Patterns of floral and chromosome evolution in Onagraceae: a fossil-calibrated supermatrix analysis

William A. Freyman[a,*]

[a] *Jepson Herbarium and Department of Integrative Biology, University of California, Berkeley*

Abstract

Onagraceae is a cosmopolitan plant family in the order Myrtales comprised of 22 genera and 657 species. I used a GenBank data-mining approach to construct an 11 gene supermatrix of 521 taxa in Onagraceae and Lythraceae. Maximum likelihood and Bayesian inference phylogenetic analyses were performed, and divergence time estimates were calibrated using 5 fossils. Chromosome number evolution was estimated using probabilistic models, and ancestral states of petal color and petal number were reconstructed using maximum likelihood. Correlated evolution between petal color and petal number was tested using Bayesian stochastic character mapping. The monophyly of all major clades in Onagraceae was supported, and Onagraceae was estimated to have diverged from Lythraceae 109 Mya. The base chromosome number of Onagraceae was inferred to be $x = 5$. Petal color and petal number were found to follow a pattern of correlated evolution ($p = 0.00$), suggesting concerted shifts in floral traits may play an important role shaping diversification of Onagraceae.

1. Introduction

Questions: are the taxonomic groups described in Wagner et al. (2007) monophyletic? When did the major clades diverge? What are the patterns of petal color and petal number evolution, and are their evolution correlated? What is

[*] Corresponding author

Email address: freyman@berkeley.edu (William A. Freyman)

the pattern of chromosome number evolution?

2. Methods

Supermatrix assembly. Lythraceae was selected to use as an outgroup since previous molecular phylogenetic analyses place it sister to Onagraceae (Conti et al., 1997; Sytsma et al., 2004). I downloaded all DNA sequences from GenBank release 200 PLN division and performed an exhaustive all-by-all BLASTn (Camacho et al., 2009) comparison of sequences in Onagraceae and Lythraceae. Using a BLASTn e-value of 1.0×10^{-10} threshold and a sequence length percent similarity cutoff of 0.5, I constructed clusters of putative homologs using a single-linkage hierarchical clustering algorithm. Subspecies names were removed from all sequences, and all but one sequence of each species was pruned from each cluster. Clusters that were not phylogenetically informative ($<$ 4 taxa) were discarded, and each cluster was aligned using MUSCLE (Edgar, 2004). The alignments were concatenated by species, and any species that was not present in at least two clusters was removed from the supermatrix. The program written to data-mine GenBank and assemble the supermatrix is available as the open source Python module SUMAC (Freyman, 2014). SUMAC can be used to assemble supermatrices for any taxonomic group recognized in GenBank, and is optimized to run on multicore processors and clusters by utilizing multiple parallel processes.

Phylogenetic analyses. Maximum likelihood (ML) analyses were performed with RAxML-HPC (Stamatakis, 2014) on the CIPRES Scientific Gateway (Miller et al., 2010) using the rapid bootstrap heuristic and the GTRCAT nucleotide substitution model. I used the ML tree to select 15 taxa phylogenetically widely distributed in Lythraceae to act as outgroup for the divergence time analysis; all other members of Lythraceae were subsequently removed from the supermatrix. Bayesian estimates of divergence times were inferred using BEAST v1.8 (Drummond and Rambaut, 2007; Suchard and Rambaut, 2009) on CIPRES and calibrated with five fossils identified with morphological synapomorphies (Table

Group	Age (Mya)	Prior Distribution	Mean	SD	Offset	Reference
Circaea (Onagraceae)	12	lognormal	0.0	2.0	12	(Grímsson et al., 2012)
Epilobium (Onagraceae)	12	lognormal	0.0	2.0	12	(Grímsson et al., 2012)
S. Pacific *Fuschia* (Onagraceae)	23	lognormal	0.0	1.0	23	(Lee et al., 2013)
Ludwigia (Onagraceae)	Paleocene	normal	60.0	3.0	-	(Zhi-Chen et al., 2004)
Lythraceae	82	lognormal	0.0	2.0	82	(Graham, 2013)

Table 1: Fossils used as priors in the Bayesian divergence time analysis.

