



Modes of speciation and the neutral theory of biodiversity

Rampal S. Etienne, M. Emile F. Apol, Han Olff and Franz J. Weissing

R. S. Etienne (r.s.etienne@rug.nl), M. E. F. Apol, H. Olff and F. J. Weissing, Centre for Ecological and Evolutionary Studies, Univ. of Groningen, PO Box 14, NL-9750 AA Haren, the Netherlands.

Hubbell's neutral theory of biodiversity has generated much debate over the need for niches to explain biodiversity patterns. Discussion of the theory has focused on its neutrality assumption, i.e. the functional equivalence of species in competition and dispersal. Almost no attention has been paid to another critical aspect of the theory, the assumptions on the nature of the speciation process. In the standard version of the neutral theory each individual has a fixed probability to speciate. Hence, the speciation rate of a species is directly proportional to its abundance in the metacommunity. We argue that this assumption is not realistic for most speciation modes because speciation is an emergent property of complex processes at larger spatial and temporal scales and, consequently, speciation rate can either increase or decrease with abundance. Accordingly, the assumption that speciation rate is independent of abundance (each species has a fixed probability to speciate) is a more natural starting point in a neutral theory of biodiversity. Here we present a neutral model based on this assumption and we confront this new model to 20 large data sets of tree communities, expecting the new model to fit the data better than Hubbell's original model. We find, however, that the data sets are much better fitted by Hubbell's original model. This implies that species abundance data can discriminate between different modes of speciation, or, stated otherwise, that the mode of speciation has a large impact on the species abundance distribution. Our model analysis points out new ways to study how biodiversity patterns are shaped by the interplay between evolutionary processes (speciation, extinction) and ecological processes (competition, dispersal).

A few years ago Hubbell (2001) presented a theory of community ecology that stirred the scientific community. He argued that the interplay between a few basic processes (speciation, birth and death, and – on a local scale – dispersal), could explain general large-scale diversity patterns, such as species-abundance distributions and species-area curves. He made three basic assumptions: 1) individuals of different species are functionally equivalent (neutrality assumption); 2) the community size is constant (zero-sum assumption); 3) speciation is comparable to mutation where each individual has an equal probability of producing mutated, i.e. speciated, offspring (point mutation assumption). Hubbell developed his theory analogous to the neutral theory of molecular evolution (Kimura 1983) with individuals, species, speciation and ecological drift replacing genes, alleles, mutation and genetic drift, respectively (Chave 2004, Hu et al. 2006). Although his theory was praised as a fresh, falsifiable

contribution to community ecology (Brown 2001, De Mazancourt 2001, Jetschke 2002) and has stimulated the development of more advanced neutral community theory (Chave and Leigh 2002, Vallade and Houchmandzadeh 2003, Etienne 2005), it has also been heavily criticized (Abrams 2001, Clark and McLachlan 2003, Fargione et al. 2003, McGill 2003a, 2003b, Ricklefs 2003, Fargione and Tilman 2004, Wootton 2005, Dornelas et al. 2006). The criticism was predominantly directed at the neutrality assumption, basically because it contradicts empirical evidence on variation among species in life history traits which is considered to cause functional differences rather than functional equivalence. However, the zero-sum assumption and the point mutation assumption have so far mainly escaped criticism. There is good reason that the zero-sum assumption has escaped criticism. It is supported by empirical evidence on community saturation, i.e. a fairly constant density of individuals in

ecological communities across sufficiently large spatial and temporal scales (Hubbell 1979, 2001). Moreover, it is a non-crucial mathematical simplification, as shown in population genetics (Watterson 1974, Etienne et al. 2006b). The point mutation assumption lacks strong arguments in its favor. In fact, Hubbell already pointed out that, in line with scientific consensus, real speciation is essentially different from point mutation and particularly should take into account the abundances of the species in the community. He suggested two other modes of speciation (Hubbell 2001, 2003, Hubbell and Lake 2003): “random fission speciation” in which populations are assumed to split randomly into two species, and “peripheral isolate speciation” in which populations are temporarily isolated to allow for more rapid speciation. Still, while speciation has received a tremendous amount of attention in evolutionary biology (Howard and Berlocher 1998, Schluter 2000, Coyne and Orr 2004, Van Doorn et al. 2004, Dieckmann et al. 2004, Gavrillets 2004), the consequences of different speciation modes for community structure have hardly been studied.

