Supplementary materials

Trait-dependent diversification in angiosperms: patterns, models and data

Andrew J. Helmstetter[1], Rosana Zenil-Ferguson[2], Hervé Sauquet[3,4], Sarah P. Otto[5], Marcos Méndez[6], Mario Vallejo-Marin[7], Jürg Schönenberger[8], Concetta Burgarella[9], Bruce Anderson[10], Hugo de Boer[11], Sylvain Glémin[12], Jos Käfer[13]

[1]*Fondation pour la Recherche sur la Biodiversité - Centre for the Synthesis and Analysis of Biodiversity, 34000 Montpellier, France. email: andrew.j.helmstetter@gmail.com* , [2]*Department of Biology, University of Kentucky, 101 T.H. Morgan Building, Lexington, KY 40506 USA. email: roszenil@uky.edu*, [3]*National Herbarium of New South Wales, Royal Botanic Gardens and Domain Trust, Sydney, New South Wales, 2000, Australia*, [4]*Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia. email: herve.sauquet@gmail.com*, [5]*Department of Zoology, 6270 University Blvd., University of British Columbia, Vancouver BC, V6T 1Z4, Canada. email: otto@zoology.ubc.ca*, [6]*Area of Biodiversity and Conservation, Universidad Rey Juan Carlos, 28933 Móstoles (Madrid), Spain. email: marcos.mendez@urjc.es*, [7]*Department of Ecology and Genetics, Uppsala University, Uppsala, 752 36, Sweden. email: mario.vallejo-marin@ebc.uu.se*, [8]*Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria. email: juerg.schoenenberger@univie.ac.at*, [9]*Department of Organismal Biology, University of Uppsala, Uppsala, 75236, Sweden. email: concetta.burgarella@gmail.com*, [10]*Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa. email: banderso.bruce@gmail.com*, [11]*Natural History Museum, University of Oslo, 0318 Oslo, Norway. email: h.d.boer@nhm.uio.no*, [12]*CNRS, Ecosystèmes Biodiversité Evolution (Université de Rennes), 35000 Rennes, France. email: sylvain.glemin@univ-rennes1.fr*, [13]*Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France; DIADE, Université Montpellier, IRD, CIRAD, Montpellier, France. email: jos.kafer@cnrs.fr*

Correspondence: Andrew J. Helmstetter, FRB-CESAB Institut Bouisson Bertrand, 5, rue de l'École de médecine - 34000 MONTPELLIER. Tel: +33 4 11 28 20 58. Email: andrew.j.helmstetter@gmail.com

Contents

1 Appendix S1: 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 1).

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 (i.e. no preprints).

sse_model The name of the state-dependent speciation and extinction (SSE) model used. There may be multiple models of the same (or different) types per study. See Fig. 1 for the range of SSE models considered.

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".

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. GeoSSE and GeoHiSSE models are specifically designed to assess diversification differences among geographic regions and here we only consider states representing the geographic regions, omitting the special state "widespread" used for taxa that are present in both regions.

putative_ancestral_state A binary column that indicates the character state that 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.

level The taxonomic level of the clade that the SSE model was run on (e.g. genus, tribe or order).

order The order the study clade belongs to. If the clade was larger than order the category "multiple" was used.

family The family the study clade belongs to. If the clade was larger than family the category "multiple" was used.

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 interpreted results based on the available information and the narrative of the study. In multi-state models we classified the state with the lowest net diversification rate as 0 and other states as 1,2,3... etc depending on their net diversification rate. If two states are explained as having comparable rates in the text they are assigned the same number.

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.

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 chloroplast 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).

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.

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 where possible.

transition_direction The direction of the transition rate (e.g. 0 to 1).

transition_rate The transition rate between the states indicated in transition direction (transitions per lineage per million years).

div_rate The net diversification rate (per lineages per million years). If hidden states were included in the model this was approximated as the mean across hidden states (e.g. $(r_1A + r_1B)/2$) where possible. Speciation rates were used for FiSSE as extinction rates were not reported.

2 Supplementary figures

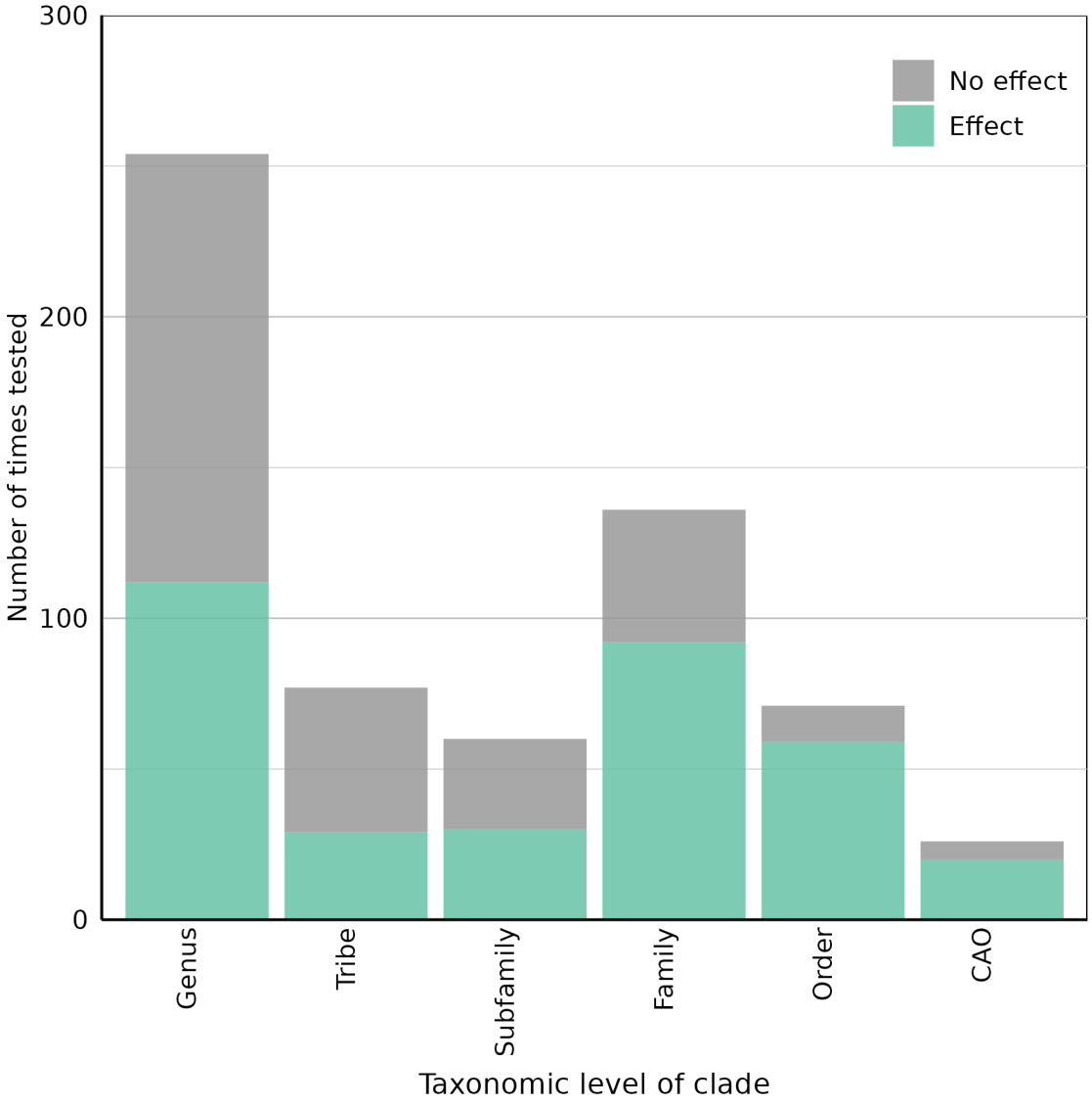

Figure S1: Stacked barplots showing the number of models run per different taxonomic levels. The smallest level is genus, increasing in scope until clades above order (abbreviated as CAO). Note that the category tribe includes subtribes. Bars are coloured based on whether trait-dependent diversification was inferred in the model.

