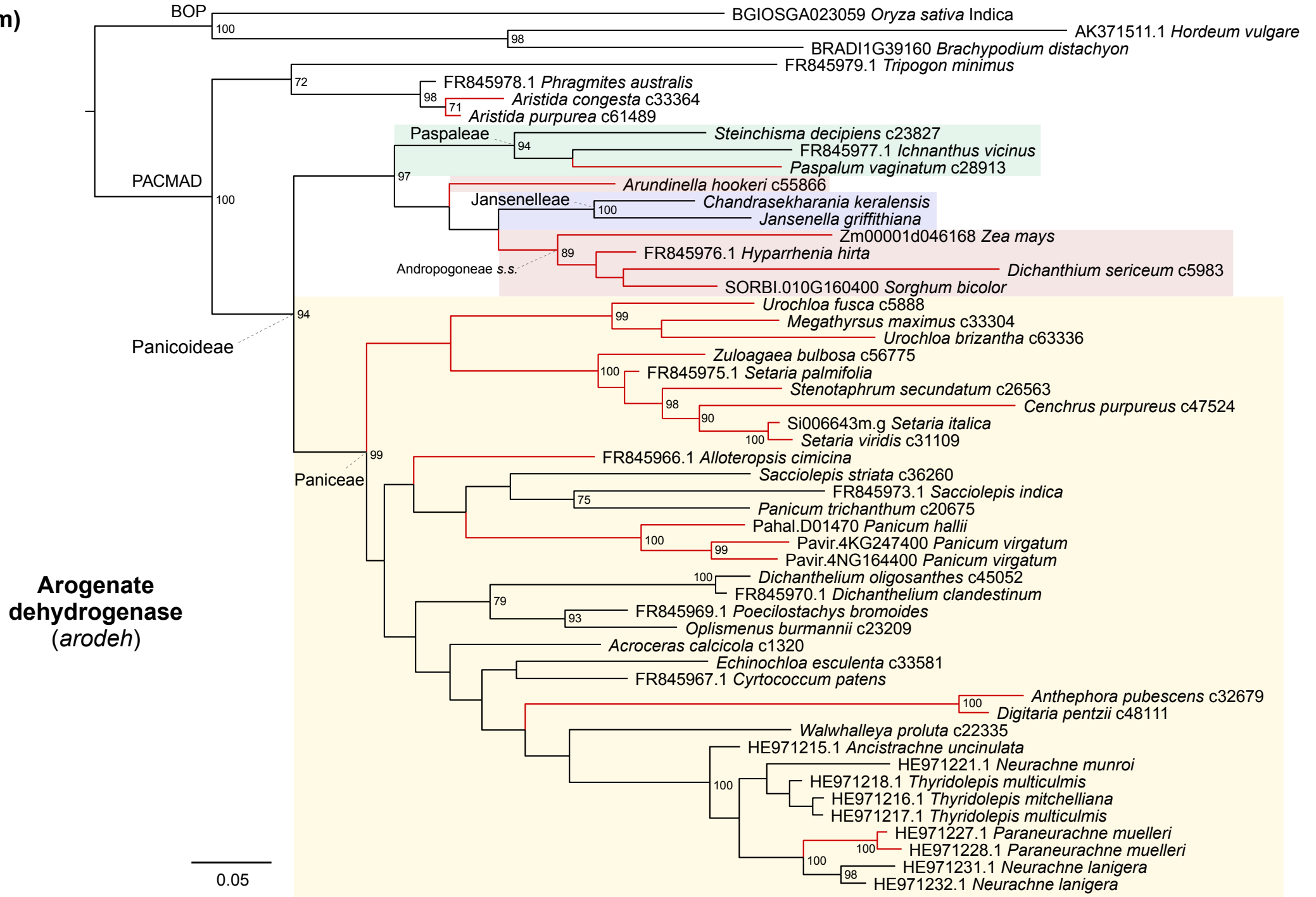
I)



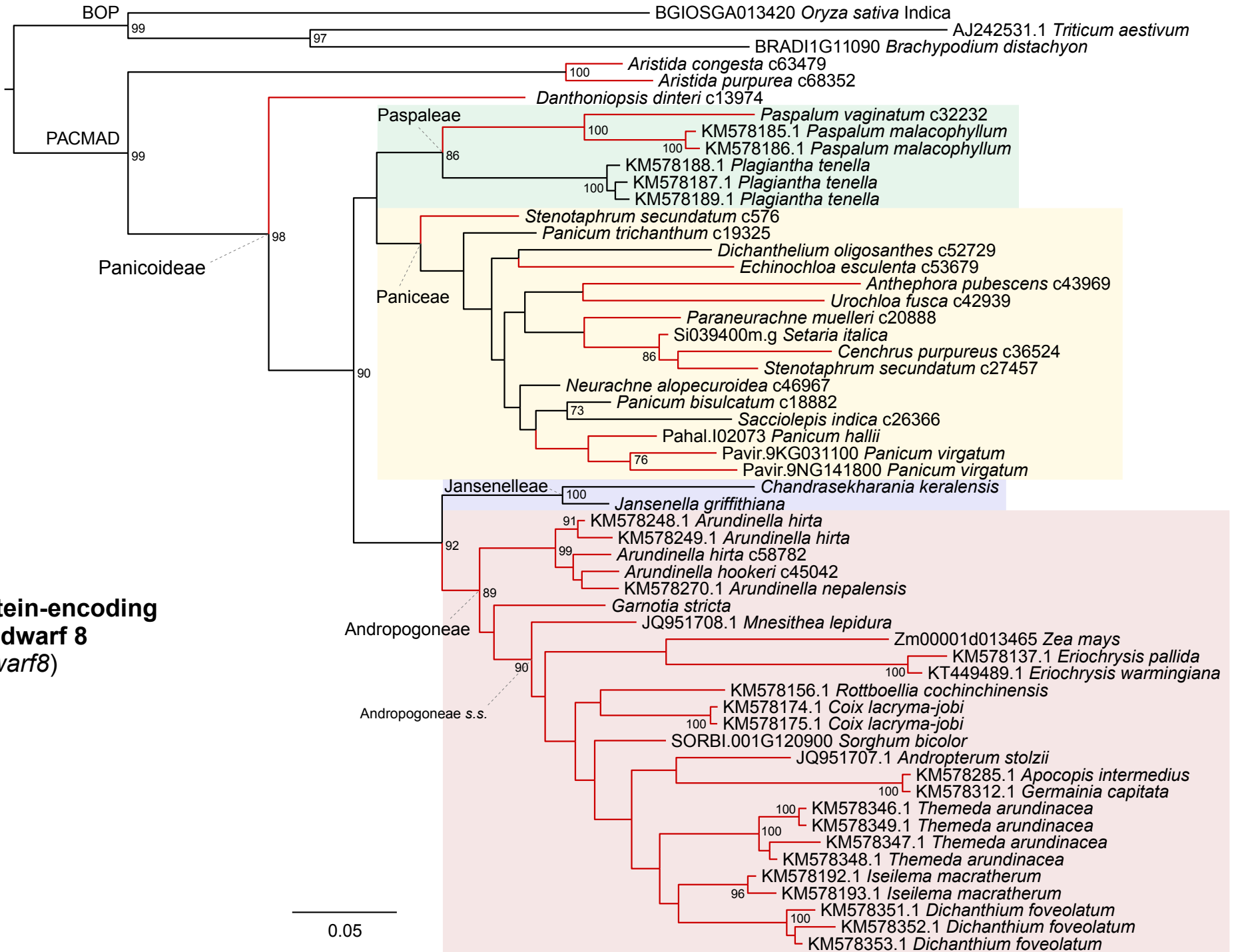
**Pyruvate, phosphate  
dikinase  
(ppdk-1P1)**

0.05

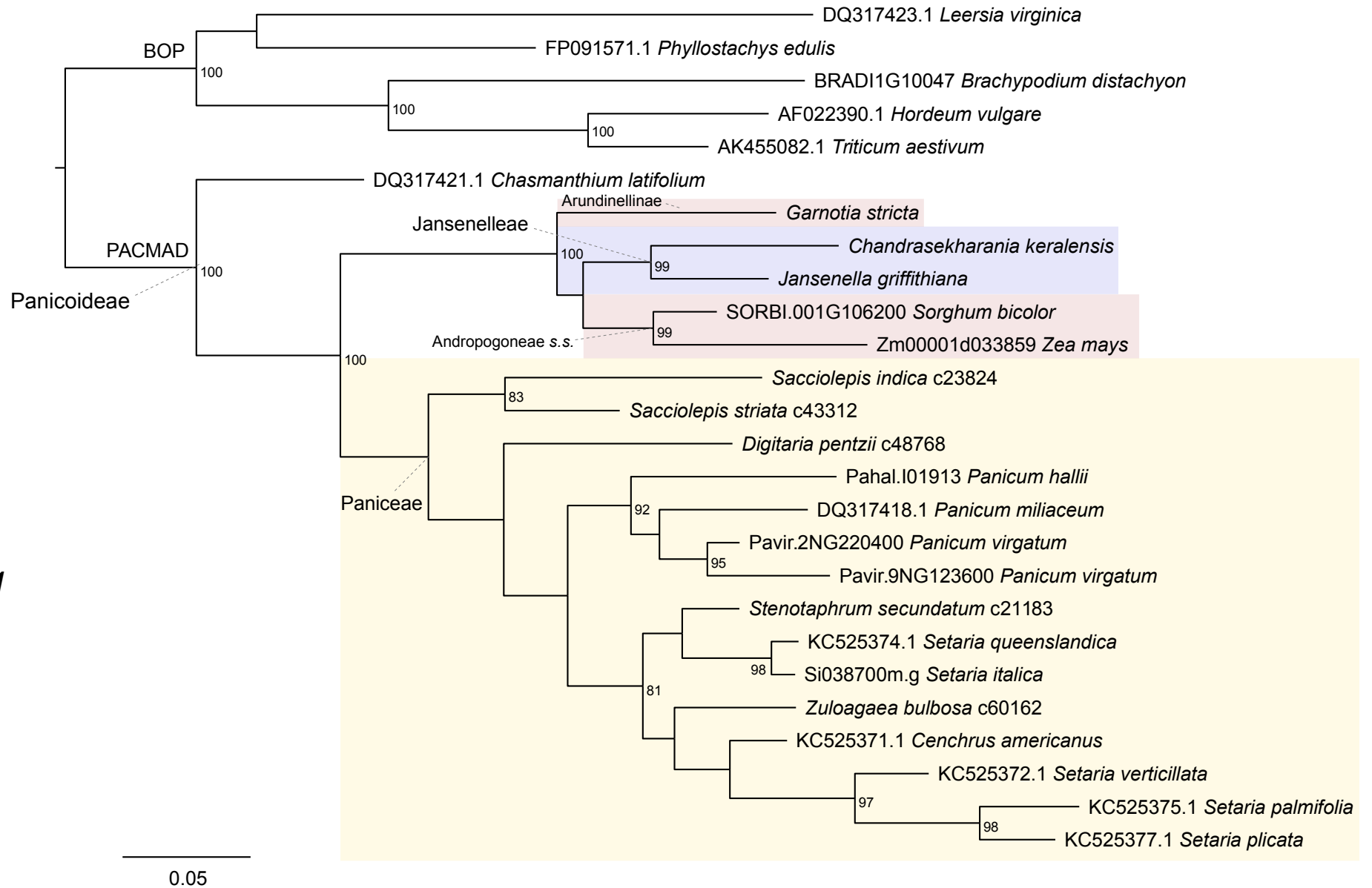
m)



n)



o)



p)

phytochrome B  
(*phyB*)