The speciation process in Hubbell’s model, as it has been mainly used so far in the literature, is an individual-level process that leads to a proportional relationship between the speciation rate of a species in a community and its abundance. Speciation by polyploidy is the only speciation mode that roughly corresponds to this pattern (Coyne and Orr 2004). In virtually all other speciation scenarios, speciation requires the evolution of reproductive isolation, which implies that speciation is not an individual-level process, but a population-level (or species-level) process. Therefore it is not self-evident at all that speciation rates should in general be proportional to species abundance. In fact, it has long been argued (Wright 1931, Mayr 1970, Bush et al. 1977) that just low density and isolation by distance (allopatric speciation) favor the emergence of new species, suggesting even an inverse relationship between speciation rate and abundance. It is far from clear what the relationship with abundance would be in the still controversial sympatric mode of speciation (Coyne and Orr 2004), and in cases where it is sexual selection or sexual conflict rather than natural selection that drives speciation. Unfortunately, data on the relationship between speciation rate and abundance are scarce. We found only one systematic review of this relationship (Makarieva and Gorshkov 2004) and although it suggests that speciation is independent of abundance, it seems irrelevant to our current analysis. Makarieva and Gorshkov (2004) relate speciation rates of 12 major taxa to their abundances (which span 20 orders of magnitude), either directly or indirectly via the rates of appearance of new genotypes per species, whereas we are interested in more restricted taxa such as tropical forests. Using indirect evidence on this relationship – for example by using range size, age

or body size of a species as a proxy for its abundance – is tricky, because speciation may be driven by these proxies themselves rather than by abundance or by other processes that determine both (Jablonski and Roy 2003). We conclude that, data being scarce, there seems no a priori reason for a proportional relationship between speciation rate and abundance, and hence we argue that for a “null model” of biodiversity, it is more natural to assume that speciation is independent of abundance than that speciation is proportional to abundance.

We therefore modify Hubbell’s metacommunity model in such a way that the overall speciation rate per species is indeed independent of abundance. We derive an analytical expression for the full likelihood of the corresponding model parameter given a data set of species abundance distributions which can be used to fit the model to such data. Evidently, a good fit will not automatically imply that the model should be accepted (because other predictions of the model may be false), but a bad fit will certainly suggest that some ingredients of the model are unrealistic. Furthermore, we construct a model where both forms of speciation (proportional vs no dependence on abundance) are possible. By confronting this model to the data, we can let the data decide to what degree speciation depends on abundance in a neutral model. Finally, to test the robustness of our results, we also construct a model with dispersal limitation, i.e. a model for a local community subject to limited migration from a metacommunity that is governed by abundance-independent speciation, and we compare it to the dispersal-limited model of Hubbell (2001).

Models

We start with Hubbell’s original discrete-time model for the metacommunity (i.e. ignoring dispersal which only appears in his local community model), with one slight difference that simplifies some expressions in the modified model and allows better comparison with population genetic models (see Supplementary material, SM). Where in Hubbell’s model individuals that die leave no offspring, we assume that dying individuals can still contribute offspring to the next generation, for example seeds. This slightly different model is known as the Moran (1958) model in population genetics, and has basically the same properties as Hubbell’s model. In the Hubbell-Moran model, there are J_M individuals in the metacommunity (resp. population), each having a probability of speciation (resp. mutation), v_1 . For a species with abundance n the total probability of speciation is $v(n) \approx v_1 n$ which is a first order dependence on abundance; hence the subscript 1. We will refer to this model as M_1 , the neutral model with a 1st

order (proportional) speciation-abundance relationship. Because $v(n)$ is a probability in a discrete-time model, it cannot exceed unity and hence the relationship is only proportional to a very good approximation when $v_1 n$ is small; however, the corresponding speciation rate in a continuous-time model is exactly proportional to abundance (see SM). Both J_M and v_1 are contained in the compound parameter θ which is, in the Moran model, defined as

$$\theta = \frac{v_1}{1 - v_1} J_M \quad (1)$$

Assume now that we have a species abundance data set \vec{D} with abundances of S species n_1, n_2, \dots, n_s summing to the sample size, J . We then have the following likelihood for the model parameter θ , which is the well-known Ewens sampling formula (ESF) in population genetics (Ewens 1972, Karlin and McGregor 1972):

$$P_{M_1}[\vec{D}|\theta, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(\theta)_J} \quad (2)$$

Here Φ_j is the observed number of species with abundance j and $(\theta)_J$ is the Pochhammer symbol defined as

$$(\theta)_J := \prod_{i=1}^J (\theta + i - 1) = \frac{\Gamma(\theta + J)}{\Gamma(\theta)} \quad (3)$$

where $\Gamma(x)$ is the Gamma function.

Now we turn to the modified model where the overall speciation probability of a species is independent of abundance. We denote this probability of speciation by $v(n) = v_0$, the subscript 0 indicating the zeroth order dependence on abundance n . We will refer to this model as M_0 , the neutral model with a 0th order speciation-abundance relationship. The likelihood of v_0 is a new sampling formula, which we derive in the Supplementary Material:

$$P_{M_0}[\vec{D}|v_0, J] = \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} v_0^{S-1} \frac{(S-1)!}{(J-1)!} \times \prod_{i=1}^S \frac{(1-v_0)^{n_i-1}}{(n_i-1)!} \quad (4)$$

Note that the metacommunity size J_M does not appear in this formula, whereas it does appear in Eq. 2, as it enters θ (Eq. 1). This reflects the assumption that speciation only depends on the number of species, not on the number of individuals. Indeed, in model M_0 the total probability of speciation in the metacommunity is proportional to species richness S , whereas in model M_1 it is proportional to metacommunity size J_M (see SM). These models thus connect to the important debate whether diversity begets diversity (Palmer and Maurer

1997, Emerson and Kolm 2005, see also Losos and Schluter 2000).