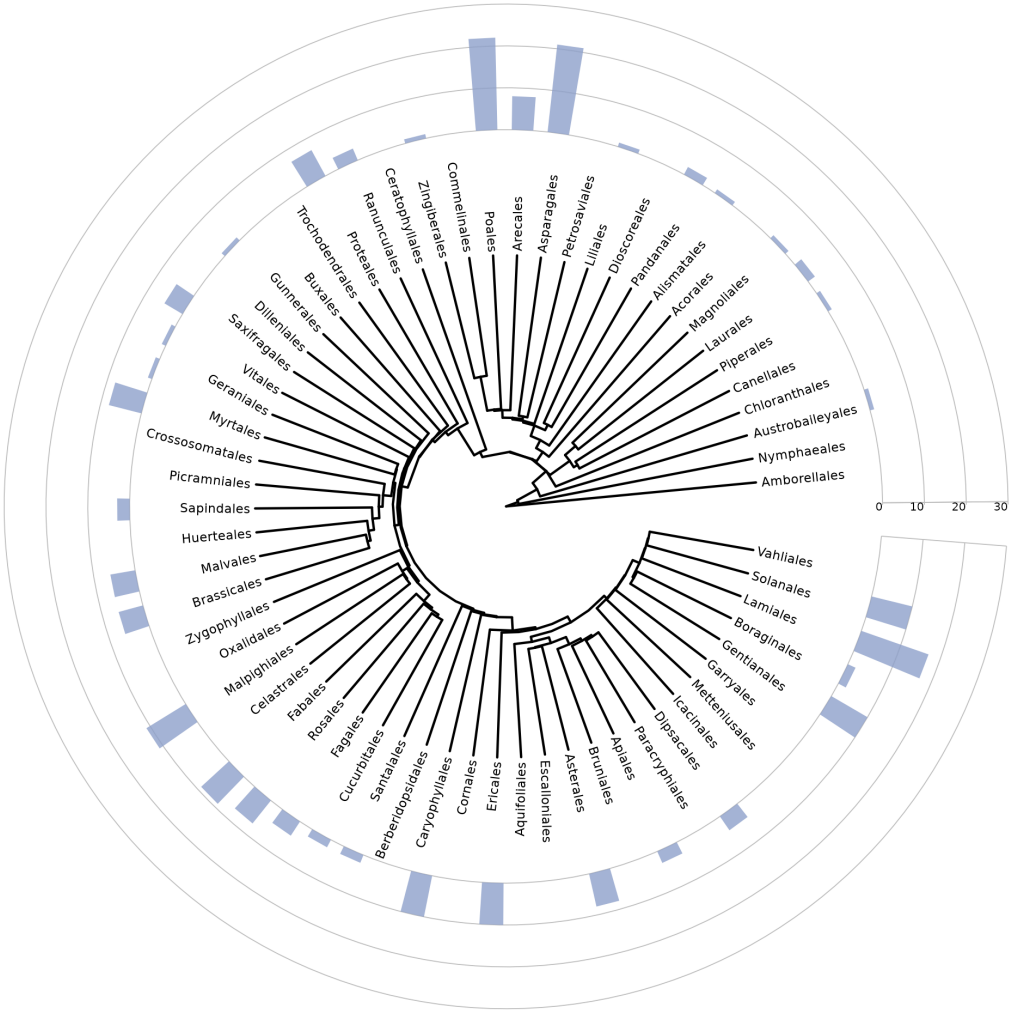

Figure S2: A phylogenetic tree of angiosperm orders taken from Li et al. (2019) annotated with bars representing the number of studies using SSE models that focused on each order.

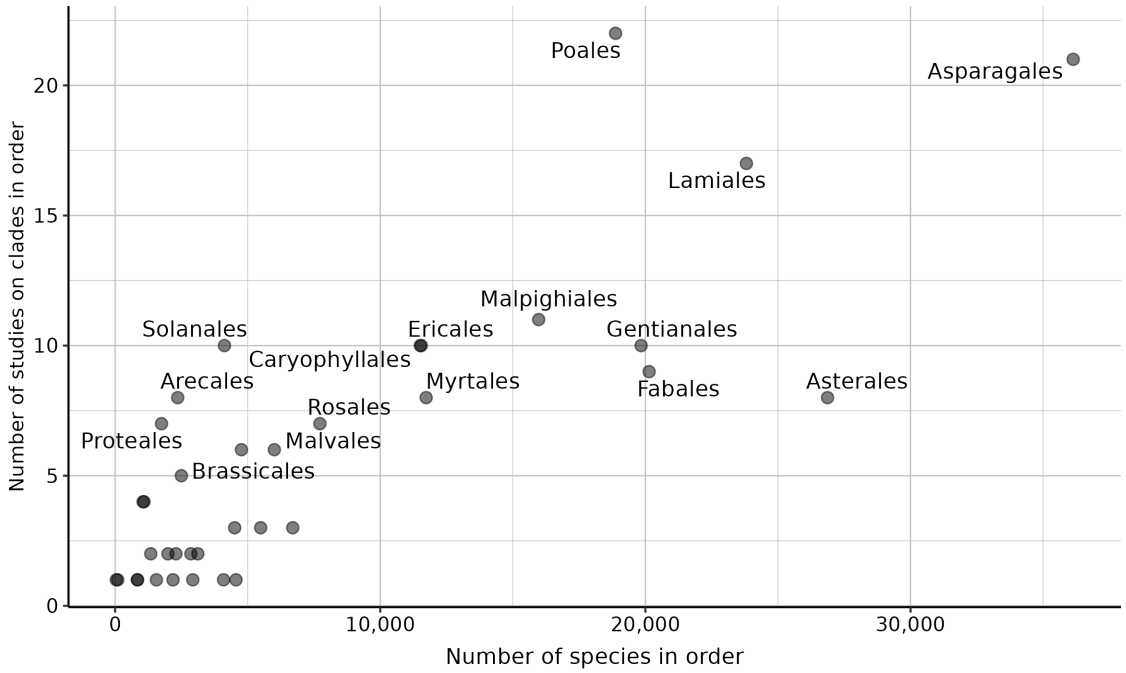

Figure S3: A scatterplot showing the relationship between the number of species in each order (taken from http://www.mobot.org/) and the number of papers that studied clades in that order. Only those orders in our dataset are included. SSE models used on clades larger than order (i.e. those coded as 'multiple') were not counted. Text labels have been added to the points corresponding to large, commonly studied orders in our dataset.

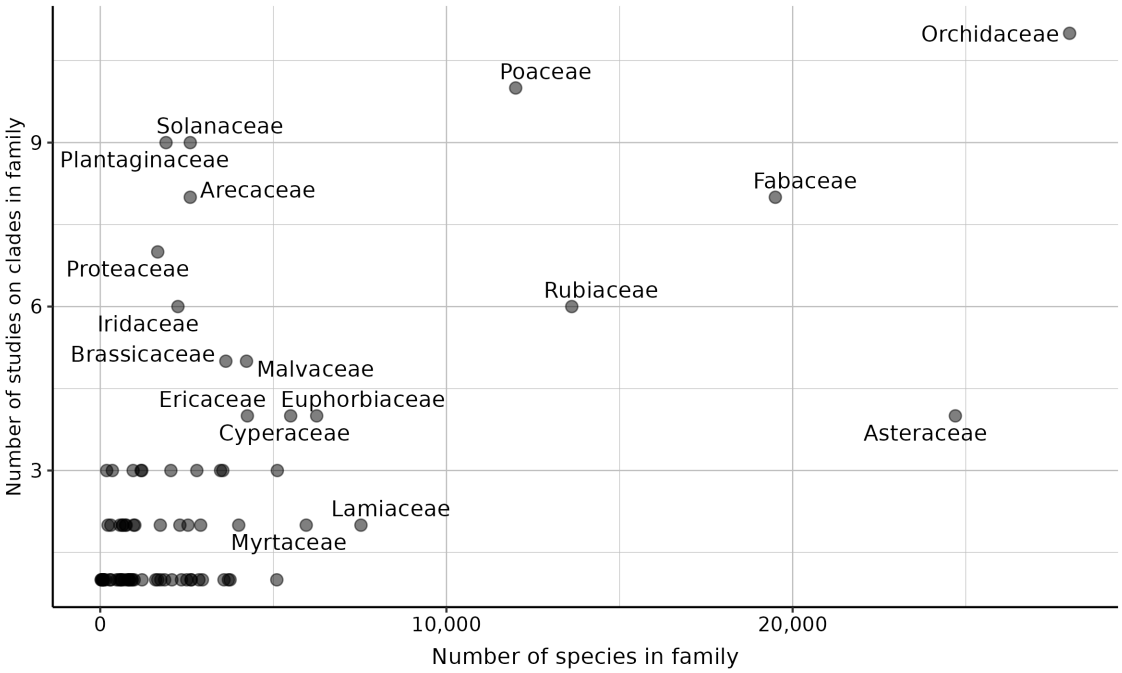

Figure S4: A scatterplot showing the relationship between the number of species in each family (taken from Christenhusz and Byng, 2016) and the number of papers that studied clades in that family. Only those families in our dataset are included. SSE models used on clades larger than family (i.e. those coded as 'multiple') were not counted. Text labels have been added to the points corresponding to large, commonly studied families in our dataset.

