New Phytologist Supporting Information

Susy Echeverría-Londoño, Tiina Särkinen, Isabel S Fenton, Sandra Knapp, Andy Purvis

Methods S1: correcting for non-random taxon sampling

We used the maximum clade credibility of the Särkinen *et al.* (2013) phylogeny as the "backbone" to define the general relationships among the main sections of the genus. Monophyletic constraints were used only for well-supported nodes (i.e., posterior probability $> 0.95\%$, see Table S1). Clades with poorly-supported nodes were left unconstrained (i.e., the Torva, African Non-spiny and Erythrotrichum clades). One of the nodes in the Särkinen *et al.* (2013) phylogeny within the Dulcamaroid clade had negative branch lengths; this can occur when a clade is poorly sampled across the posterior distribution (i.e., when there is a low support of a direct ancestor-descendant relationship). Since setting negative branch lengths to zero would create a non-ultrametric tree, we decided to drop all the tips from this weakly supported node from the backbone and include them in the PASTIS analysis as unconstrained. We then assigned each accepted name of *Solanum* to one of the clades or sections which are shown in Table S1, following the taxonomic treatment of *Solanum* species in Solanaceae Source and expert opinion (S. Knapp, pers. comm.).

For each clade or subclade, PASTIS creates an output file in a nexus format, which contains the full set of tree constraints ready to be executed in MrBayes. Therefore, posterior distributions of phylogenies for each clade were then inferred in MrBayes 3.2.3 (Ronquist & Huelsenbeck, 2003) using a relaxed clock model (independent branch rates – igr prior), with the default (exponential) prior on the distribution of branching rates. Two species from the closest sister clade were constrained as outgroups in each *Solanum* clade. Four chains, four independent runs and 400 million generations were run for clades with more than 150 species and 100 million generations for the rest of the clades. MrBayes was run using the Cipres gateway (Miller *et al.*, 2010)(http://www.phylo.org).

We assessed the convergence, mixing and burn-in of all the parameters for the posterior distribution of each clade by visual examination using Tracer v 1.6.0 (Rambaut *et al.*, 2009). To create the complete species phylogeny of *Solanum*, a sample from the posterior distribution of the phylogeny or subtree of each group of *Solanum* was selected randomly and then grafted to the backbone following the methodology in Jetz & Fine (2012). To do that, we first rescaled the depth of each subtree to 1. Then we computed the proportional depth of the crown group of each subtree, and grafted the subtree into the backbone at its original position keeping its ultrametricity (see Figure S1 for a graphic explanation). This procedure was repeated 100 times for all subtrees (each representing a group of

Solanum) until a distribution of 100 complete species phylogenies (i.e., containing nearly all described species, 1169 species in total) was produced.

Methods S2: diversification analyses

A macroevolutionary cohort analysis sensu (Rabosky *et al.*, 2014) and (Shi & Rabosky, 2015) was performed across all the 100 BAMM runs to visualise complex mixtures of dynamics of diversification. This approach calculates the pairwise probability that any two lineages share a common macroevolutionary regime. For a given sample from the posterior distributions, a pair of lineages is assigned a value of 1 when the species inherit a common rate regime and a value of 0 when the rate dynamics are completely decoupled (see Figure S2). The mean of these values is then calculated over all the posterior distribution samples in each of the 100 runs of BAMM. In this case, an average value of 0.1 between a pair of lineages implies that in 10% of the samples across the posterior distributions these two lineages support a common macroevolutionary rate.

BAMM sensitivity analysis

(1) Reliability of diversification rate estimates: Using an incorrect likelihood function could produce unreliable estimates of speciation and extinction rates and therefore, an unreliable number and location of shifts. To demonstrate the accuracy of the estimation and location of rates inferred from BAMM, it is necessary to perform an extensive series of simulations under different diversification models, something which is outside the scope of this analysis. However, to corroborate the results drawn from the BAMM analysis, we implemented two alternative approaches to estimate rates of speciation and extinction and identify shifts of diversification through time and across branches. With the first approach, we tested whether there is evidence of an episodic tree-wide increase in diversification rates through time using the R package TESS (Höhna, 2015). This package implements an approach which fits a birth-death model with episodically varying rates (i.e., identifies discrete tree-wide changes in speciation and extinction rates) assuming that the diversification among lineages at any point in time is constant (Höhna, 2015). As in other approaches that model diversification-rate through time such as TreePar (Stadler, 2011), TESS divides the tree into equal time intervals and tests whether there are significant changes in speciation and extinction rates among these intervals. For comparison purposes, we first performed this analysis using the Särkinen *et al.* (2013) phylogeny and then using the 100 complete species trees of *Solanum* obtained in PASTIS. For both datasets, we ran the function "tess.analysis" which uses a reversible-jump MCMC algorithm to estimate the number and magnitude of rate shifts. The number of events is assumed to follow a Poisson distribution and the estimates of speciation and extinction are assumed to follow a lognormal distribution with a fixed mean calculated empirically by an initial MCMC analysis

under a constant birth-death model. All the analyses were run until the effective sample size reached 500 or the time reached a maximum of 24 hours.

The second approach we tested was the RevBayes program (Höhna *et al.*, 2016), which estimates species-specific rates of speciation and extinction and locates branches with significant shifts in diversification. Unlike BAMM, RevBayes does not model the rates of diversification from a continuous distribution directly, but instead it divides the probability distribution into discrete rate categories. Then within each of the N quantiles of the distribution, in this case a lognormal, Revbayes integrates over all possible rates of speciation. The mean of this lognormal distribution is fixed to represent the mean of the expected diversification rate given each tree, which is equal to ln(ln(N taxa)/age)), with a fixed standard deviation of (0.587405 x 2) to represent two orders of magnitude of variance in the rates. In this analysis, extinction rates were assumed to be equal in all the rate categories and the number of rate categories was set to 10. Without any prior information about the rate-shift events, we used a conservative prior distribution of the number of shifts defined by an exponential distribution with mean equal to one, giving the highest prior probability to a zero-shift model. We ran a pre-burn-in analysis using 10,000 generations to obtain starting values from the posterior distribution and improve the mixing of the MCMC analysis. We then ran the analysis using 100,000 generations or more until convergence was reached. Every 200 generations a tree was printed with the average parameters values of speciation and extinction at the branches and nodes. At the end of the analysis, all the trees were used to calculate the posterior distribution of the rates at the branches through a maximum a posteriori tree. Due to computational resource availability, we ran this analysis using only the first 20 trees from the pool of complete species phylogenies. The results from these two approaches were contrasted with those obtained using the BAMM analysis.

By using the *Solanum* phylogeny from Särkinen *et al.* (2013) and assuming an even sampling of species throughout the phylogeny, the TESS approach identifies significant shifts in diversification (Figure S5). However, when the distribution of complete species phylogenies of *Solanum* produced in PASTIS was used, the signal of diversification shifts through time is no longer supported (Figure S6).

Overall, the analysis of diversification rates in RevBayes showed similar results to those from the BAMM analysis (i.e., a strong signal of a diversification shift in the node supporting the crown group of the Old world clade, see Figure S3 and S4). However, the RevBayes analysis revealed an overall greater heterogeneity of rates compared with BAMM and identified other signal of diversification in groups such as Torva which were not found in BAMM but were found in MEDUSA (see Figure S3 and S4)

(2) Effects of the model prior: BAMM assumes that the expected number of shifts of diversification follows

a Poisson prior distribution with an exponentially distributed hyperprior. This distribution can be simplified as a geometric distribution with mean γ (i.e., the expected number of shifts). According to (Mitchell & Rabosky, 2016) the prior distribution automatically set in BAMM is a conservative way to define the number-of- shifts prior since the zero-shift model is the most likely outcome. To test the sensitivity of the number of shifts found in BAMM to the prior distribution, we ran the analysis using five different prior expectations of $\gamma = 0.5, 1, 2, 10, 100$ using the original (Särkinen *et al.*, 2013) *Solanum* tree (due to computational limitations). The value of γ defines the shape of the geometric distribution with small values (e.g., 0.5) defining a strong prior with a skewed probability distribution towards a zero-shift model. In contrast, large values of γ (e.g., 100) represent liberal priors with a relatively flat probability distribution. Each γ treatment was run for 4 million generations and the first 20% of the samples were discarded as burn-in. We then plotted the marginal posterior distribution of the number of shifts obtained in each γ treatment and the prior distribution for each tree. A significant change in the marginal posterior distribution with different values of γ would demonstrate a strong sensitivity of the BAMM results to the prior distribution.