We now consider the generalized model where the speciation probability is a linear combination of the speciation rate of the two previous models $v(n) \approx v_0 + v_1 n$. We refer to this model as M_c , the neutral model with combined speciation-abundance relationship. For this case the likelihood is another new sampling formula (see SM):

$$P_{M_c}[\vec{D}|v_0, v_1, J_M, J] = P_{M_1}[\vec{D}|\theta, J] \prod_{i=1}^{S-1} \left(1 + \frac{i v_0}{J_M v_1}\right) \times \prod_{i=1}^S \prod_{j=1}^{n_i-1} \frac{1 - v_1(1 + \frac{i v_0}{J_M v_1})}{1 - v_1} \quad (5)$$

which reduces to Eq. 2 and Eq. 4 for $v_0 = 0$ and $v_1 = 0$ respectively. Note that v_1 and J_M enter this formula not only through θ , but also separately.

Finally, we build two models for a dispersal-limited local community which has no local speciation, but which receives immigrants from a metacommunity that is in speciation-extinction balance governed by M_0 or M_1 . We will refer to these dispersal-limited models as DLM_0 and DLM_1 respectively. We denote by m the probability of immigration (Hubbell 2001). In the Moran version of Hubbell's dispersal-limited model, we have a fundamental dispersal number I that is related to m by

$$I = \frac{m}{1 - m} J_L \quad (6)$$

where J_L is the local community size (which is usually set equal to the sample size J). Just as for θ , Hubbell's model has a slightly different I , i.e. $I = \frac{m}{1-m}(J_L - 1)$, which is a negligible difference. If the metacommunity has speciation proportional to abundance (M_1), we have the dispersal-limited version of Eq. 2, which is given by the sampling formula (Etienne 2005),

$$P_{DLM_1}[\vec{D}|\theta, I, J] = P_{M_1}[\vec{D}|\theta, J] \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(I)_J} \frac{(\theta)_J}{(\theta)_A} \quad (7)$$

where $K(\vec{D}, A)$ is a coefficient determined by the data (see Etienne 2005 and SM). If, instead, the metacommunity has speciation independent of abundance (M_0), we have the dispersal-limited version of Eq. 4, which is given by (see SM):

$$P_{DLM_0}[\vec{D}|v_0, I, J] = P_{M_0}[\vec{D}|v_0, J] \sum_{A=S}^J L(\vec{D}, v_0, A) \frac{I^A}{(I)_J} \frac{(J-1)!}{(A-1)!} \quad (8)$$

where the coefficients $L(\vec{D}, v_0, A)$ are determined by the data and the parameter v_0 (see SM).

Results

We confronted the three metacommunity models (M_1 , M_0 and M_c) with 20 large (i.e. sample size larger than 5 000) species abundance data sets of tree communities taken from the literature (Fig. 1; see SM for data set selection and references) and obtained maximum likelihood estimates for the model parameters. The results are listed in Table 1. In all cases the model with first-order dependence of the overall speciation rate on abundance (M_1) gives a much better fit than the model with independence of abundance (M_0). Moreover, the model combining both speciation mechanisms (M_c) does not perform better than the model with proportional dependence of overall speciation rate on abundance (M_1), indicated by the Akaike weights (see SM) for this model in Table 1. Most of these weights are smaller than those for M_1 , as the maximum likelihood for M_c is equal to the maximum likelihood for M_1 . This may seem odd, as a model with more parameters should always have a higher maximum likelihood. However, this is only true for unconstrained likelihood maximization. The parameter values are here constrained by $0 < \frac{v_0}{n} + v_1 < 1$ for all n , whereas the higher maximum likelihood values are found at biologically impossible parameter values ($v_0 < 0$).

We also confronted the dispersal-limited versions of DLM_0 and DLM_1 with the same data. The parameter estimates, loglikelihoods and Akaike weights are given in Table 2. The expected and observed number of species in each abundance class for six data sets studied previously in a neutral context (Volkov et al. 2005, Chave et al. 2006) are shown in Fig. 2. In most cases DLM_1 again outperforms DLM_0 , but there are several

data sets that favor DLM_0 as indicated by their higher Akaike weights. However, the estimates of the speciation probability v_0 are unrealistically high for all data sets, and the corresponding immigration probabilities are unreasonably low. This forces us to reject DLM_0 . Put in Bayesian terms (Etienne and Olff 2005), the prior probability of these estimated values is very small which reduces the posterior probability enormously and makes DLM_0 the inferior model. Furthermore, the poorer fit of DLM_1 for some data sets is caused by the huge underestimation of the number of very abundant species. This becomes clear when we plot the number of species logarithmically (Fig. 3). According to M_1/DLM_1 , such species should not exist (because they speciate directly due to the proportional speciation-abundance relationship in this model). To check this, we removed the two most abundant species from the data sets and we found that indeed DLM_1 fitted the data much better than DLM_0 (results not shown).

