Supplementary materials for : The state of state-dependent speciation and extinction models : patterns in angiosperm macroevolution

Andrew J. Helmstetter*, Sylvain Glemin, Jos Käfer, Rosana Zenil-Ferguson, Hervé Sauquet, Hugo de Boer

**Corresponding author : Andrew J. Helmstetter, FRB-CESAB, Montpellier, 34000, France. andrew.j.helmstetter@gmail.com*

Contents

1 Data collection

1.1 Dataset characteristics

Here we give details of all of the different dataset characteristics collected from studies that used SSE models to investigate diversification in angiosperm groups. These characteristics correspond to the column headers in our synthesised dataset as well as the template for reporting SSE results (Supplementary Data).

entrant The name of the person who entered the data into the spreadsheet.

study This is the unique name given to the study, this was taken from the text file that the PDF version of the article was converted to. It is in the format "AuthorYearTitle.txt".

year The year in which the study was published. Only peer-reviewed, published studies were included (no preprints).

sse_model The name of the state-dependent speciation and extinction model used. There may be multiple models of the same (or different) types per study.

model_no A numeric identifier given to each model in each study. This number is repeated for the number of states included in each model. If a study reports results from a BiSSE model and MuSSE model with three states, numeric identifiers would be as follows: "1,1,2,2,2".

order The angiosperm order that the study group belongs to. If the scope of the study includes multiple orders (e.g. a study at the angiosperm level) then "multiple" is used.

trait_level_1-trait_level_6 Character states are classified into trait types at different levels, with level 1 being the most broad, and level 6 being the most specific. This classification follows the trait ontology, which can be found in Table S1.

character_state The character states used in a given model. There may be one or more character states per model.

putative_ancestral_state A binary column that indicates the character state which is supposed to be ancestral (1) in the analysis. We use "putative" because in many studies the ancestral state was not explicitly reported. In studies that assumed an ancestral state, we treated this as the ancestral state. In studies that performed ancestral state reconstruction and reported results, we chose the state that was most likely at the root of the tree used with SSE model. In studies that did not report either of these, we searched the text for statements related to which trait was ancestral or examined the distribution of tip states to identify which state is putatively ancestral. Therefore in some cases this characteristic is somewhat subjective and should not be considered as accurate evidence for the ancestral state of the trait, but instead a hypothesis about what the ancestral state is likely to be.

clade The name of the clade that the SSE model was run on.

div_inc A binary column indicating when net diversification rate was higher (1) for a character state. We initially wanted to include only significant results here, but in many cases significance was not reported, so we considered all cases in which net diversification was reported to be higher in one trait than the other as trait-dependent diversification. In multi-state models we classified the state with the lowest net diversification rate rate as 0 and all other states as 1.

sp_inc A binary column indicating when speciation rate was higher (1) for a character state.

ext_inc A binary column indicating when extinction rate was higher (1) for a character state.

tips The number of tips used with in the SSE model. In some cases only the number of tips in the tree was reported. We tried to remove outgroups/those taxa included in the phylogenetic tree bit not included in the SSE model where possible, but this was not always evident.

no_markers The total number of nuclear, plastid and mitochondrial markers used to build the phylogenetic tree that was used with the SSE model. Equal weight was given to each marker type in this column (even though plastid markers, for example, are not entirely independent).

no_plastid The total number of plastid markers used to build the phylogenetic tree that was used with the SSE model.

no_nuclear The total number of nuclear markers used to build the phylogenetic tree that was used with the SSE model.

no_mito The total number of mitochondrial markers used to build the phylogenetic tree that was used with the SSE model.

age The age of the root of the phylogenetic tree in million years. As with tips, we tried to get the age of the tree that was used with the SSE model if this differed from the age of the tree reported in the main text of the study. However, this was not always possible. In cases where ages were not reported/data was not available we attempted to estimate ages from figures.

age_inferred In some cases the phylogenetic tree was not time-calibrated (e.g. the root of the tree was set to a fixed age of 1). This binary column indicates whether age was inferred (1) or not (0).

perc_sampling The global sampling fraction for taxa used in the SSE model. In cases where this was not reported we acquired estimates for the number of species in the clade of interest from http://www.theplantlist.org/ and used this to calculate a sampling fraction.

sampling_per_state The sampling fraction per state, as reported in the study. If this was not reported we did not try to calculate it based on other sources of information as it requires specialist knowledge about which species possess each character state.

samples_per_state The number of tips that belong to each character state in an SSE model. In cases where this was not reported we counted tip states from figures.

2 Supplementary figures

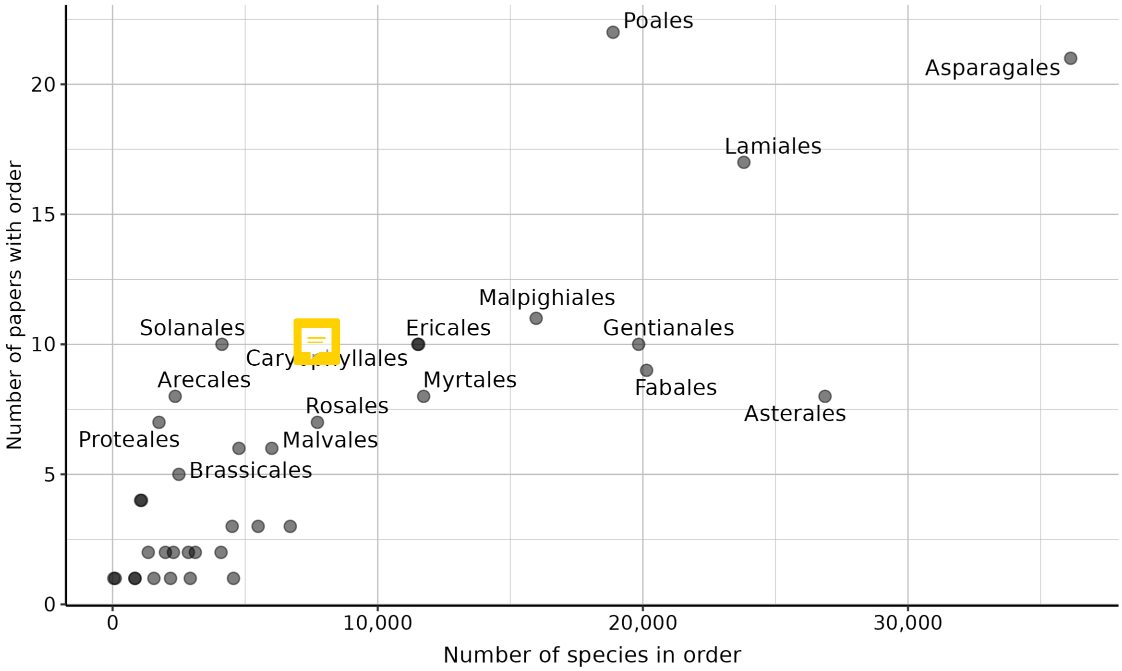

Figure S1: A scatterplot showing the relationship between the number of species in each order (taken from hrefhttp://www.mobot.org/http://www.mobot.org/) and the number of papers that studied that order. Only those orders in our dataset are included. Text labels have been added to the points corresponding to the most commonly studied orders in our dataset.

