Supporting Information

Guilhaumon *et al.* 10.1073/pnas.0803610105

SI Materials and Methods

Information Theoretical Analysis of SARs. Over the last decade, statistical science has been moving away from the classical Null Hypotheses Testing (NHT, refs. 1, 2) framework. In the present study, nonnested SAR models are subject to selection which raises methodological issues: hypothesis testing is problematic in this context, and the use of classical tools such as the adjusted coefficient of multiple determination (R_a^2) is not advocated as it has no strong theoretical justification (3). The information-theoretic framework for model-selection departs from the NHT paradigm as it is based on the evaluation of multiple working hypotheses. This evaluation of concurring hypotheses each represented by a different model is achieved through the estimation of Kullback-Leibler (K-L) information (or distance, ref. 4):

$$I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x|\theta)}\right)dx \qquad [1]$$

is the "information" lost when the model g (with parameters θ) is used as an approximation of the full reality or truth f. Similarly, $I(f,g)$ is interpreted as the distance from the approximating model to full reality (5). The principle of information theoretical model selection is to find within a predefined set of models the one that minimizes $I(f,g)$. However, because it requires knowledge both of full reality and the parameters for each candidate model, the calculation of the K-L distance remains intractable. Akaike (6) developed a method to approximate $I(f,g)$ based on the empirical log-likelihood function: the Akaike Information Criterion (AIC):

$$AIC = -2\log(L(y|\hat{\theta})) + 2p \qquad [2]$$

where $\hat{\theta}$ is the maximum likelihood estimate of the vector of parameters, $L(y \mid \hat{\theta})$ is the logarithm of the likelihood of the data evaluated at $\theta = \hat{\theta}$, and p is the number of estimated parameters in the model (which includes the estimated variance).

Although AIC constitutes the foundation of the information-theoretic model selection framework, it may perform poorly when the sample size n is small (more precisely when $n/P < 40$, ref. 5). To account for this potential source of bias, Sugiura (7) derived the corrected Akaike's information criterion (*AICc*):

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}. \qquad [3]$$

Equivalently, in the context of nonlinear regression (under assumption of normality of residuals and homoscedasticity, ref. 5):

$$AICc = n\log\left(\frac{RSS}{n}\right) + 2np\frac{n}{n-np-1} \qquad [4]$$

where RSS is the residual sum of squares.

Information-theoretic Criteria (IC such as AIC and AICc) are built such that the first term, representing the lack of fit of the model to the observed data, is penalized by the second term, which captures model complexity. The lower the IC associated with a model, the better this model is considered in explaining the data. In the present study, we used AIC or AICc when appropriate. Because AIC and AICc produce relative measures, absolute values are not relevant to compare models and the selection is usually based on Akaike weights. For a fitted model i, its weight w_i is given by:

$$w_i = \frac{e^{-1/2\Delta_i}}{\sum_{r=1}^{M} e^{-1/2\Delta_r}} \qquad [5]$$

where M is the number of models in the set and Δ_i is defined as $\Delta_i = IC_i - IC_{\min}$ with $IC_{\min}$ the IC value for the best model. Akaike's weights are interpreted in terms of probabilities of a given model being the best in explaining the data within a predefined set of alternative models. When the data support more than a single model (i.e., no w_i is higher than 0.9; ref. 5), robust inferences can be carried out by averaging inferences within the set of models with respect to their w_i. As advocated for non-nested models, we obtained multimodel SARs by averaging the model predictions with respect to their weight:

$$\bar{\hat{S}} = \sum_{i=1}^{M} \hat{S}_i w_i \qquad [6]$$

where $\bar{\hat{S}}$ is the multimodel averaged species richness and $\hat{S}_i$ is the vector of species richness inferred from model i.