For M_1 and DLM_1 it can be shown that the parameter estimates that maximize the likelihood also make the expected number of species equal to the observed number of species. For M_1 , the observed number of species is even a sufficient statistic for θ , that is, information on species abundances is not needed to estimate θ ; species richness suffices (Tavaré and Ewens 1997). In contrast, for M_0 and DLM_0 the expected number of species is substantially smaller than the observed number of species, particularly in the non-dispersal-limited model M_0 . This is a clear indicator of the lack of fit of $(DL)M_0$ to the data.

The general conclusion is therefore that the data do not favor independence between speciation rate and abundance ($(DL)M_0$), but a proportional relationship ($(DL)M_1$).

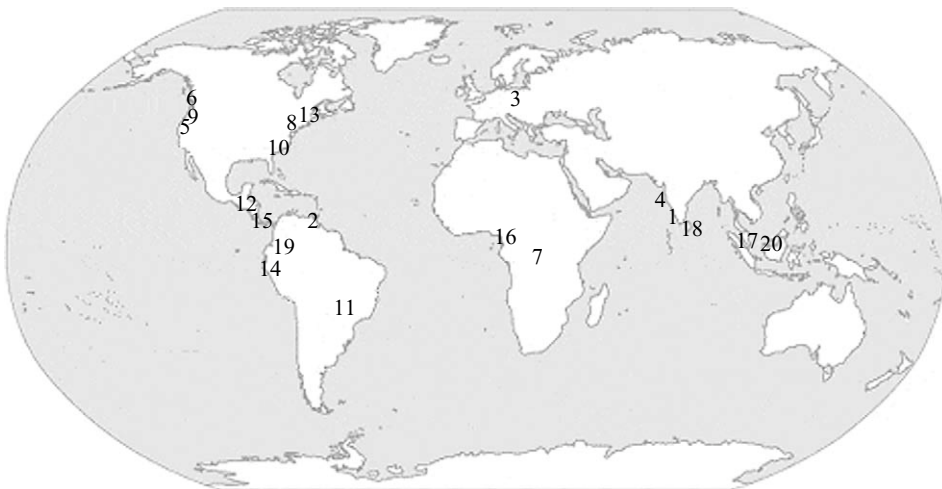


Fig. 1. Location of the 20 tree communities of Table 1 and 2.

Table 1. Parameter estimates obtained by likelihood maximization of the three neutral metacommunity models with different speciation mechanisms (M_0 : speciation rate is independent of abundance; M_1 : speciation rate is proportional to abundance, M_c : the combination of M_0 and M_1) for 20 species abundance data sets of tree communities (Fig. 1).