The number of diversification shifts found in the BAMM analysis is largely robust to the defined prior distribution. Figure S7–S10 show that across all 100 BAMM runs, the distribution of estimated shifts (posterior samples) differs from the distribution of the expected number of shifts (prior). Although the prior distribution (γ=1) applied in this analysis was strong and conservative (i.e., the zero-shift model was set to be the most likely outcome), the zero-shift model was never sampled in the posterior for any of the BAMM runs showing an overwhelming evidence of the heterogeneity of shifts found in this analysis. Moreover, the number of shifts found in the Särkinen *et al.* (2013) phylogeny was not sensitive to different priors of diversification rates (i.e., different values of γ, 0.5,1,2,10,100) as shown in Figure S11.

Geographical patterns of diversification

To determine which regions have accumulated and are accumulating a higher or lower number of *Solanum* lineages across the globe, the average diversification rates per species were displayed to a 1 x 1 degree grid scale map.

84,606 botanical records from 1005 *Solanum* species were extracted from the Solanaceae Source database `http://solanaceaesource.org/` on 13 March 2016; and 34,462 records from 215 species were compiled from the Australian National Herbarium, accessed through the Atlas of Living Australia (ALA) website `http://www.ala.org.au/`, on 16 April 2016. We applied a series of quality control filters to discard the following:

1. Records with latitude and longitude coordinates of 0° , 0° .

2. Records matching coordinates from major herbaria or political centroids at all administrative divisions

extracted from Edwards *et al.* (2015).

3. Any records with the described country conflicting with the country extracted from their coordinates. This step was performed by overlaying the records with a global administrative polygons extracted from the R package "rworldmap" (South, 2011) using the function *"over"* from the R package "sp" v 1.1-0 (Pebesma & Bivand, 2005).

4. Records with identical coordinates for a given species

5. Records from non-native, cultivated or naturalized species

Taxonomic names were updated for all records to correct for synonymy using as a reference the accepted name list from http://solanaceaesource.org/. After cleaning, we were left with 64,826 unique records for 1,096 taxa.

Species occurrences were then converted into a presence-absence matrix of a 110 x 110 km equal area grid using the function "lets.presab.points" in the R package letsR v 2.1 (Vilela & Villalobos, 2015). Using the mean species-specific rates, we estimated the mean assemblage diversification rates as the geometric mean of all species' rates present in a grid cell. We also computed a weighted version of this, dividing the mean species-specific diversification rates by the inverse of their range size — log of the area (sqm) occupied by each species, to correct for the overestimation of rates in an area as a result of the occurrence of widespread species.

Sections	Crown age (Mya)	PP	Expected number	Species in phylogeny	% in phylogeny	Undersampled
Acanthophora	6.0 - 3.6	1	19	11	57.89	
African non spiny	8.6 - 3.1	0.86	17	2	11.76	*
Anarrhichomenum	4.0 - 0.8	1	11	2	18.18	*
Androceras-Crinitum	7.7 - 5.2	1	36	17	47.22	
Archaesolanum	6.1 - 2.7	1	8	8	100.00	
Asterophorum		1	2	1	50.00	
Bahamense	3.8 - 0.8	1	4	1	25.00	*
Basarthrum	6.4 - 2.2	1	19	4	21.05	*
Brevantherum	10.1 - 5.9	1	95	10	10.53	*
Carolinense	6.7 - 4.1	1	9	4	44.44	
Crotonoides		1	1	1	100.00	
Cyphomandra	7.8 - 4.5	1	49	31	63.27	
Dulcamaroid	9.5 - 5.7	1	44	11	25.00	*
Elaeagnifolium	5.6 - 2.3	1	7	3	42.86	
Erythrotrichum	5.2 - 2.9	0.8	31	11	35.48	
Etuberosum	9.3 - 4.4	1	3	2	66.67	
Gardneri	5.2 - 2.6	1	9	7	77.78	
Geminata	10.1 - 5.9	1	150	10	6.67	*
Lasiocarpa	5.1 - 2.5	1	13	12	92.31	
Mapiriense-Clandestinum	12.2 - 5.8	1	3	2	66.67	
Micracantha	5.5 - 2.6	1	17	7	41.18	
Morelloid	11.5 - 7.8	1	75	17	22.67	*
Multispinum		1	1	1	100.00	
Nemorense	10.3 - 2.4	1	3	3	100.00	
Normania	5.1 - 1.9	1	3	2	66.67	
Old World	6.4 - 4.4	1	296	119	40.20	
Petota	8.5 - 5.9	1	137	45	32.85	*
Pteroidea-Herpystichum	9.0 - 5	1	18	10	55.56	
Regmandra	6.6 - 1.9	1	11	3	27.27	*
Sisymbriifolium	4.6 - 1.3	1	2	2	100.00	
Solanum hieronymi		1	1	1	100.00	
Thelopodium	10.6 - 3	1	3	2	66.67	
Thomasiifolium	6.5 - 3.3	1	7	4	57.14	
Tomato	7.4 - 4.5	1	17	14	82.35	
Torva	4.1 - 2.6	0.84	54	22	40.74	
Valdiviense		1	1	1	100.00	
Wendlandii-Allophyllum	11.2 - 6.2	1	10	4	40.00	

Table S1. Sampling proportions for the main subclades of the *Solanum* phylogeny defined by Särkinen *et al.* (2013). The proportions were calculated dividing the number of species included in the phylogeny by the expected number of extant species within each section or clade. PP = Posterior probability. Undersampled groups are those that have less than 30% of the expected number of species.

Models	lnL	*k*	*d*	*e*	*j*	AIC	Δ AIC
DEC+J M1	-302.53	3	0.017	0	0.003	611.06	0
DEC M1	-303.81	2	0.019	0	–	611.62	0.55
BAYAREALIKE+J M1	-305.92	3	0.013	0.002	0.009	617.84	6.77
DIVALIKE M1	-307.94	2	0.02	0	–	619.89	8.82
DIVALIKE+J M1	-306.96	3	0.019	0	0.003	619.92	8.85
BAYAREALIKE+J M0	-319.65	3	0.006	0.001	0.005	645.31	34.24
DEC+J M0	-320.19	3	0.008	0	0.002	646.39	35.33
DEC M0	-322.08	2	0.009	0	–	648.16	37.09
DIVALIKE+J M0	-327.68	3	0.009	0	0.002	661.36	50.30
DIVALIKE M0	-329.10	2	0.01	0	–	662.21	51.15
BAYAREALIKE M1	-364.11	2	0.017	0.019	–	732.23	121.17
BAYAREALIKE M0	-378.47	2	0.008	0.019	–	760.94	149.88

Table S2. Estimated parameters, log-likelihood and AIC values of the biogeographic models tested in BioGeoBEARS. Models were ranked based on their AIC values. +j models allowed founder-events. M1 models included a matrix that weight the dispersal probability of adjacent areas. lnL= log-likelihood, *k*= number of parameters, *d*=rate of range expansion, *e*=rate of range contraction, *j*= rate of jump dispersals.

Event	Type	Mean	SD	Percentage
Sympatric speciation	within-area speciation	354.6	3.65	80.0
	subset (peripatric speciation)	17.36	4.16	3.9
Dispersal	Range expansions	57.52	1.19	13.0
	Range contractions	0	0	
	Founder events	0	0	
Vicariance	Vicariance	13.99	1.17	3.2
Total events		443.47		100.0

Table S3. Mean number of biogeographic events estimated across the 100 biogeographical stochastic simulations using the DEC M1 model. No range contractions or founder events were estimated since the inclusion of these parameters did not improve the model significantly.