Confidence Intervals and Ecoregion Ranking. The biological richness (generally expressed as the number of species, of endemic species or of threatened species) of regions with varying size should be compared by controlling for the effect of area. Since the beginning of the study of the SAR, richness has been recognized to increase with area at a decreasing rate (8). Thus, to control for the effect of area, the use of species-area ratios (e.g., ref. 9) has been found to be problematic (10, 11) as this method implicitly assumes a linear relationship between richness and area and thus produces over-estimated relative diversity for the smallest areas (12). Indeed, accounting for the non-linearity of the SAR through the use of a log-linear power SAR rescaling (a linear relationship is assumed between the logarithm of richness and the logarithm of area) dramatically changes the ranking of regions with respect to their biological richness. Furthermore, this prioritization scales in better agreement with previous studies and knowledge of global biodiversity patterns (10, 11). Although the power model is generally assumed, it may not hold at all scales (13) and its use may not be ubiquitously appropriate (14–16). Moreover, it has been shown that the choice of the SAR model affects the identification of hotspots (17–19). Consequently, one step further in the rescaling of biological richness with respect to area is the incorporation of the uncertainty about the best fitting SAR model (17, 18).

The detection of biodiversity hotspots by SARs is achieved by ranking regions with respect to their displacement above the regression line (17–20). How to quantify the displacement above the SAR is still controversial. The residuals of the SAR have been used repeatedly (17–20) but fail to provide a formal criterion for when to select a region as being a hotspot (21). The use of the 95% confidence limits of the intercept of the log-linear power SAR (21) is also problematic as it relies on the assumption that the dataset could adequately be described by a linearized power SAR.

Devising a ranking methodology robust to the processes that underlie species-area patterns we compared the regions with respect to their position in a confidence interval of the multimodel SAR. The confidence interval was constructed such that it incorporates uncertainty regarding both model selection and parameter estimation.

We used the percentile method following a non parametric bootstrap scheme (22, 23) to generate a high number of resamples in the following manner:

1. One of the SAR models included in the analysis was selected with a probability equal to its weight as calculated from Eq. **6** on the observed dataset.

2. The selected model was fitted to the observed dataset under study.

3. The vectors of inferred species richness (regression line) and residuals were obtained from the regression and the residuals were standardized in the sense of ref. 24.

4. The residuals were sampled completely at random with replacement until sample size reached that of the dataset to form a vector of modified residuals.

5. The vector of modified residuals was added to the vector of inferred species richness to form the resample (bootstrap set of pseudo responses).

For each observed dataset, we obtained a collection of 9,999 multimodel SARs inferred from each of the resamples by applying the whole procedure of model selection and averaging. For each ecoregion in the dataset the 9,999 bootstrap estimates of species richness were sorted in ascending order to provide the percentile confidence intervals (23): the limits of an approximate $(1 - \alpha)100\%$ confidence interval are given by choosing the rth and sth values in the ordered vector of bootstrap estimates such that $r = (b + 1)\alpha$ and $s = (b + 1)(1 - \alpha)$. In the present study, the limits of a 95% confidence interval for a point estimate of species richness are given by the 250th and the 9,750th values.

In so doing, we were able to rank the ecoregions of a dataset with respect to their biological richness by positioning their observed richness in the associated vectors of ordered bootstrap species richness estimates: the higher the position of the observed species richness in the vector of bootstrap estimates the higher the ecoregion in the ranking. Note that when several ecoregions fell in the same position, they were ranked using the vertical distance (expressed as species richness) to the closest inferior bootstrap resample: the higher the distance the higher the ecoregion in the ranking.