Site ^b	J^c	S^d	M_1		M_0		M_c				Comparison ^a			
			θ	ML ^e	v_0	ML ^e	J_M	v_1	v_0	θ^f	ML ^e	w_{M_1}	w_{M_0}	w_{M_c}
1	13383	148	23.19	-241.85	0.246	-290.27	$3.08 \cdot 10^7$	$7.52 \cdot 10^7$	$2.22 \cdot 10^{-13}$	23.19	-241.85	0.731	$6.82 \cdot 10^{-22}$	0.269
2	13045	103	15.17	-216.70	0.224	-255.30	$3.09 \cdot 10^7$	$4.90 \cdot 10^7$	$1.29 \cdot 10^{-14}$	15.17	-216.70	0.731	$1.26 \cdot 10^{-17}$	0.269
3	9897	12	1.28	-65.86	0.145	-71.47	$1.34 \cdot 10^6$	$9.54 \cdot 10^7$	$2.48 \cdot 10^{-14}$	1.28	-65.86	0.729	$2.66 \cdot 10^{-3}$	0.268
4	5298	158	30.51	-161.72	0.309	-200.25	$6.22 \cdot 10^7$	$4.90 \cdot 10^7$	$1.76 \cdot 10^{-13}$	30.51	-161.72	0.731	$6.75 \cdot 10^{-33}$	0.269
5	7536	20	2.42	-72.67	0.200	-78.42	$3.27 \cdot 10^6$	$7.39 \cdot 10^7$	$1.29 \cdot 10^{-12}$	2.42	-72.67	0.729	$2.31 \cdot 10^{-3}$	0.268
6	6799	16	1.89	-68.30	0.179	-74.18	$2.66 \cdot 10^6$	$7.09 \cdot 10^7$	$5.89 \cdot 10^{-13}$	1.89	-68.30	0.730	$2.04 \cdot 10^{-3}$	0.268
7	6687	149	26.90	-163.77	0.326	-183.11	$2.65 \cdot 10^7$	$7.02 \cdot 10^7$	$9.88 \cdot 10^{-2}$	18.59	-162.07	0.330	$1.32 \cdot 10^{-9}$	0.670
8	10358	26	3.14	-104.45	0.212	-110.64	$3.18 \cdot 10^6$	$9.88 \cdot 10^7$	$7.87 \cdot 10^{-13}$	3.14	-104.45	0.730	$1.49 \cdot 10^{-3}$	0.269
9	5313	11	1.25	-45.83	0.185	-48.93	$1.67 \cdot 10^6$	$7.50 \cdot 10^7$	$3.27 \cdot 10^{-15}$	1.25	-45.83	0.708	$3.20 \cdot 10^{-2}$	0.260
10	18744	34	3.96	-154.84	0.166	-170.66	$5.38 \cdot 10^6$	$7.35 \cdot 10^7$	$4.85 \cdot 10^{-13}$	3.96	-154.84	0.731	$9.84 \cdot 10^{-8}$	0.269
11	6507	153	27.96	-161.67	0.326	-182.83	$2.78 \cdot 10^7$	$7.23 \cdot 10^7$	$9.04 \cdot 10^{-2}$	20.07	-160.23	0.392	$2.52 \cdot 10^{-1}$	0.608
12	12804	140	21.89	-214.86	0.255	-255.11	$2.96 \cdot 10^7$	$7.39 \cdot 10^7$	$5.30 \cdot 10^{-13}$	21.89	-214.86	0.731	$2.41 \cdot 10^{-18}$	0.269
13	6765	12	1.34	-53.25	0.174	-57.25	$1.83 \cdot 10^6$	$7.35 \cdot 10^7$	$2.27 \cdot 10^{-13}$	1.34	-53.25	0.721	$1.31 \cdot 10^{-2}$	0.265
14	9088	701	176.99	-202.07	0.383	-337.56	$2.27 \cdot 10^7$	$7.72 \cdot 10^6$	$4.59 \cdot 10^{-3}$	174.86	-202.06	0.729	$1.05 \cdot 10^{-59}$	0.271
15	21457	225	34.96	-318.85	0.245	-392.41	$4.65 \cdot 10^6$	$7.52 \cdot 10^6$	$3.42 \cdot 10^{-13}$	34.96	-318.85	0.731	$8.25 \cdot 10^{-33}$	0.269
16	24591	308	49.52	-318.67	0.284	-377.58	$4.01 \cdot 10^6$	$1.11 \cdot 10^5$	$2.89 \cdot 10^{-2}$	44.35	-318.31	0.655	$1.71 \cdot 10^{-26}$	0.345
17	26554	678	126.62	-392.51	0.280	-617.66	$1.72 \cdot 10^7$	$7.35 \cdot 10^6$	$1.32 \cdot 10^{-14}$	126.62	-392.51	0.731	$1.21 \cdot 10^{-98}$	0.269
18	16936	167	25.64	-253.78	0.258	-297.31	$5.23 \cdot 10^6$	$4.90 \cdot 10^6$	$3.02 \cdot 10^{-12}$	25.64	-253.78	0.731	$9.09 \cdot 10^{-20}$	0.269
19	17546	821	178.43	-307.58	0.340	-492.93	$2.30 \cdot 10^7$	$7.75 \cdot 10^6$	$3.21 \cdot 10^{-13}$	178.43	-307.58	0.731	$2.33 \cdot 10^{-81}$	0.269
20	33175	1004	195.18	-437.89	0.290	-763.80	$3.98 \cdot 10^7$	$4.90 \cdot 10^6$	$1.01 \cdot 10^{-13}$	195.18	-437.89	0.731	$2.24 \cdot 10^{-142}$	0.269

Notes

a. Comparison between M_0 , M_1 and M_c expressed in terms of Akaike weights w_i (see also Supplementary material (SM))

b. For references see SM

c. Sample size

d. Number of species observed

e. The logarithm of the maximum likelihood

f. The value of θ is fully determined by J_M and v_1 , only given here for better comparison with θ in M_1

Table 2. Parameter estimates obtained by likelihood maximization of the two neutral local community models (DLM₀: the dispersal-limited version of M₀, the model with speciation rate independent of abundance; DLM₁: the dispersal-limited version of M₁, the model with speciation rate proportional to abundance) for 20 species abundance datasets of tree communities (Fig. 1).