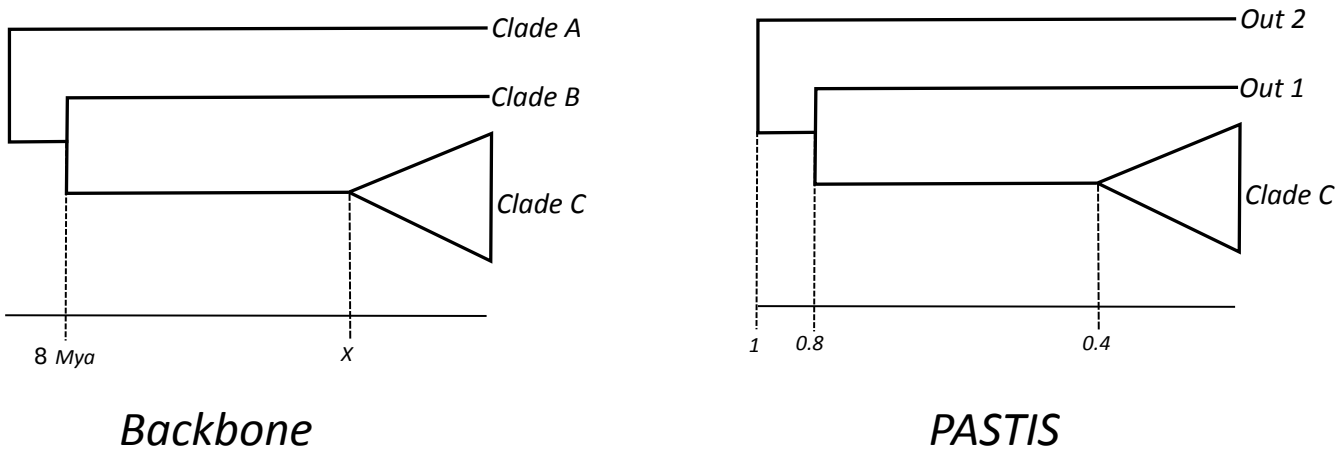

Figure S1. ***Solanum*** **phylogeny backbone-clade grafting** The depth of each PASTIS subtree, which represents the PASTIS runs for each clade/section of *Solanum*, was scaled to 1.0 in order to substitute the ingroup into the backbone tree to replace the single branch that the clade represents. The two outgroups of the clade were dropped and then the clade was grafted into the backbone inferring the depth of the crown group in the backbone from the depth of the stem group of the PASTIS subtree and the depth of the node in the backbone that supports the clade and its outgroup. For instance, if the depth of the node linking the clade C and the clade containing clade C's outgroup is 8 Ma, the crown age of Clade C in the backbone is set to $8 * 0.4/0.8 = 4Ma$.

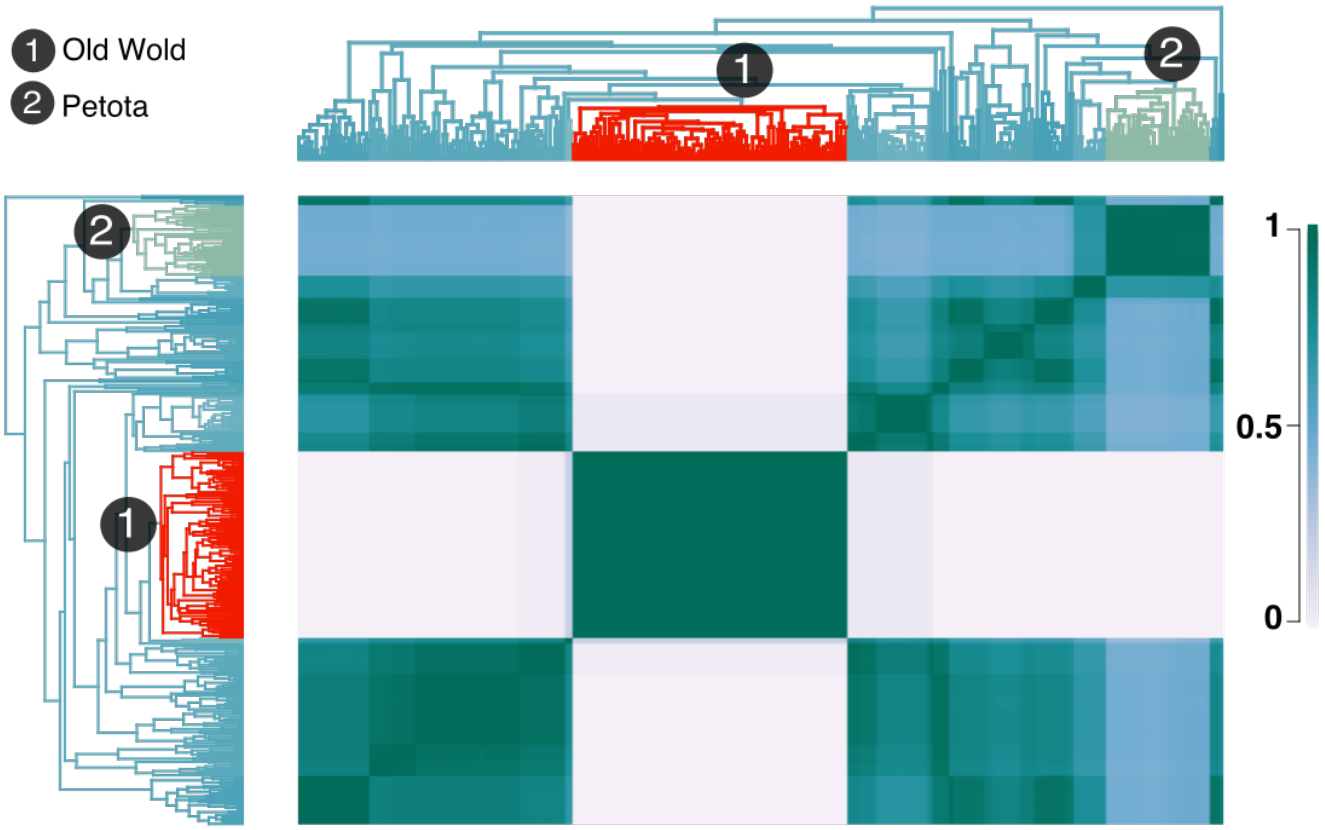

Figure S2. Average macroevolutionary cohort across the 100 trees obtained by PASTIS using the *Solanum* species in the Särkinen *et al.* (2013) phylogeny. Macroevolutionary cohort is the distance of diversification rate parameters across the taxa. Lineages are considered to be part of the same macroevolutionary cohort when there is an elevated pairwise probability (> 0.5) and completely decoupled when the probability is 0. For reference, the phylogeny is shown to the right and the top of the matrix.