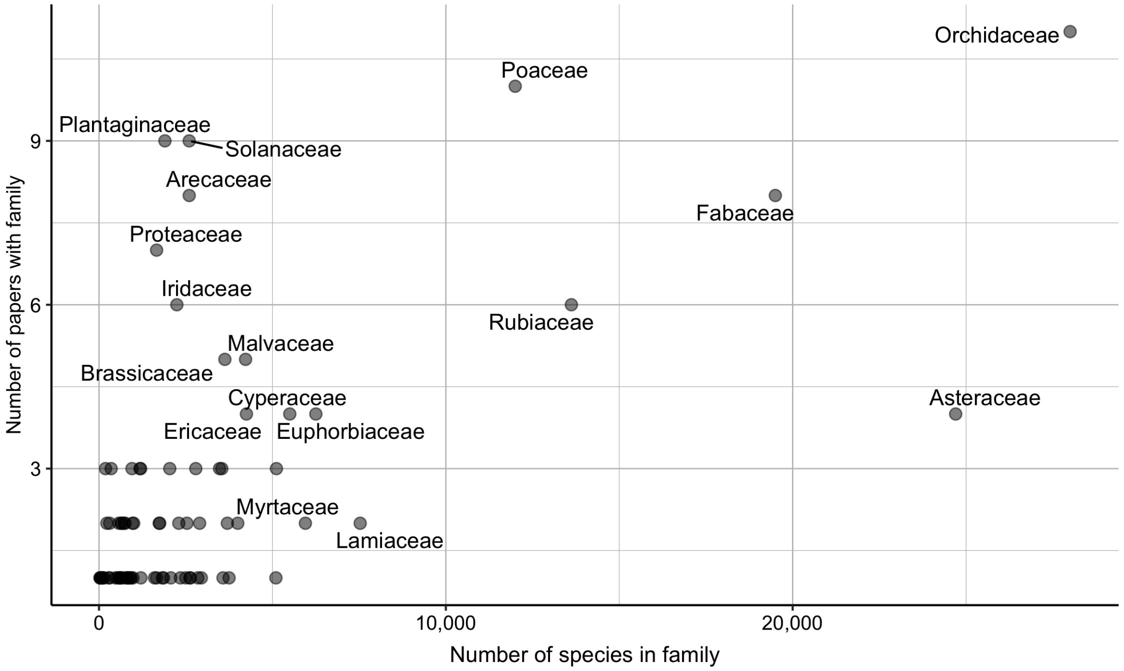

Figure S2: A scatterplot showing the relationship between the number of species in each family (taken from Christenhusz & Byng 2016) and the number of papers that studied that family. Only those orders in our dataset are included. Text labels have been added to the points corresponding to the most commonly studied orders in our dataset.

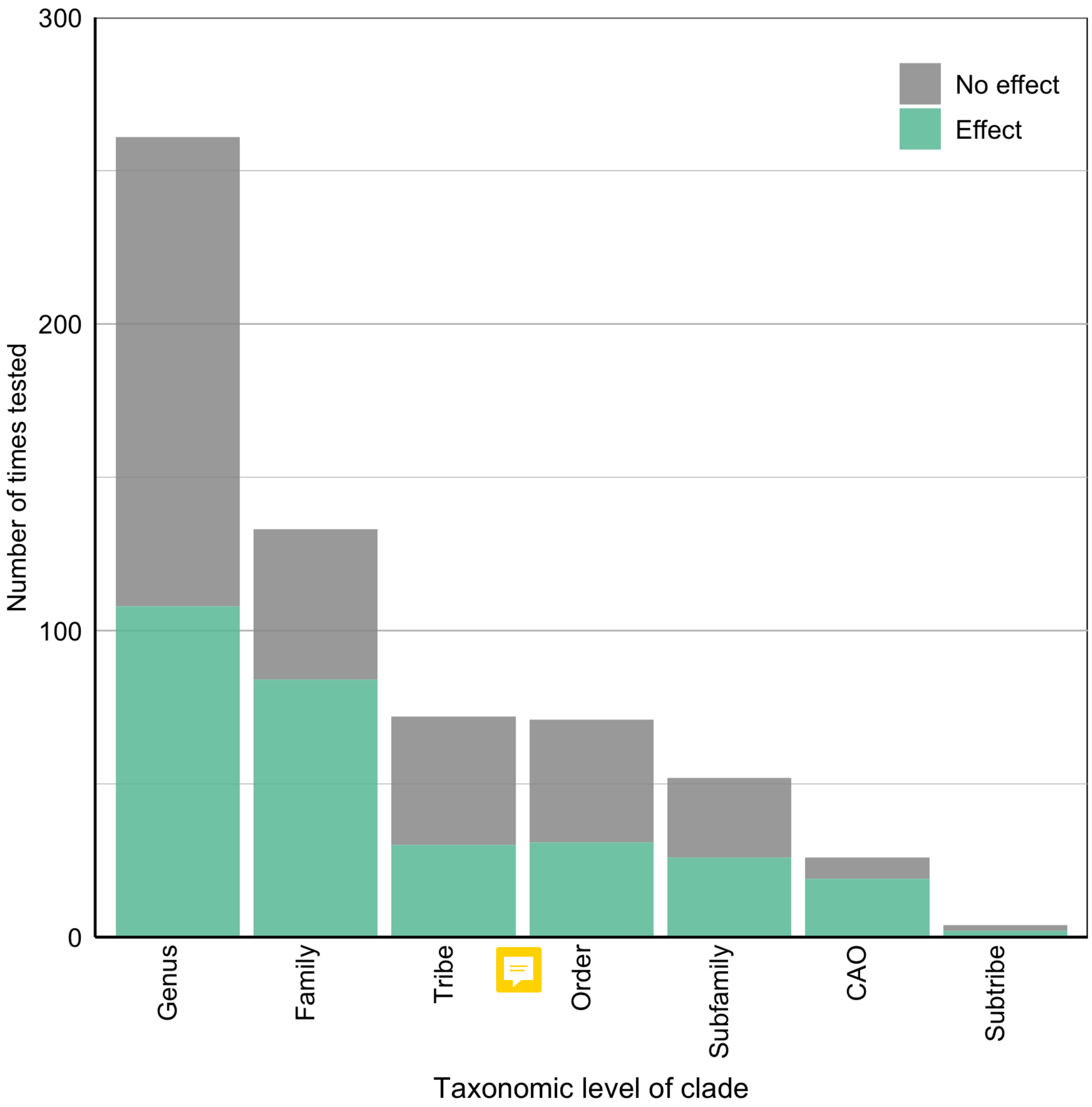

Figure S3: Stacked barplots showing the number of models run per different taxonomic levels. The smallest level is genus, increasing in scope until clades above order (abbreviated as CAO). Bars are coloured based on whether trait-dependent diversification was inferred in the model.