1. Stephens PA, Buskirk SW, del Rio CM (2007) Inference in ecology and evolution. *Trends Ecol Evol* 22:192–197.
2. Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N (2006) Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv Biol* 20:1539–1544.
3. Ratkowsky DA (1983) *Non Linear Regression Modelling. A Unified Practical Approach* (Dekker, New York).
4. Kullback S, Leibler RA (1951) On information and sufficiency. *Anal Math Stat* 22:79–86.
5. Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, New York), 2nd Ed.
6. Akaike H (1973) *Proceedings of the Second International Symposium on Information Theory*, eds Petro BN, Csaki F (Akademiai Kiado, Budapest), pp 267–281.
7. Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun Sta Theor Methods A* 7:13–26.
8. Arhennius O (1921) Species and area. *J Ecol* 9:95–99.
9. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature* 403:853–858.
10. Ovadia O (2003) Ranking hotspots of varying sizes : a lesson from the nonlinearity of the species-area relationship. *Conserv Biol* 17:1441.
11. Brummitt N, Lughadha E (2003) Biodiversity: Where's hot and where's not. *Conserv Biol* 17:1448.
12. Connor EF, McCoy ED (1979) The statistics and biology of the species-area relationship. *Am Nat* 113:791–833.
13. He F, Legendre P (1996) On species-area relations. *Am Nat* 148:719–737.
14. Lomolino MV (2000) Ecology's most general, yet protean pattern: the species-area relationship. *J Biogeogr* 27:17–26.
15. Tjørve E (2003) Shapes and functions of species-area curves: A review of possible models. *J Biogeogr* 30:827–835.
16. Stiles A, Scheiner SM (2007) Evaluation of species-area functions using Sonoran Desert plant data: not all species-area curves are power functions. *Oikos* 116:1930–1940.
17. Veech JA (2000) Choice of species-area function affects identification of hotspots. *Conserv Biol* 14:140–147.
18. Fattorini S (2006) Detecting biodiversity hotspots by species-area relationships: a case study of Mediterranean beetles. *Conserv Biol* 20:1169–1180.
19. Fattorini S (2007) To fit or not to fit? A poorly fitting procedure produces inconsistent results when the species-area relationship is used to locate hotspots. *Biodivers Conserv* 16:2531–2538.
20. Hobohm C (2003) Characterization and ranking of biodiversity hotspots: Centres of species richness and endemism. *Biodivers Conserv* 12:279–287.
21. Werner U, Buszko J (2005) Detecting biodiversity hotspots using species-area and endemics-area relationships: the case of butterflies. *Biodivers Conserv* 14:1977–1988.
22. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7:1–26.
23. Buckland ST, Burnham KP, Augustin NH (1997) Model selection: An integral part of inference. *Biometrics* 53:603–618.
24. Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Application* (Cambridge Univ Press, Cambridge, UK).