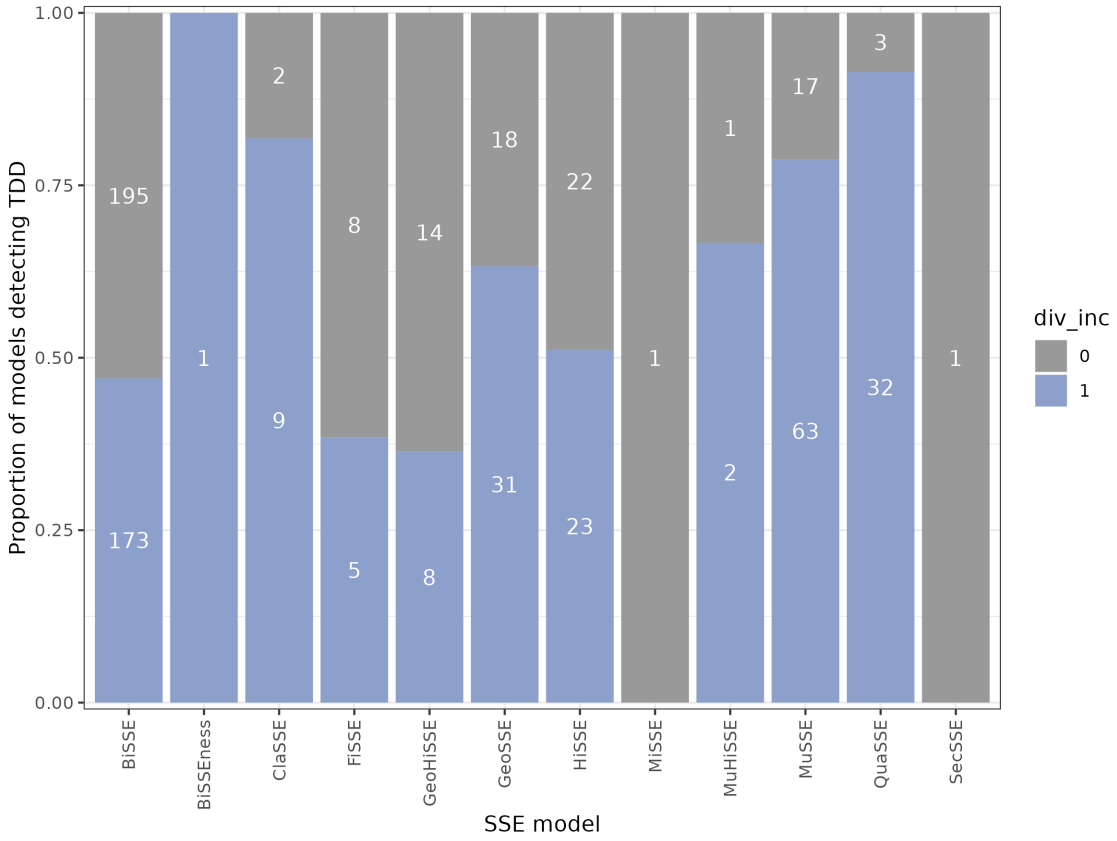

Figure S5: Stacked barplot showing the relative frequency of state-dependent diversification (TDD) per model type. Each bar represents an SSE model type and bars are coloured by the result of each model where blue indicates TDD and grey indicates no effect detected.

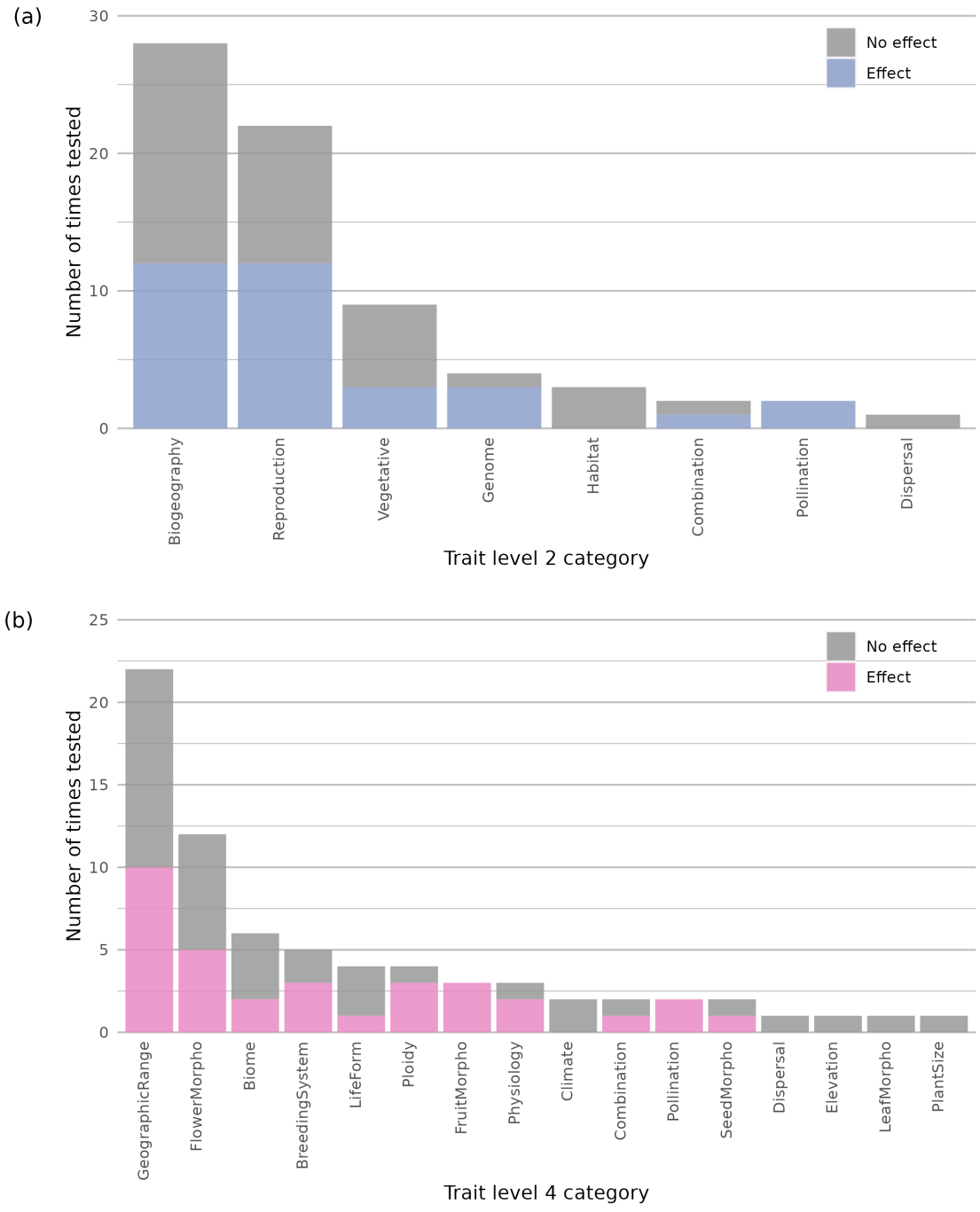

Figure S6: Stacked barplots showing how often particular trait types were tested, for models with hidden states only. Bars are coloured to depict how often trait-dependent diversification was detected per trait type. If multiple state-dependent speciation and extinction (SSE) models were used in a single study they were considered cumulatively. Two plots are shown, (a) one with broad level 2 trait categories and (b) one with more narrow level 4 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.