0.05

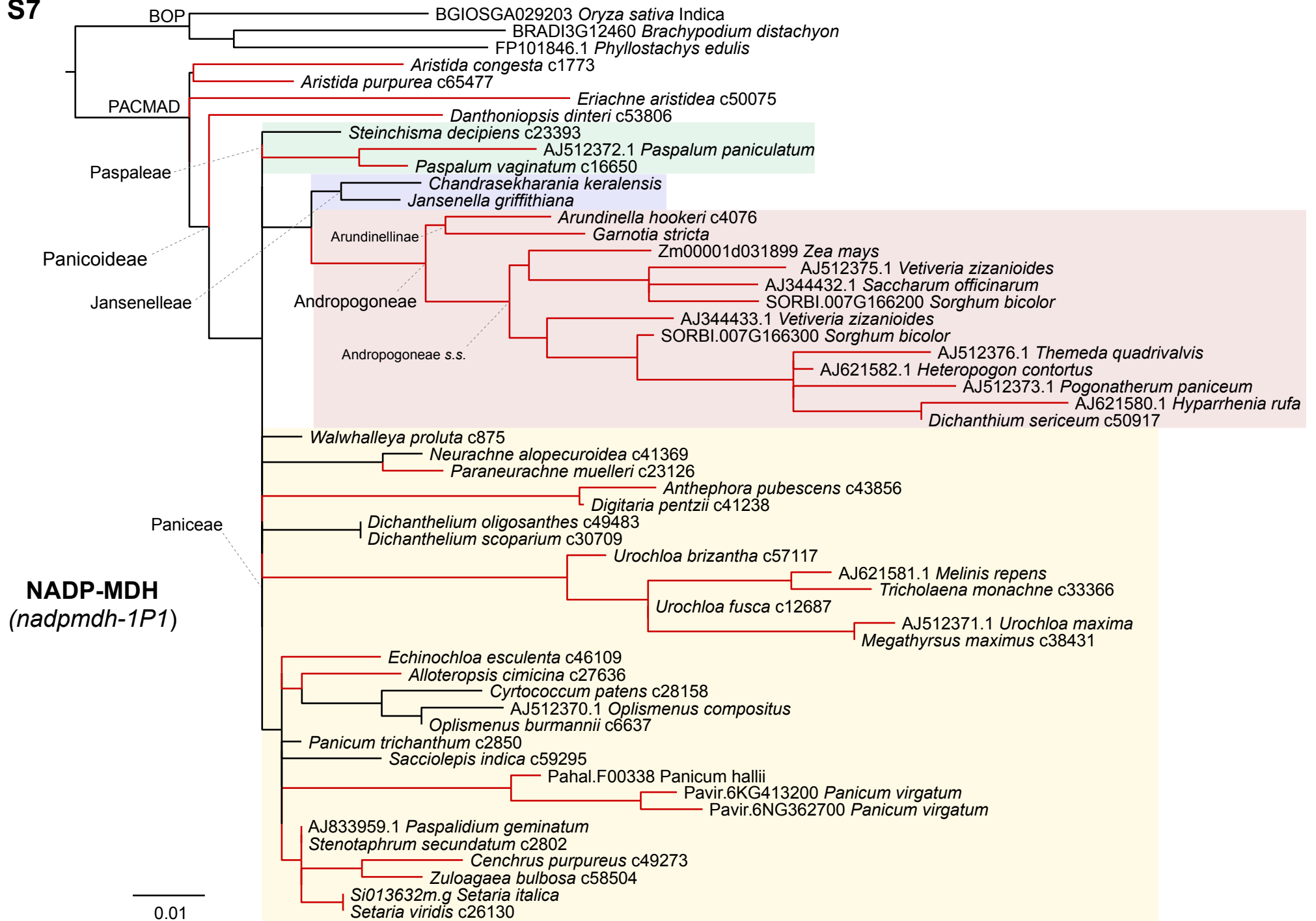
q)

**Granule-bound starch synthase 1 (waxy)**



Fig. S7

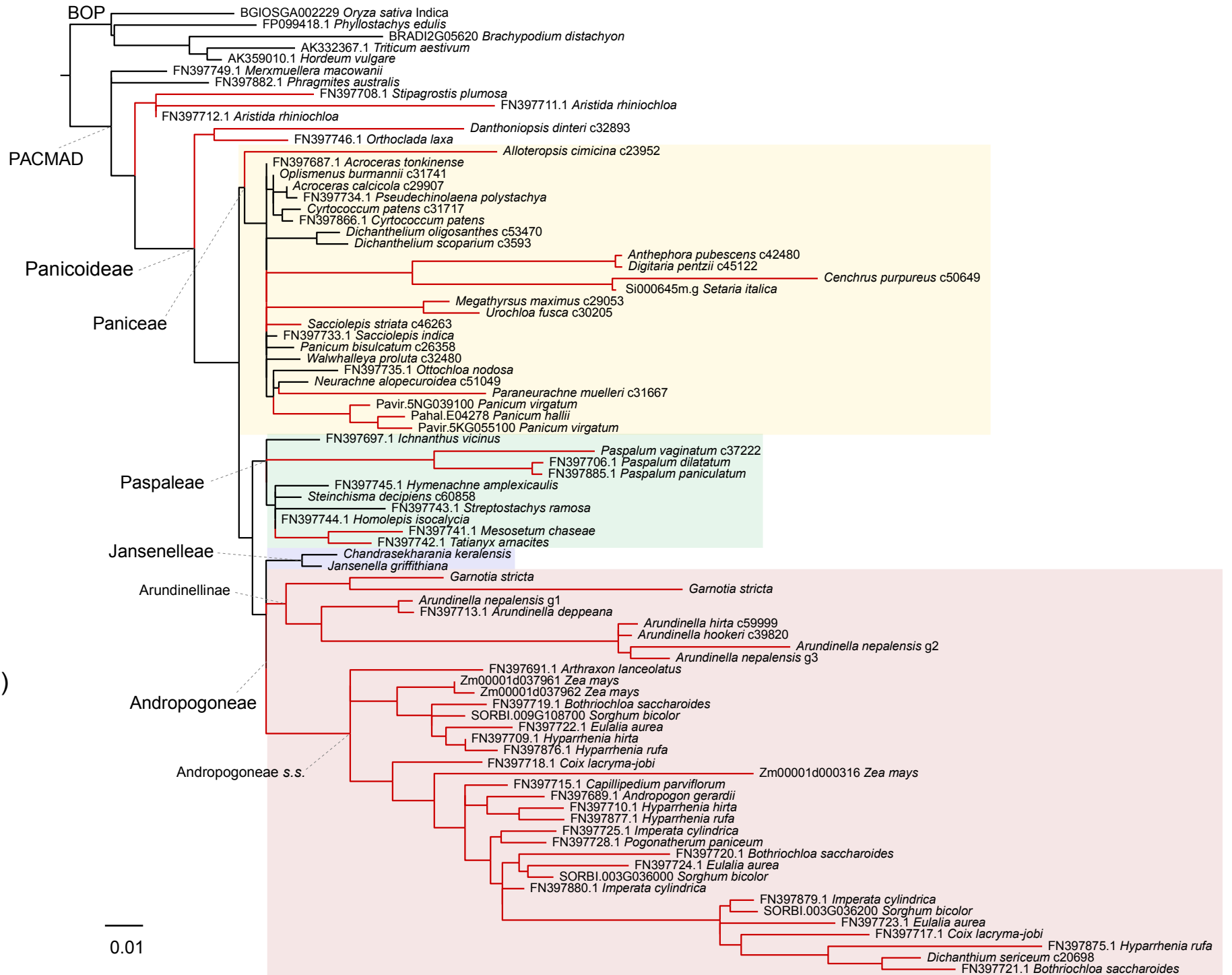
a)