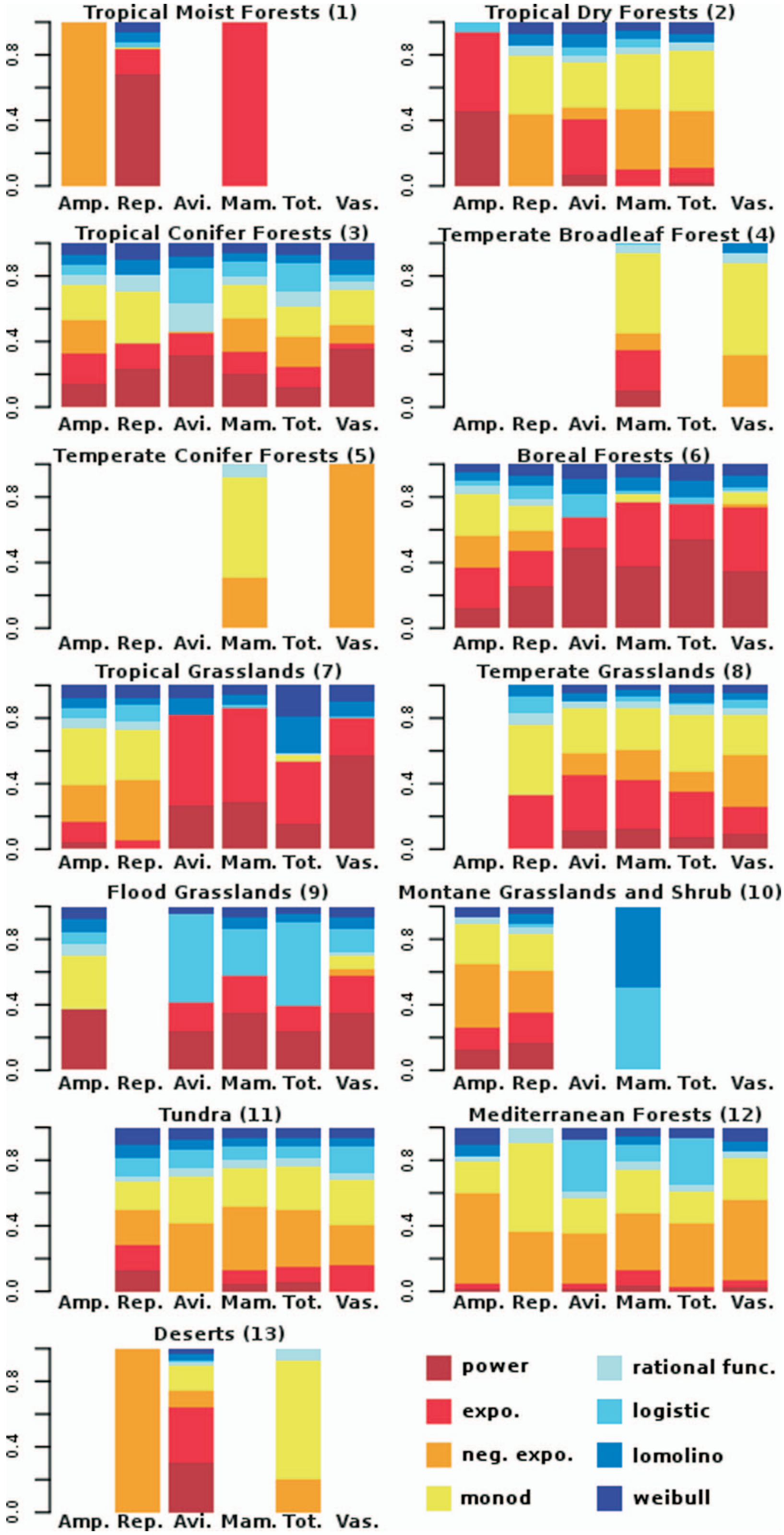

Fig. S1. SAR model selection patterns for the BIC analysis. Patterns of model selection are presented in each biome for amphibians (Amp.), reptiles (Rep.), birds (Avi.), mammals (Mam.), total vertebrates (Tot.), and vascular plants (Vas.). The height of each fraction of the colored band is proportional to the probability (Akaike weight) that each model [see color legend, exponential (expo.), negative exponential (neg. expo.), rational function (rational func.)] is the best in explaining the dataset. A lack of colored band means that none of the eight SAR models was statistically valid for the corresponding dataset.

Table S1. Description of the dataset

Biome	Number of ecoregions	Area, km^2	Amphibians	Reptiles	Mammals	Birds	Total vertebrates	Vascular plants
1	228	14.4–746,652.7	0–227	0–257	0–272	0–785	1–1,413	30–10,000
2	54	101.5–318,937.1	0–68	0–217	0–183	0–491	6–745	110–4,300
3	17	39.4–222,334	0–114	0–214	0–309	0–470	180–918	167–4,900
4	83	145.9–850,317.2	0–67	0–162	0–145	0–585	1–833	225–5,000
5	53	2,901.9–358,833.4	0–70	0–84	0–169	0–673	80–947	459–5,000
6	28	2,035.4–3,922,554.7	0–11	0–7	4–79	70–270	75–366	258–1,600
7	47	14.7–3,042,451.4	0–84	0–191	0–241	4–712	4–1225	30–6,500
8	42	167.4–997,072.7	0–42	0–131	0–134	6–458	6–585	90–3,500
9	25	628.4–178,951.6	0–36	0–90	14–143	139–508	230–767	80–1,700
10	49	1,238.7–629,190.4	0–93	0–118	0–203	0–538	50–786	400–3,800
11	34	877–71,040,235.6	0–5	0–3	0–50	0–208	2–261	24–1,000
12	39	2,867.9–358,243.2	0–31	0–136	10–121	103–337	125–517	400–6,300
13	93	7.7–4,629,416.3	0–41	0–191	0–162	0–405	4–717	80–4,850

Number of ecoregions, area range, and richness ranges for each taxa are presented for all biomes studied. See Fig. S1 for biome names.