1). The *Ludwigia* fossil pollen was dated broadly to the Paleocene (Grímsson et al., 2012), so I set the prior to a normal distribution with a wide standard deviation to cover the entire time period. For all other calibration points I used a lognormal prior distribution with the offset (the minimum age of the node) corresponding to the fossil age. The BEAST analysis utilized the GTR+Γ nucleotide substitution model with a relaxed molecular clock (uncorrelated lognormal model) and a Yule process tree prior. The Markov Chain Monte Carlo (MCMC) was run for 100 million generations, sampling every 10 thousand generations. Tracer v1.6 (Rambaut et al., 2013) was used to assess the MCMC output for parameter convergence and ensure that the effective sample size for all parameters was above 200. The first 1000 trees were discarded as burn-in, and the remaining 9000 trees were summarized as a maximum clade credibility (MCC) tree with mean divergence times.

Character state reconstruction. I scored six characters, including chromosome number, floral merosity, petal color, and self-compatibility/incompatibility. Character data was assembled from the comprehensive Wagner et al. (2007) Onagraceae monograph. Ancestral chromosome numbers were inferred using maximum likelihood and Bayesian methods as implemented in ChromEvol 2.0 (Glick and Mayrose, 2014). Eight different models of chromosome evolution were fit to the Bayesian MCC phylogeny using ChromEvol, and the best fit model was selected using Akaike's information criterion (Akaike, 1981). Ancestral character state reconstructions of petal number and petal color were performed using Mesquite v2.75 (Maddison and Maddison, 2011) over the Bayesian MCC tree.

Characters were treated as unordered categorical data, and optimized using maximum likelihood with the Markov k-state 1 parameter (Mk1) model (Lewis, 2001). Additionally, Bayesian stochastic character mapping (Huelsenbeck et al., 2003) was used to test whether petal color and petal number covaried over the phylogeny. Models of petal color and petal number evolution were configured in the program SIMMAP v1.5 (Bollback, 2006) using unordered states, gamma rate priors, and equal state bias priors. The correlation analysis used SIMMAP's default predictive sampling configuration and calculated the following correlation statistics: D the overall association between the two characters, and d_{ij} the association between the individual states of each character (Huelsenbeck et al., 2003).

3. Results

Supermatrix assembly. SUMAC evaluated 5571 Onagraceae and 2832 Lythraceae nucleotide sequences to construct the supermatrix. The completed supermatrix consisted of 11 clusters of homologous sequences (Table 2). As used in the maximum likelihood analyses (before pruning the number of outgroup taxa), the supermatrix contained 521 taxa, was 31862 nucleotides long, and contained 93.0% missing data.

Phylogeny and divergence time estimates. The topologies of the ML and Bayesian phylogenies were identical for all major clades within Onagraceae, so only the Bayesian MCC tree (Figures 1 and 2) is shown here. All Onagraceae genera described in Wagner et al. (2007) were recovered as monophyletic clades with posterior probabilities of > 0.95 except for sister genera *Neoholmgrenia* and *Camissoniopsis* (posterior = 0.31) (Figure 1). Onagraceae was found to diverge from Lythraceae at 109 Mya (Figure 2). Divergence time estimates of other major clades and 95% highest posterior density (HPD) intervals can be seen in Table 3.

Character evolution. blah blah

DNA Region	# of Taxa	Aligned Length	Missing data (%)	Taxon Coverage Density
ITS	453	1746	13.2	0.87
trnL	234	1429	55.2	0.45
rpl16	91	1414	82.6	0.17
rbcL	77	1474	85.2	0.15
rps16	74	1016	85.8	0.14
rbcL	64	1310	87.7	0.12
PgiC2	47	4028	91.0	0.09
matK	37	921	92.9	0.07
ndhF	37	2063	92.9	0.07
pgiC	26	14709	95.0	0.05
R5	18	3129	96.6	0.03

Table 2: Clusters of homologous sequences used to assemble the supermatrix.