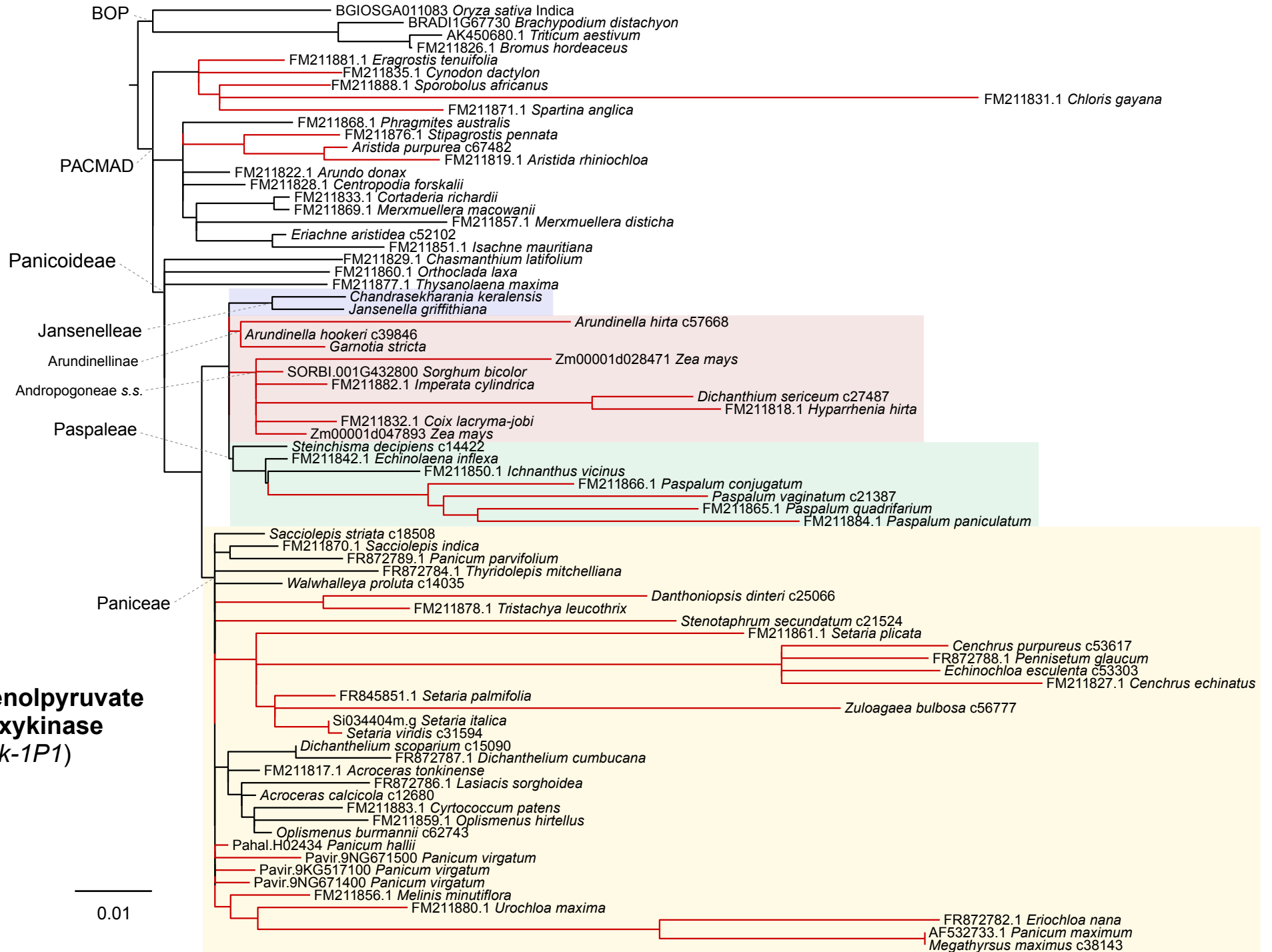
b)

**NADP-malic enzyme  
(nadpme-1P4)**



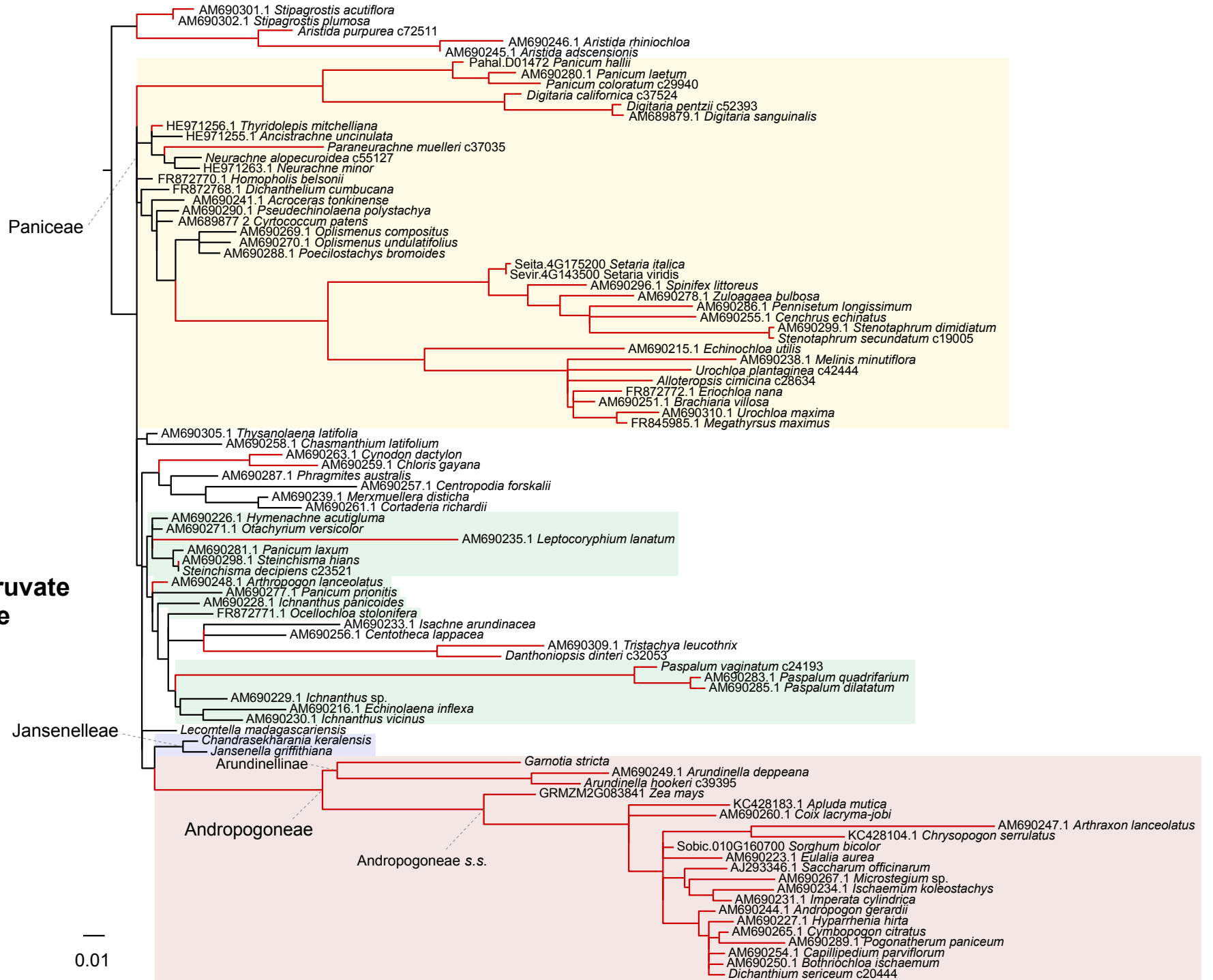
0.01

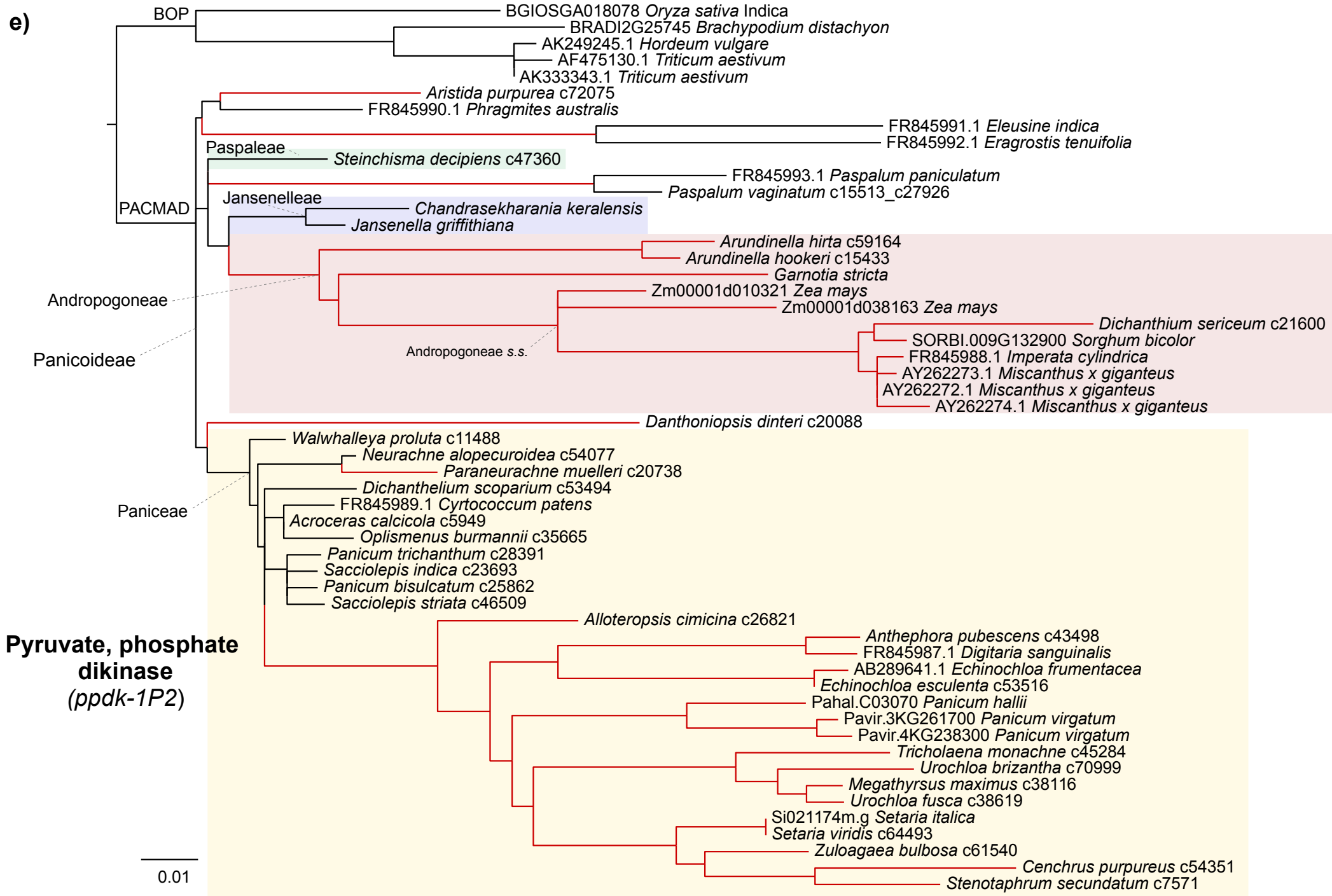
c)



d)