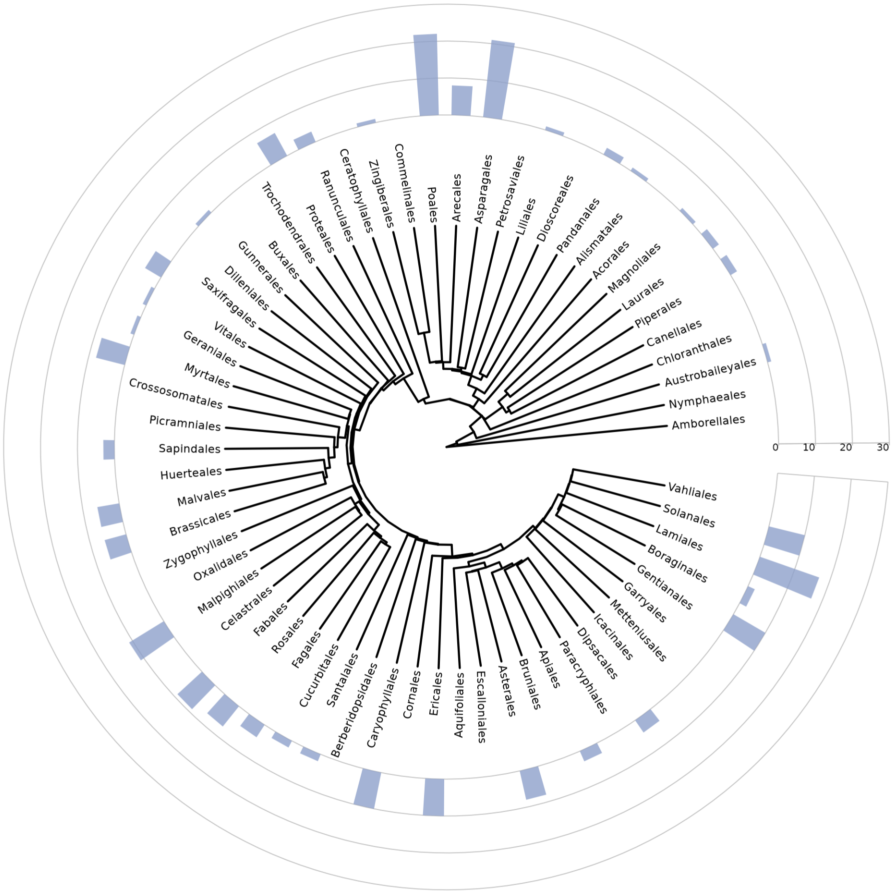

Figure S4: A phylogenetic tree of angiosperm orders taken from Li et al. 2019 [1] annotated with bars representing the number of studies using SSE models that focused on the order.

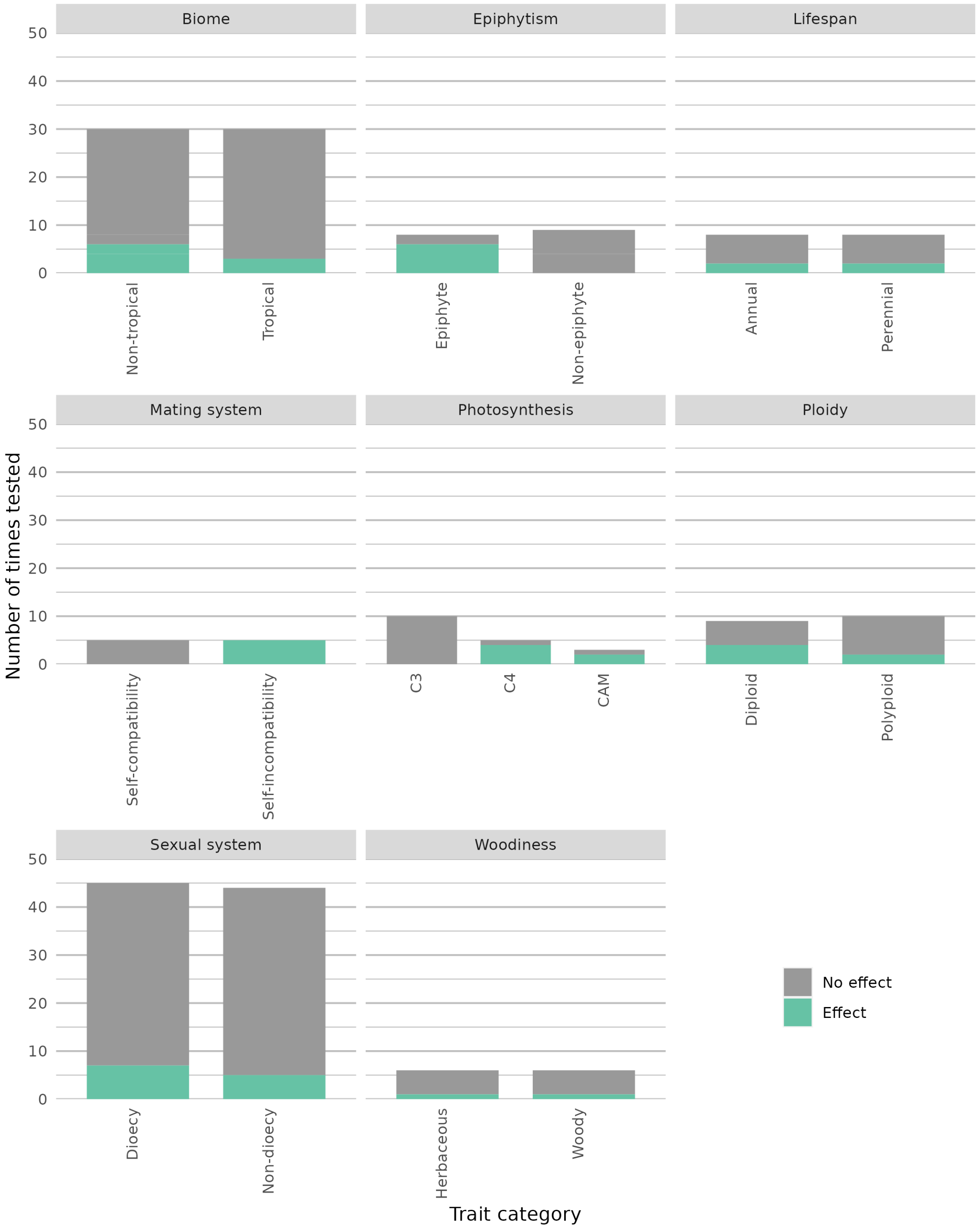

Figure S5: Stacked barplots for eight different character state groups. For each character state there is a bar. The coloured portion of the bar represents the number of models where that state was associated with increased diversification. If there was no effect of either state in the model, this was counted as no effect for both states. Though the photosynthesis category has three states, these were only tested with binary-state models (e.g. C3 vs C4 or CAM).

.