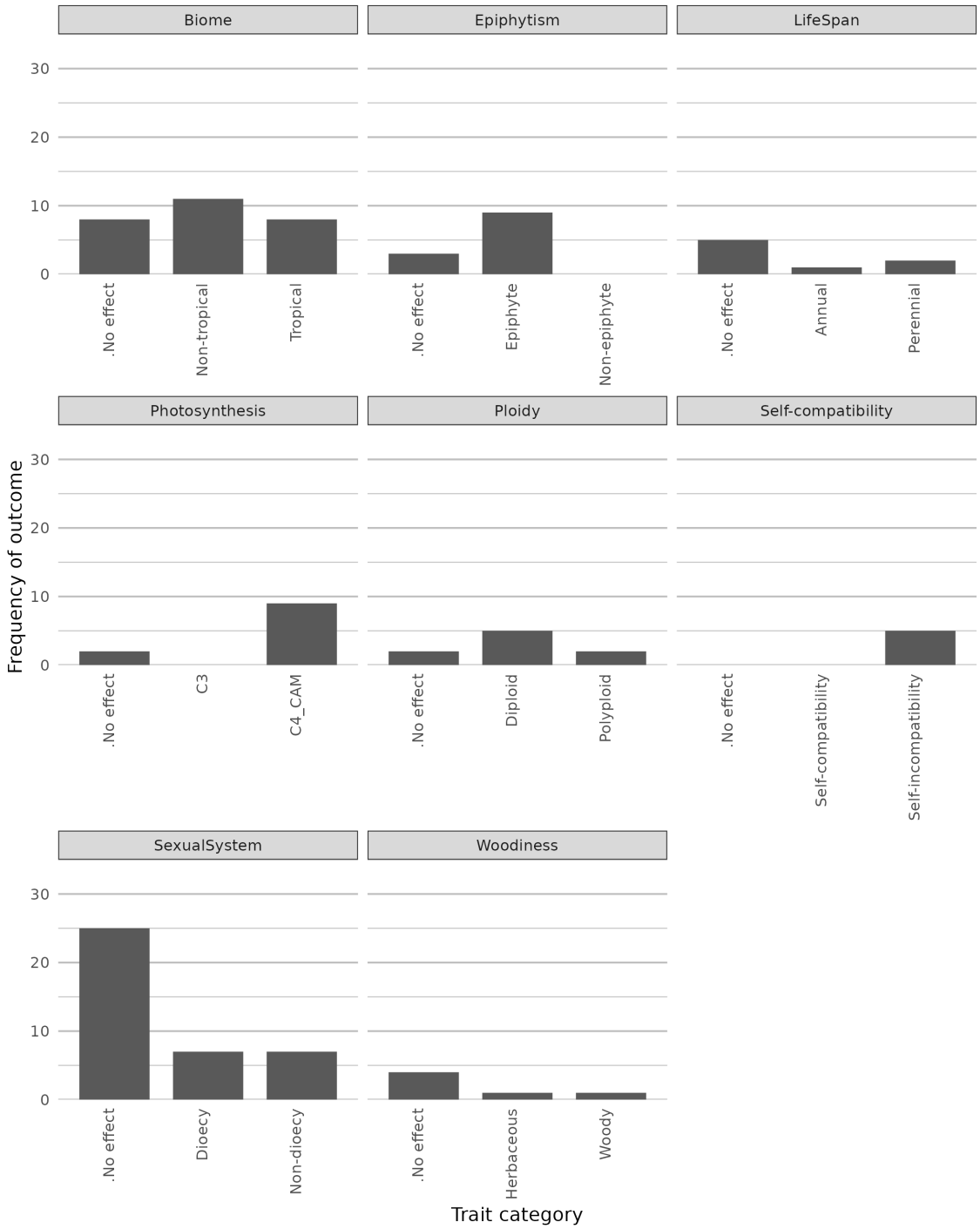

Figure S7: Barplots for eight different character state groups. For each character state there is a bar representing the number of models that state was found to increase diversification rate. If there was no effect of either state in a model, this was counted in the no effect bar. To aid comparisons here and in Fig. S8, some state names were changed: for example for the photosynthesis category, 'C4' and 'CAM' results were combined to 'C4_CAM', 'Temperate' was changed to 'Non-tropical' and 'Terrestrial' and 'Large tree' were changed to 'Non-epiphyte'.

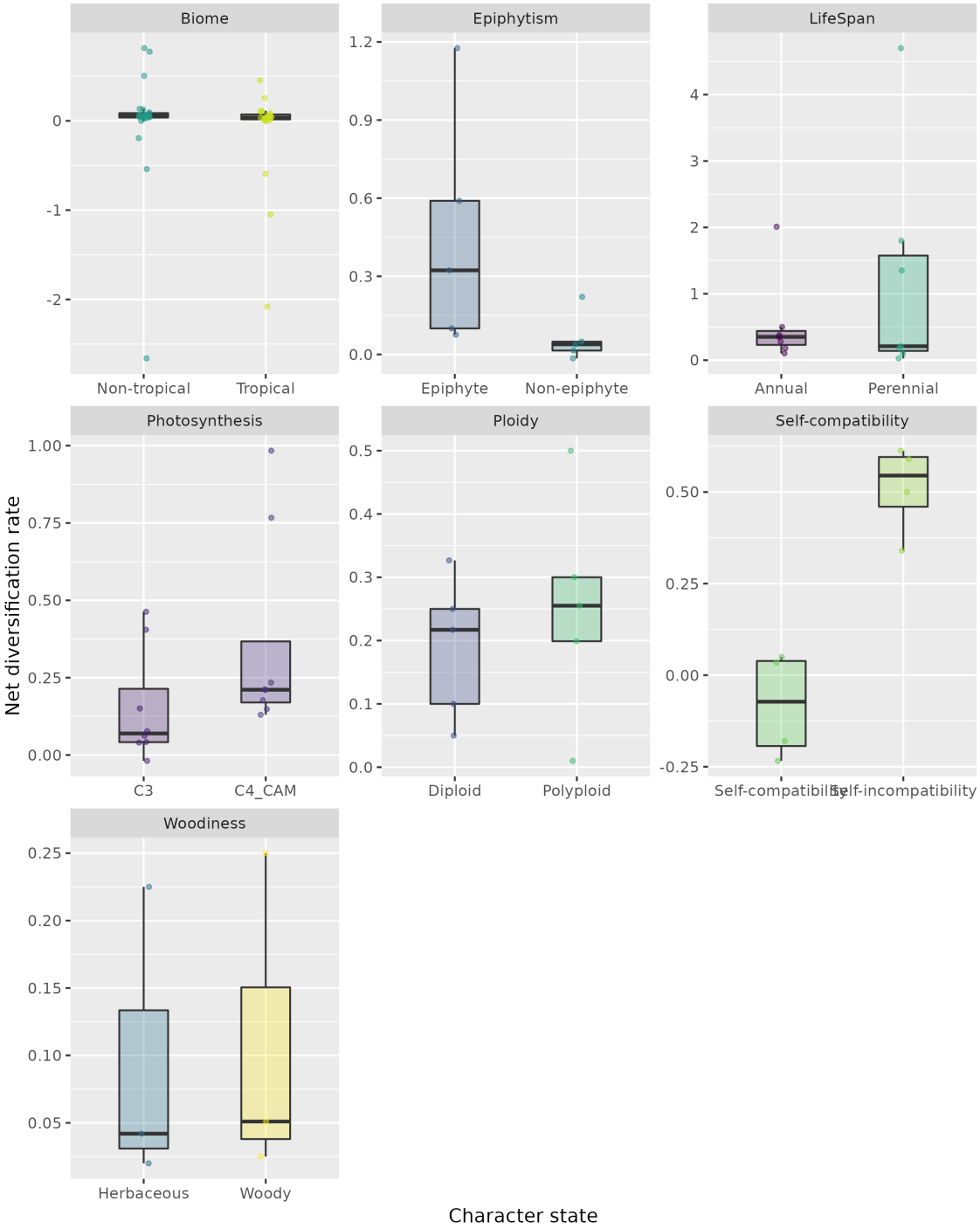

Figure S8: Boxplots showing the net diversification rates for seven traits and their associated states. Jittered points overlain on the boxplots indicate the mean net diversification rate values recovered from each model. Statistical tests were conducted but significant differences ($p < 0.05$) were not found except for "Photosynthesis" ($p = 0.0142$). Rates appeared to differ substantially in "Self-compatibility" and "Epiphytism" but the paired Wilcoxon tests run on these traits appeared not to have enough power to attain the significance threshold.

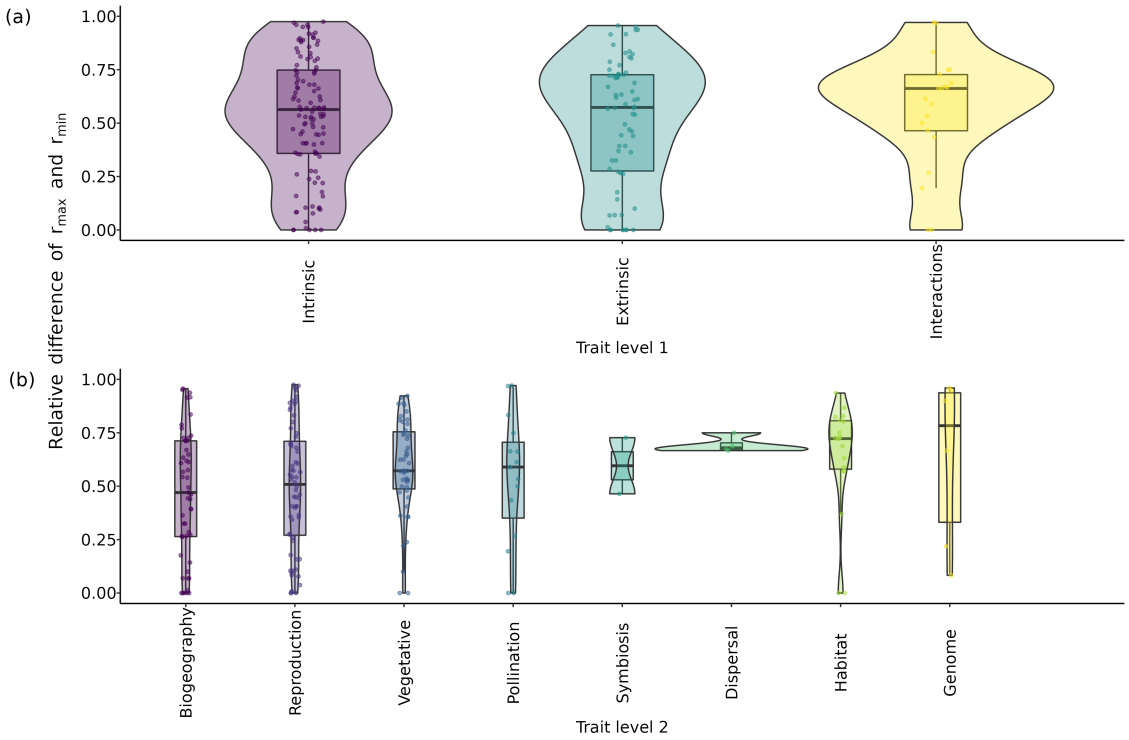

Figure S9: Violin plots showing relative difference between minimum and maximum net diversification rates for models belonging to (a) trait level 1 categories and (b) trait level 2 categories. Combination categories were not considered and we were unable to find any diversification rates for the defense category. We calculated the relative differences using the formula $(r_{max} - r_{min})/r_{max}$. Only those models in which all rates were positive were used (n = 208). Jittered points overlain on the violin plots indicate the ratio values recovered from each model. For (a) trait level 1, we found relative rates were generally similar across categories (Kruskal-Wallis chi-squared = 0.652, df = 2, p-value = 0.722). For (b) trait level 2, traits related to the genome had the largest median relative difference but categories were not significantly different (Kruskal-Wallis chi-squared = 11.581, df = 7, p-value = 0.115).