Clade	Mean Age (Mya)	95% HPD Min	95% HPD Max
Onagraceae / Lythraceae	109	88	131
Ludwigia	97	76	118
Hauya	49	35	64
Circaea / Fuchshia	37	28	47
Lopezia	71	55	68
Gongylocarpus	60	45	77
Epilobium	49	38	60
Chamerion	47	36	57
Xylonagra	43	33	52
Clarkia	40	32	48
Terapteron	19	10	29
Camissoniopsis / Neoholmgrenia	14	5	23
Eremothera / Camissonia	24	16	33
Taraxia	30	22	38
Chylismiella / Gayophytum	20	10	30
Eulobus	26	19	34
Chylismia / Oenothera	25	18	31

Table 3: Bayesian divergence time estimates of major clades.

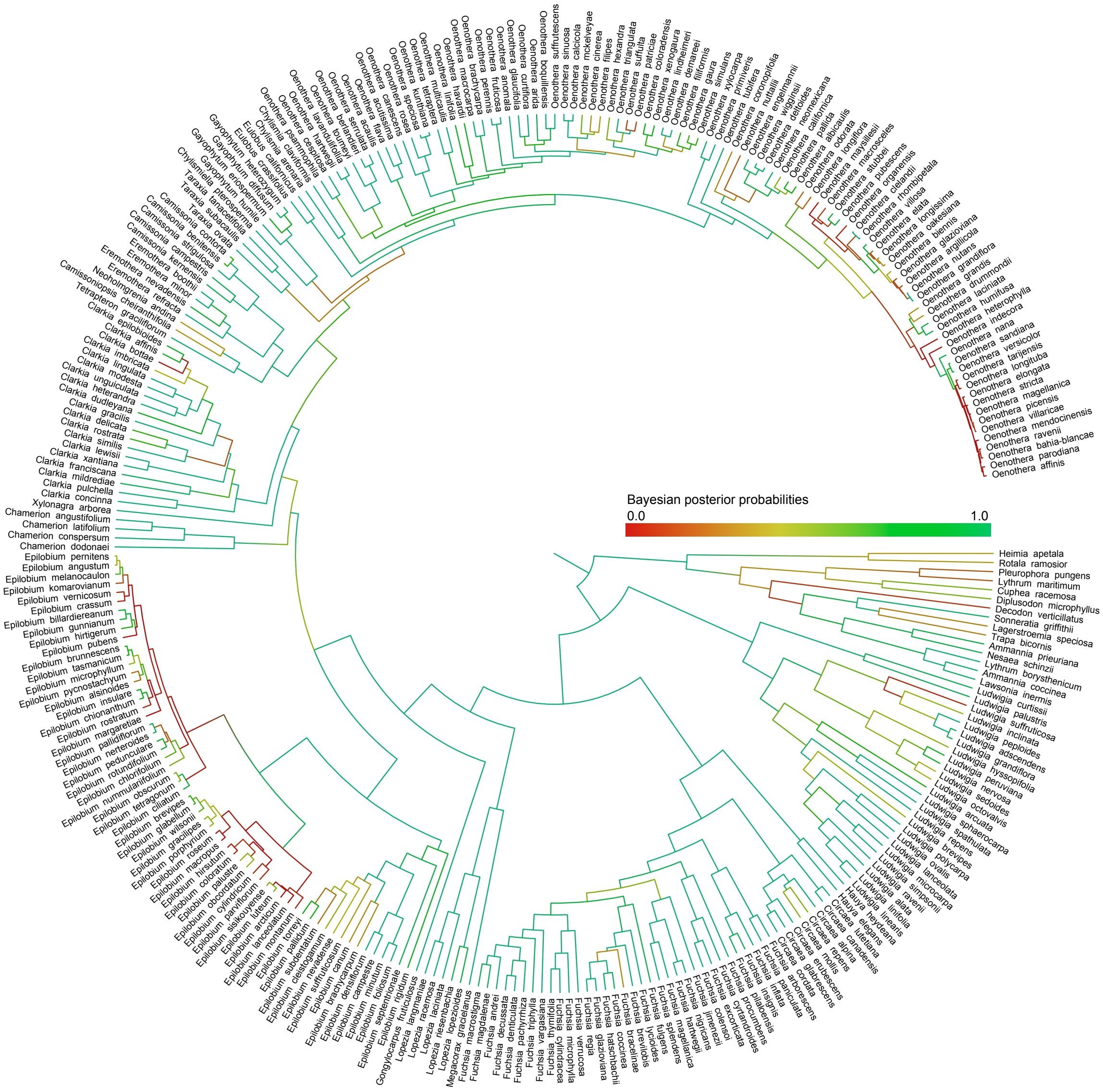

Figure 1: Bayesian maximum clade credibility phylogeny of 280 Onagraceae taxa and 15 Lythraceae taxa. Estimated posterior probabilities close to 1.0 are shown in green. All genera described in Wagner et al. (2007) were found to be monophyletic with posterior probabilities of > 0.95 except for sister genera *Neoholmgrenia* and *Camissoniopsis* (posterior = 0.31).