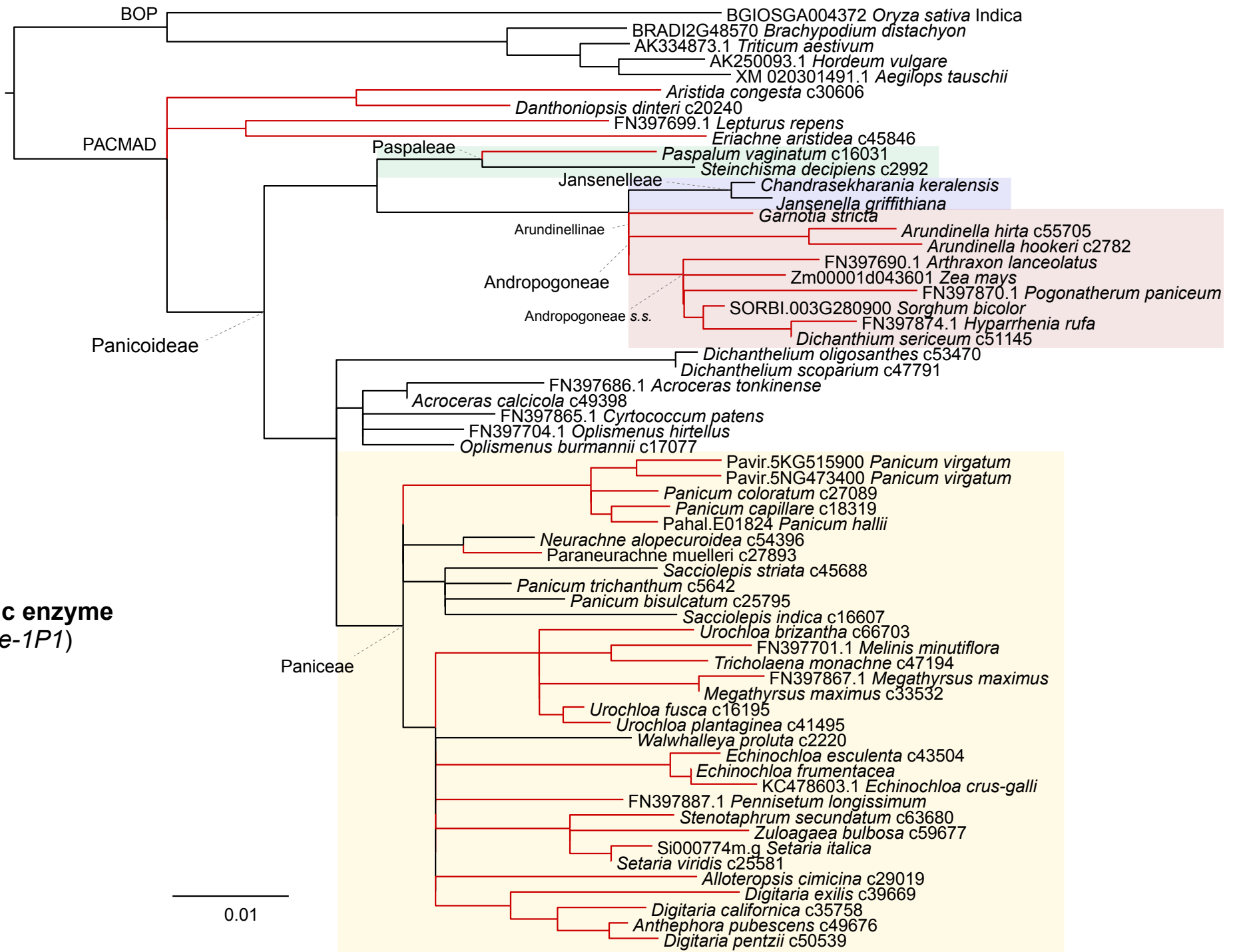
**Phosphoenolpyruvate  
carboxylase  
(ppc-1P3)**





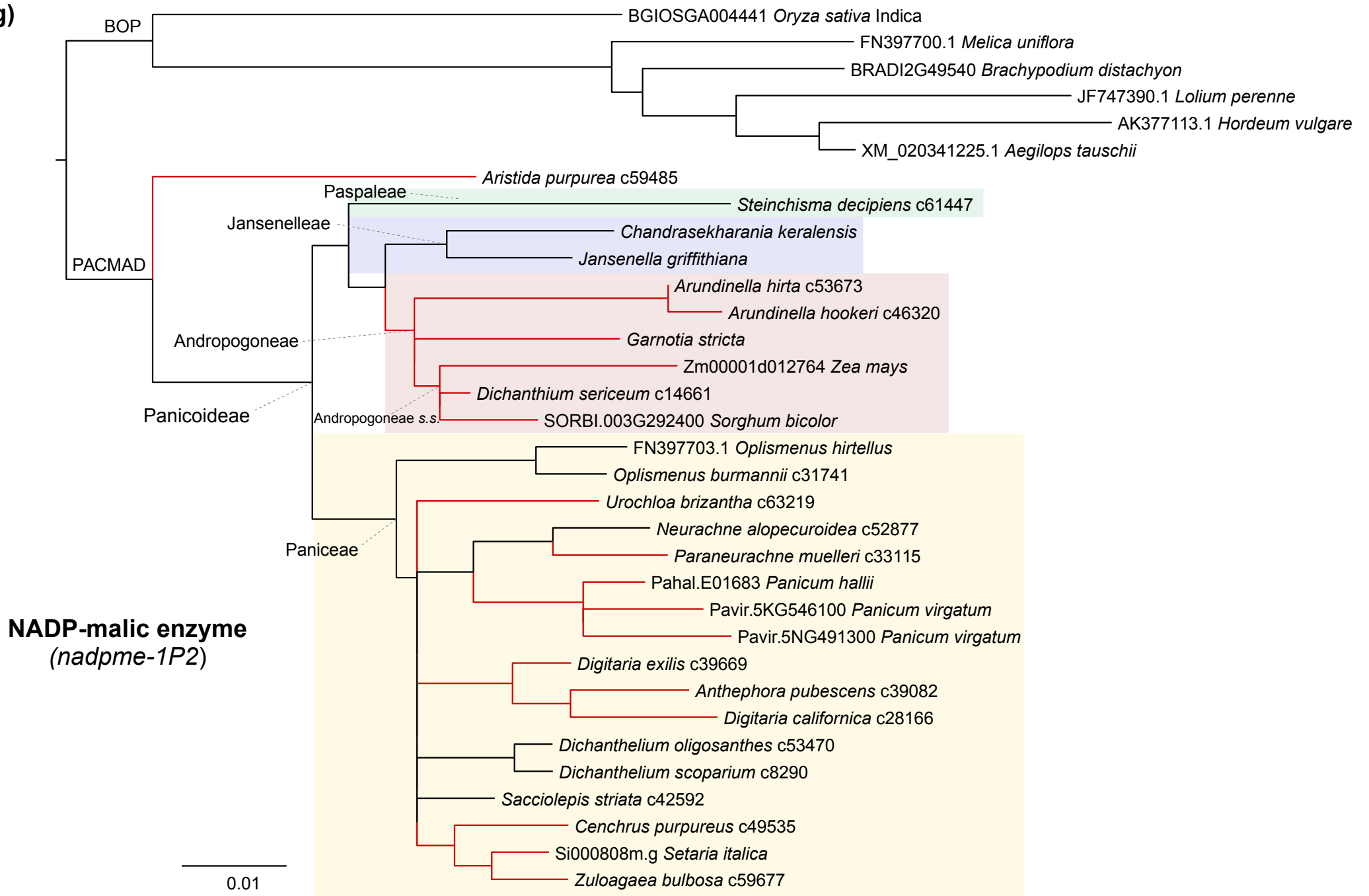
f)