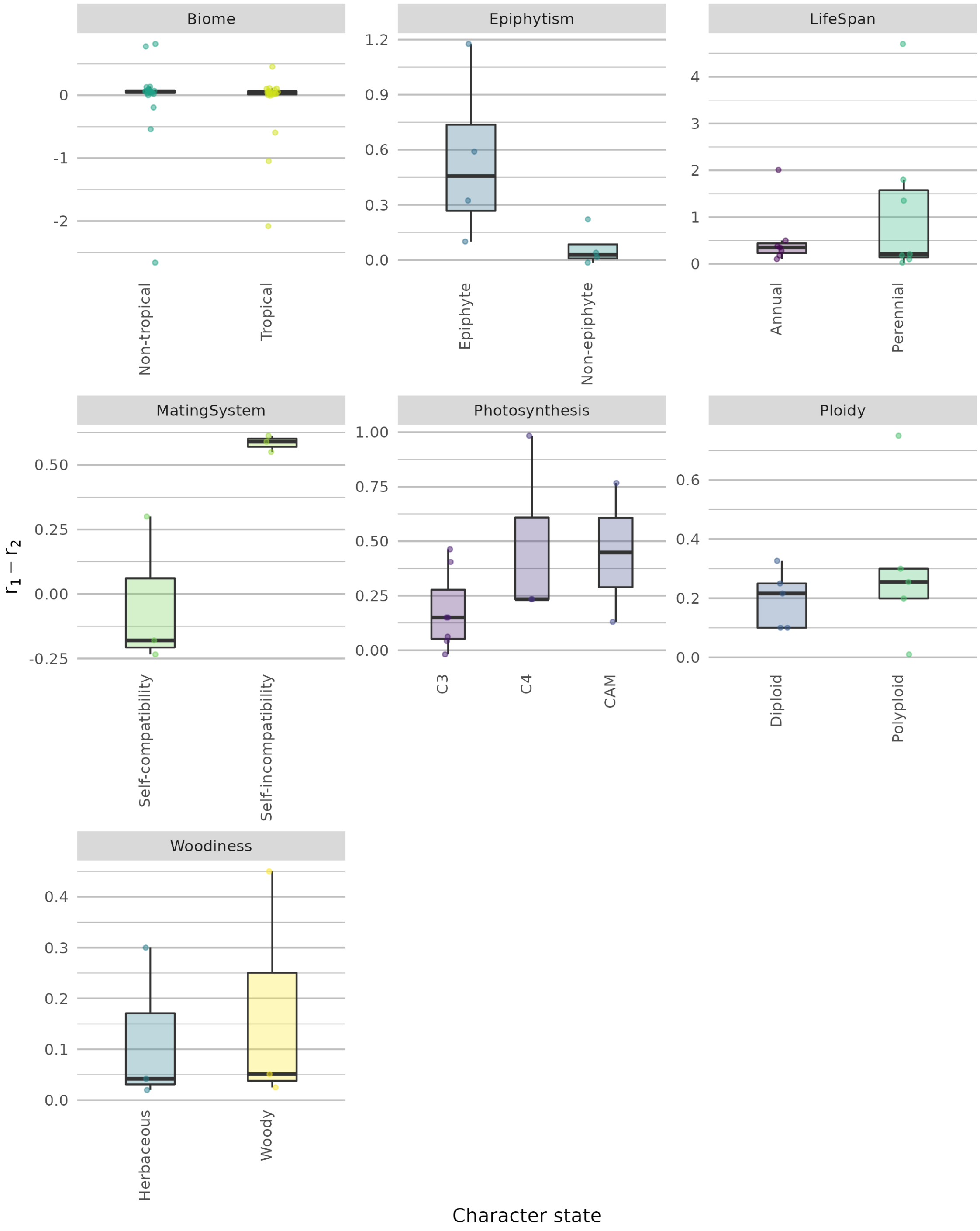

Figure S6: Boxplots showing the net diversification rates for seven traits and their associated states. Jittered points overlain on the boxplots indicate the mean net diversification rate values recovered from each model. The ”LifeForm” trait level 6 category was separated into ”Epiphytism” and ”Woodiness” for presentation purposes.

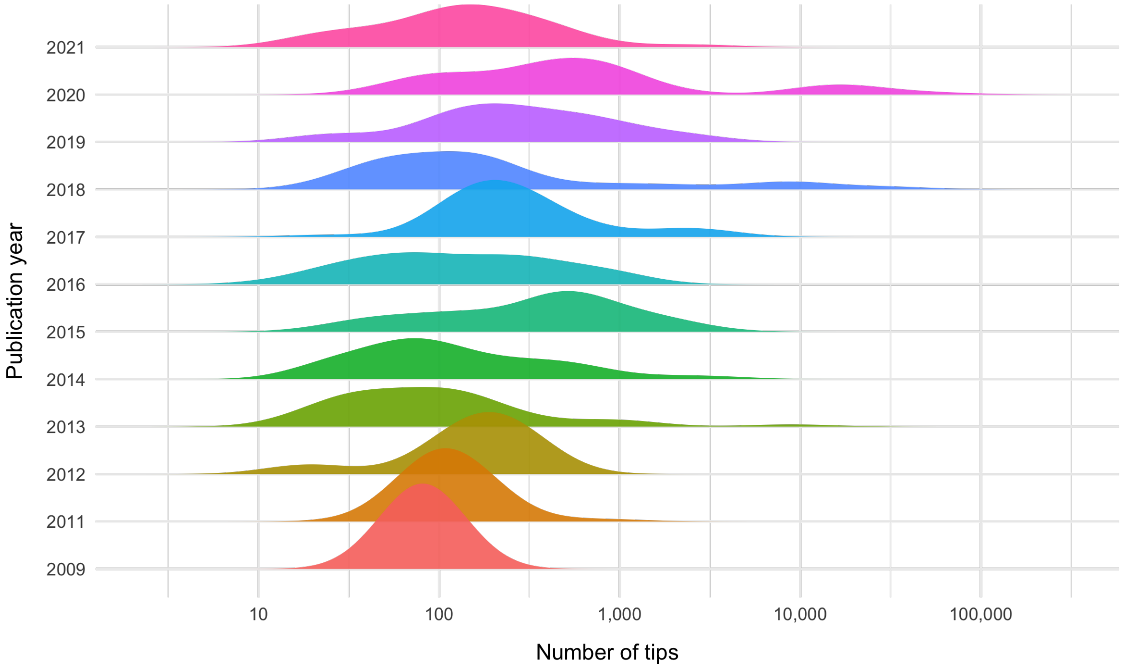

Figure S7: A ridgeplot showing how the number of tips on trees used with SSE models have changed over time. Each ridge displays a density plot corresponding to a single publication year (2009-2021) with the most recent year at the top of the plot. The x-axis is on a log scale. There was not enough data from 2010 to calculate a density so this year was removed from the plot.

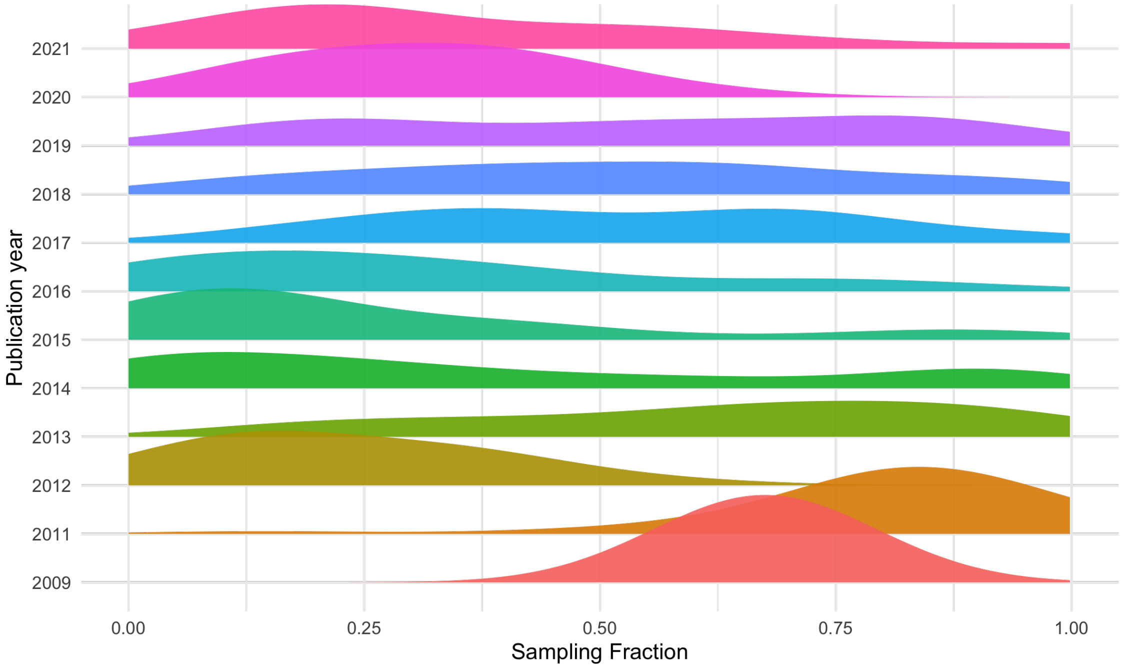

Figure S8: A ridgeplot showing how sampling fractions of trees used in SSE models have changed over time. Each ridge displays a density plot corresponding to a single publication year (2009-2021) with the most recent year at the top of the plot. The x-axis is on a log scale. There was not enough data from 2010 to calculate a density so this year was removed from the plot.