Table S2. Forms of SAR used in the study

Name	Formula	Number of parameters	Shape	Asymptotic nature
Power	$S = cA^{z}$	2	Convex	No
Exponential	$S = c + z\log(A)$	2	Convex	No
Negative exponential	$S = c(1 - \exp(-zA))$	2	Convex	Yes
Monod	$S = (cA) / (z + A)$	2	Convex	Yes
Rational function	$S = (c + zA) / (1 + fA)$	3	Sigmoid	Yes
Logistic	$S = c / (1 + \exp(-zA + f)$	3	Sigmoid	Yes
Lomolino	$S = c / 1 + (z^{\log(f/A)})$	3	Sigmoid	Yes
Cumulative Weibull	$S = c(1 - \exp(-zA^{f}))$	3	Sigmoid	Yes

Table S3. R^2 values for multimodel inferences

	Biomes													
	1	2	3	4	5	6	7	8	9	10	11	12	13	Means
Amp	0.17	0.02	0.36	-	-	0.34	0.3	-	0.11	0.06	-	0.21	-	0.2
Rep	0.3	0.06	0.56	-	-	0.15	0.43	0.1	-	0.04	0.25	0.21	0.06	0.22
Avi	-	0.51	0.47	-	-	0.45	0.58	0.44	0.33	-	0.29	0.39	0.1	0.4
Mam	0.46	0.33	0.4	0.45	0.22	0.52	0.38	0.46	0.24	0.02	0.21	0.38	-	0.34
Tot	-	0.45	0.37	-	-	0.49	0.69	0.56	0.42	-	0.23	0.47	0.14	0.42
Vas	-	-	0.69	0.25	0.14	0.47	0.41	0.27	0.23	-	0.26	0.16	-	0.32
Means	0.31	0.27	0.48	0.35	0.18	0.4	0.47	0.37	0.27	0.04	0.25	0.3	0.1	0.32

See Fig. S1 for biomes names and taxa description. Dash cells correspond to biome-taxon datasets that could not be fit by any of the models.