**NADP-malic enzyme  
(nadpme-1P1)**

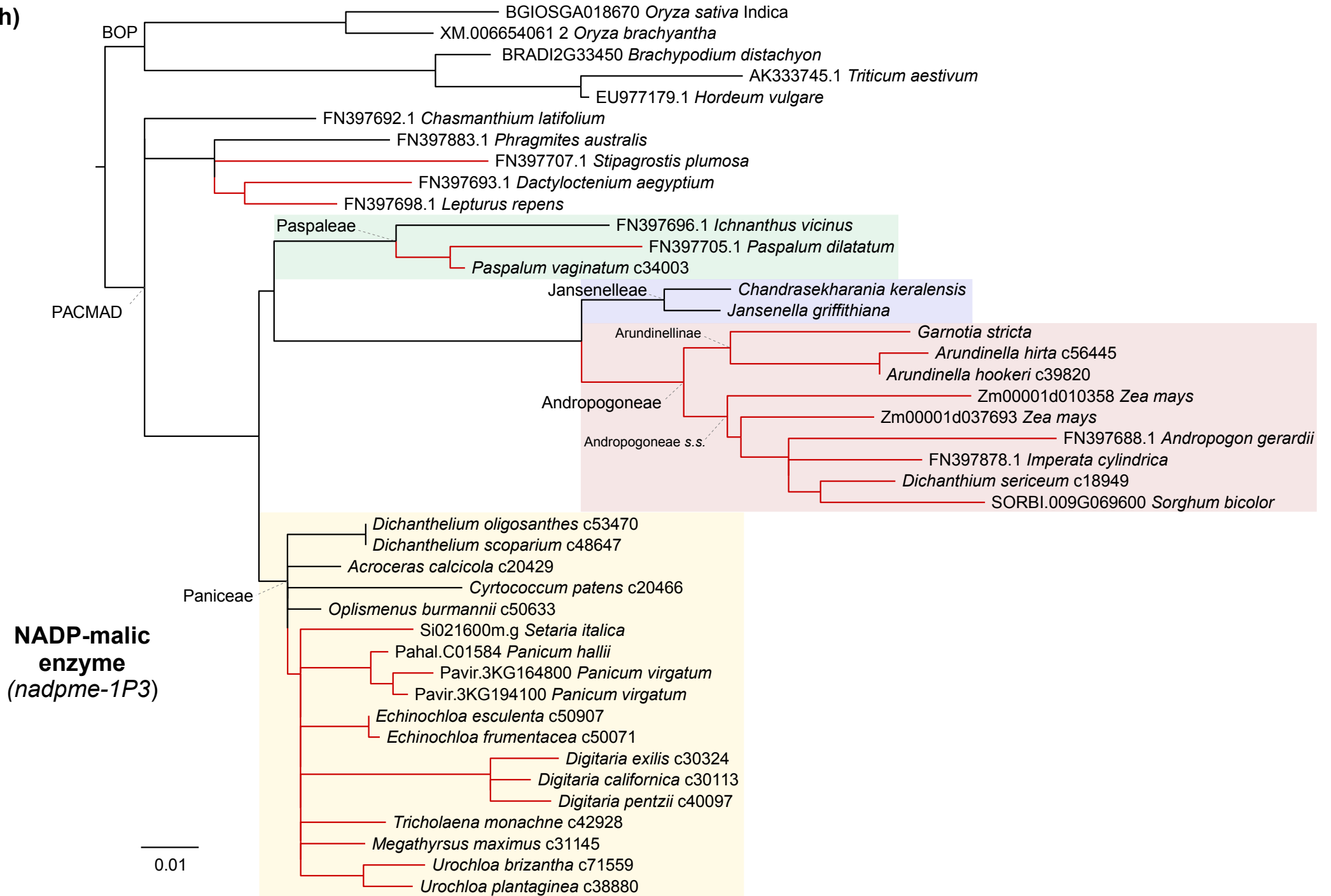


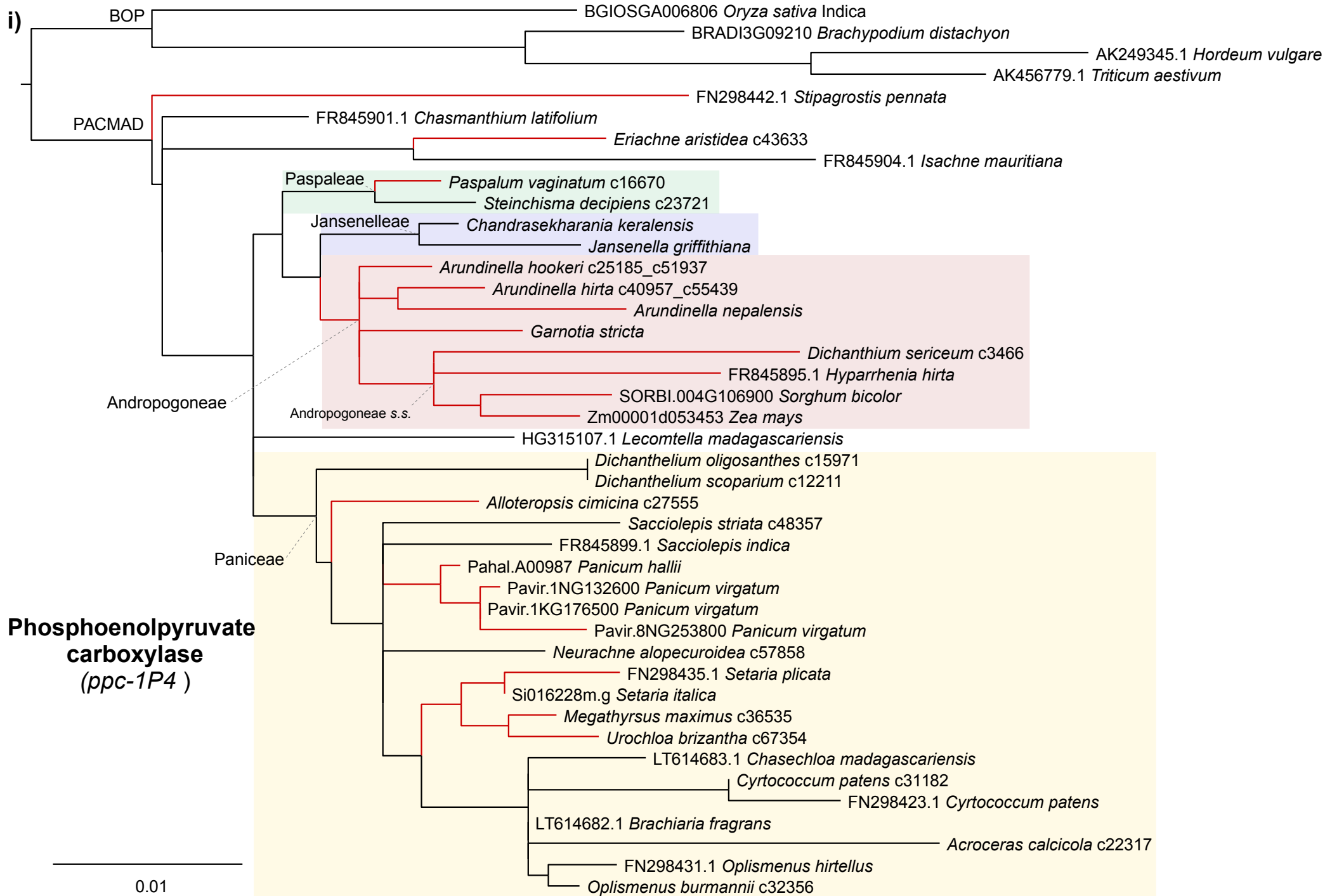
0.01

g)



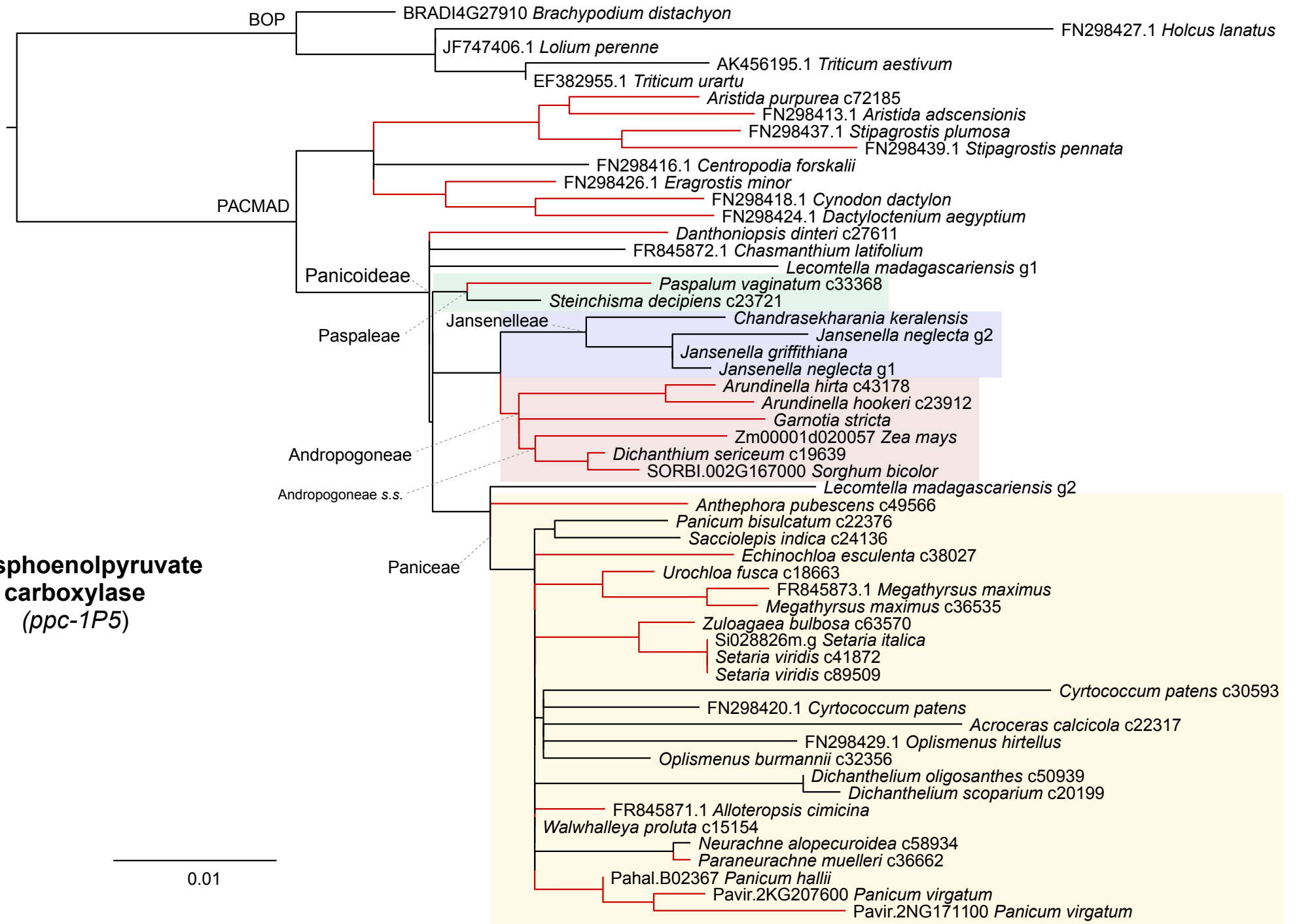
h)