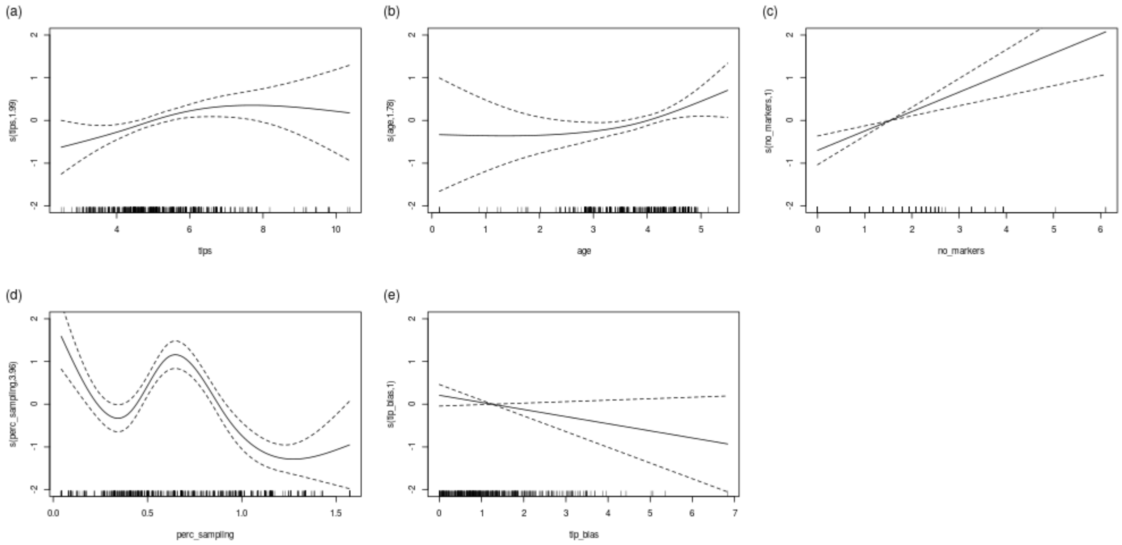

Figure S9: Panels (a-e) show the relationships inferred with generlized additive models (GAM) between each continuous dataset property in Figure 4 and SSE model outcome (1 = trait dependent diversification inferred in model, 0 = no effect of traits in model). Cubic regression splines with five dimensions were fitted to each property independently.

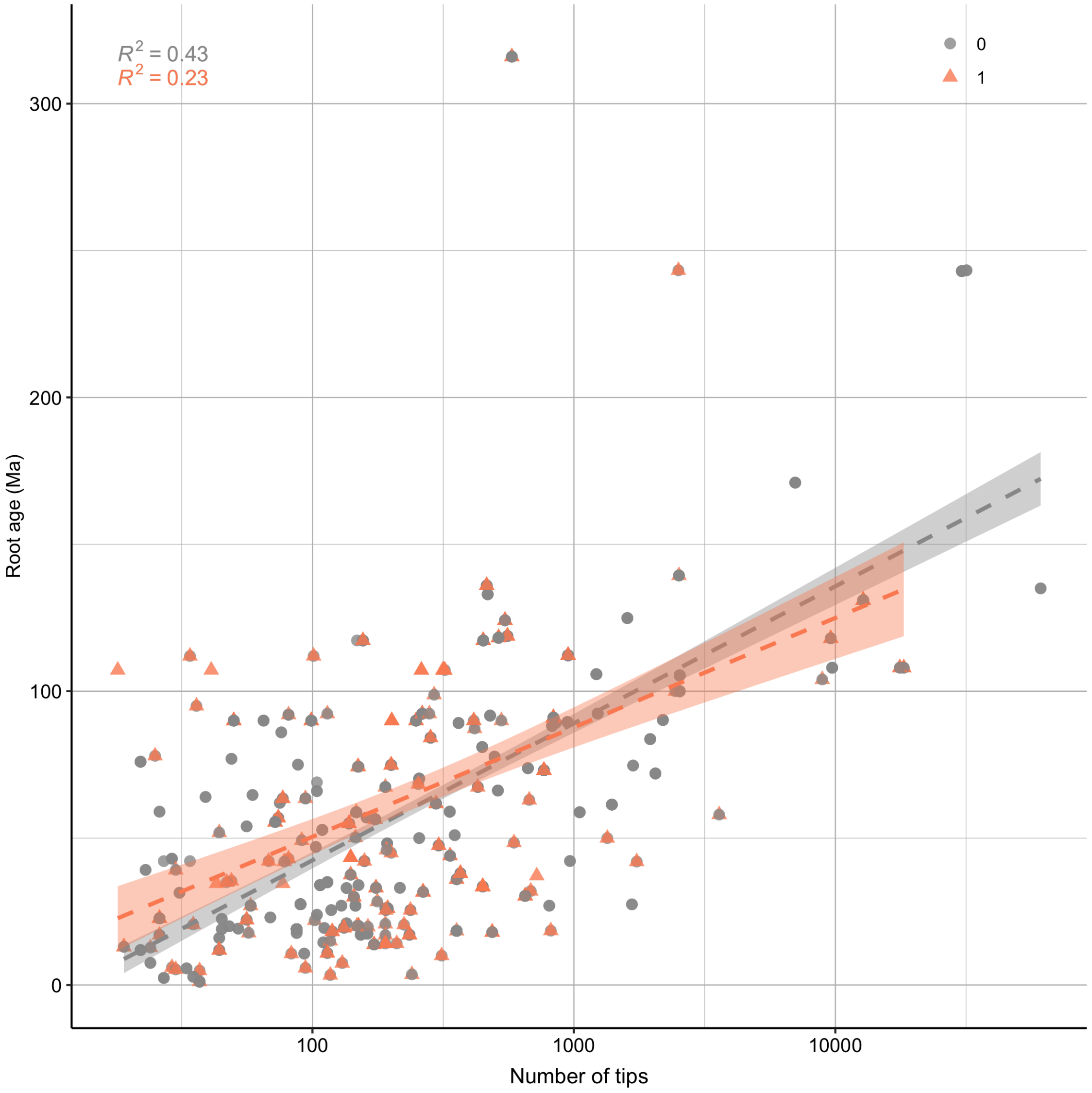

Figure S10: A scatterplot showing the relationship between the root age of the trees used with SSE models in our dataset and the number of tips in the trees. Points are coloured based on whether trait-dependent diversification was inferred (red) or not (grey) when the associated model was run.

