Supporting Online Material for

Rates of Molecular Evolution Are Linked to Life History in Flowering Plants

Stephen A. Smith* and Michael J. Donoghue

*To whom correspondence should be addressed. E-mail: stephen.smith@yale.ed

Published 3 October 2008, *Science* **322**, 86 (2008)
DOI: 10.1126/science.1163197

This PDF file includes:

Supporting Online Material (SOM)

Supporting Text

Matrix assembly

Data on the plant clades Apiales, Commelinidae (sensu S*1*), Dipsacales, Moraceae +Urticaceae, and Primulales were assembled from DNA sequences deposited in GenBank. Because multiple sequence alignments across large clades can be inaccurate we identified large alignable clusters for each gene region, favoring alignable groups corresponding to named clades. Phylota (rel. 1.01; S*2*) was used to circumscribe most clusters; however blastclust (vers. 2.2.18; S*3*) and N x N sequence comparisons of corrected and uncorrected distances were also used to identify problematic sequences. To minimize missing data, only phylogenetically informative clusters (with at least 4 taxa) were used. The average gene region in our analyses contained 305 species; the smallest contained just 10 species.

Many gene regions are sequenced using primers that cover large overlapping sections (e.g., *trn*L-*trn*T, *trn*T-*trn*D, *trn*L-*trn*F). In some cases, sections of adjacent gene regions have been sequenced in addition to the target regions. Clusters with similar sequences spanning more than one gene region were not divided into individual regions. ITS1 and ITS2 were combined when they had been submitted to online databases (i.e. NCBI GenBank, EMBL) as separate entries. Small gene regions were kept separate if they were sampled densely within a taxonomic cluster (i.e., *trn*L).

Once clusters for each taxonomic group and each gene region were established, multiple sequence alignments for each cluster were carried out using Dialign (vers. 2.2.1; S*4*) when clusters contained fewer than 100 sequences and Muscle (vers. 3.6; S*5*) for

larger clusters. Sites with fewer than two nucleotides were trimmed out of each cluster with Phyutility (S*6*). After aligning each cluster, clusters of the same or similar gene regions were aligned together using group-to-group profile alignments (S*5*). Profile alignment procedures align two multiple sequence alignments together preserving the columns of each individual alignment. Attempts were made to profile align more closely related clusters first where information was available on the broader phylogenetic relationships of each clade. Each aligned gene region is shown in Figs. S1 and S2. These gene regions were then concatenated with Phyutility (S*6*) and the concatenated matrices were subjected to phylogenetic analyses.

Phylogenetic inference

Maximum-likelihood trees were inferred with RAxML (vers. 7.0.0; S*7*). All analyses were partitioned into gene-regions, allowing for parameter estimates on each partition; branch length estimates were optimized across all gene regions. All analyses employed a GTRMIX nucleotide substitution model for each partition, which conducts a phylogenetic search using the faster GTRCAT method followed by a final parameter optimization step using the GTR substitution model and Γ model of among-site rate variation. Other parametric phylogenetic reconstruction methods were found to be impractical for the exceptionally large datasets used here. Our matrices and trees can be obtained from http://blackrim.org/data or from TreeBASE (www.treebase.org).

The Apiales were rooted along the *Pennantia* branch (see S*8*). Commelinidae were rooted along the branch connecting Arecales (the palms) and the rest of the Commelinidae (see S*9*, S*10*). Dipsacales were rooted on the branch connecting

Adoxaceae and Caprifoliaceae (see S*11*). We used *Maesa tenera* (Maesaceae, Ericales) as an outgroup for Primulales (see S*12*, S*13*). We used *Humulus lupulus* as an outgroup for the Moraceae-Urticaceae analyses (S*14*, S*15*).

Bootstrapping was not possible due to the large size of the datasets. In any case, bootstrapping on data matrices with large amounts of missing data typically yields values that are small and may not accurately reflect node support (S*16*). Bayesian analyses might provide more accurate measures of support; however, analyses of our smallest datasets never converged, making run-times impractical. For these reasons, particular relationships recovered in our analyses must be treated with caution. However, we note that the results obtained generally correspond well with previously published studies and that our conclusions regarding the difference between life history categories should be robust to many phylogenetic rearrangements.

Dating analyses

The sizes of our datasets narrowed the options for dating analyses. Each phylogeny presented here deviates from a molecular clock on at least two levels: (1) rate heterogeneity among closely related species, and (2) large rate differences among clades. For all clades except Commelinidae (which failed to run to completion), we used non-parametric rate smoothing (NPRS; S*17*), as implemented in r8s (vers. 1.7; S*18*). NPRS analyses with TreeEdit (vers. 1.10; S*19*) did run to completion but internal nodes cannot be calibrated with this implementation, and rate heterogeneity is smoothed over the entire tree (rendering lineages with slow rates younger and lineages with fast rates older). PATHD8 (S*20*) was used to obtain dates for the Commelinidae. PATHD8 is a non-

parametric dating method based on mean path lengths that allows an arbitrary number of calibrated nodes. It addresses deviations from the molecular clock locally, allowing very fast run times for large datasets.

To obtain absolute rates of molecular evolution, fossil calibrations were used in each phylogeny. Owing to the nature of the rate heterogeneity reported here, multiple calibrations were used whenever possible, with special attention given to sections of the tree likely to include large rate differences. All calibrations are from previously published estimates and are provided in Table S1. For Apiales, we calibrated three nodes using estimates from (S*21*). We calibrated 23 nodes within Dipsacles using estimates from (S*22*). For Commelinidae, we calibrated 14 nodes from (S*23*), and for Primulales we used two estimates from (S*21*). Three nodes of Moraceae+Uritcaceae were calibrated with estimates from (S*15*).

Ancestral state reconstruction

Ancestral state reconstructions for the “tree/shrub” versus “herb” character were conducted with a likelihood method. Specifically, we calculated global marginal reconstructions at each node with a non-symmetric model of evolution as implemented in Lasrdisc (vers. 1.0; S*24*).

Substitutions per site per million years

The measure of substitutions per site per million years represents the absolute rate of evolution of each lineage given the data and the model of phylogenetic reconstruction, the branch length estimations, and the dating method. A branch's

substitutions/site/million years was calculated by dividing the substitutions per site estimated from the molecular phylogenetic analyses by the time spent in that lineage estimated from the dated phylogeny. For these calculations, zero branch lengths were given a value of 10^{-6}.