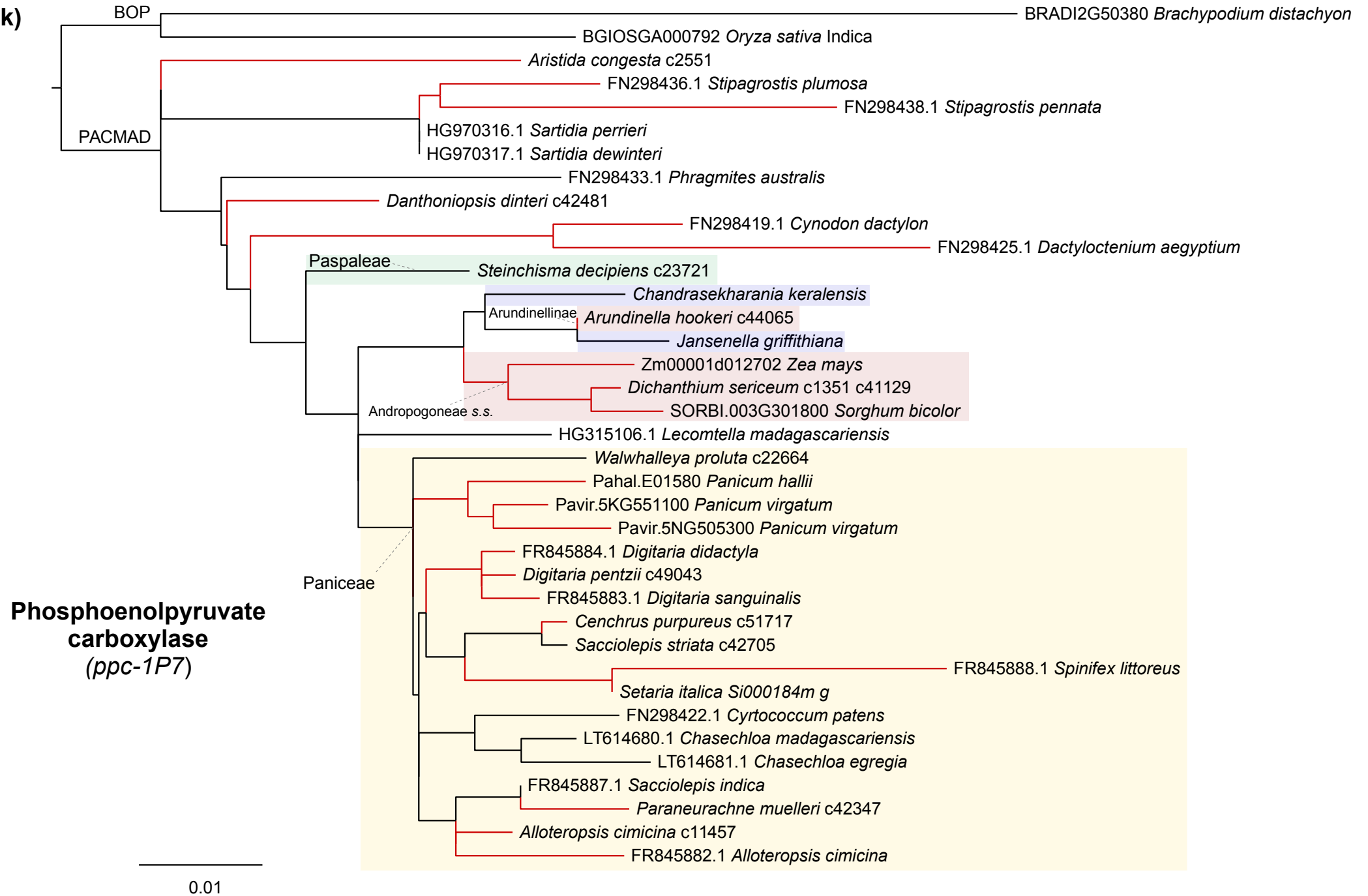




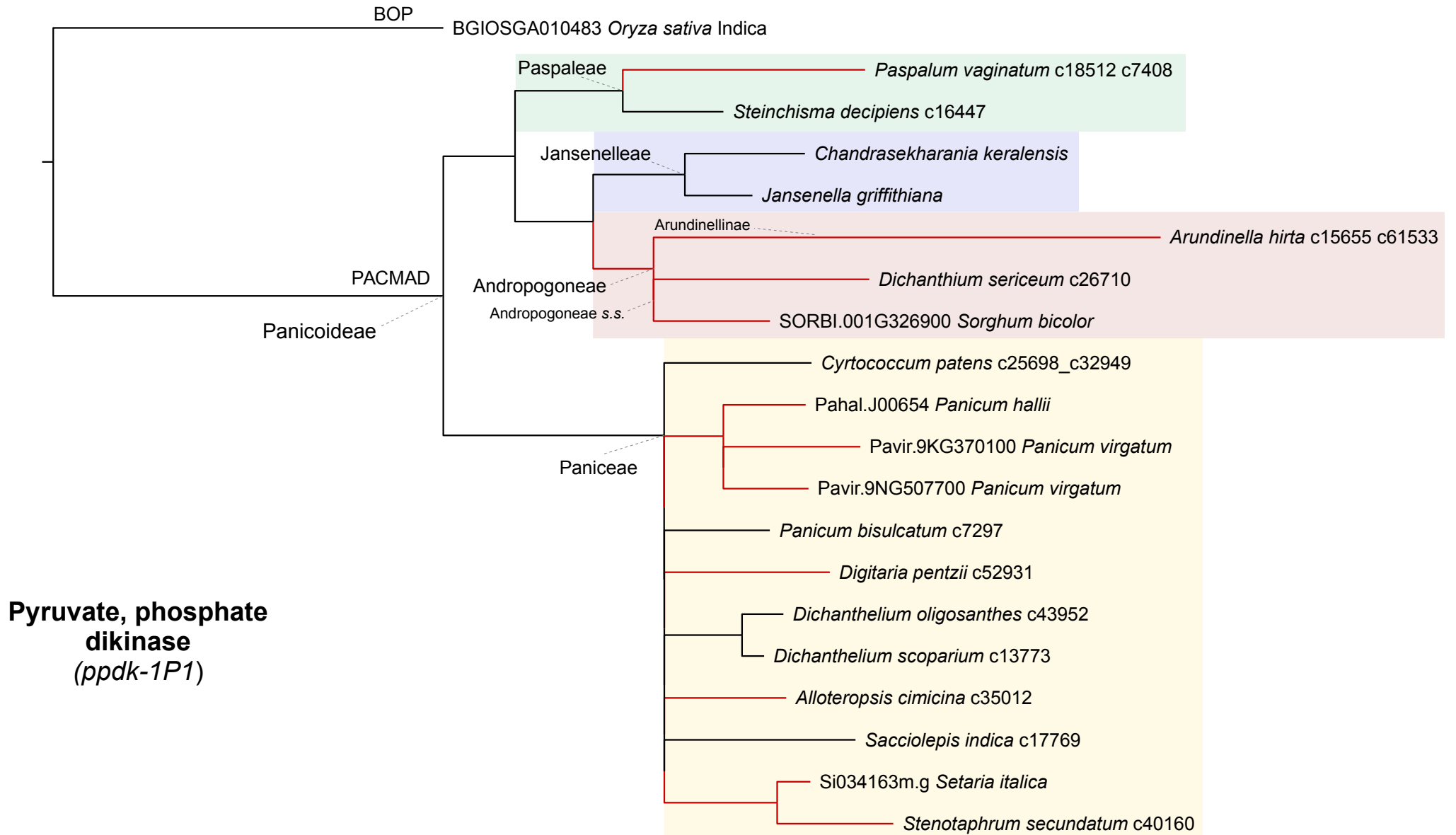
j)



k)



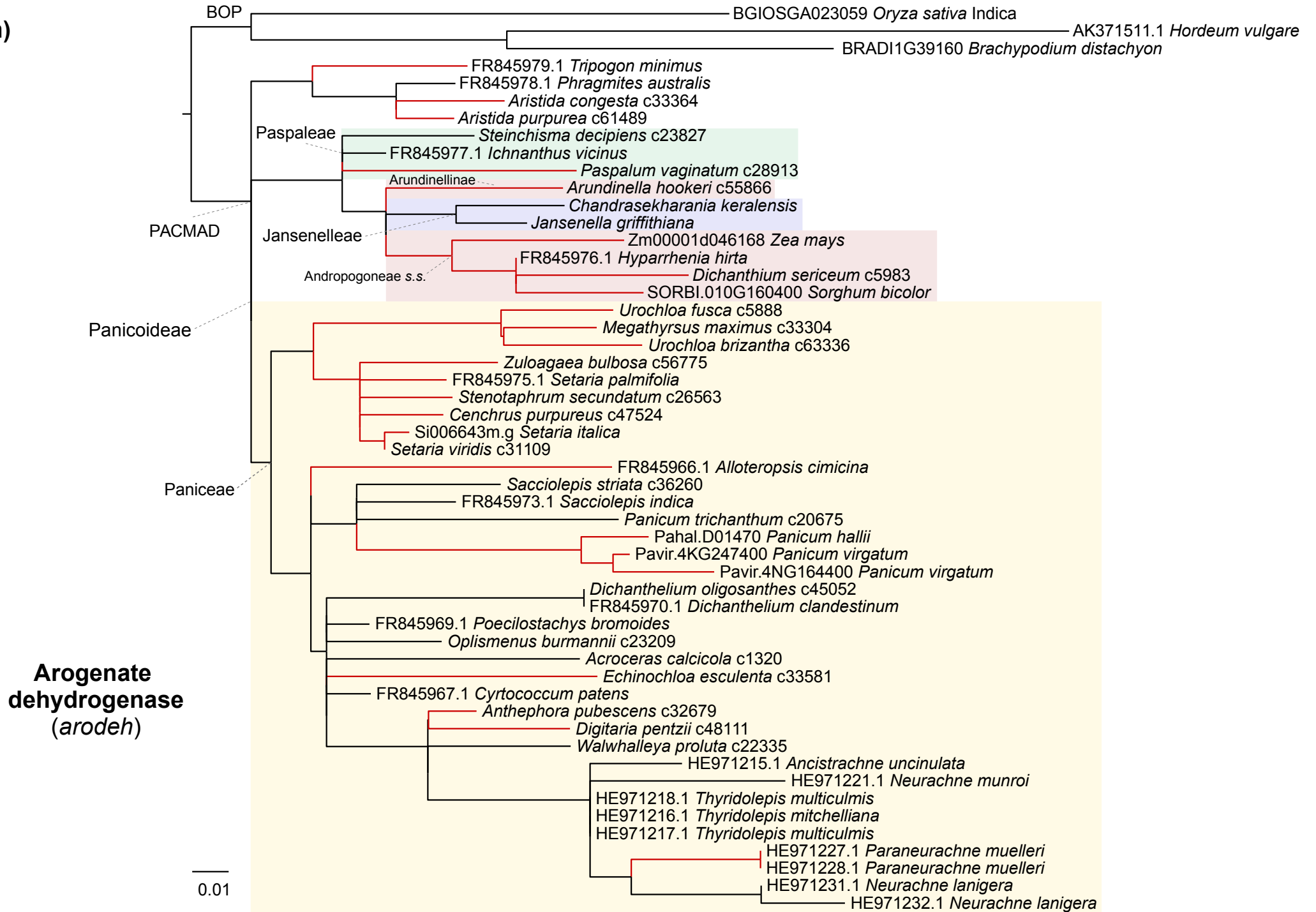
I)