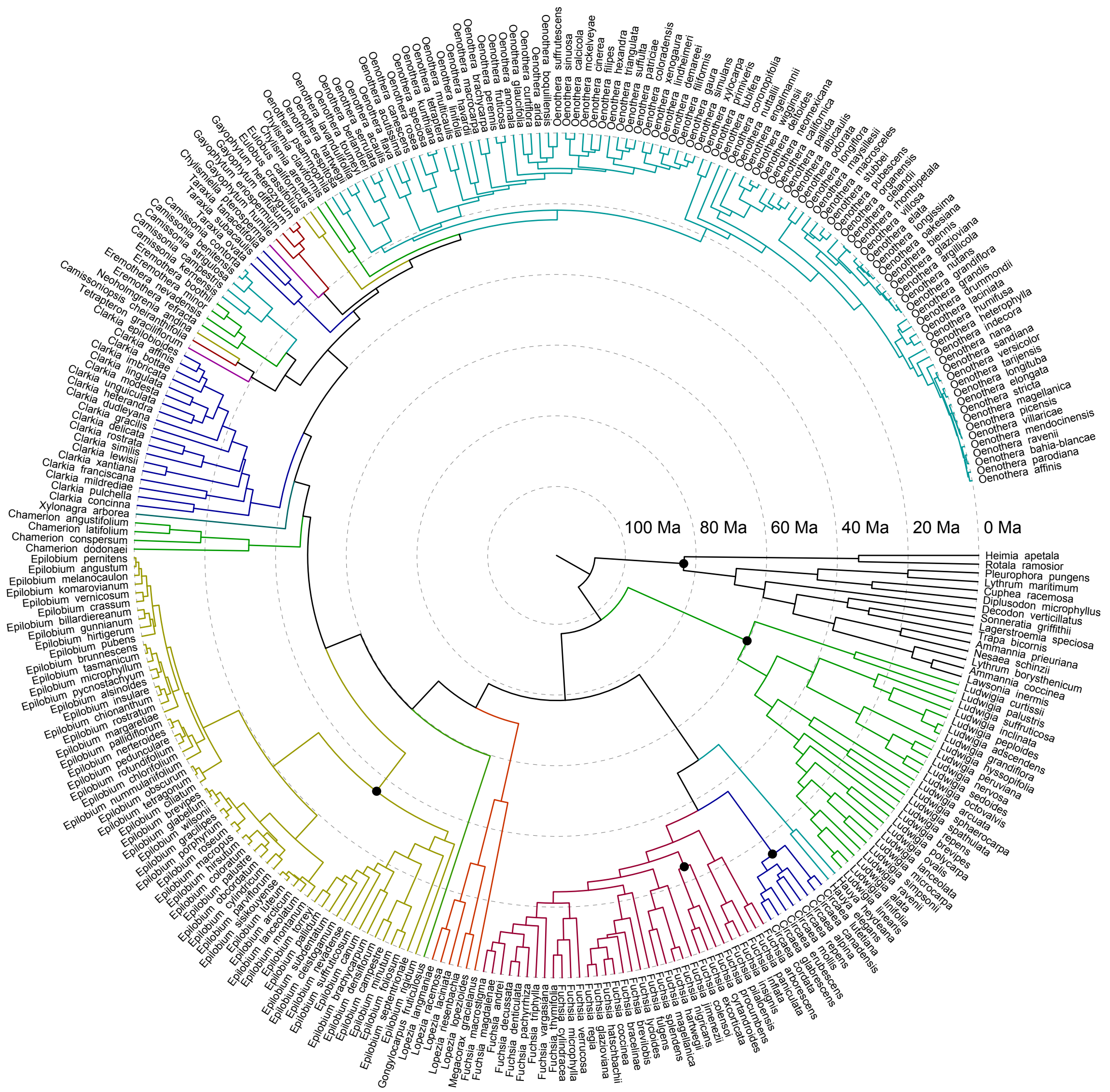

Figure 2: Bayesian chronogram of 280 Onagraceae taxa and 15 Lythraceae taxa. Approximate positions of fossil calibration points are shown as black circles. All genera described in Wagner et al. (2007) are colored, and their divergence time estimates and %95 HPD intervals can be seen in Table 3.