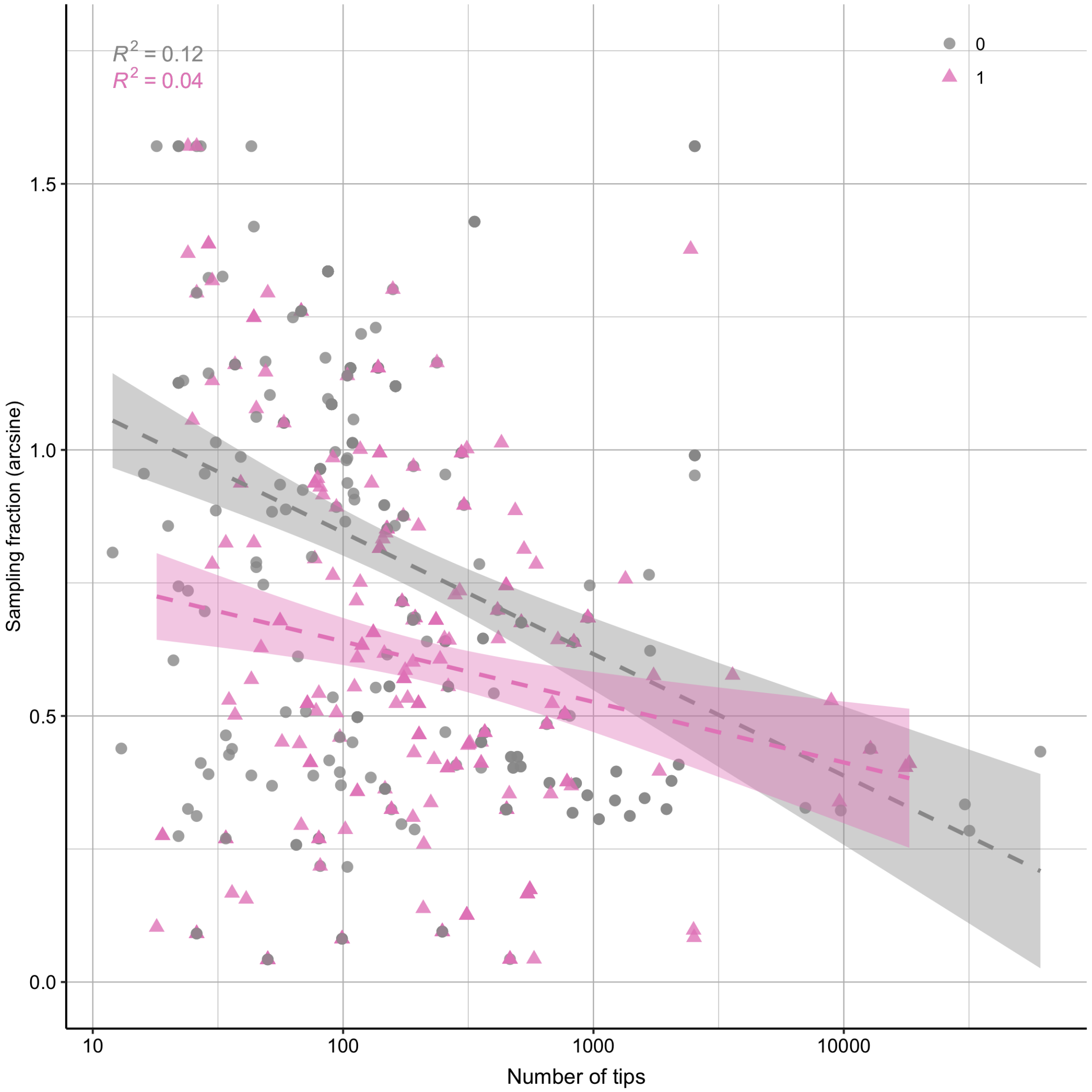

Figure S11: A scatterplot showing the relationship between sampling fraction of the tree used with an SSE model, and the number of tips in the tree. Coloured points are models for which trait-dependent diversification was detected, and grey points are models where it was not detected. Lines were fitted using linear models to these two groups with 95% confidence intervals estimated.

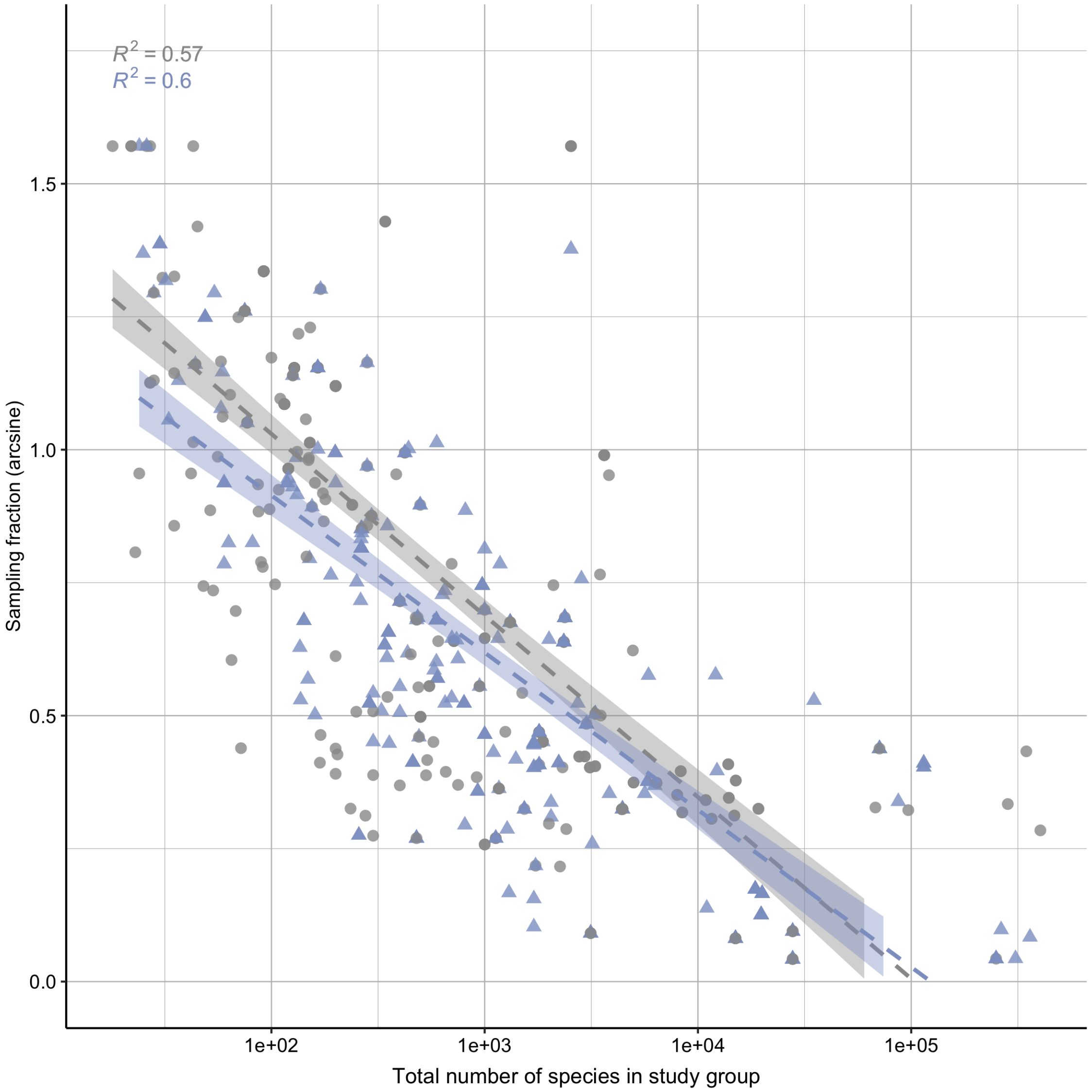

Figure S12: A scatterplot showing the relationship between sampling fraction of the tree used with an SSE model, and the total number of species in the study group the tree represents. Coloured points are models for which trait-dependent diversification was detected, and grey points are models where it was not detected. Lines were fitted using linear models to these two groups with 95% confidence intervals estimated.