Phylogenetic contrasts

Previous studies of lineage specific rate heterogeneity have relied mainly on relative rates tests in which multiple measurements are made across the same phylogenetic branch. We avoided this problem by focusing on phylogenetic contrasts with non-overlapping nodes. If we assume a molecular clock from a given node, then, on average, 50% of the substitutions are expected to occur along each of the descendant lineages arising from that node. Deviations should be randomly distributed with respect to the life history difference if it had no effect on the rate of molecular evolution.

We calculated the accumulation of substitutions/site in each woody clade versus its herbaceous sister clade in each phylogeny. Contrasts were only considered where reliable information on habit was available, and when the tree/shrub and herbaceous clades consisted of at least two tips. 13 sister-clade contrasts were identified: three from Apiales, one from Commelinidae, three from Moraceae+Urticaceae, two from Primulales, and four from Dipsacales (Figs. S3-7). To explore the robustness of the conclusions from these contrasts we considered two alternative hypotheses within Dipsacales. One of these (contrast 14) compared woody Linnaeeae not just to Morinaceae but to the entire herbaceous Valerina clade, including Morinaceae, Valerianaceae, and Dipsacaceae (see S*11*). The other (contrast 15) treated all *Sambucus* species as

"trees/shrubs," and contrasted this clade with herbaceous Adoxina, including *Adoxa*, *Tetradoxa*, and *Sinadoxa* (S*11*). In an additional test we omitted the questionable contrasts 11-15.

For each contrast we calculated the average substitutions/site from tips scored as trees/shrubs to the base of the tree/shrub clade, and from herbaceous tips to the base of the herbaceous sister clade (S*25*,S*26*). In the case of nested contrasts (e.g., *Dorstenia* versus *Brosimum* and relatives within Moraceae, and Moraceae versus Urticacaeae), the phylogenetically less inclusive (shallower) contrast was carried out first, and then was removed in carrying out the more inclusive (deeper) contrast. We tested whether herbaceous clades had higher rates more often than expected by chance using a sign test (Table S2).

_rbc_L in Commelinidae

We examined rate heterogeneity in codons in Commelinidae using the chloroplast gene *rbcL*, the largest coding region in our datasets. We assumed that the phylogeny estimated from the entire Commelinidae dataset was more accurate than the one produced from *rbc*L alone, and estimated parameters and branch lengths on a pruned version of the larger commelinid phylogeny that included only the 1208 species with *rbc*L sequences in GenBank. Four sets of branch length estimates were contrasted between palms and the remainder of the commelinds: (1) 1st and 2nd positions together, (2) 3rd positions, (3) 1st, 2nd, and 3rd positions, and (4) sequences translated to amino acids. All DNA sequence analyses were run in RAxML (vers. 7.0.0) with the GTR substitution model with among site rate heterogeneity modeled with Γ; the amino acid sequence analysis was

run using the WAG substitution model and Γ rate heterogeneity among sites. The results are presented in Table 2.

Supporting References

S1. Cantino, *et al.*, *Taxon* **56**, 822-846 (2007)

S2. M. J. Sanderson, D. Boss, D. Chen, K. A. Cranston, A. Wehe, *Syst. Biol.* **57**, 335-346 (2008)

S3. S. F. Altschul, *et al.*, *Nucleic. Acids. Res*. **25**, 3389-3402 (1997)

S4. B. Morgenstern, *Bioinformatics* **15**, 211-218 (1999)

S5. R. C. Edgar, *Nucleic. Acids. Res*. **32**, 1792-1797 (2004)

S6. S. A. Smith, C. D. Dunn, *Bioinformatics* **24**, 714-716 (2008)

S7. A. Stamatakis, *Bioinformatics* **22**, 2688-2690 (2006)

S8. G. T. Chandler. G. M. Plunkett, *Bot. J. Linn. Soc.* **144**, 123-147 (2004)

S9. M. W. Chase, *Am. J. Bot.*, **91**, 1645-1655 (2004)

S10. M. W. Chase, *et al. Aliso* **22**, 63-75 (2006)

S11. M. J. Donoghue, C. D. Bell, R. C. Winkworth, *Int. Jour. Plant Sci.* **164**, S453- S464 (2003)

S12. D. E. Soltis, P. S. Soltis, P. K. Endress, M. W. Chase, *Phylogeny and Evolution of Angiosperms* (Sinaur Associates, Sunderland, MA, 2005)

S13. P. F. Stevens, Angiosperm Phylogeny Website, http://www.mobot.org/MOBOT/research/APweb/ (2008)

S14. Sytsma et al., *American Journal of Botany*, **89**, 1531-1546 (2002)

S15. N. J. C. Zerega, W. L. Clement, S. L. Datwyler, G. D. Weiblen, *Mol. Phylogenet. Evol.* **37**, 402-416 (2005)

S16. M. M. McMahon, M. J. Sanderson, *Syst. Biol.* **55**, 818-836 (2006)

S17. M. J. Sanderson, *Mol. Biol. Evol*. **14**, 1218-1231 (1997)

S18. M. J. Sanderson, *Bioinformatics* **19**, 301-302 (2003)

S19. A. Rambaut, M. Charleston, TreeEdit version v1.0a8, http://evolve.zoo.ox.ac.uk/software/TreeEdit/main.html (2001)

S20. T. Britton, C. L. Anderson, D. Jacquet, S. Lundqvist, K. Bremer, *Syst. Biol*. **56**, 741-752 (2007)

S21. K. Bremer, E. M. Friis, B. Bremer *Syst. Biol*. **53**, 496-505 (2004)

S22. C. D. Bell, M. J. Donoghue, *Am. J. Bot.* **92**, 284-296 (2005)

S23. T. Janssen, K. Bremer, *Bot. J. Linn. Soc.* **146**, 385-398 (2004)

S24. V. K. Jackson, LASRDisc: Likelihood Ancestral State Reconstruction for Discrete Characters. Version 1.0., http://ceb.csit.fsu.edu/lasrdisc/ (2004)

S25. T. G. Barraclough, P. H. Harvey, S. Nee, *Proc. R. Soc. Lond. B*, **263**, 589- 591 (1996)

S26. T. J. Davies, V. Savolainen, *Evolution* **60**, 476-483 (2006)

Supporting Tables

Table S1. Fossil calibrations; clade names, phylogeny name (see Fig. 1 and Figs. S3-S7), calibration age, and literature source (*S15,S21-S23*).