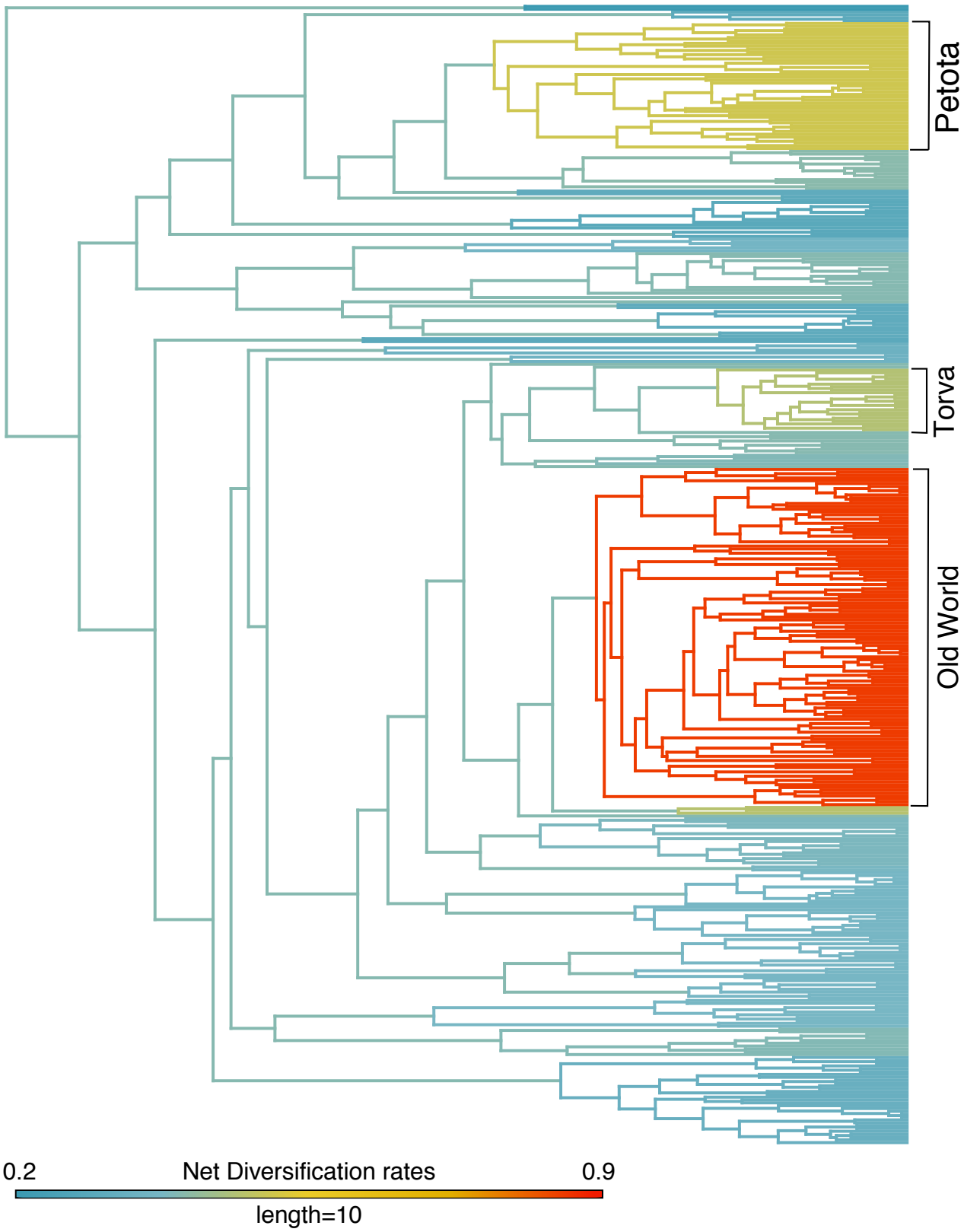

Figure S3. Net diversification rates of *Solanum* estimated by RevBayes. RevBayes results (using 20 out of the 100 trees created by PASTIS) supporting the distinctive radiation of the Old world clade and the signal of other potential radiations such as the Petota clade within the non-spiny solanums and Torva clade within the spiny solanums.

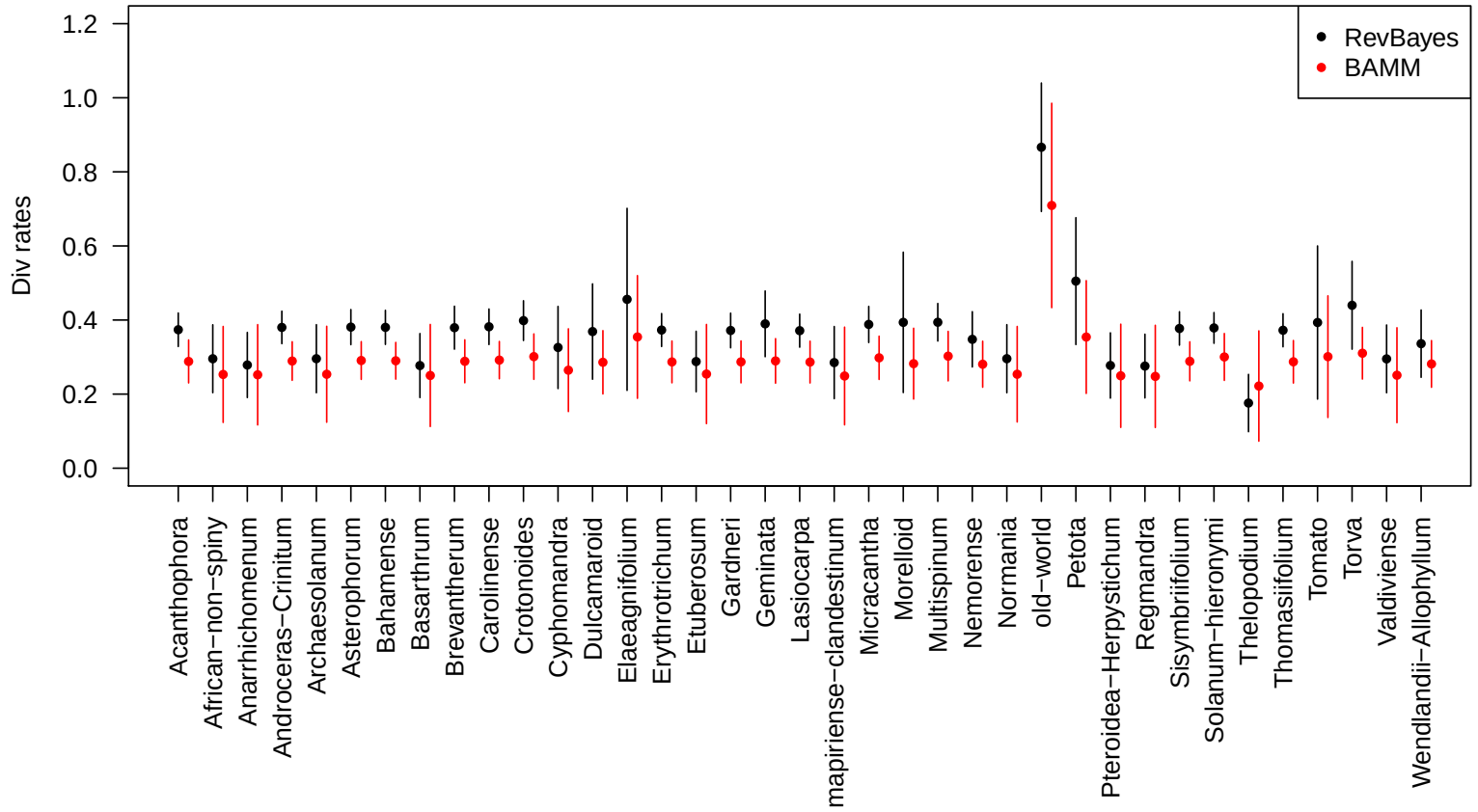

Figure S4. Mean net diversification rates in each of the *Solanum* sections estimated using BAMM and RevBayes. Average net diversification rates under BAMM and RevBayes analyses using a distribution of 20 trees obtained by the polytomy resolver PASTIS. Bars represent the 95% confidence intervals for the diversification estimates.