Site ^b	J ^c	S ^d	DLM ₁			DLM ₀			Comparison ^a	
			θ	m	ML ^e	v ₀	m	ML ^e	w _{DLM₁}	w _{DLM₀}
1	13383	148	32.58	0.082	-235.98	0.84	0.0030	-241.75	0.9969	0.0031
2	13045	103	20.38	0.090	-206.38	0.64	0.0056	-214.60	0.9997	0.0003
3	9897	12	2.37	0.0058	-63.02	0.55	0.00052	-65.44	0.9179	0.0820
4	5298	158	36.34	0.300	-160.03	0.86	0.0109	-160.48	0.6116	0.3884
5	7536	20	2.76	0.215	-71.77	1.00	0.00032	-72.67	0.7103	0.2897
6	6799	16	9.23	0.00084	-67.02	1.00	0.00028	-68.30	0.7832	0.2168
7	6687	149	26.90	1.000	-163.77	1.00	0.00401	-163.77	0.5000	0.5000
8	10358	26	3.37	0.408	-103.52	0.28	0.155	-106.90	0.9671	0.0329
9	5313	11	13.31	0.00038	-45.55	1.00	0.00024	-45.83	0.5713	0.4287
10	18744	34	6.70	0.012	-148.61	0.62	0.00075	-153.78	0.9944	0.0056
11	6507	153	28.19	0.929	-161.63	0.67	0.0283	-163.85	0.9023	0.0977
12	12804	140	28.28	0.130	-208.47	0.70	0.00824	-208.70	0.5589	0.4411
13	6765	12	8.89	0.00046	-52.57	1.00	0.00020	-53.25	0.6638	0.3362
14	9088	701	188.43	0.701	-199.96	0.90	0.0320	-193.74	0.0020	0.9980
15	21457	225	47.67	0.093	-308.73	0.773	0.0040	-317.72	0.9999	0.0001
16	24591	308	52.73	0.547	-317.04	0.682	0.0139	-310.56	0.0015	0.9985
17	26554	678	190.93	0.093	-359.38	0.894	0.0066	-391.59	1.0000	0.0000
18	16936	167	436.76	0.0019	-252.93	0.827	0.0032	-253.93	0.7311	0.2689
19	17546	821	204.17	0.429	-297.15	0.859	0.0219	-284.09	0.0000	1.0000
20	33175	1004	285.58	0.115	-386.38	0.802	0.0129	-427.22	1.0000	0.0000

Notes

- a. Comparison between DLM₁ and DLM₀ expressed in terms of Akaike weights w_i (see also SM)
- b. For references see SM
- c. Sample size
- d. Number of species observed
- e. The logarithm of the maximum likelihood

Discussion

We have shown that the original neutral model with a proportional relationship of speciation with abundance provides the statistically best fit to species abundance data of 20 large tree communities. At least three different conclusions are possible. First, speciation is really proportional to abundance in natural ecological communities. Rather than metacommunity diversity, metacommunity size promotes speciation, so diversity does not seem to beget diversity. However, metacommunity size and diversity are difficult to separate, as they are obviously correlated. Second, the sampling formula corresponding to (DL)M₁ indeed describes the data best, but this does not mean that the model used here to generate it is the most realistic one. Several different mechanisms could (approximately) lead to this formula, because pattern does not equal process (Cohen 1968, Purves and Pacala 2005). For example, in population genetics, overdominant selection as well as neutral mutations with genetic drift have been shown to lead to the Ewens sampling formula (ESF) (Joyce et al. 2003), and the maximum likelihood estimate of θ provided by the ESF also follows from non-neutral models (Joyce 1995). Hence, the better fit of the sampling formula corresponding to (DL)M₁ does not

imply that the “true” model must involve neutrality with a speciation rate that is proportional to abundance. If we want to maintain abundance-independent speciation because we believe it to be more realistic, our results force us to reject neutrality. This is further supported by increasing evidence on adaptive inter-specific differences leading to strong niche differentiation in tree communities (Wright et al. 2004, Poorter et al. 2005). However, a third conclusion is that the models are still too simple: a more sophisticated speciation model with a more sophisticated implementation of neutrality (e.g. the saturating speciation-abundance relationship mentioned below, a speciation model where new species do not start with a single individual but with many more, a spatial model, or symmetric species subject to density-dependence) could lead to a sampling formula similar to that of (DL)M₁. Hubbell (2001, 2003) suggests that this is indeed the case for the random fission and peripheral isolate modes of speciation, but this has not been fully explored, and no analytical treatment exists. Our paper represents a first strike at a more realistic and analytically tractable alternative neutral speciation model, but so far fails to provide a better description of community structure.