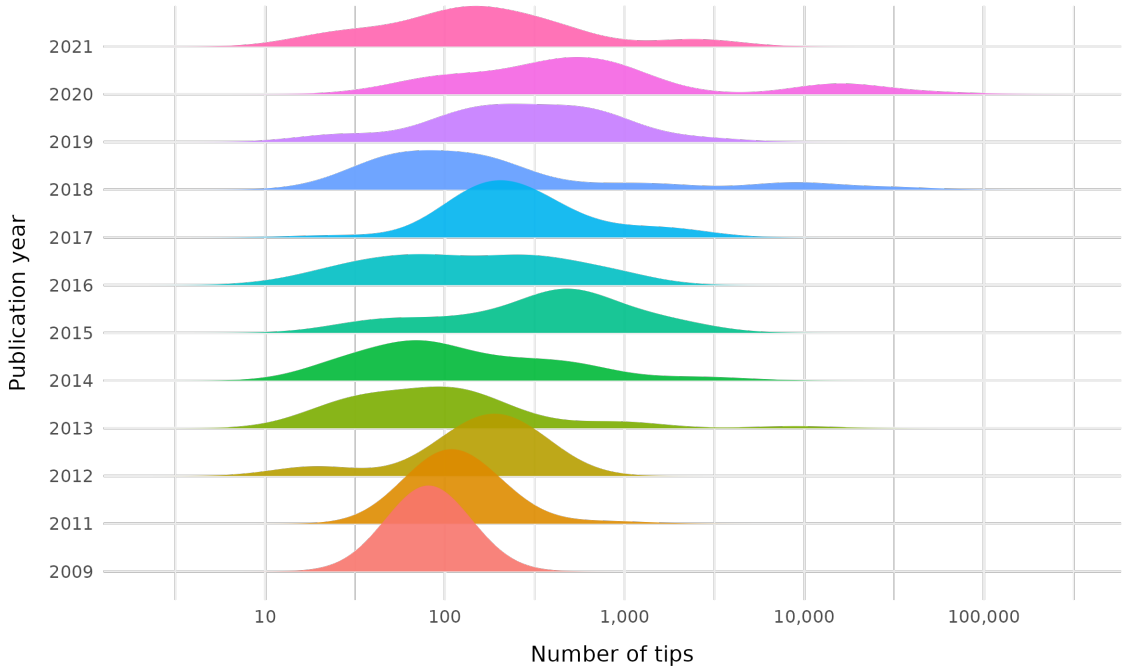

Figure S10: A ridgeplot showing how the number of tips on trees used with SSE models have changed over time. Each ridge displays a density plot corresponding to a single publication year (2009-2021) with the most recent year at the top of the plot. The x-axis is on a log scale. There was not enough data from 2010 to calculate a density so this year was removed from the plot.

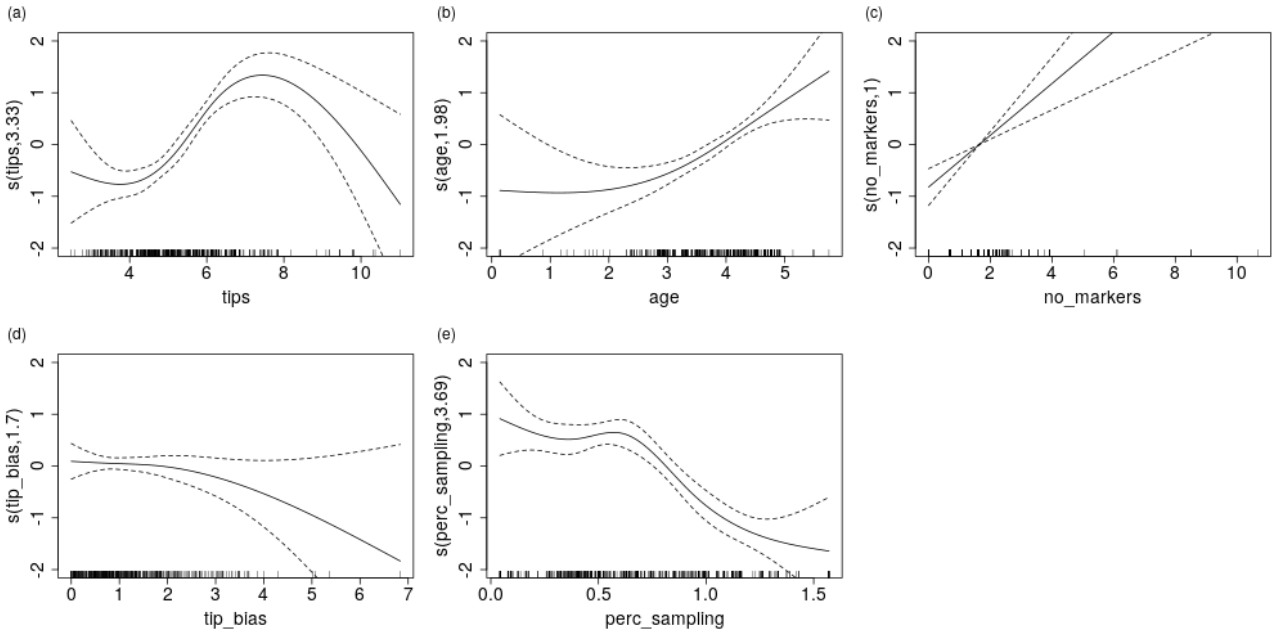

Figure S11: Panels (a-e) show the relationships inferred with generalized additive models (GAM) between each continuous dataset property in Figure 4 and SSE model outcome. Values that are positive on the y-axis indicate that the dataset property at this x-axis value tends to be associated with trait-dependent diversification. Negative y-axis values indicate when dataset properties are associated with no effect. Cubic regression splines were fitted to each property independently.