Clade	Phylogeny	Age (mya)	Source
Apiaceae	Apiales	65	S21
Araliaceae	Apiales	75	S21
Pittosporaceae	Apiales	65	S21
Adoxa	Dipsacales	16	S22
Adoxa+Sambucus	Dipsacales	46	S22
Sambucus	Dipsacales	7-20	S22
Adoxaceae	Dipsacales	65-75	S22
Viburnum	Dipsacales	45-75	S22
Diervilleae	Dipsacales	53	S22
Valeriana+Centranthus	Dipsacales	20	S22
Centranthus+Nardostachys	Dipsacales	42	S22
Patrinia+Valeriana	Dipsacales	51	S22
Paramo Valeriana	Dipsacales	5-7	S22
Sixalix+Pterocephalus	Dipsacales	22	S22
Patrinia+Pterocephalus	Dipsacales	62	S22
Morina+Cryptothladia	Dipsacales	17	S22
Morina+Acanthocalyx	Dipsacales	31	S22
Acanthocalyx+Valeriana	Dipsacales	69	S22
Abelia+Dipelta	Dipsacales	35	S22
Dipelta+Linnaea	Dipsacales	45	S22
Lonicera+Triosteum	Dipsacales	46	S22
Lonicera+Heptacodium	Dipsacales	76	S22
Lonicera	Dipsacales	37-44	S22
Linnaeeae+Heptacodium	Dipsacales	84	S22
Diervilleae+Heptacodium	Dipsacales	85	S22
Dipsacales	Dipsacales	103	S22
Commelinidae	Commelinidae	120	S23
Arecaceae	Commelinidae	110	S23

Zingiberales	Commelinidae	88	S23
Commelinales	Commelinidae	110	S23
Typhaceae+Sparganium	Commelinidae	89	S23
Bromeliaceae	Commelinidae	96	S23
Rapaceae	Commelinidae	79	S23
Cyperaceae+Juncaceae	Commelinidae	88	S23
Cyperaceae	Commelinidae	76	S23
Juncaceae	Commelinidae	74	S23
Eriocaulaceae+Xyridaceae	Commelinidae	105	S23
Centrolepidaceae+ Anarthriaceae+Restionaceae	Commelinidae	96	S23
Poaceae+Flagellariaceae	Commelinidae	108	S23
Poaceae	Commelinidae	83	S23
Myrsinaceae	Primulales	45	S21
Primulaceae	Primulales	45	S21
Theophrastaceae	Primulales	65	S21
Urticaceae+Moracaee	Urticales	105	S15
Moraceae	Urticales	89	S15
Urticaceae	Urticales	76	S15

Table S2. Datasets used in phylogenetic analyses; phylogeny name (see Fig. 1 and Figs. S3-S7), gene region, description of the region, number of sequences in the alignment, and the length of the alignment in nucleotide sites.

Phylogeny	Gene Region	Description	Number of Sequences	Length of Alignment
Apiales	5S	5S ribosomal RNA	113	341
Apiales	ITS	Internal transcribed spacer	1427	934
Apiales	matK	Maturase K	22	1910
Apiales	rbcL	Ribulose bisphosphate carboxylase, large chain	123	1382
Apiales	rpl16	Ribosomal protein L16	128	1208
Apiales	rps16	Ribosomal protein S16	325	1303
Apiales	trnL	Chloroplast trnL gene	86	522
Apiales	trnLtrnF	trnL-trnF intergenic spacer	363	1064
Apiales	trnLtrnT	trnT-trnL intergenic spacer	86	850
Commelinidae	atpB	Atp synthase beta chain	85	1538
Commelinidae	ETS	External transcribed spacer	339	621
Commelinidae	ITS	Internal transcribed spacer	2419	900
Commelinidae	matK	Maturase K	922	2490
Commelinidae	ndhF	NADH-plastoquinone oxidoreductase	649	2217
Commelinidae	prK	Phosphoribulokinase	333	676
Commelinidae	psbAtrnH	psbA-trnH integenic spacer	10	617
Commelinidae	rbcL	Ribulose bisphosphate carboxylase, large chain	1219	2323
Commelinidae	rpb2	RNA polymerase II second largest subunit	292	892
Commelinidae	rpoC2	RNA polymerase beta" subunit	188	649
Commelinidae	rps16	Rps16 intron	596	1037
Commelinidae	S16	Ribosomal protein S16	89	1145
Commelinidae	trnD	Chlorplast trnD gene	165	900
Commelinidae	trnK	Chloroplast trnK gene	403	2825
Commelinidae	trnL	Chloroplast trnL gene	106	499

Commelinidae	trnLtrnF	trnL-trnF intergenic spacer	1939	2351
Commelinidae	trnTtrnL	trnT-trnL intergenic spacer	24	695
Dipsacales	atpBrbcL	atpB-rbcL intergenic spacer	114	901
Dipsacales	ITS	Internal transcribed spacer	334	794
Dipsacales	petNpsbM	petN-psbM intergenic spacer	55	1314
Dipsacales	psbA	photosystem II protein	129	548
Dipsacales	psbMtrnD	psbM-trnD intergenic spacer	55	1209
Dipsacales	rbcL	Ribulose bisphosphate carboxylase, large chain	74	1439
Dipsacales	rpoBtrnC	rpoB-trnC intergenic spacer	55	1349
Dipsacales	trnK	Chloroplast trnK gene	182	1388
Dipsacales	trnLtrnF	trnL-trnF intergenic spacer	179	1698
Dipsacales	trnStrnG	trnS-trnG intergenic spacer	121	726
Mora-Urti	26S	Ribosomal protein 26S	91	989
Mora-Urti	atpBrbcL	atpB-rbcL intergenic spacer	11	940
Mora-Urti	ETS	External transcribed spacer	107	482
Mora-Urti	ITS	Internal transcribed spacer	299	872
Mora-Urti	ndhF	NADH-plastoquinone oxidoreductase	104	2036
Mora-Urti	rbcL	Ribulose bisphosphate carboxylase, large chain	39	1408
Mora-Urti	trnLtrnF	trnL-trnF intergenic spacer	161	1076
Primulales	ITS	Internal transcribed spacer	276	775
Primulales	matK	Maturase K	171	1607
Primulales	ndhF	NADH-plastoquinone oxidoreductase	118	1965
Primulales	rbcL	Ribulose bisphosphate carboxylase, large chain	62	1408
Primulales	rpl16	Ribosomal protein L16	192	1196
Primulales	rps16	Ribosomal protein S16	97	939
Primulales	trnL	Chloroplast trnL gene	423	1560
Primulales	trnLtrnF	trnL-trnF intergenic spacer	176	556
Primulales	trnStrnG	trnS-trnG intergenic spacer	42	778
Primulales	trnTtrnL	trnT-trnL intergenic spacer	60	913