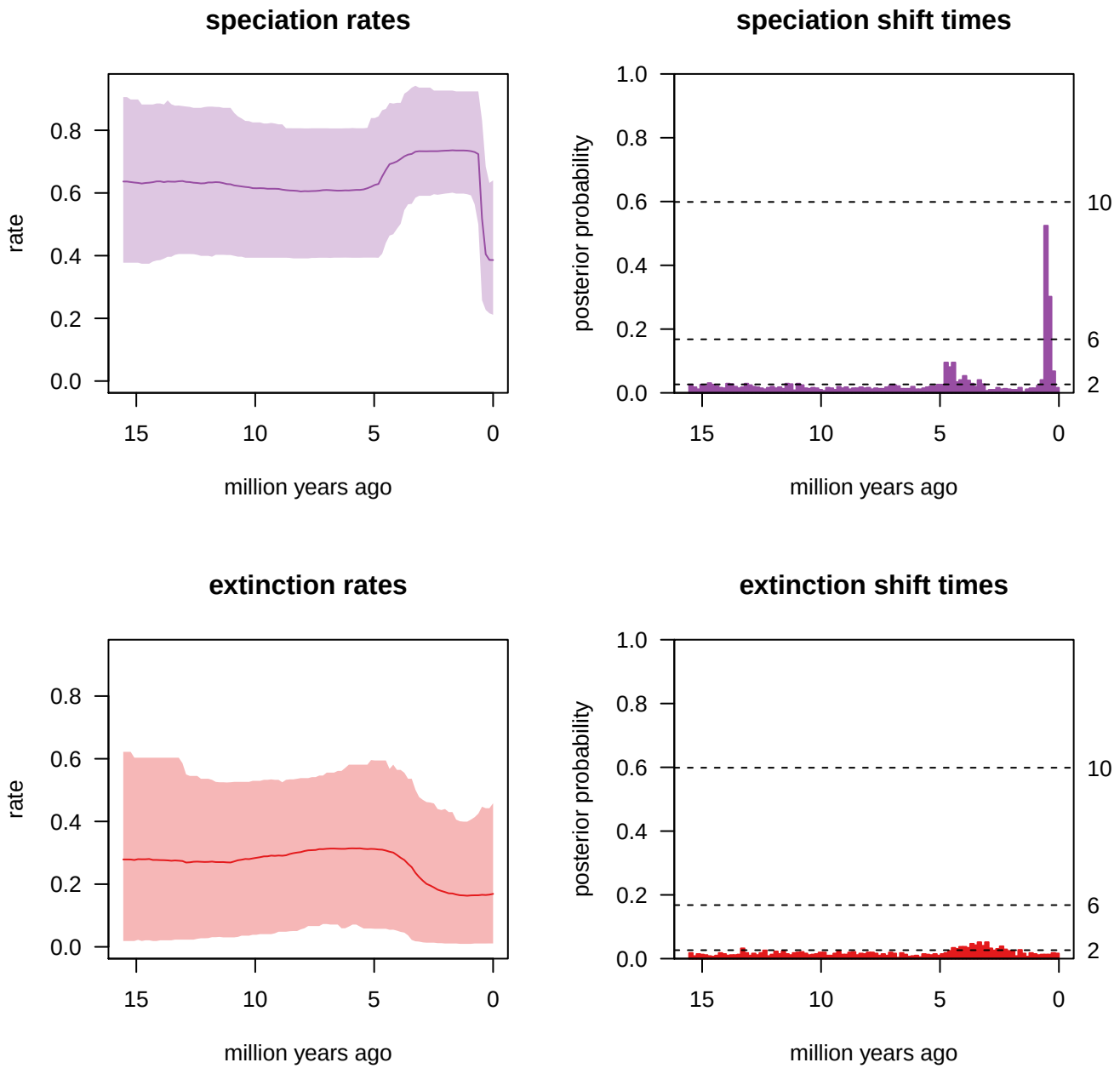

Figure S5. Estimates of the rates and shifts in lineages diversification through time under the TESS approach using the Särkinen *et al.* (2013) phylogeny. Plots in the left show the posterior mean and 95% confidence intervals for speciation and extinction rates. Plots in the right show the temporal significant shifts estimated by Bayes factors (lnBF, numbers in the right axes). Bars indicate the posterior probability of shifts in the time slide. Significant shifts exceed the specified significant threshold (2 ln BF $>$ 6).

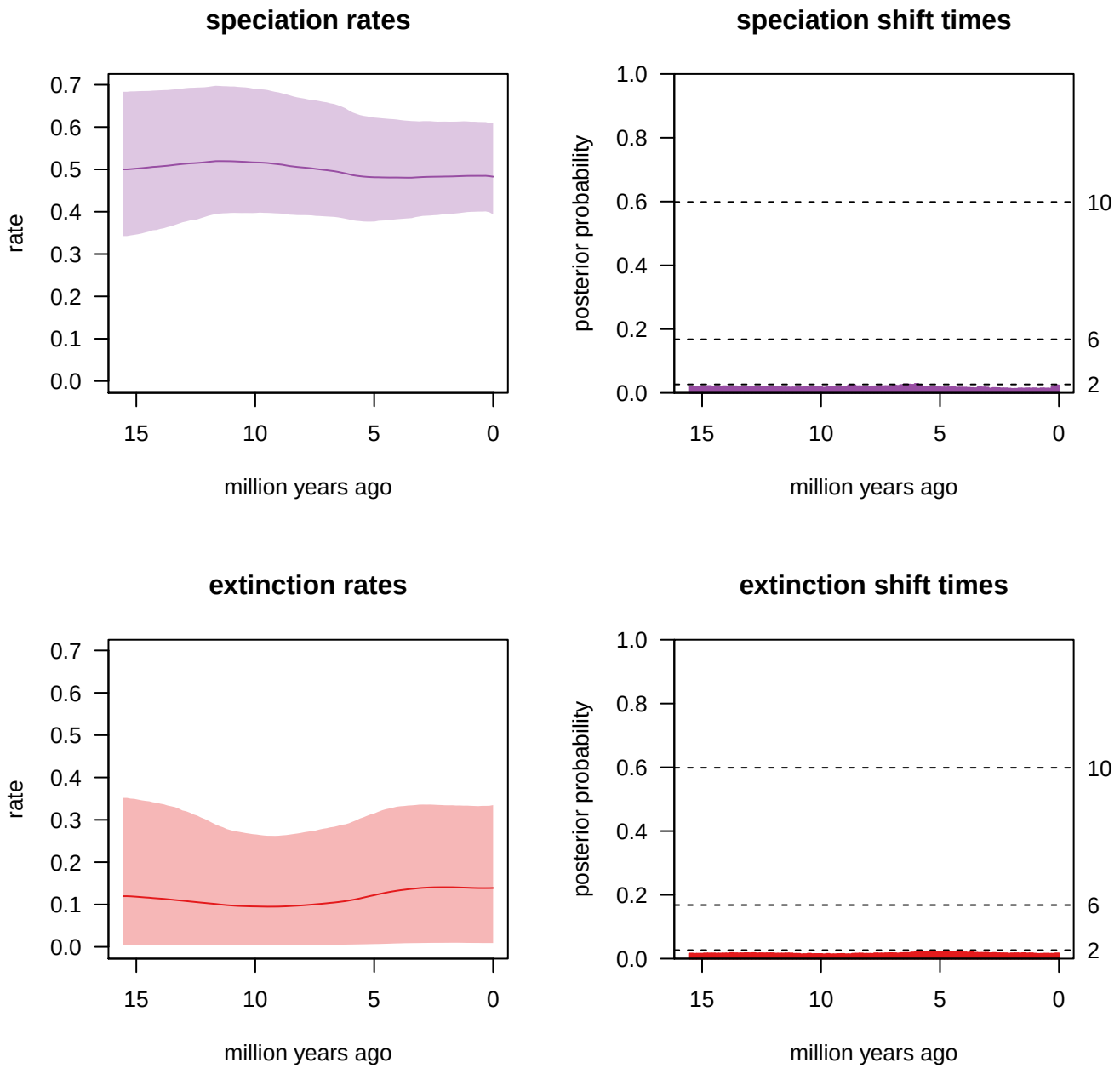

Figure S6. Estimates of the rates and shifts in lineages diversification through time under the TESS approach using a distribution of 100 trees created by the polytomy resolver PASTIS. Plots in the left show the posterior mean and 95% confidence intervals for speciation and extinction rates. Plots in the right show the temporal significant shifts estimated by Bayes factors (lnBF, numbers in the right axes). Bars indicate the posterior probability of shifts in the time slide. Significant shifts exceed the specified significant threshold ($2 \ln BF > 6$).