The model with speciation independent of abundance, (DL)M₀, produces fewer species than observed

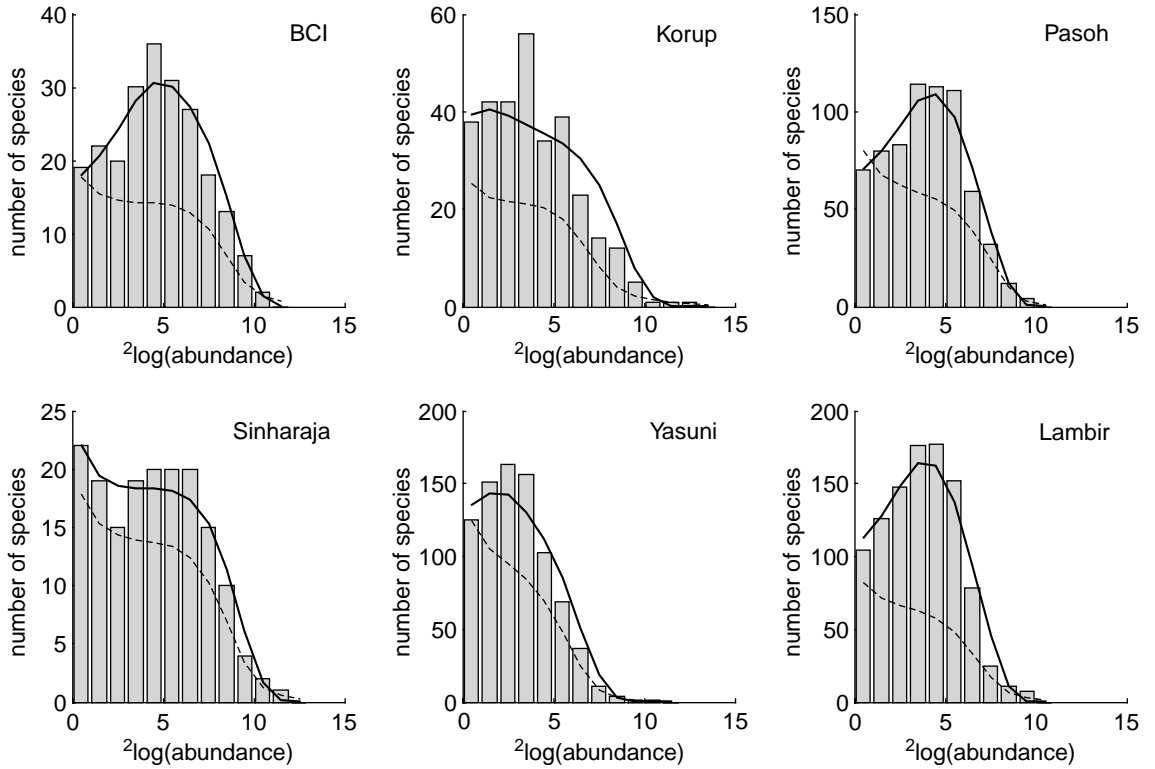


Fig. 2. Observed (bars) and expected (curves) number of species, according to M_0 , the neutral model with speciation rate independent of abundance (dotted curve) and M_1 , the neutral model with speciation rate proportional to abundance (solid curve), for six data sets (15–20 in Tables 1 and 2). The binning method is similar to that proposed by Pueyo (2006) but without converting number of species into densities. This method is essentially different from the Preston binning method as used by, for example, Volkov et al. (2005). The Preston binning method was designed to obtain a lognormal shape, whereas the method used here is a general-purpose method (Pueyo 2006).

which explains most of the bad fit to the data. An inverse relationship between speciation rate and abundance, as mentioned in the introduction, will probably produce even fewer species, and consequently it will even have a worse fit to the data. The only characteristic in which $(DL)M_0$ performs better than $(DL)M_1$ is that it explains the existence of highly abundant species, which is highly improbable under $(DL)M_1$. A saturating relationship of speciation with abundance will have the best of both worlds (proportional to abundance for small abundances and independent of abundance for large abundance). In the supplementary material we specify such a model, M_S . Although we have not found the sampling formula for this model (or any other with such a saturating relationship), we have found an exact expression for the abundance distributions, for a random and a dispersal-limited sample from a meta-community with infinite J_M :

$$E_{M_S}[S_n|x_s, \theta, J] = \binom{J}{n} \int_0^1 x^n (1-x)^{J-n} \Omega(x) dx \quad (9a)$$

$$E_{DLM_S}[S_n|I, x_s, \theta, J] = \frac{1}{(I)_J} \binom{J}{n} \int_0^1 (Ix)_n (I(1-x))_{J-n} \Omega(x) dx \quad (9b)$$

where x_s is a characteristic relative abundance at which the speciation-abundance curve starts to saturate and $\Omega(x)$ is given by

$$\Omega(x) = \frac{\theta}{x} (1-x)^{\theta \frac{x_s}{1+x_s} - 1} \left(\frac{x_s}{x+x_s} \right)^{\theta \frac{x_s}{1+x_s} + 1} \quad (10)$$

This indeed suggests a better fit at high abundances (Fig. 4). However, there are also alternative explanations of highly abundant species. For example, Magurran and Henderson (2003) hypothesize a two-component ecological community: on the one hand there are core species that are highly adapted to their habitat, and on the other hand there are occasional species that largely obey neutrality.

It has been argued that species abundance data cannot discriminate between different models of community structure (McGill 2003a,b, Volkov et al. 2005).