Supplemental Figures

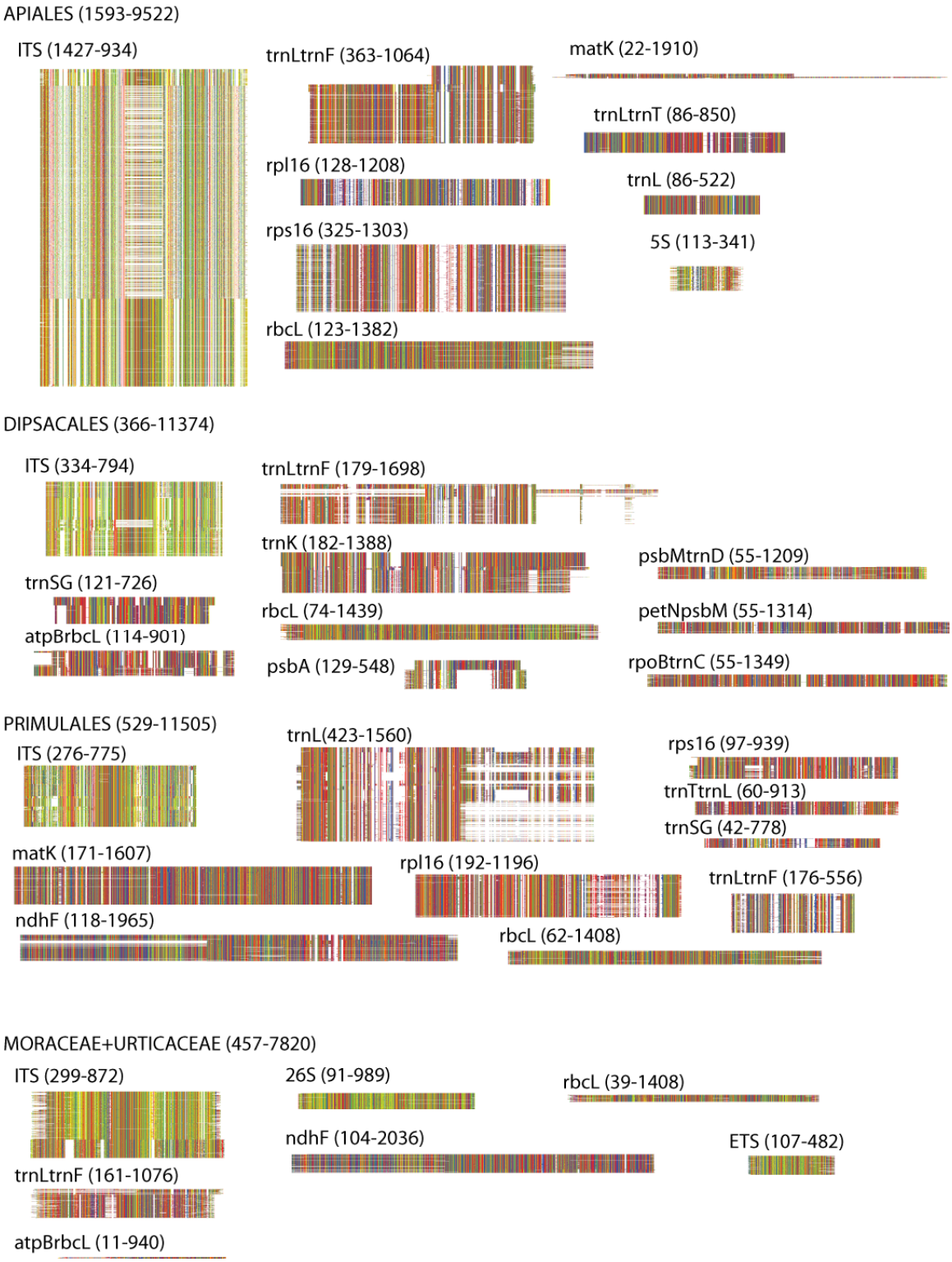

Fig. S1. Visualization of individual alignments combined in profile alignment for phylogenetic analysis of Apiales, Dipsacales, Primulales, and Moraceae + Uritcaceae.