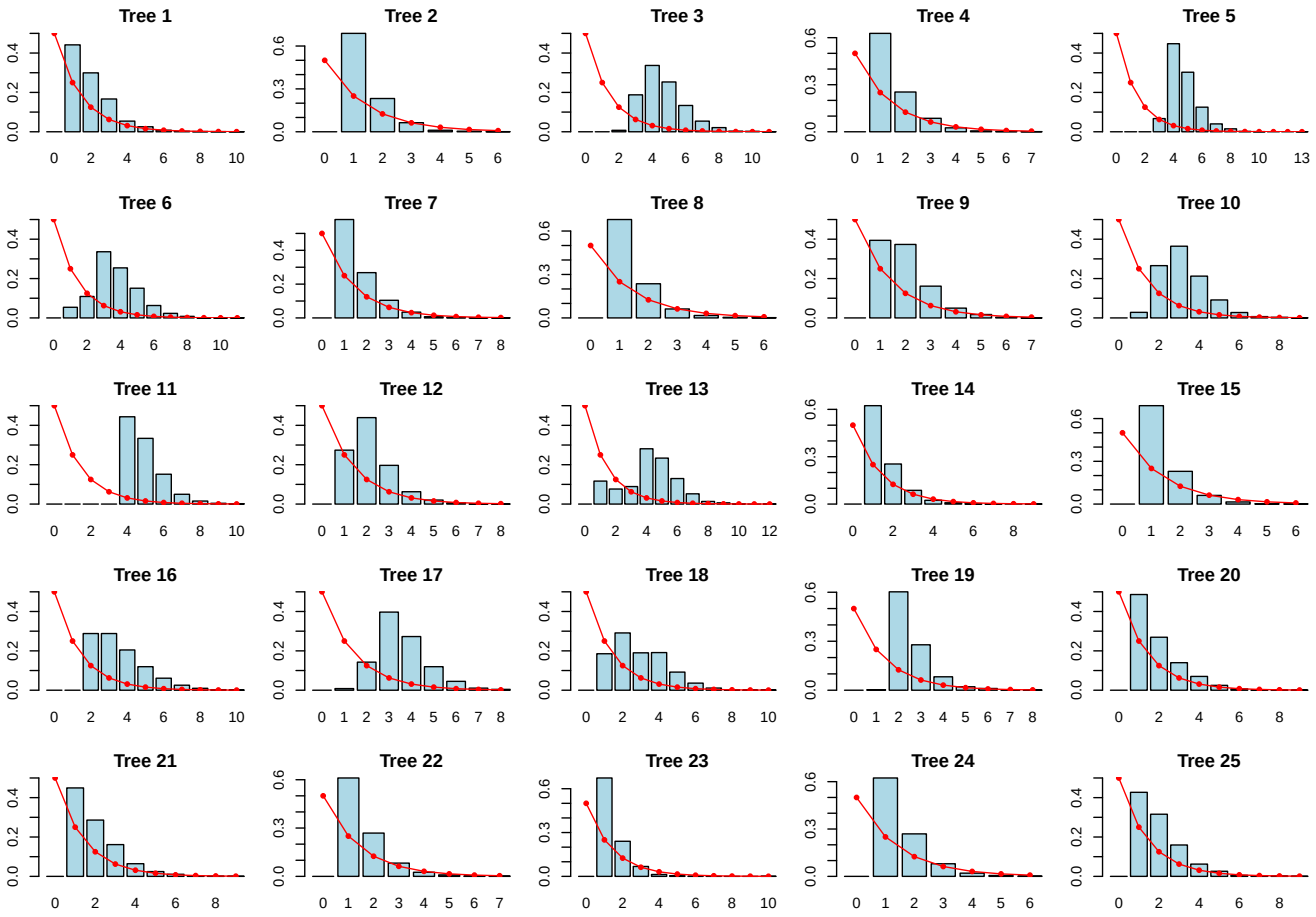

Figure S7. Marginal posterior probability distributions on the number of shifts for 1-25 trees used in the BAMM analysis (filled histograms). The prior distribution on the number of shifts for each tree is illustrated in red.

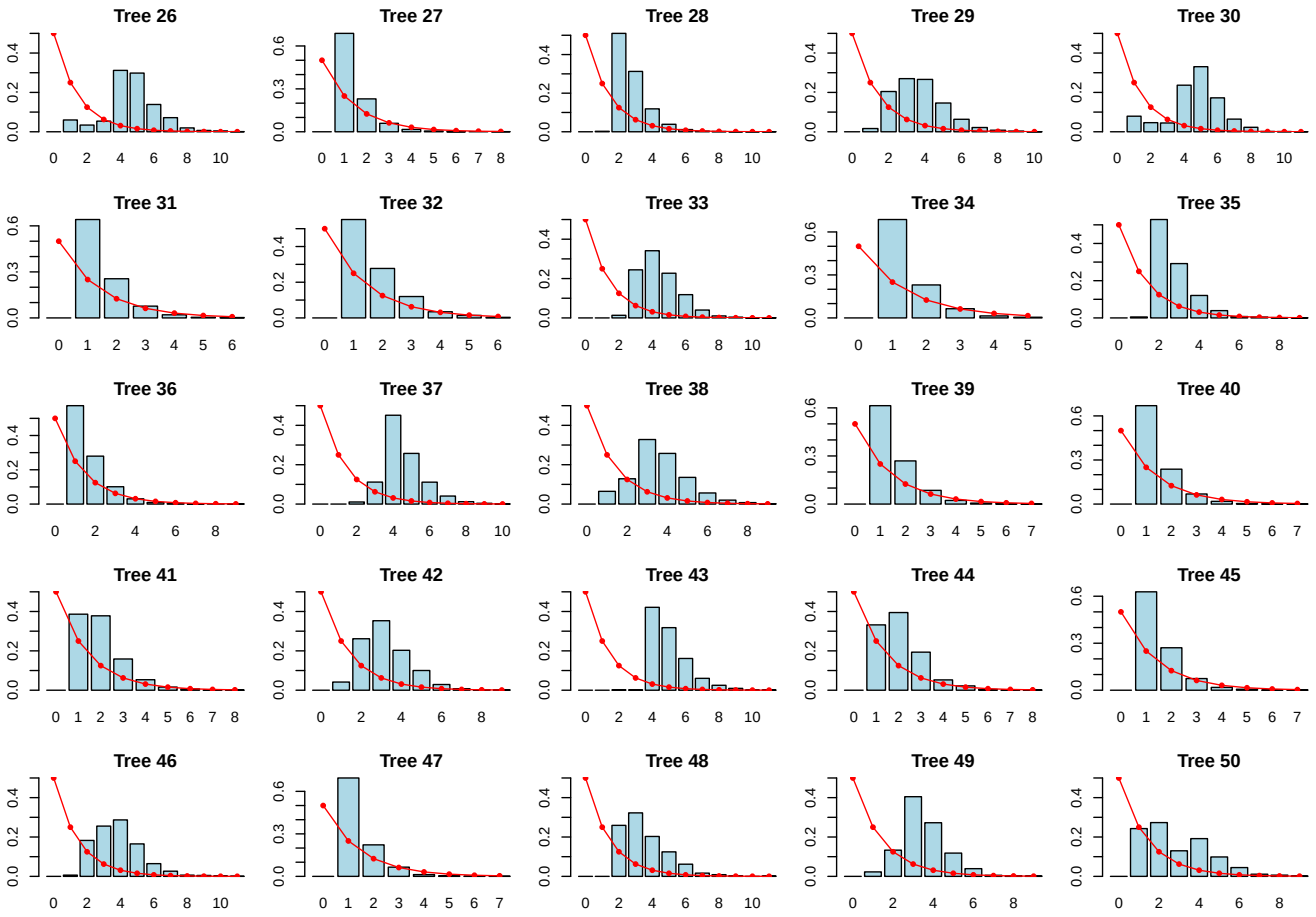

Figure S8. Marginal posterior probability distributions on the number of shifts for 26-50 trees used in the BAMM analysis (filled histograms). The prior distribution on the number of shifts for each tree is illustrated in red.