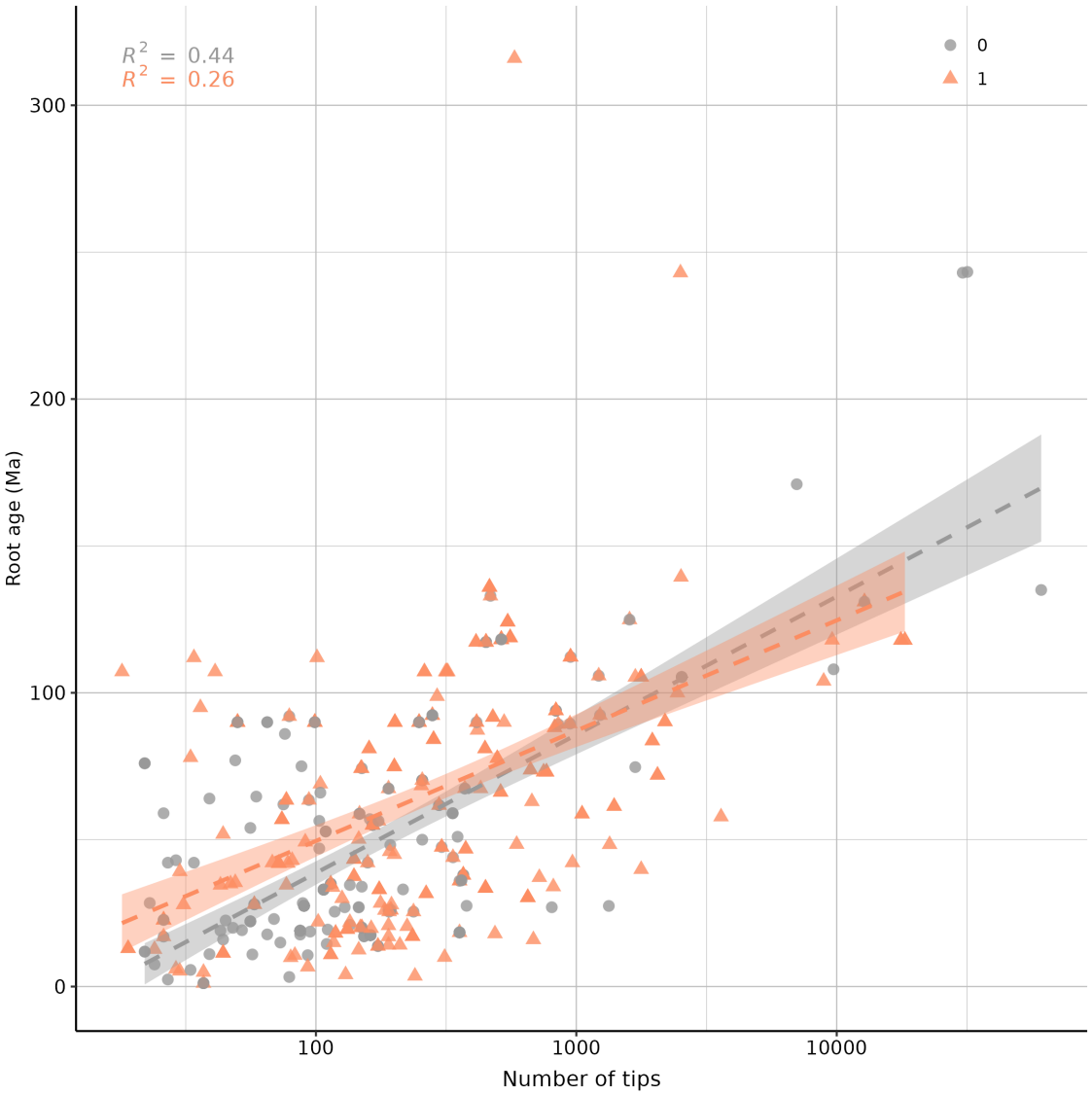

Figure S12: A scatterplot showing the relationship between the root age of the trees used with SSE models in our dataset and the number of tips in the trees. Points are coloured based on whether trait-dependent diversification was inferred (coloured) or not (grey) when the associated model was run. Lines were fitted using linear models to these two groups with 95% confidence intervals estimated.

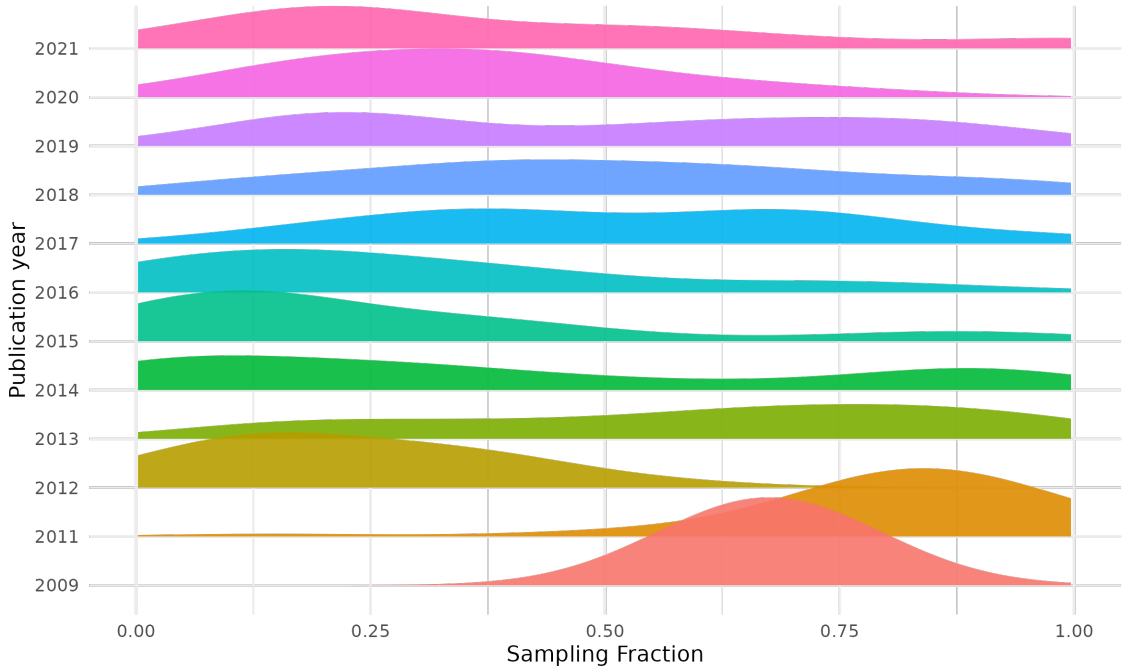

Figure S13: A ridgeplot showing how sampling fractions of trees used in SSE models have changed over time. Each ridge displays a density plot corresponding to a single publication year (2009-2021) with the most recent year at the top of the plot. The x-axis is on a log scale. There was not enough data from 2010 to calculate a density so this year was removed from the plot.

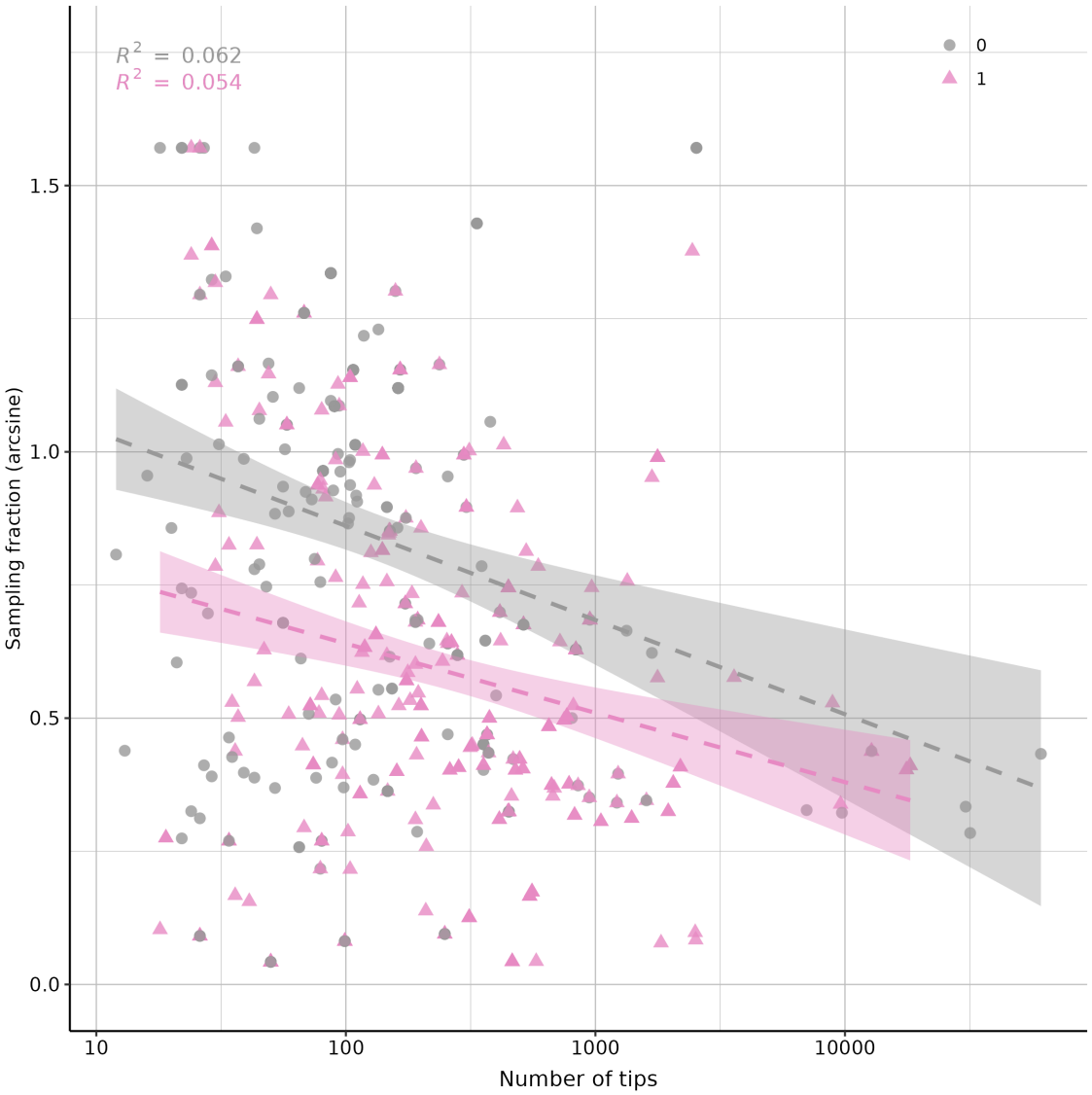

Figure S14: A scatterplot showing the relationship between sampling fraction of the tree used with an SSE model and the number of tips in the tree. Coloured points are models for which trait-dependent diversification was detected, and grey points are models where it was not detected. Lines were fitted using linear models to these two groups with 95% confidence intervals estimated.