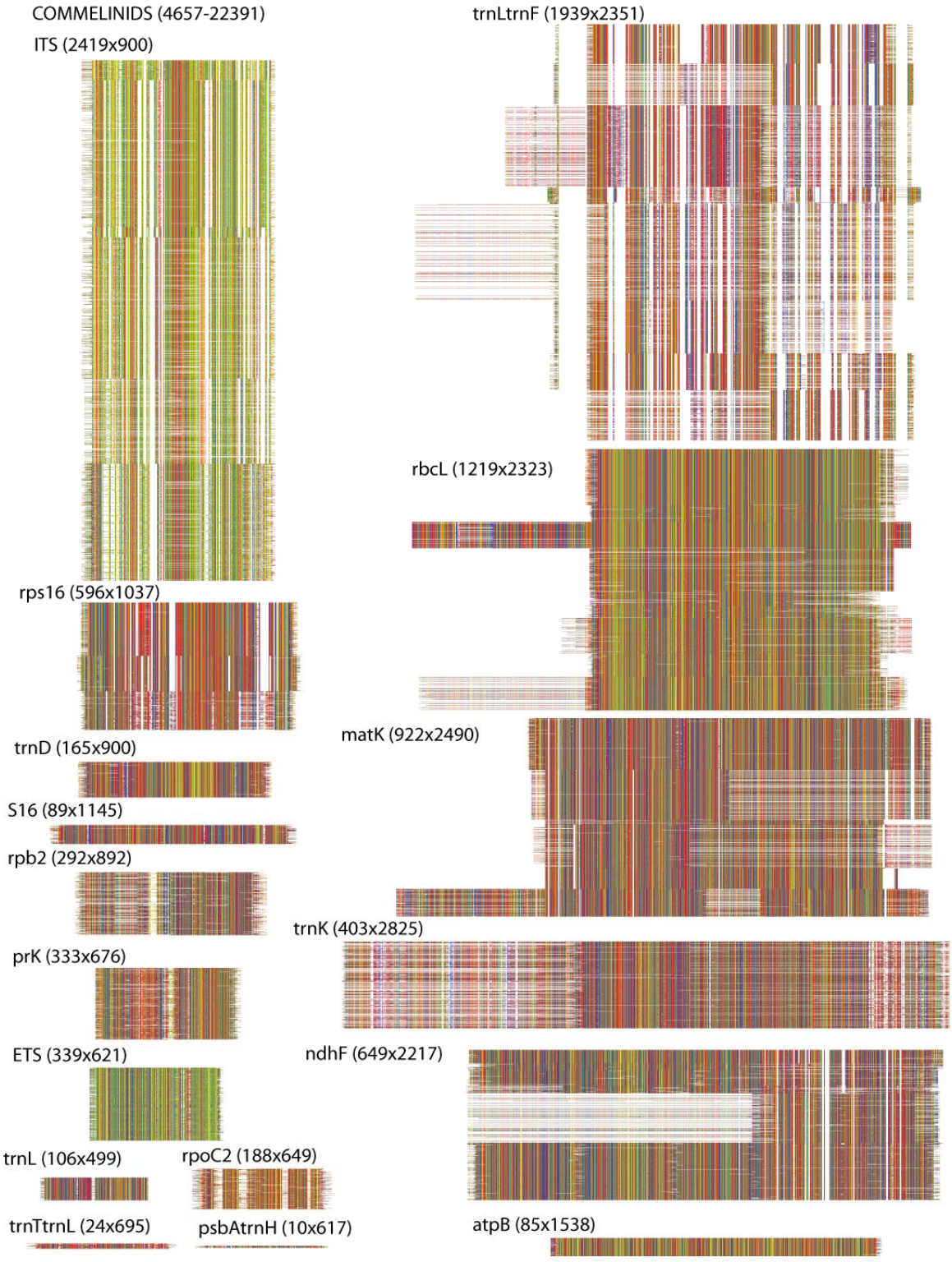

Fig. S2. Visualization of individual alignments combined in profile alignment for phylogenetic analysis of Commelinidae.

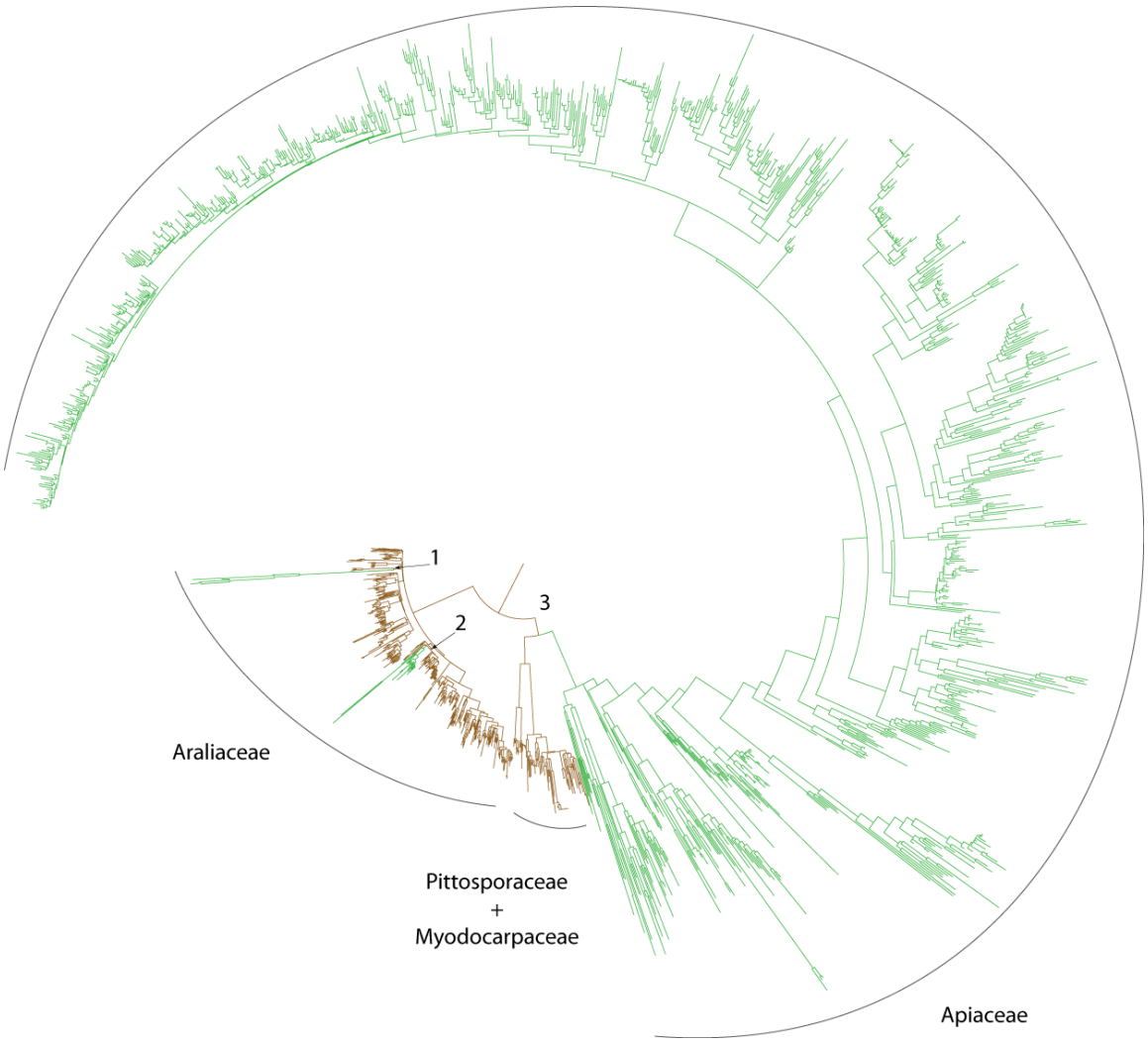

Fig. S3. Apiales phylogeny with major clades labeled, showing the location of contrasts 1-3 in Table 1.