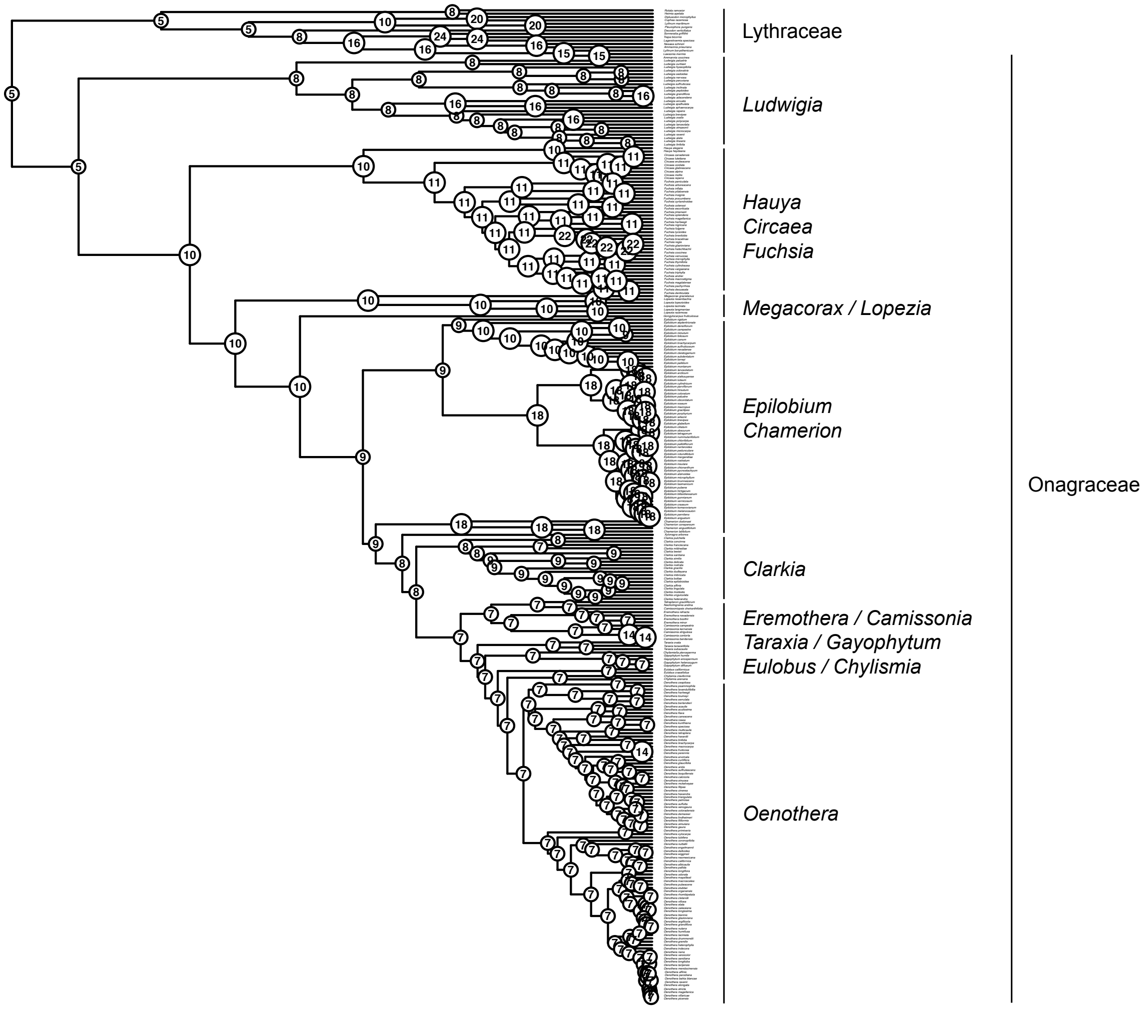

Figure 3: Maximum likelihood estimates of ancestral chromosome numbers over the Bayesian MCC phylogeny using the 4 parameter (λ, δ, μ, ρ) model of chromosome evolution selected by Akaike's information criterion. The base number of Onagraceae is inferred to be $x = 5$.

	Number of Petals			
	2	4	5	6
Petal Color				
Pink	-0.005	.011	*ns*	*ns*
Yellow	-0.008	.021	*ns*	*ns*
White	0.008	0.013	-0.006	-0.006
Green	*ns*	-0.011	*ns*	*ns*
Red	*ns*	*ns*	*ns*	*ns*

Table 4: Test statistics for the correlation between flower color evolution and petal number evolution. d values are shown for the pairwise comparison of states with $p < 0.01$. ns indicates no significant association. The overall D value was 0.263 ($p = 0.00$).

4. Conclusion

blah blah

References

Akaike, H., 1981. Likelihood of a model and information criteria. Journal of econometrics 16, 3–14.

Bollback, J.P., 2006. Simmap: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7, 88.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Conti, E., Litt, A., Wilson, P.G., Graham, S.A., Briggs, B.G., Johnson, L., Sytsma, K.J., 1997. Interfamilial relationships in Myrtales: molecular phylogeny and patterns of morphological evolution. Systematic Botany 22, 629–647.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7, 214.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797.

Freyman, W.A., 2014. SUMAC: Supermatrix Constructor. `https://github.com/wf8/sumac`.

Glick, L., Mayrose, I., 2014. Chromevol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. Molecular biology and evolution , msu122.

Graham, S.A., 2013. Fossil records in the Lythraceae. The Botanical Review 79, 48–145.