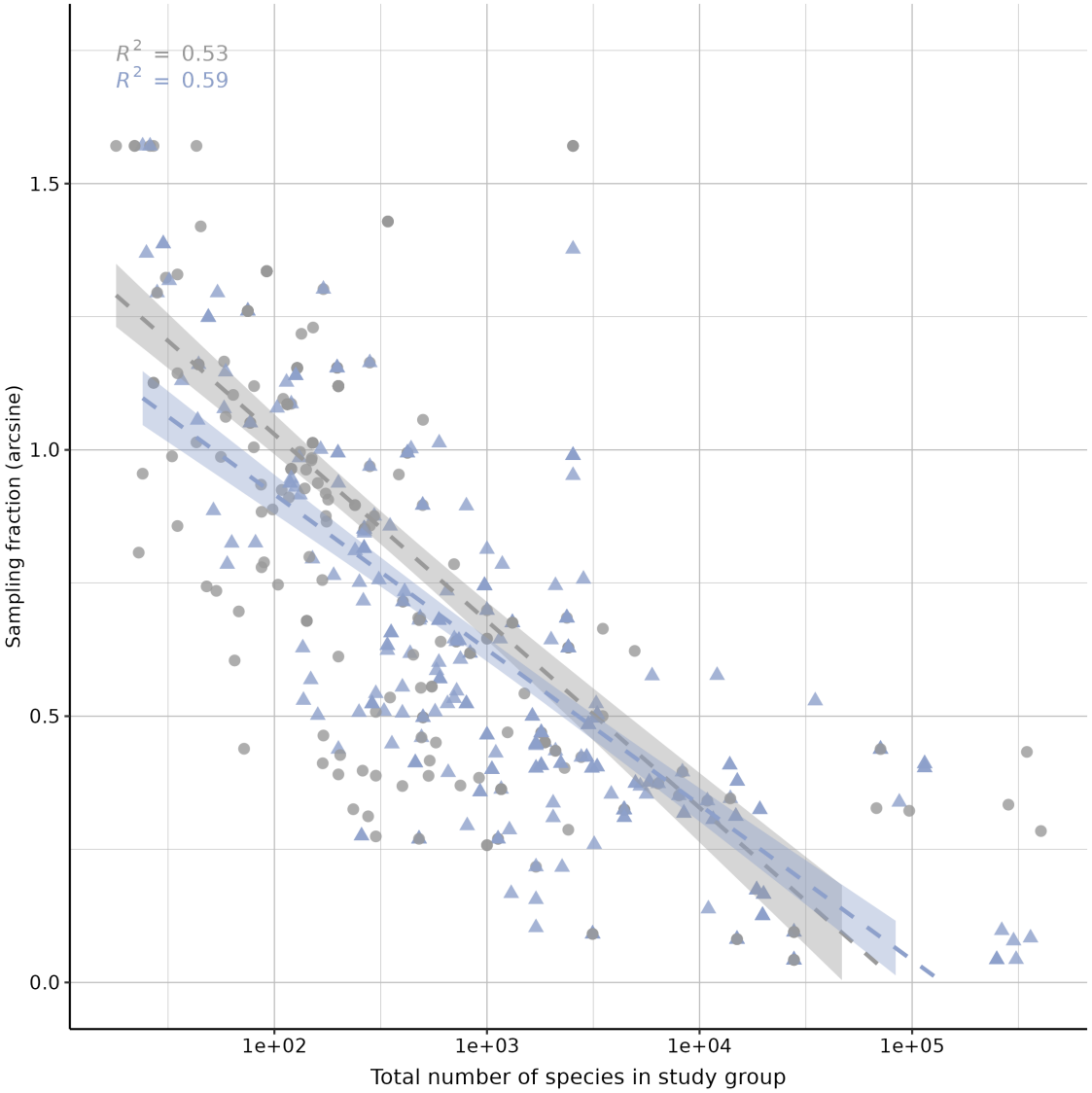

Figure S15: A scatterplot showing the relationship between sampling fraction of the tree used with an SSE model and the total number of species in the study group the tree represents. Coloured points are models for which trait-dependent diversification was detected, and grey points are models where it was not detected. Lines were fitted using linear models to these two groups with 95% confidence intervals estimated.

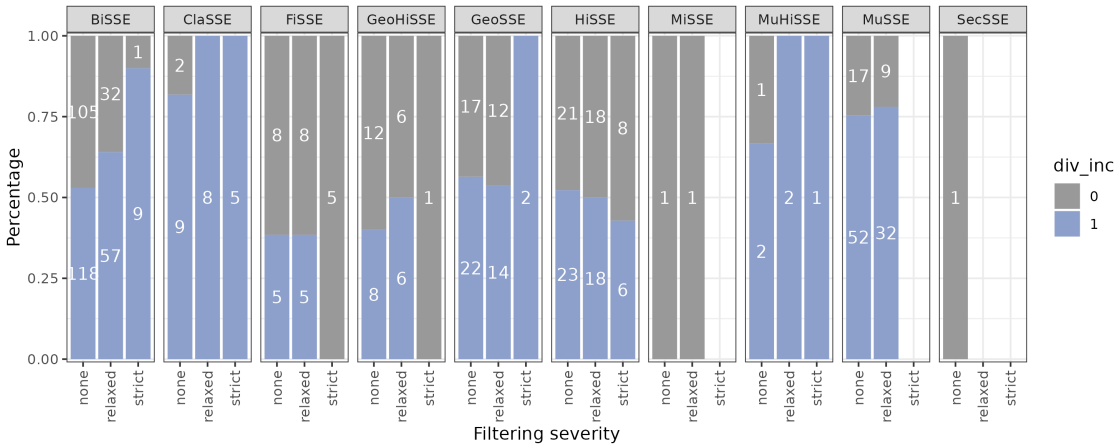

Figure S16: Stacked barplots showing the relative frequency of trait-dependent diversification (TDD) at three different levels of dataset filtering. Bars are grouped by SSE model type. Within each group, each bar represents a filtering approach with increasing severity. 'none' indicates that all models were kept. 'relaxed' only includes models where number of tips > 100, global sampling fraction $>10\%$ and tip bias ratio is 20:1 or less. 'strict' only includes models where number of tips > 300, global sampling fraction $>25\%$ and tip bias ratio is 10:1 or less. Models were only included if each of the three dataset properties were recovered. QuaSSE is not included as a tip bias could not be calculated. SSE model type and bars are coloured by the result of each model where blue indicates TDD and grey indicates no effect detected.

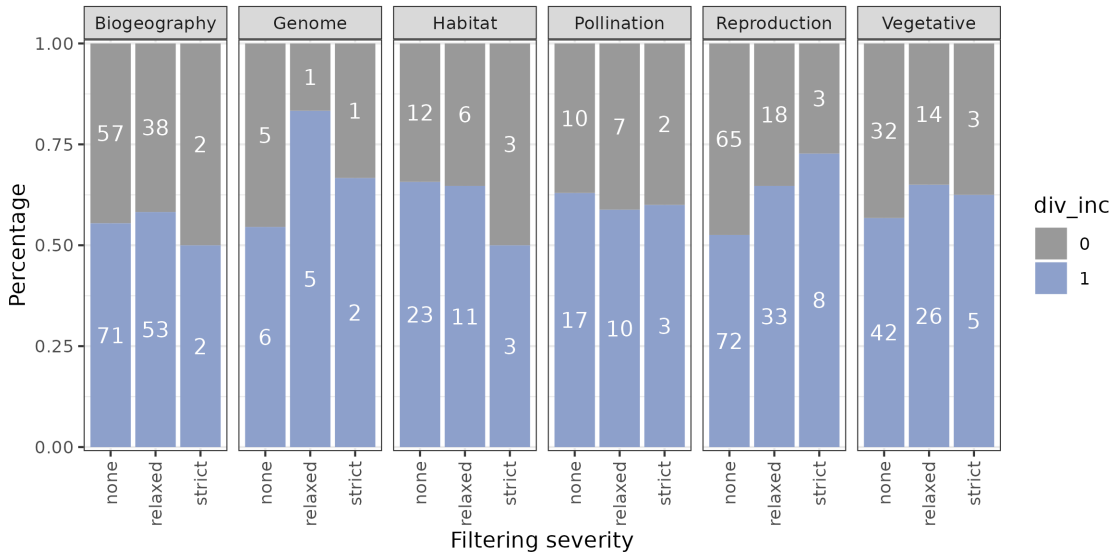

Figure S17: Stacked barplots showing the relative frequency of trait-dependent diversification (TDD) at three different levels of dataset filtering. Bars are grouped by trait category (level 2), for commonly tested trait categories only. Within each group, each bar represents a filtering approach with increasing severity. 'none' indicates that all models were kept. 'relaxed' only includes models where number of tips > 100, global sampling fraction >10% and tip bias ratio is 20:1 or less. 'strict' only includes models where number of tips > 300, global sampling fraction >25% and tip bias ratio is 10:1 or less. Models were only included if each of the three dataset properties were recovered. SSE model type and bars are coloured by the result of each model where blue indicates TDD and grey indicates no effect detected.