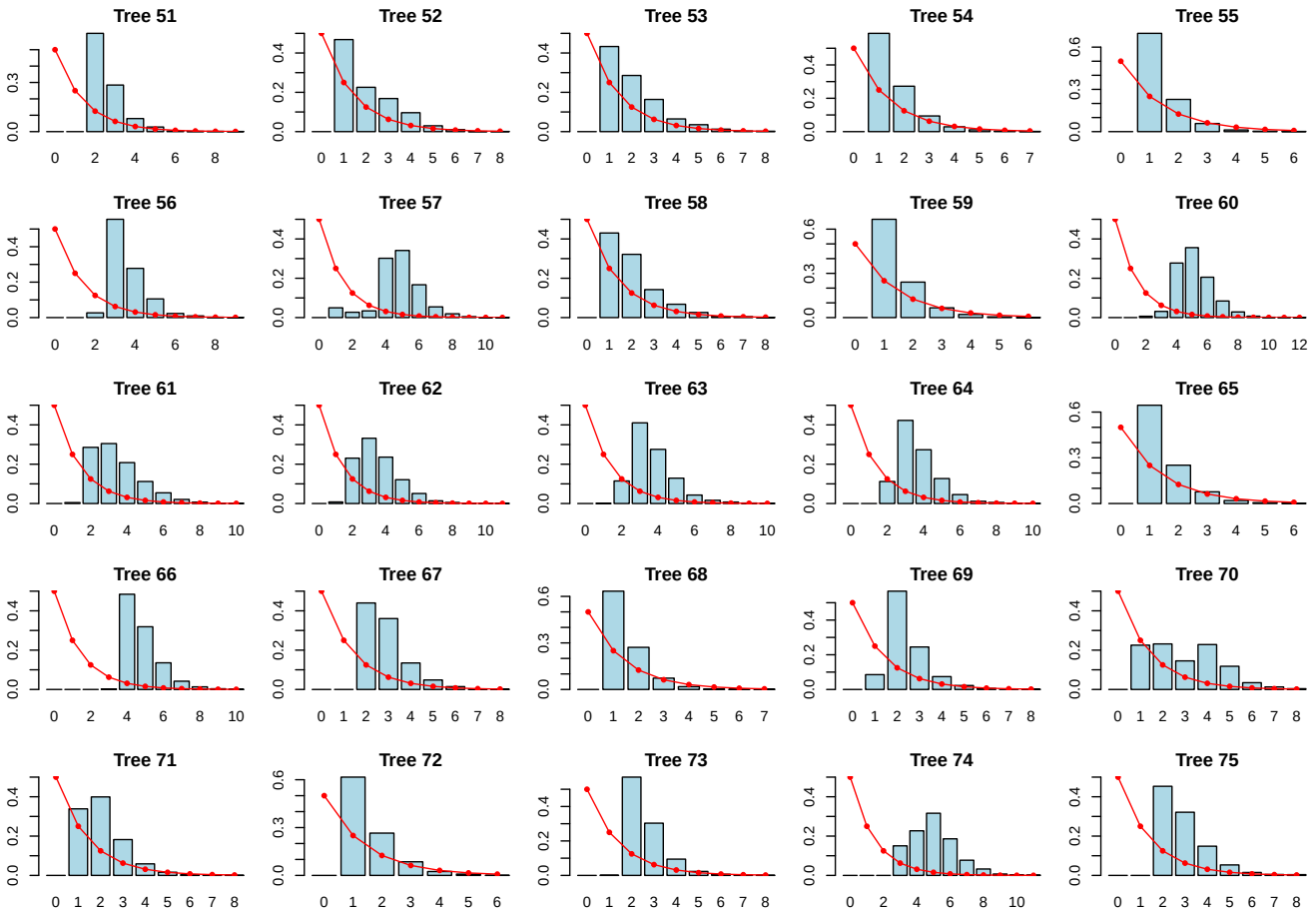

Figure S9. Marginal posterior probability distributions on the number of shifts or 51-75 trees used in the BAMM analysis (filled histograms). The prior distribution on the number of shifts for each tree is illustrated in red.

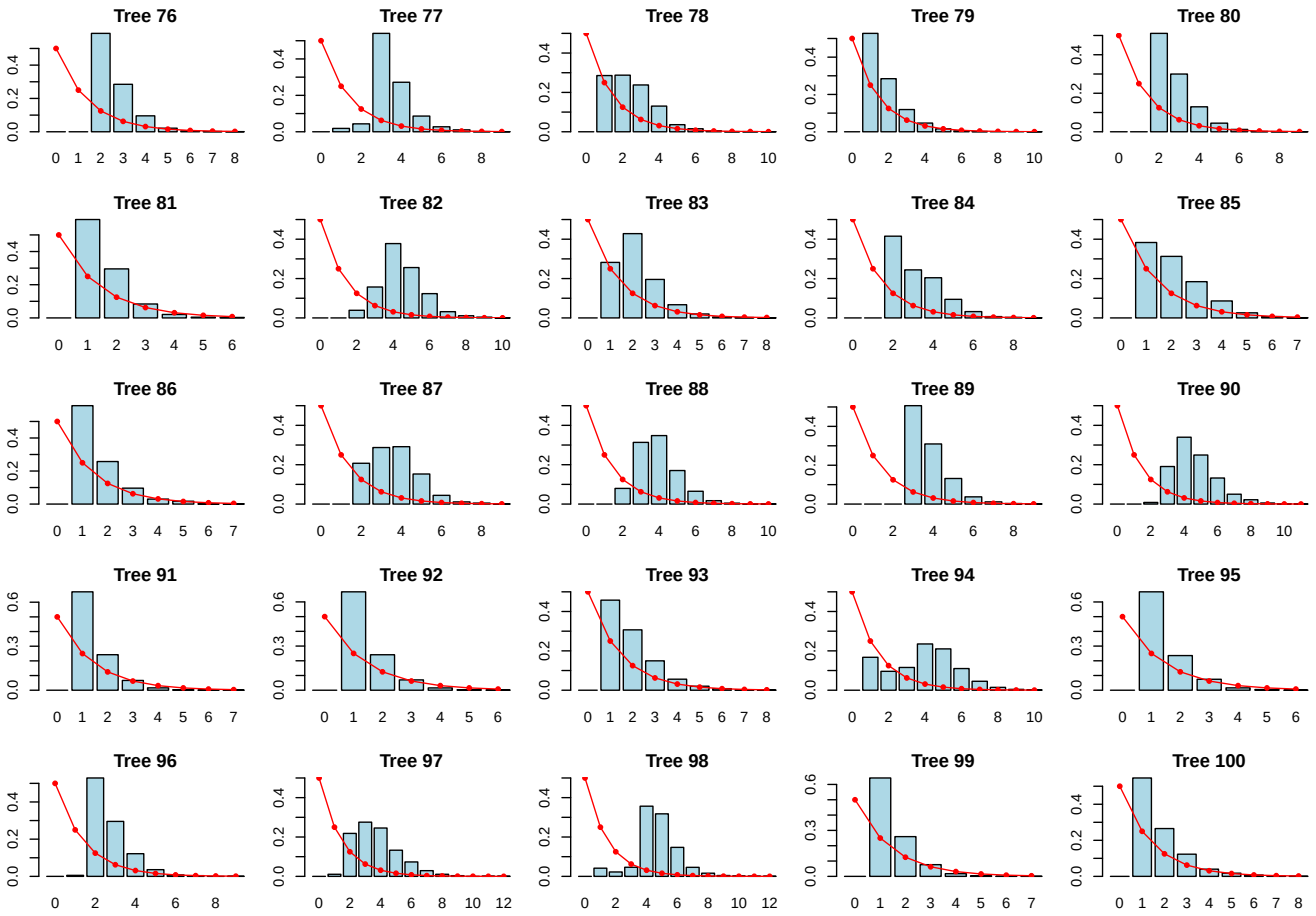

Figure S10. Marginal posterior probability distributions on the number of shifts for 76-100 trees used in the BAMM analysis (filled histograms). The prior distribution on the number of shifts for each tree is illustrated in red.

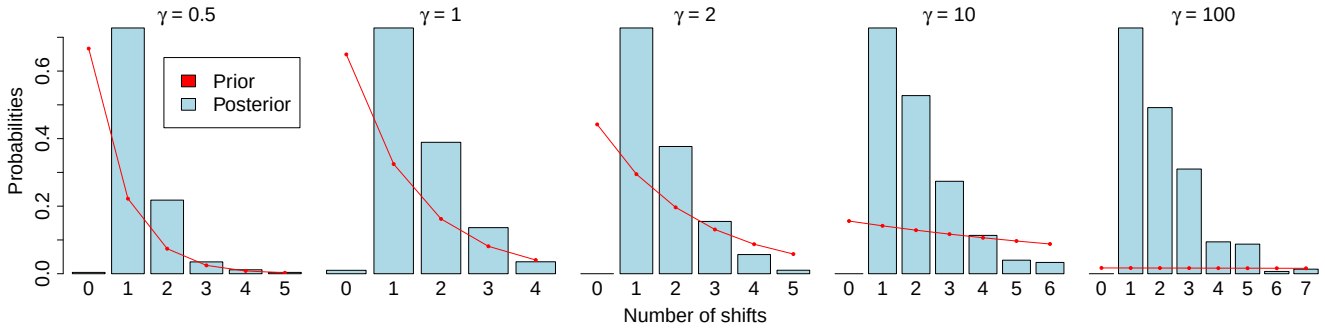

Figure S11. Effect of the prior on the marginal posterior distribution of rate shifts in *Solanum* using the Särkinen *et al.* (2013) phylogeny. Histograms represent the marginal posterior probability distribution on the number of shifts and lines correspond to the prior distribution on the expected number of shifts. Although there is a slight change in the distribution on the probabilities with different priors, there is no evidence that the results are unusually sensitive to the priors used. Note that some models were not observed in the different treatments (e.g., models with 0 shifts were usually not observed showing an overwhelming evidence of rate heterogeneity in both trees).