Table S4. Model selection procedure results

Biome	Higher taxa	power	expo	neg. expo.	Monod	rational func.	logistic	Lomolino	Weibull
1	Amp	-	-	1	-	-	-	-	-
1	Rep	0.454	0.106	wi < 10e-4	0.006	0.002	0.078	0.177	0.176
1	Avi	-	-	-	-	-	-	-	-
1	Mam	-	1	-	-	-	-	-	-
1	Tot	-	-	-	-	-	-	-	-
1	Vas	-	-	-	-	-	-	-	-
2	Amp	0.437	0.467	-	-	-	0.096	-	-
2	Rep	-	-	0.377	0.313	0.09	-	0.111	0.109
2	Avi	0.05	0.26	0.054	0.212	0.066	0.102	0.135	0.122
2	Mam	-	0.079	0.298	0.266	0.082	0.093	0.085	0.097
2	Tot	0.014	0.079	0.282	0.303	0.094	-	0.096	0.133
2	Vas	-	-	-	-	-	-	-	-
3	Amp	0.181	0.237	0.261	0.267	0.013	0.012	0.014	0.014
3	Rep	0.302	0.192	-	0.414	0.03	-	0.03	0.031
3	Avi	0.387	0.171	0.007	-	0.141	0.168	0.063	0.063
3	Mam	0.223	0.146	0.231	0.215	0.04	0.066	0.039	0.04
3	Tot	0.137	0.135	0.212	0.206	0.072	0.142	0.039	0.057
3	Vas	0.384	0.039	0.118	0.232	0.041	0.031	0.077	0.078
4	Amp	-	-	-	-	-	-	-	-
4	Rep	-	-	-	-	-	-	-	-
4	Avi	-	-	-	-	-	-	-	-
4	Mam	0.089	0.22	0.095	0.437	0.145	0.014	-	-
4	Tot	-	-	-	-	-	-	-	-
4	Vas	-	-	0.253	0.445	0.148	-	0.154	-
5	Amp	-	-	-	-	-	-	-	-
5	Rep	-	-	-	-	-	-	-	-
5	Avi	-	-	-	-	-	-	-	-
5	Mam	-	-	0.281	0.549	0.17	-	-	-
5	Tot	-	-	-	-	-	-	-	-
5	Vas	-	-	1	-	-	-	-	-
6	Amp	0.122	0.237	0.183	0.249	0.059	0.036	0.06	0.054
6	Rep	0.285	0.249	0.137	0.167	0.025	0.053	0.042	0.042
6	Avi	0.442	0.165	-	-	-	0.169	0.112	0.112
6	Mam	0.357	0.361	0.007	0.041	0.01	0.027	0.1	0.098
6	Tot	0.505	0.193	-	-	-	0.046	0.128	0.128
6	Vas	0.332	0.362	0.018	0.067	0.017	0.024	0.09	0.089
7	Amp	0.037	0.1	0.183	0.288	0.084	0.103	0.086	0.119
7	Rep	0.007	0.034	0.304	0.245	0.072	0.152	0.074	0.112
7	Avi	0.229	0.456	-	-	-	-	0.171	0.144
7	Mam	0.253	0.51	-	-	-	0.024	0.108	0.105
7	Tot	0.111	0.257	-	0.031	0.009	wi < 10e-4	0.314	0.277
7	Vas	0.47	0.19	-	-	-	0.015	0.162	0.163
8	Amp	-	-	-	-	-	-	-	-
8	Rep	-	0.278	-	0.362	0.103	0.152	0.104	-
8	Avi	0.102	0.302	0.12	0.241	0.071	-	0.09	0.073
8	Mam	0.111	0.263	0.167	0.231	0.068	0.046	0.07	0.043
8	Tot	0.068	0.23	0.111	0.303	0.089	0.013	0.103	0.083
8	Vas	0.08	0.147	0.271	0.213	0.063	0.076	0.071	0.08
9	Amp	0.372	-	-	0.329	0.07	0.07	0.079	0.079
9	Rep	-	-	-	-	-	-	-	-
9	Avi	0.212	0.157	-	-	-	0.581	-	0.051
9	Mam	0.326	0.209	-	-	-	0.308	0.078	0.078
9	Tot	0.218	0.13	-	-	-	0.547	0.052	0.052
9	Vas	0.331	0.211	0.041	0.075	0.018	0.164	0.079	0.079
10	Amp	0.121	0.125	0.361	0.231	0.063	-	-	0.099
10	Rep	0.156	0.159	0.238	0.198	0.055	0.037	0.088	0.07
10	Avi	-	-	-	-	-	-	-	-
10	Mam	-	-	-	-	-	0.501	0.499	-
10	Tot	-	-	-	-	-	-	-	-
10	Vas	-	-	-	-	-	-	-	-
11	Amp	-	-	-	-	-	-	-	-
11	Rep	0.196	0.221	0.324	0.259	wi < 10e-4	wi < 10e-4	wi < 10e-4	wi < 10e-4
11	Avi	-	-	0.362	0.243	0.065	0.152	0.072	0.105
11	Mam	0.046	0.072	0.337	0.216	0.058	0.113	0.067	0.091

Biome	Higher taxa	power	expo	neg. expo.	Monod	rational func.	logistic	Lomolino	Weibull
11	Tot	0.05	0.084	0.305	0.233	0.064	0.107	0.07	0.086
11	Vas	-	0.133	0.205	0.227	0.062	0.214	0.066	0.093
12	Amp	0.017	0.024	0.486	0.169	0.047	0.001	0.108	0.148
12	Rep	-	-	0.343	0.513	0.144	-	-	-
12	Avi	0.012	0.021	0.228	0.164	0.047	0.427	-	0.101
12	Mam	0.034	0.072	0.285	0.227	0.065	0.162	0.067	0.088
12	Tot	0.007	0.015	0.297	0.15	0.043	0.394	-	0.094
12	Vas	0.027	0.033	0.428	0.22	0.063	0.003	0.097	0.128
13	Amp	-	-	-	-	-	-	-	-
13	Rep	-	-	1	-	-	-	-	-
13	Avi	0.247	0.281	0.081	0.131	0.044	0.038	0.091	0.087
13	Mam	-	-	-	-	-	-	-	-
13	Tot		–	0.177	0.616	0.207	-	-	-
13	Vas	-	-	-	-	-	-	-	-

For each biome, values correspond to model weights (wi) that are equivalent to model probabilities in being the best to fit the dataset (see Fig. S1 for biome names, taxa, and model description). Dash cells correspond to biome-taxon datasets that could not be fit by any of the models.