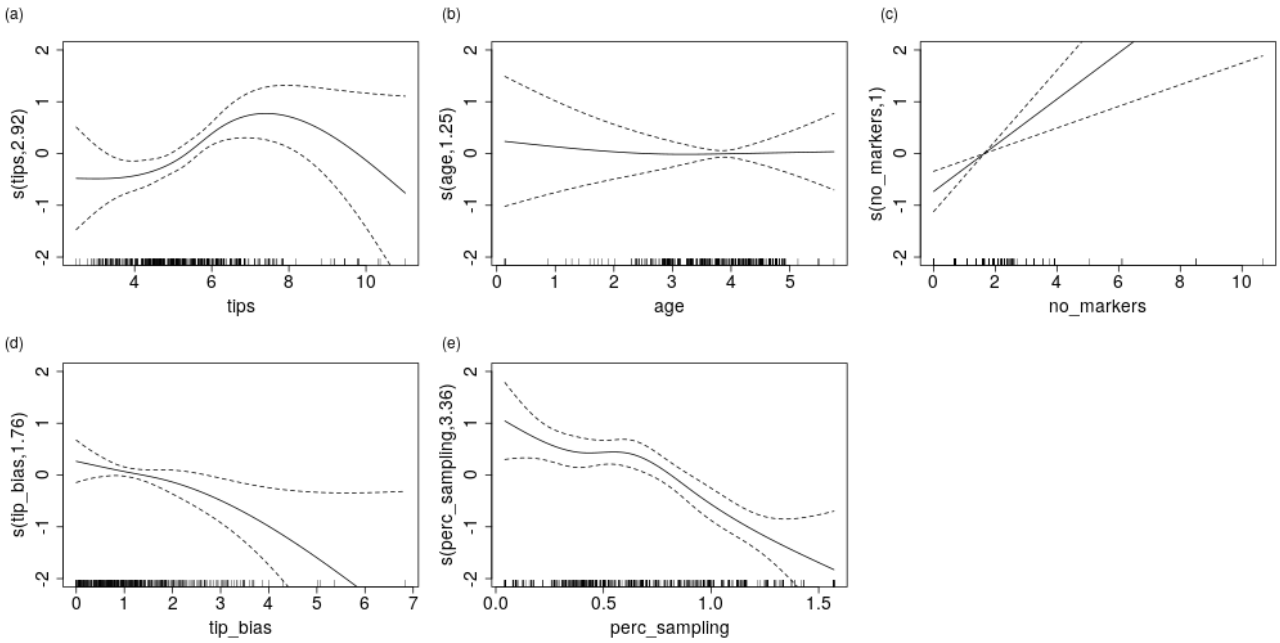

Figure S18: Panels (a-e) show the relationships in a generalized additive model that included all five of the continuous dataset properties in Figure 4 and SSE model outcome as the response variable. Missing values were replaced with column means. Values that are positive on the y-axis indicate that the dataset property at this x-axis value tends to be associated with trait-dependent diversification. Negative y-axis values indicate when dataset properties are associated with no effect. All variables were significant terms in the model except age, see Table S2 for full results.

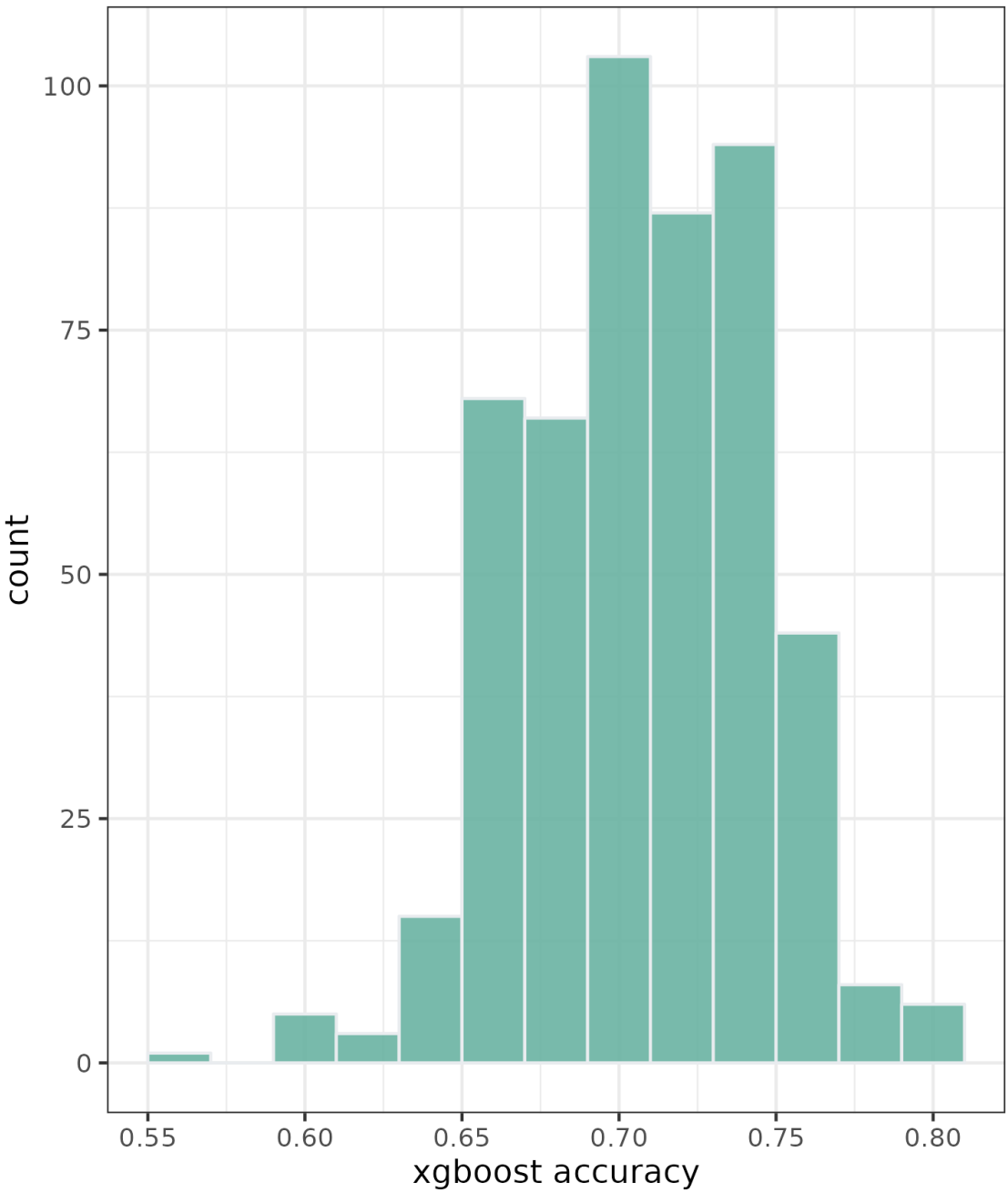

Figure S19: A histogram showing the distribution of accuracies (whether the correct SSE model outcome is predicted) after running 500 iterations of xgboosts. Each run used a different training (80%) and test (20%) dataset to capture the stochasticity in the dataset partitioning.

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). If a state does not fall into a more specific classification then the higher level classification it previously belonged to is kept.

Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Extrinsic	Biogeography	Biome			
		GeographicRange			
	Habitat	Soil			
		Climate			
		Elevation			
		Vegetation			
Intrinsic	Vegetative	Growth	LifeSpan		
			LifeForm		
		Morphology	PlantSize		
			LeafMorpho		
			PlantArchitecture	NrOfAxisCategories	
			MorphoOther		
		Physiology	Photosynthesis		
			Fire		
			Dormancy		
			NutrientAcquisition		
	Reproduction	Pre-mating	BreedingSystem	SexualSystem	
				MatingSystem	
				SexAsex	
			FlowerMorpho	Inflorescence	
				FlowerGeneral	FlowerSize
					FlowerSymmetry
					FlowerShape
					FlowerColor
					Reward
				Male	Anthers
					Pollen
				Female	Pistil
		Post-mating	FruitMorpho	FruitSize	
				FruitType	
				FruitColor	
			SeedMorpho	SeedShape	
				SeedSize	
				SeedWings	
	Genome	Ploidy			
		ChromosomeNumber			
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. 'edf' is the estimated degrees of freedom for the model terms. 'Ref.df' is the reference degrees of freedom. 'Chi.sq' is the test statistic for assessing the significance of model smooth term.

term	**edf**	**Ref.df**	**Chi.sq**	**p-value**
s(tips)	2.924	3.406	13.729	0.005 **
s(age)	1.245	1.448	0.114	0.898
s(no_markers)	1.000	1.000	14.165	<0.001 ***
s(tip_bias)	1.763	2.173	7.044	0.032 *
s(perc_sampling)	3.362	3.775	43.687	<2e-16 ***

References

Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase [Number: 3]. Phytotaxa, 261(3), 201–217. https://doi.org/10.11646/phytotaxa.261.3.1

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M. A., Fritsch, P. W., Cai, J., Luo, Y., Wang, H., van der Bank, M., Zhang, S.-D., Wang, Q.-F., Wang, J., Zhang, Z.-R., Fu, C.-N., Yang, J., Hollingsworth, P. M., ... Li, D.-Z. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. Nature Plants, 5(5), 461–470. https://doi.org/10.1038/s41477-019-0421-0