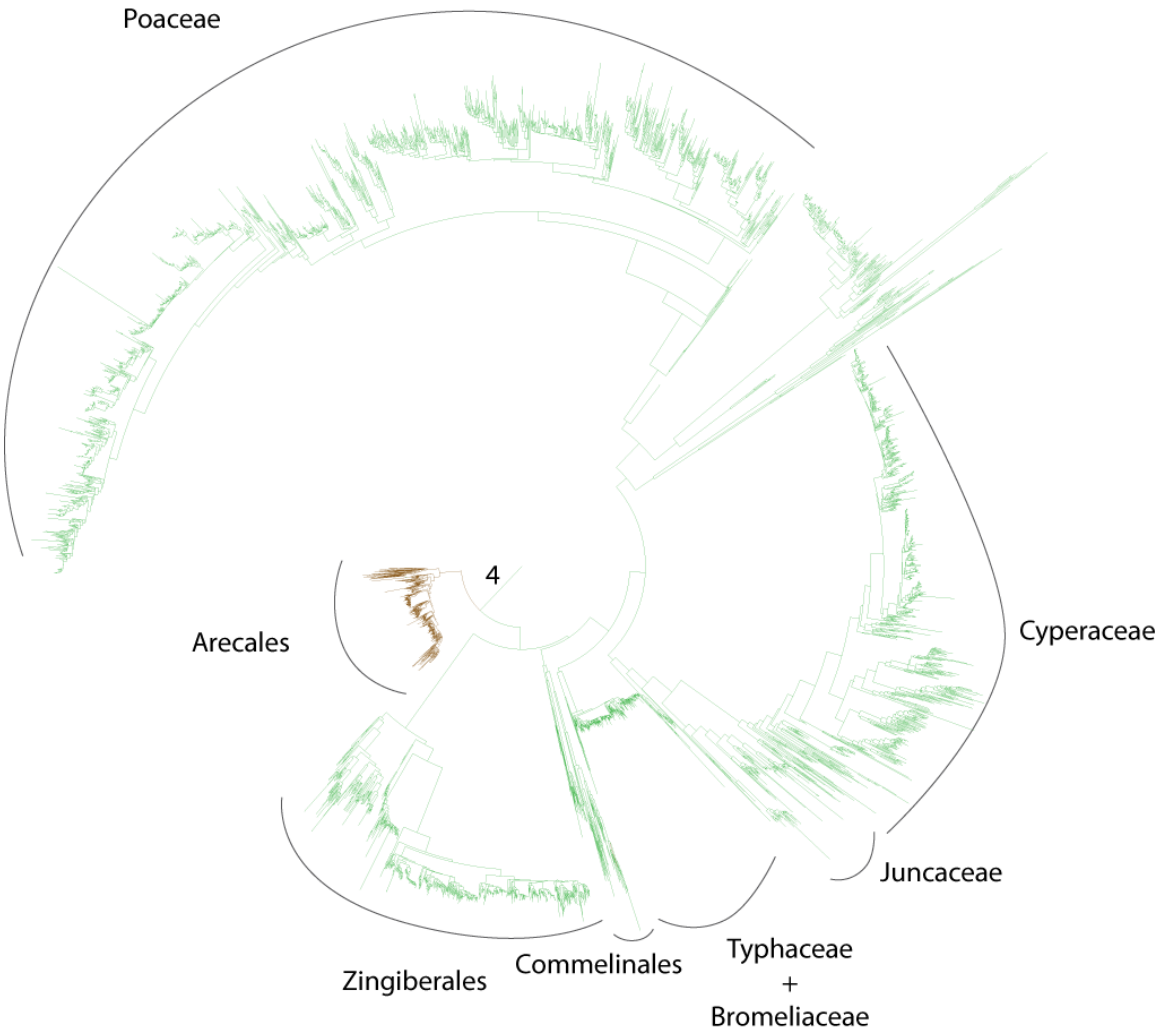

Fig. S4. Commelinidae phylogeny with major clades labeled, showing the location of contrast 4 in Table 1.

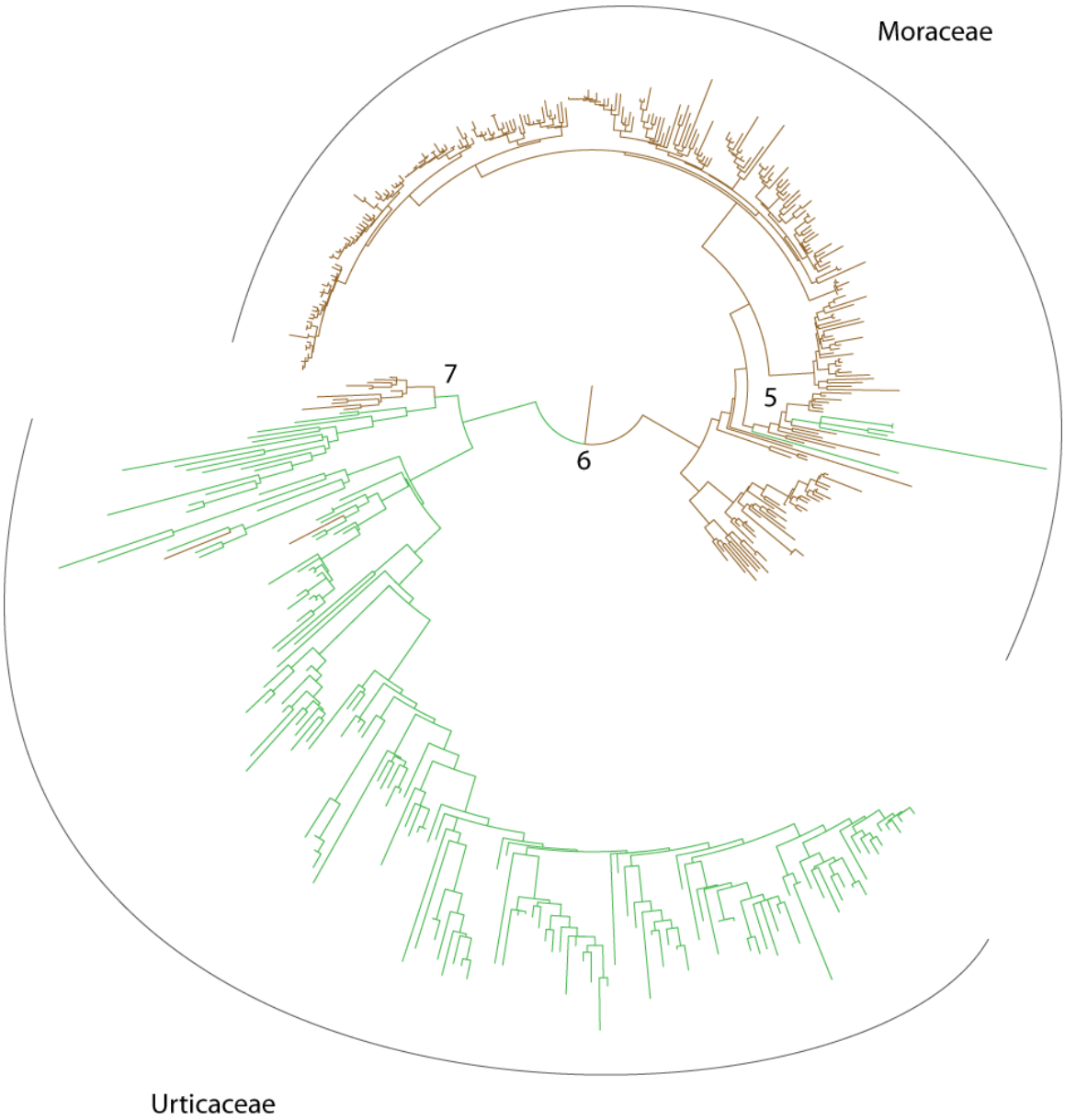

Fig. S5. Moraceae+Urticaceae phylogeny with major clades labeled, showing the location of contrasts 5-7 in Table 1.