Grímsson, F., Zetter, R., Leng, Q., 2012. Diverse fossil Onagraceae pollen from a Miocene palynoflora of north-east China: early steps in resolving the phytogeographic history of the family. Plant Systematics and Evolution 298, 671–687.

Huelsenbeck, J.P., Nielsen, R., Bollback, J.P., 2003. Stochastic Mapping of Morphological Characters. Systematic Biology 52, 131–158.

Lee, D.E., Conran, J.G., Bannister, J.M., Kaulfuss, U., Mildenhall, D.C., 2013. A fossil *Fuchsia* (Onagraceae) flower and an anther mass with in situ pollen from the early Miocene of New Zealand. American Journal of Botany 100, 2052–2065.

Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50, 913–925.

Maddison, W.P., Maddison, D., 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. `http://mesquiteproject.org`.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees, in: Gateway Computing Environments Workshop (GCE), 2010, IEEE. pp. 1–8.

Rambaut, A., Suchard, M., Drummond, A.J., 2013. Tracer v1.6. `http://tree.bio.ed.ac.uk/software/tracer/`.

Stamatakis, A., 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics .

Suchard, M.A., Rambaut, A., 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25, 1370–1376.

Sytsma, K., Litt, A., Zjhra, M., 2004. Clades, clocks, and continents: Historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern Hemisphere source. International Journal of Plant Sciences 165, S85–S105.

Wagner, W.L., Hoch, P.C., Raven, P.H., 2007. Revised classification of the Onagraceae. Systematic Botany Monographs 83.

Zhi-Chen, S., Wei-Ming, W., Fei, H., 2004. Fossil pollen records of extant angiosperms in China. The Botanical Review 70, 425–458.