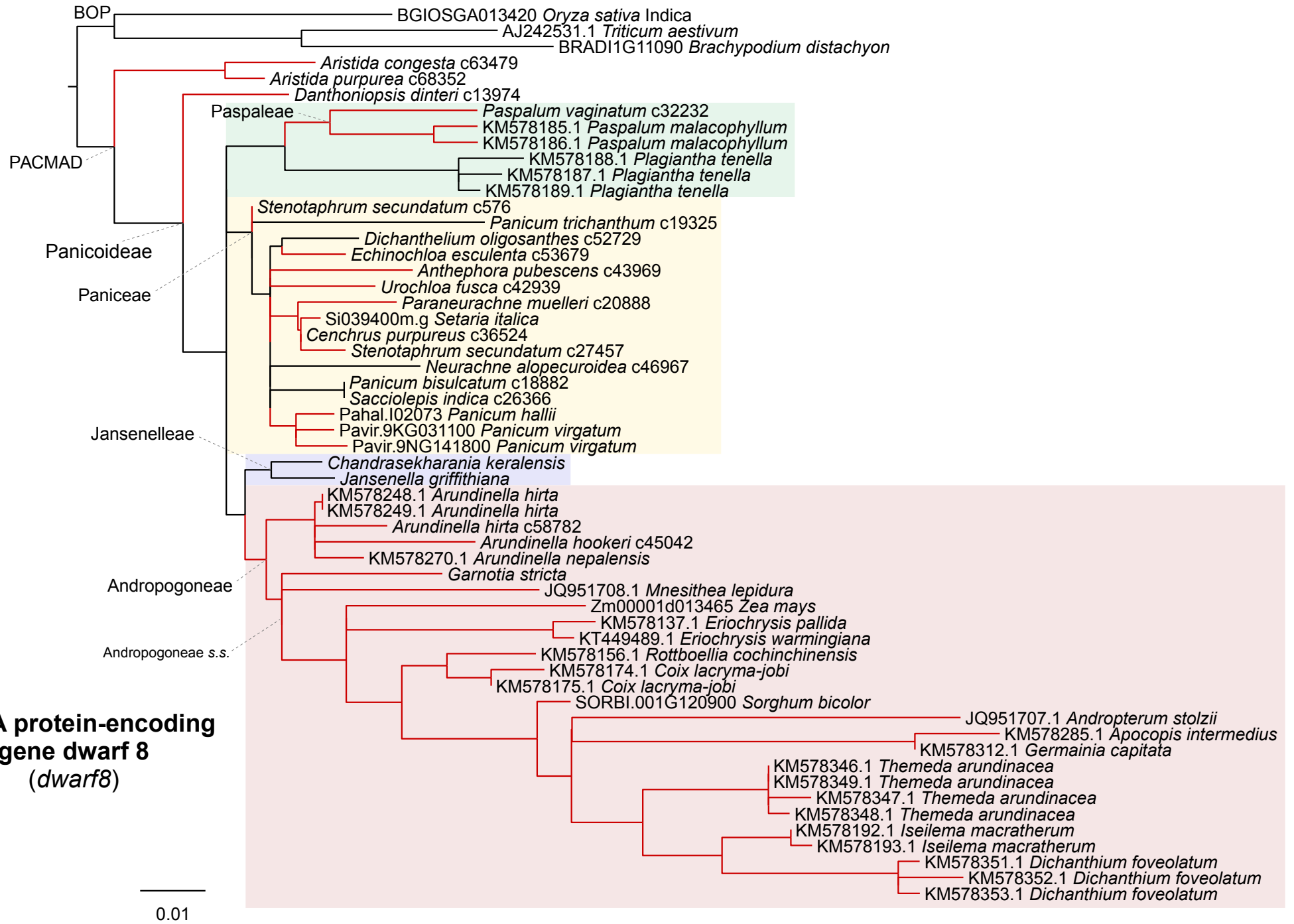
**Pyruvate, phosphate  
dikinase  
(*ppdk-1P1*)**

0.01

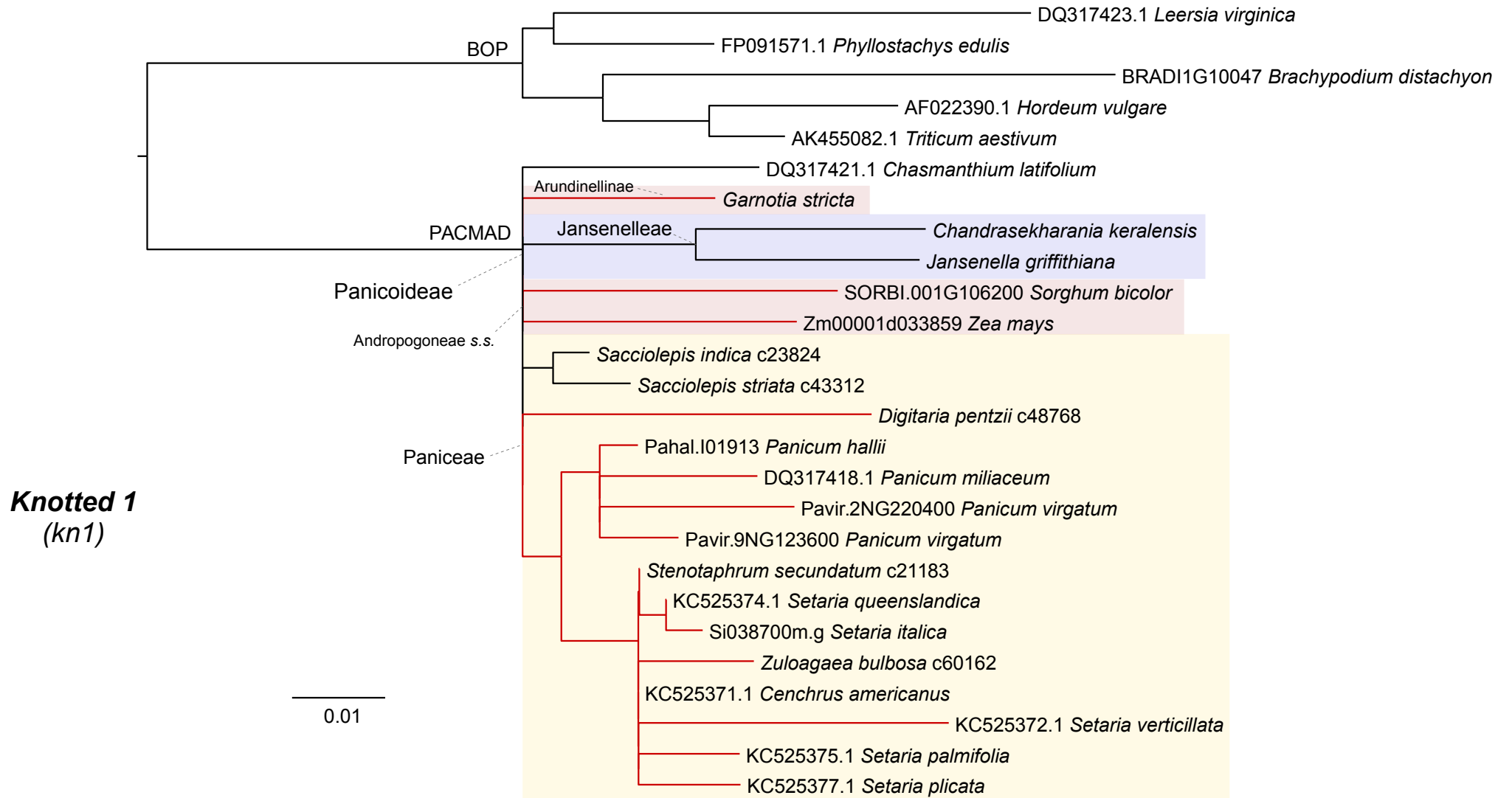
m)



n)



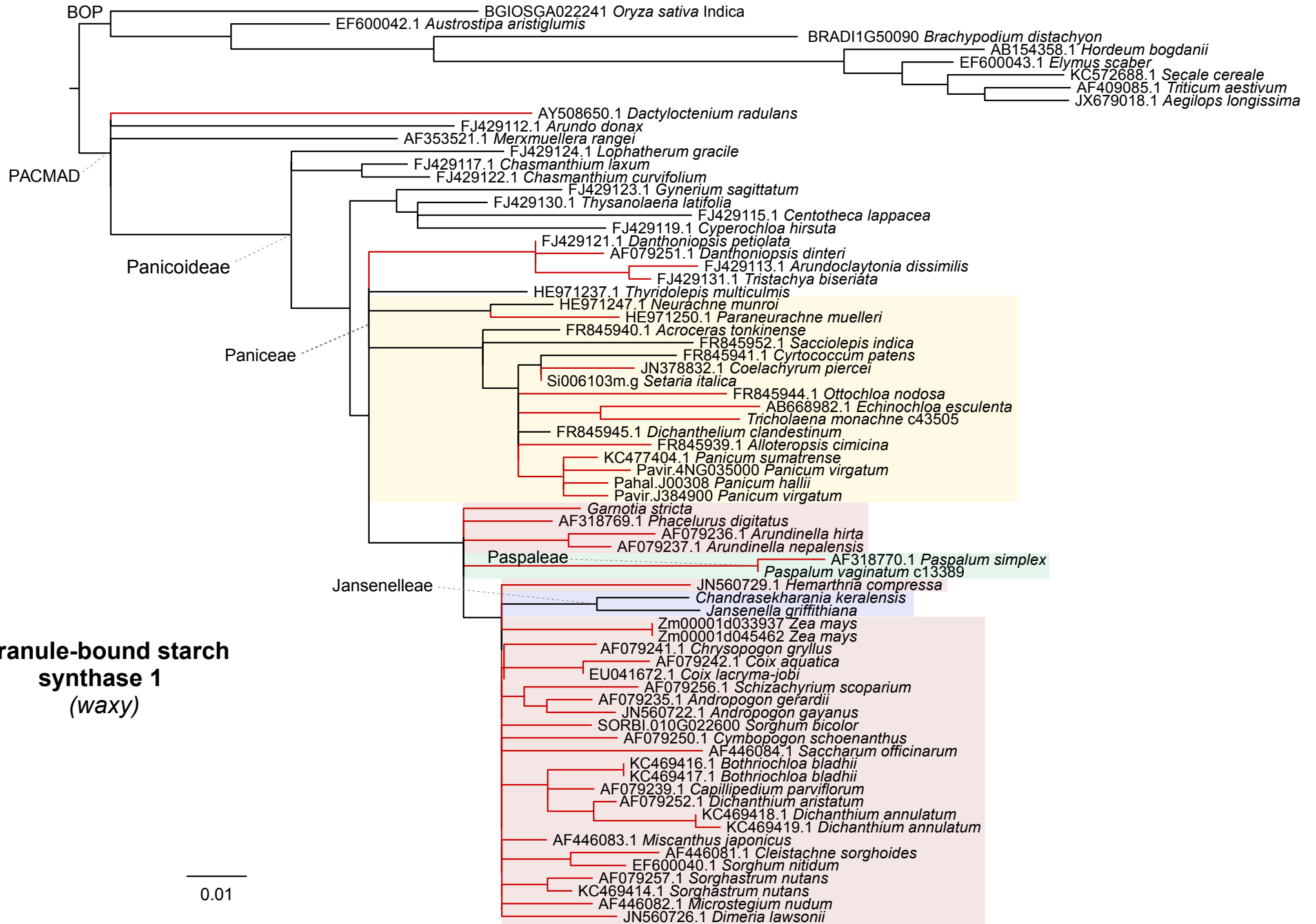
o)



p)



q)





**Table S5.** Summary of multigene coalescent species trees inferred from nuclear data under different filtering stringencies<sup>a</sup>.