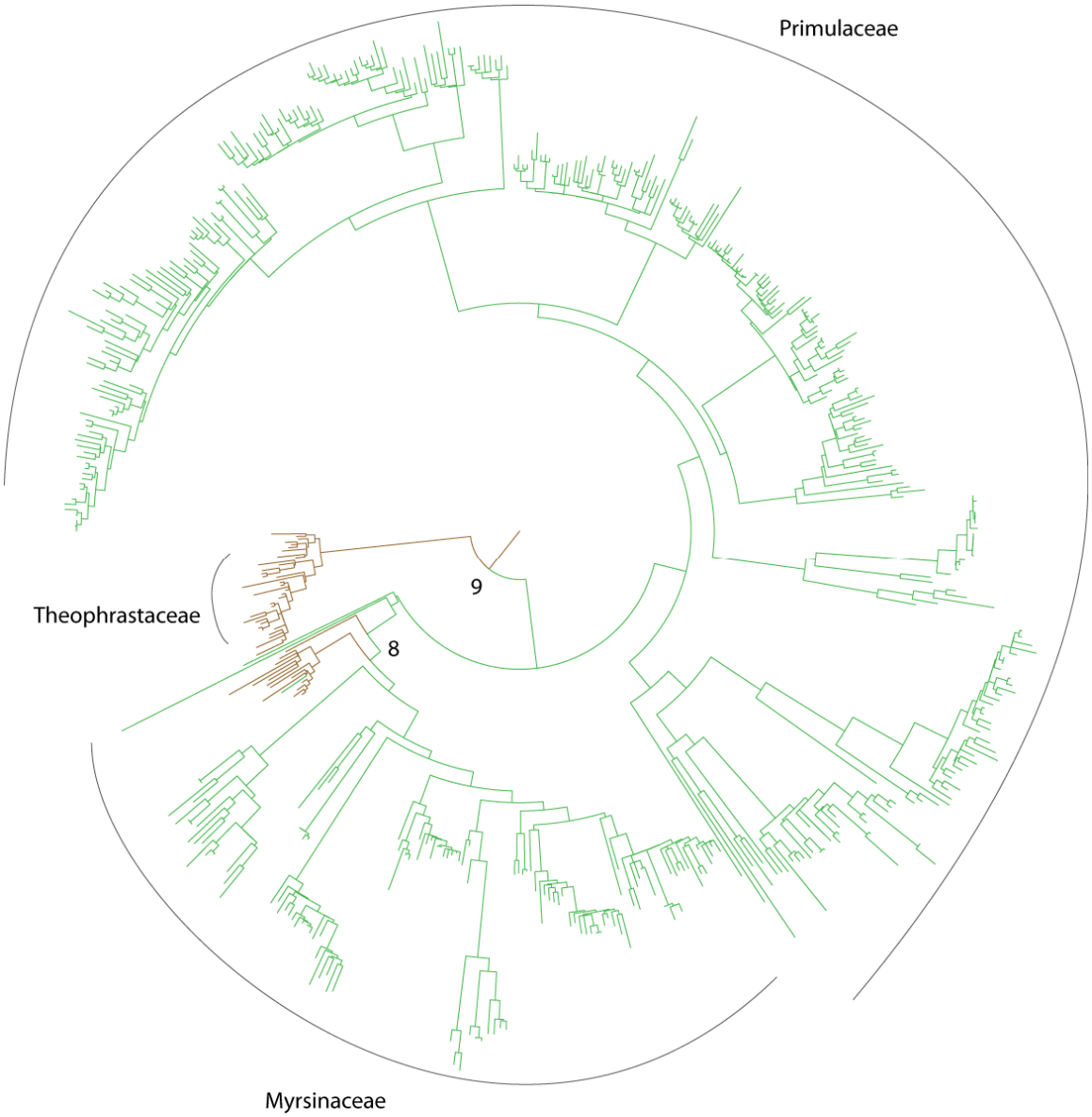

Fig. S6. Primulales phylogeny with major clades labeled, showing the location of contrasts 8 and 9 in Table 1.