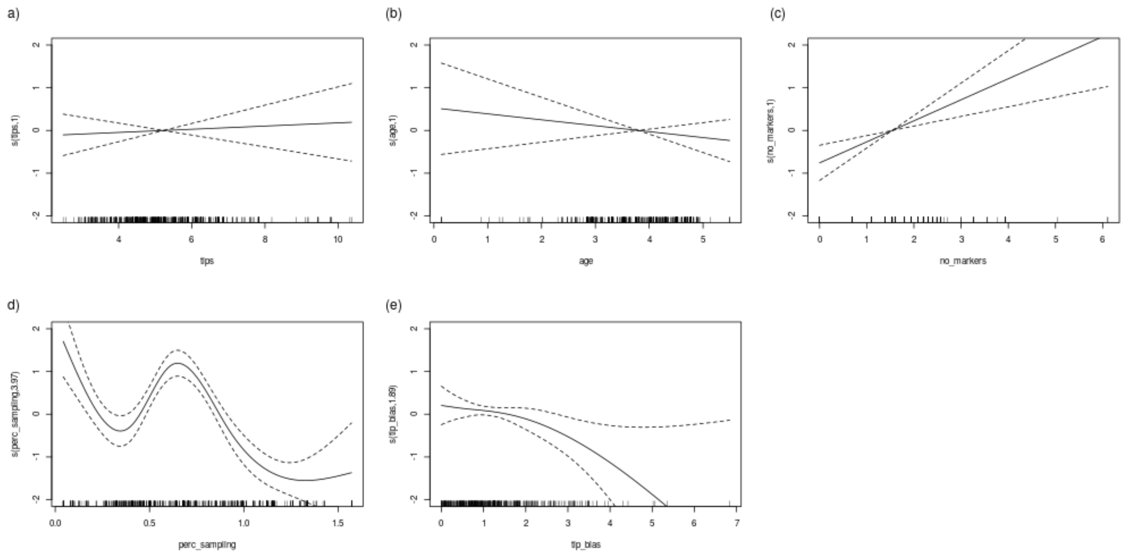

Figure S13: Panels (a-e) show the relationships in a generalized additive model that included all five of the continuous dataset properties in Figure 4 and SSE model outcome (1 = trait dependent diversification inferred in model, 0 = no effect of traits in model) as the response varible. Number of genetic markers (c), percentage sampling (d) and tip bias (e) were all significant terms in the model (n=620, R-sq.(adj) = 0.206, see Table S2 for full results). Cubic regression splines with five dimensions were fitted to each property. Missing values were replaced with column means.

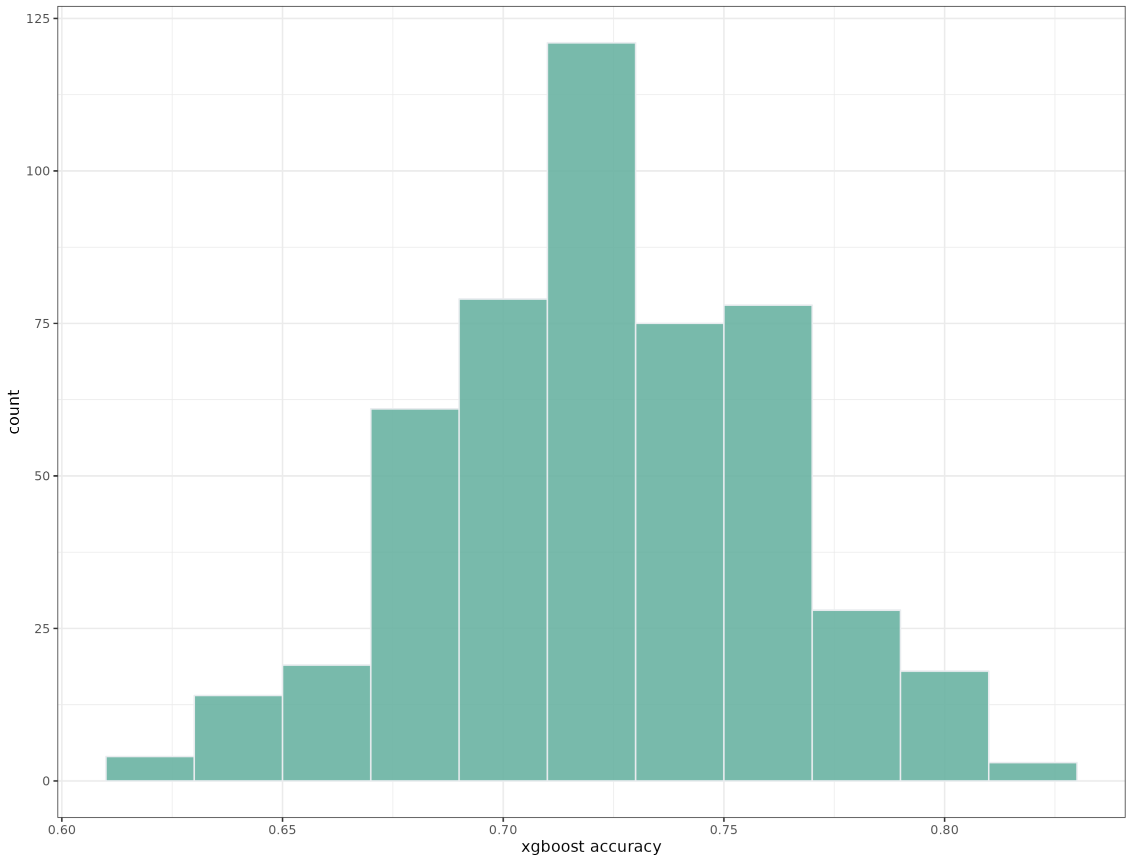

Figure S14: A histogram showing the distribution of accuracies after running 500 iterations of xgboosts, each with a different training (80%) and test (20%) datasets.

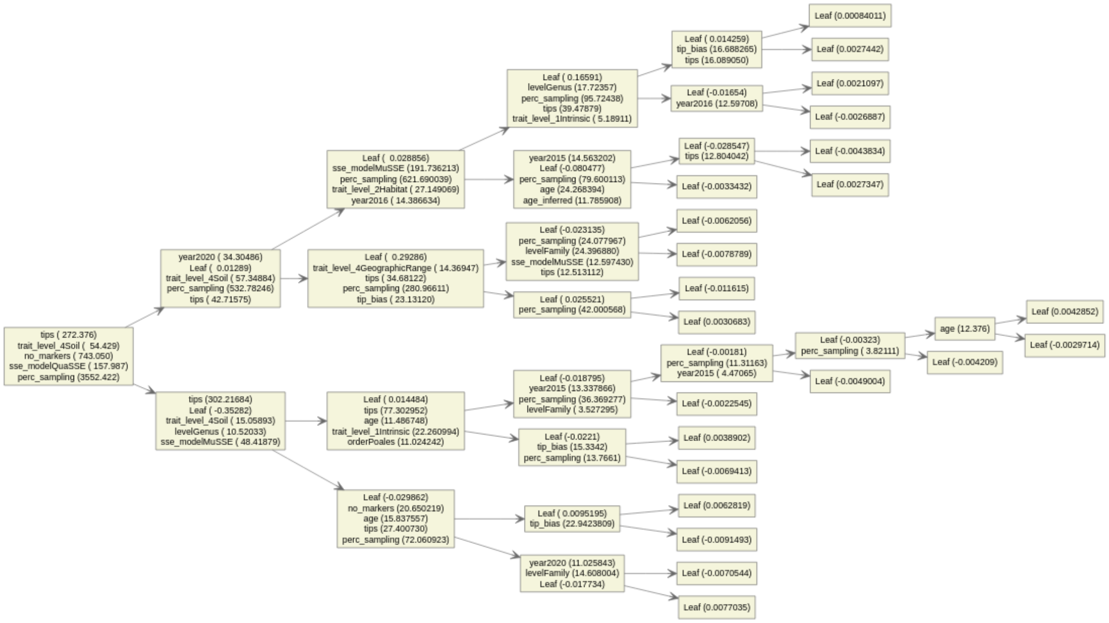

Figure S15: An example decision tree from a single run of the xgboost model used to predict the outcome of SSE models with dataset characteristics. Decisions are made from left to right and features that are used for each decision are shown in boxes, with values corresponding to their importance. ”Leaf” corresponds to when a classification is made.