N species <sup>b</sup>	Trim <sup>c</sup>	Set of trees <sup>d</sup>	Ref. species <sup>e</sup>	Clades <sup>f</sup>						
				Androp. + Jans.	Jans.	Androp.	Arundin.	Androp. s.s.	Panicaceae	Paspaleae
37	50%	bs ≥ 50	<i>Setaria</i>	0.94 (1)	0.92 (1)	0.84 (1)	Paraphyl.	0.94 (1)	0.8 (1)	0.93 (1)
			<i>Sorghum</i>	0.91 (1)	0.94 (1)	0.94 (1)	0.48 (1)	0.94 (1)	0.94 (1)	0.91 (1)
		All	<i>Setaria</i>	0.92 (1)	0.92 (1)	0.77 (1)	0.42 (0.96)	0.89 (1)	0.77 (1)	0.91 (1)
			<i>Sorghum</i>	0.88 (1)	0.95 (1)	0.72 (1)	0.49 (1)	0.9 (1)	0.88 (1)	0.89 (1)
	70%	bs ≥ 50	<i>Setaria</i>	-	-	-	Paraphyl.	-	-	-
			<i>Sorghum</i>	0.92 (1)	0.94 (1)	0.76 (1)	0.39 (0.45)	0.94 (1)	0.95 (1)	0.91 (1)
		All	<i>Setaria</i>	0.94 (1)	0.92 (1)	0.68 (1)	0.47 (0.89)	0.87 (1)	0.8 (1)	0.93 (1)
			<i>Sorghum</i>	0.89 (1)	0.97 (1)	0.67 (1)	0.45 (0.96)	0.89 (1)	0.89 (1)	0.92 (1)
	No	bs ≥ 50	<i>Setaria</i>	0.95 (1)	0.93 (1)	0.78 (1)	Paraphyl.	0.92 (1)	0.86 (1)	0.92 (1)
			<i>Sorghum</i>	0.95 (1)	0.95 (1)	0.8 (1)	0.47 (1)	0.8 (1)	0.94 (1)	0.9 (1)
		All	<i>Setaria</i>	0.93 (1)	0.91 (1)	0.76 (1)	Paraphyl.	0.9 (1)	0.8 (1)	0.89 (1)
			<i>Sorghum</i>	0.9 (1)	0.95 (1)	0.91 (1)	0.5 (1)	0.91 (1)	0.89 (1)	0.88 (1)
66	50%	bs ≥ 50	<i>Setaria</i>	-	-	-	-	-	-	-
			<i>Sorghum</i>	-	-	-	-	-	-	-
		All	<i>Setaria</i>	0.97 (1)	0.96 (1)	0.83 (1)	0.37 (0.57)	0.92 (1)	0.93 (1)	0.87 (1)
			<i>Sorghum</i>	0.94 (1)	0.98 (1)	0.76 (1)	0.66 (1)	0.93 (1)	0.93 (1)	0.92 (1)
	70%	bs ≥ 50	<i>Setaria</i>	-	-	-	-	-	-	-
			<i>Sorghum</i>	-	-	-	-	-	-	-
		All	<i>Setaria</i>	0.99 (1)	0.99 (1)	0.6 (0.86)	0.48 (0.57)	0.99 (1)	0.56 (0.48)	0.76 (0.88)
			<i>Sorghum</i>	0.91 (0.99)	0.99 (1)	0.57 (0.67)	0.72 (0.91)	0.86 (0.98)	0.98 (1)	0.49 (0.43)
	No	bs ≥ 50	<i>Setaria</i>	-	-	-	-	-	-	-
			<i>Sorghum</i>	1 (1)	0.98 (1)	0.8 (1)	0.39 (0.55)	1 (1)	1 (1)	0.82 (1)
		All	<i>Setaria</i>	0.97 (1)	0.95 (1)	0.81 (1)	Paraphyl.	0.93 (1)	0.93 (1)	0.84 (1)
			<i>Sorghum</i>	0.96 (1)	0.97 (1)	0.78 (1)	0.6 (1)	0.96 (1)	0.93 (1)	0.84 (1)

<sup>a</sup> Values are the proportion of quartets supporting the group, with local posterior probabilities between parentheses. Paraphyl. = paraphyletic groups. Missing values refer to datasets with less than 50 genes recovered after filtering; <sup>b</sup> Number of species in the dataset; <sup>c</sup> Alignment trimming threshold: 50% and 70% = sites available for less than 50% and 70% of sequences, respectively, were removed; No = no trimming; <sup>d</sup> Set of gene trees used for inferring the multigene coalescent species tree: All = all gene trees retained after filtering; bs ≥ 50 = only gene trees retained after filtering with at least 50% of branches with bootstrap support value ≥ 50%; <sup>e</sup> Reference species used for read mapping of genomic datasets: *Setaria* = *Setaria italica*, *Sorghum* = *Sorghum bicolor*; <sup>f</sup> Selected groups of grasses: Androp. = Andropogoneae, Jans. = Jansenelleae, Arundin. = Arundinellinae.

**Table S6.** Summary of branch model comparisons performed using the species tree topology.

Branch models <sup>a</sup>									
Scenarios of adaptive evolution		Single episode (Preceding C <sub>4</sub> evolution)			Single episode (During C <sub>4</sub> evolution, at the base of Andropogoneae)		Two episodes (During C <sub>4</sub> evolution, at the base of Arundinellinae and Andropogoneae <i>s.s.</i> )		dN/dS <sup>d</sup>
Gene <sup>b</sup>	N <sup>c</sup>	Null model	Internal branch only	Sustained	Internal branch only	Sustained	Internal branch only	Sustained	
Core C <sub>4</sub> genes									
<i>nadpmdh-1P1</i> (NADP-MDH)	10	<b>2.67</b>	4.67	3.71	4.64	0.00	4.32	0.17	0.11
<i>nadpme-1P4</i> (NADP-ME)	22	89.56	77.97*	38.10*	91.47	<b>0.00*</b>	90.78	1.23*	0.09 – 0.33
<i>pck-1P1</i> (PCK)	11	<b>3.15</b>	5.15	0.00	4.59	1.93	5.14	1.67	0.03
<i>ppc-1P3</i> (PEPC)	11	51.04	51.47	11.72*	38.93*	<b>0.00*</b>	46.32	28.06*	0.03 – 0.13
<i>ppdk-1P2</i> (PPDK)	9	47.65	49.66	39.12	44.00	35.36*	<b>0.00*</b>	39.61	0.11 – 0.41
Paralogs of core C <sub>4</sub> genes									
<i>nadpme-1P1</i>	8	<b>1.39</b>	2.72	0.00	3.19	1.77	3.38	2.02	0.06
<i>nadpme-1P2</i>	8	<b>5.12</b>	0.00	7.09	7.10	6.41	6.88	6.45	0.08
<i>nadpme-1P3</i>	9	<b>9.63</b>	11.14	3.84	0.00	4.75	9.27	8.80	0.05
<i>ppc-1P4</i>	12	<b>0.00</b>	1.66	1.88	0.25	2.00	0.16	1.82	0.05
<i>ppc-1P5</i>	11	<b>0.00</b>	0.31	1.93	0.79	1.97	1.63	1.78	0.07
<i>ppc-1P7</i>	9	<b>8.93</b>	10.70	5.01	7.74	0.00	-	-	0.08
<i>ppdk-1P1</i>	4	-	-	-	-	-	-	-	0.16
Other nuclear genes									
<i>arodeh</i>	6	<b>1.27</b>	3.27	3.27	0.00	2.94	-	-	0.10
<i>dwarf8</i>	11	<b>8.34</b>	10.34	6.88	10.04	0.76	10.19	0.00	0.07
<i>knotted1</i>	7	<b>2.04</b>	3.99	2.61	3.08	4.04	0.00	4.03	0.08
<i>phyB</i>	12	<b>9.35</b>	11.30	0.00	3.63	7.45	10.22	9.40	0.09
<i>waxy</i>	14	<b>0.00</b>	1.98	0.69	2.00	1.01	1.22	1.01	0.04

<sup>a</sup> dAIC values relative to the best-fit model for each gene are shown. The best-fit model is highlighted in bold. Asterisks indicate significant likelihood ratio tests (LRT) against the null model after Bonferroni correction. Two hypotheses of potential enzyme adaptation were tested for each scenario, the first assuming a shift in selective pressure only in the internal branch of the group specified, the second assuming a sustained shift from the internal branch including all descendant branches. The tree topology inferred from whole genome nuclear data (66 species dataset; Fig. S3) was used for these tests. Missing values correspond to trees in which Andropogoneae was not monophyletic, trees in which Arundinellinae was not represented or only a single representative of Andropogoneae *s.s.* was represented; <sup>b</sup> C<sub>4</sub> gene annotation following Moreno-Villena et al. (2018); <sup>c</sup> Number of sequences in the alignment; <sup>d</sup> dN/dS ratios of background and foreground branches, respectively, estimated for the best-fit model, except in cases where the null was the best-fit model, for which there was a single dN/dS estimate for all branches.