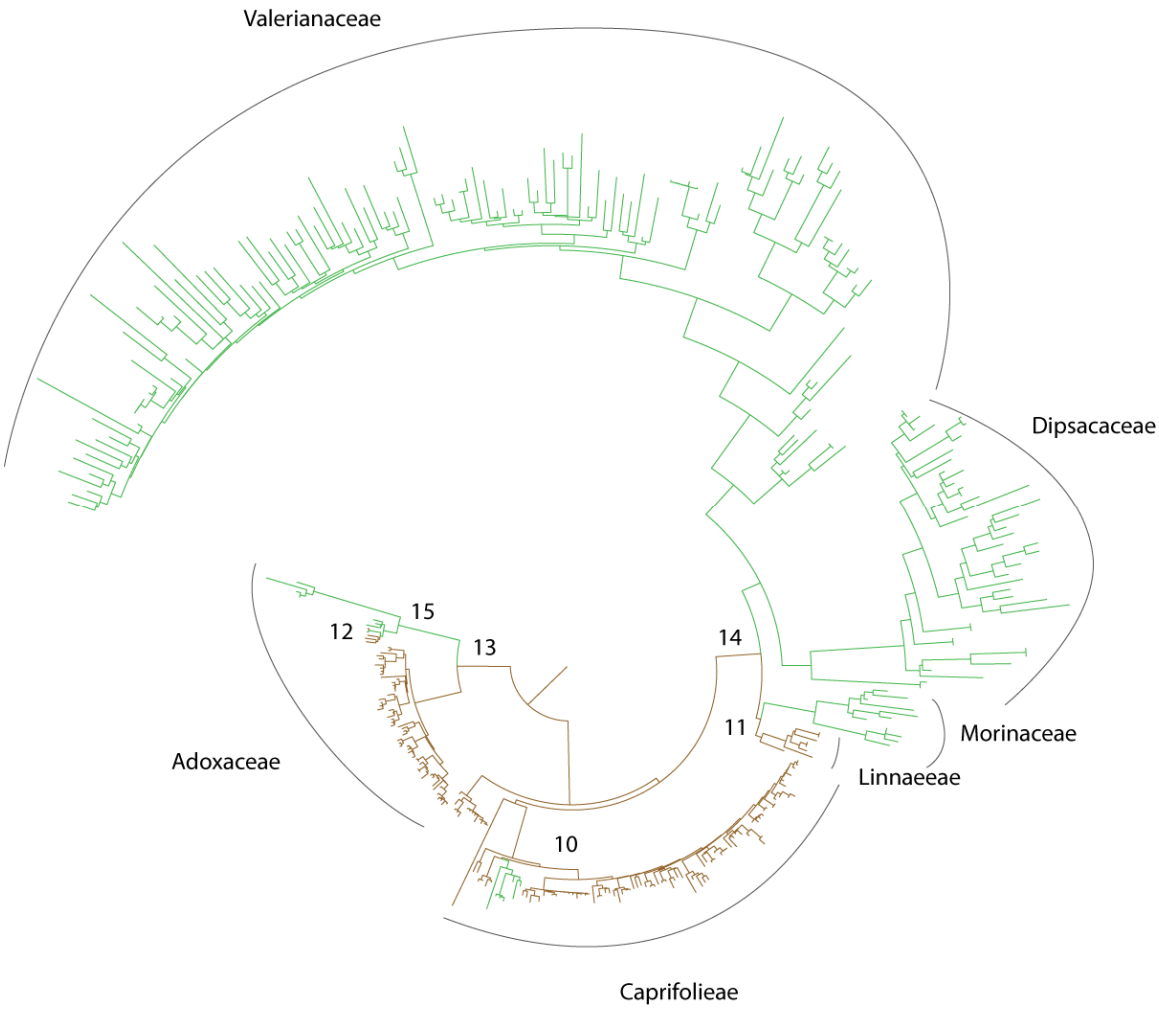

Fig. S7. Dipsacales phylogeny with major clades labeled, showing the location of contrasts 10-13 and alternatives 14 and 15 in Table 1.

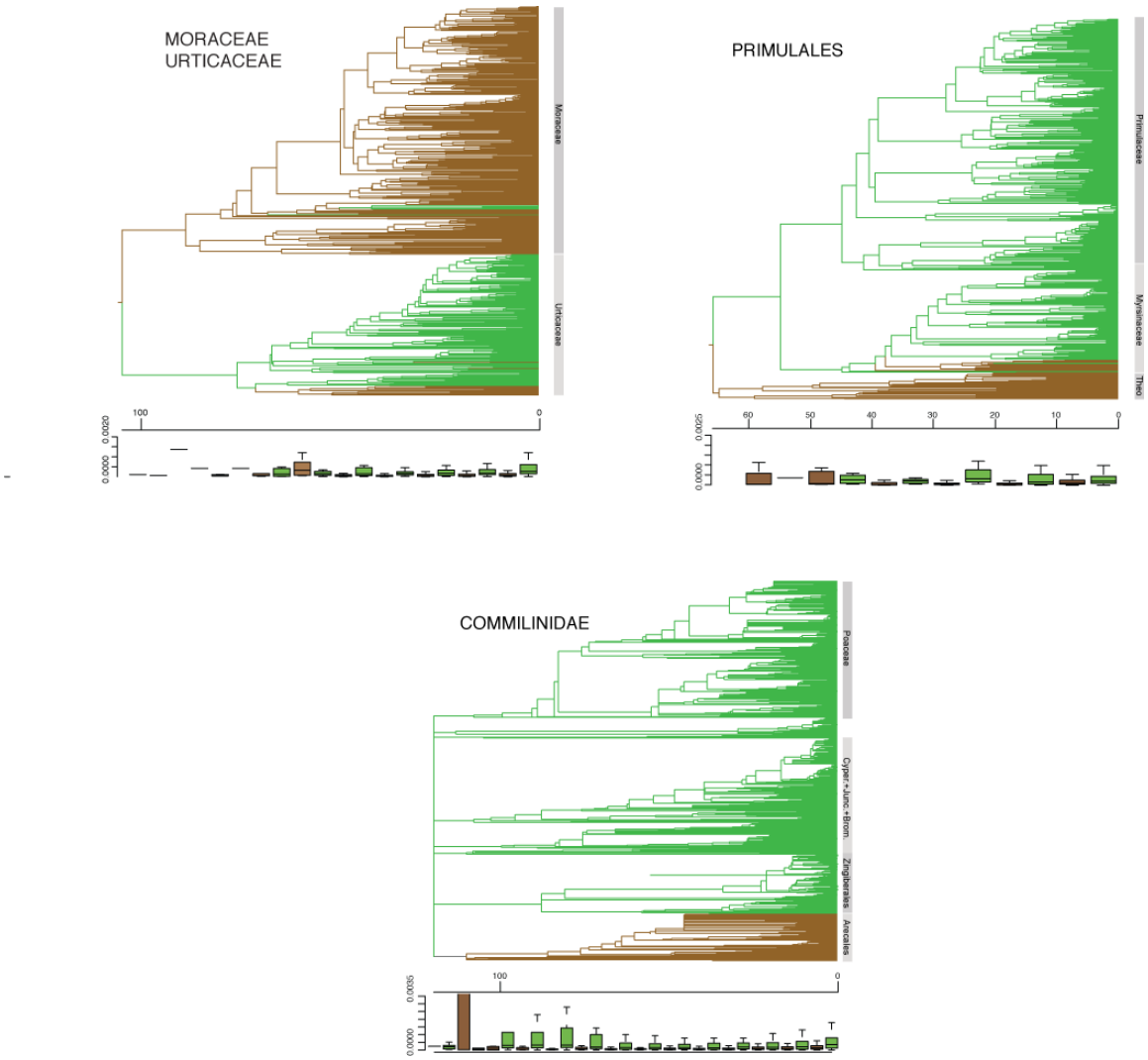

Fig. S8. Dated phylogenies for Moraceae+Urticaceae, Primulales, and Commelinidae with substitutions/site/million years plotted for 10-million year intervals through the life of the clade. Branch colors represent inferred life history states (brown for trees/shrubs; green for herbs). Box-plots as in Fig. 1.