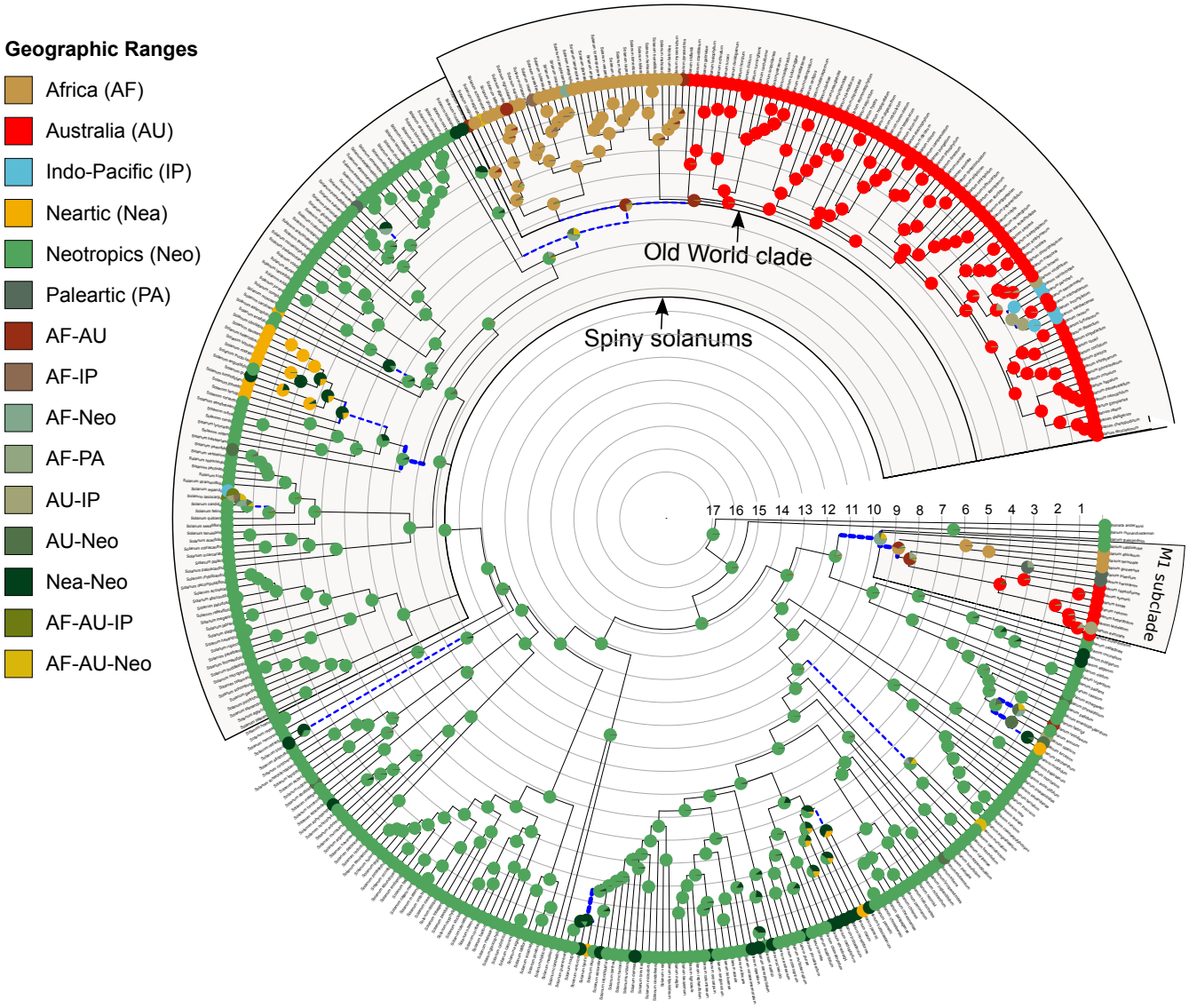

Figure S12. Reconstruction of the historical biogeography of *Solanum* under the DEC model implemented in BioGeoBEARS. Pies at each node represent the probability that each region (or the combination of regions) is, according to the model, the ancestral range distribution. The highlighted branches represent the dispersals events inferred in at least the 50% of BSM simulations. Thicker branches show dispersals inferred more than the 95% of the BSM simulations. The timescale is given in Ma.

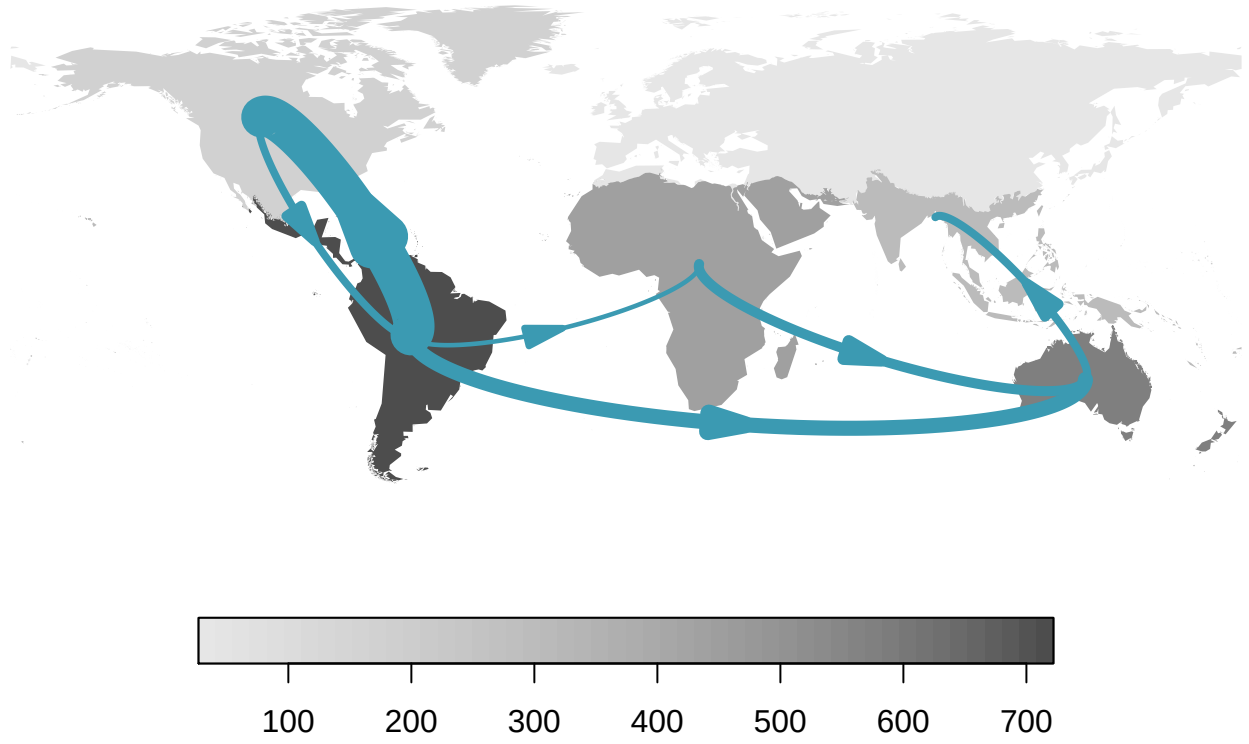

Figure S13. Summary of dispersal events within the main biogeographic regions of *Solanum*. The arrows between regions represent the frequency and direction of dispersal events. The bar represents the total number of species within each region. Only dispersal events with a mean of two or more counts are shown. The thick of the arrows describes the frequency of the events.

References

Edwards EJ, de Vos JM , Donoghue MJ. 2015. Doubtful pathways to cold tolerance in plants. *Nature*, **521**: E5–E6.

Höhna S. 2015. The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *Journal of theoretical biology*, **380**: 321–331.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP , Ronquist F. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, p. syw021.

Jetz W , Fine PV. 2012. Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. *PLoS Biol*, **10**: e1001292.

Miller MA, 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.

Mitchell JS , Rabosky DL. 2016. Bayesian model selection with bamm: effects of the model prior on the inferred number of diversification shifts. *Methods in Ecology and Evolution*.

Pebesma EJ , Bivand RS. 2005. Classes and methods for spatial data in R. *R News*, **5**: 9–13.

Rabosky DL, Donnellan SC, Grundler M , Lovette IJ. 2014. Analysis and visualization of complex macroevolutionary dynamics: an example from australian scincid lizards. *Systematic biology*, **63**: 610–627.

Rambaut A, Suchard M, Xie D , Drummond A. 2009. Mcmc trace analysis tool. version v1. 6.0. *Institute of Evolutionary Biology, University of Edinburgh. Available: a. rambaut@ ed. ac. uk. Accessed Jan*, **16**: 2011.

Ronquist F , Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**: 1572–1574.

Särkinen T, Bohs L, Olmstead RG , Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC evolutionary biology*, **13**: 214.

Shi JJ , Rabosky DL. 2015. Speciation dynamics during the global radiation of extant bats. *Evolution*, **69**: 1528–1545.

South A. 2011. rworldmap: A new r package for mapping global data. *The R Journal*, **3**: 35–43.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, **108**: 6187–6192.

Vilela B , Villalobos F. 2015. letsr: a new r package for data handling and analysis in macroecology. *Methods in Ecology and Evolution*, **6**: 1229–1234.