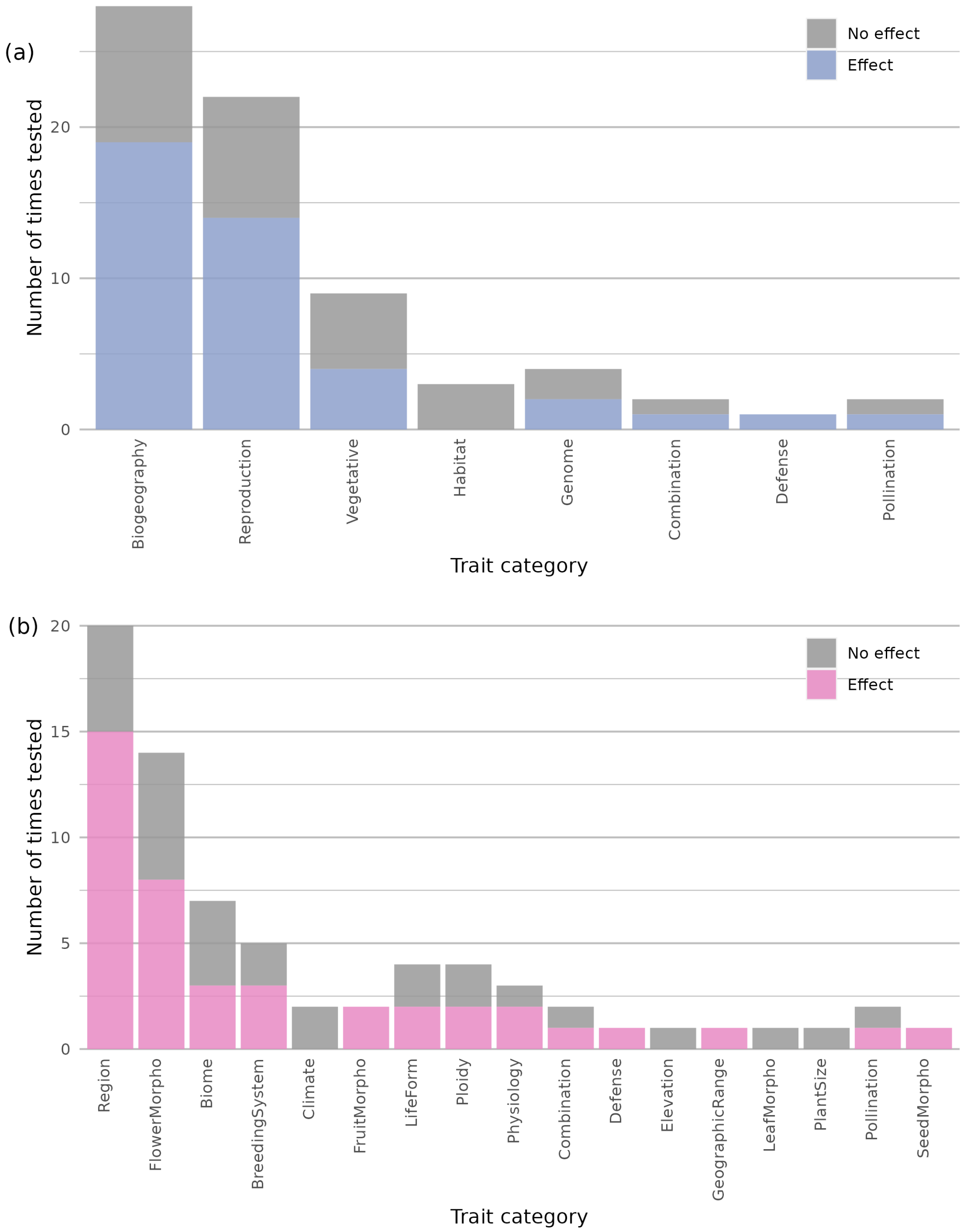

Figure S16: Stacked barplots showing how often particularly trait types were tested, for models with hidden states only. Bars are coloured to depict how often trait-dependent diversification was detected per trait type. If multiple state-dependent speciation and extinction (SSE) models were used in a single study they were considered cumulatively. Two plots are shown, (a) one with relatively narrow trait categories and (b) one with more broad categories. An ontology depicting how different trait classification levels are connected can be found in Table S1 and a similar figure including information from all SSE models in our dataset can be found in Figure 2 in the main text.

3 Supplementary tables

Table S1: A table showing the trait type ontology used to classify character states in state-dependent speciation and extinciton (SSE) models at six different levels. From left to right the classification becomes more specific. If a classification is not written at a given level, then the most specific classification that is written was used as the trait category (e.g. sexual system is used at level 5 and level 6).

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Extrinsic	Biogeography	Biome			
		Region			
	Habitat	Soil			
		Climate			
		Elevation			
		Vegetation			
	Elevation				
Intrinsic	Vegetative	Growth	LifeSpan		
			LifeForm		
		Morphology	PlantSize		
			LeafMorpho		
			PlantArchitecture	NrOfAxisCategories	
			MorphoOther		
		Physiology	Photosynthesis		
			Fire		
			Dormancy		
	Reproduction	Pre-mating	BreedingSystem	SexualSystem	
				MatingSystem	
				SexAsex	
			FlowerMorpho	Inflorescence	
				FlowerGeneral	FlowerSize
					FlowerSymmetry
					FlowerShape
					FlowerColor
					Reward
				Male	Anthers
					AntherGlands
					Pollen
				Female	Pistil
		Post-mating	FruitMorpho	FruitSize	
				FruitType	
				FruitColor	
			SeedMorpho	SeedShape	
				SeedSize	
				SeedWings	
	Genome	Ploidy			
		ChromosomeNumber			

Table S1: A table showing the trait type ontology used to classify character states in state-dependent speciation and extinciton (SSE) models at six different levels. From left to right the classification becomes more specific. If a classification is not written at a given level, then the most specific classification that is written was used as the trait category (e.g. sexual system is used at level 5 and level 6).

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Interactions	Defense				
	Symbiosis				
	Pollination				
	Dispersal				

Table S2: Results of the full generalized additive model (GAM) including five continuous dataset properties with SSE model outcome (trait-dependent diversification vs no effect) as the response.

term	**edf**	**Ref.df**	**Chi.sq**	**p-value**
s(perc_sampling)	3.969	3.999	93.410	<2e-16 ***
s(tip_bias)	1.885	2.313	6.496	0.044664 *
s(age)	1.000	1.000	0.905	0.341526
s(tips)	1.000	1.001	0.178	0.672966
s(no_markers)	1.000	1.000	13.563	0.000231 ***

References

1. Li, H.-T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. en. Nature Plants **5.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biodiversity;Phylogenetics;Plant evolution Subject_term_id: biodiversity;phylogenetics;plant-evolution, 461–470. ISSN: 2055-0278. https://www.nature.com/articles/s41477-019-0421-0 (2022) (May 2019).