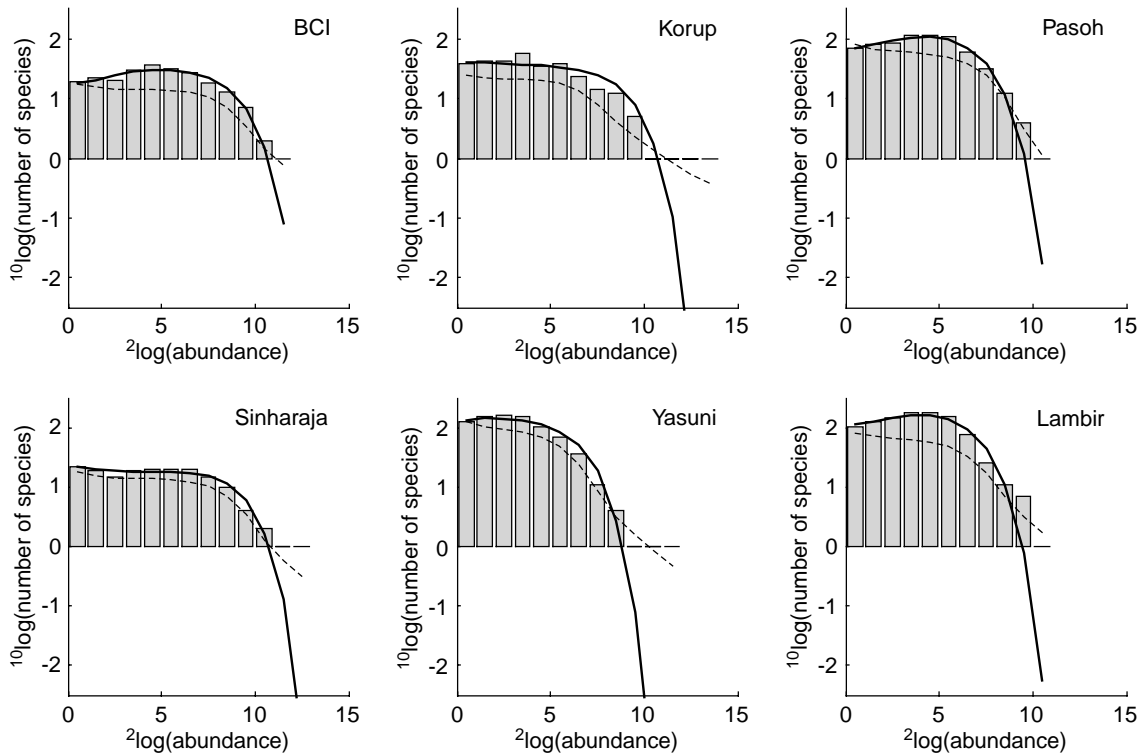


Fig. 3. Same as Fig. 2 but with logarithmic y-axis.

Our results strongly suggest that species abundance data do have discriminative power, in agreement with previous assertions (Etienne and Olff 2005, Chave et al. 2006). It appears that species abundance data can be informative about some processes (e.g. mode of speciation) and uninformative about others (e.g. neutrality). At any rate, we stress that goodness-of-fit alone is an insufficient criterion; the plausibility of the parameter estimates must also be taken into account (the Bayesian view). This also implies that descriptive statistical models having parameters without a clear interpretation (e.g. the Poisson lognormal) have very limited value. Furthermore, visual inspection of goodness-of-fit may be misleading, as indicated by the comparison of linear and log-transformed abundance distributions. The statistical likelihood analysis was able to detect differences between model predictions and observations that are hidden in standard linear species abundance plots.

Although our results suggest that species abundance data do have discriminative power with respect to speciation modes, a completely different interpretation of our results exists that does not involve speciation. The metacommunity abundance distribution of the original neutral model M_1 can also be derived by assuming a very low level of immigration into the (non-zero-sum) metacommunity with immigrants coming from some abstract infinite species pool where all relative abun-

dances are equal (Etienne et al. 2006b). This was in fact the original explanation of Fisher's logseries by Kendall (1948). One may be tempted to conclude from the good performance of M_1 and DLM_1 relative to M_0 and DLM_0 that the probability of immigration of a new species into the metacommunity does not increase with the number of species in the metacommunity, a reasonable explanation; it also makes the relatively high values of the "speciation" parameters more likely. However, M_0 and DLM_0 were built upon assumptions on speciation and, unlike $(DL)M_1$, do not seem to allow for a (simple) reinterpretation in terms of immigration. Moreover, the immigration interpretation only pushes the problem to another level: it still requires an explanation of the diversity and abundance distribution in the abstract species pool (which acts as a metacommunity), ultimately in terms of speciation and extinction. The metacommunity was precisely defined to play this role (Etienne et al. 2006b); immigration is treated in local community models (DLM). In this sense, the species pool in Bell's (2001) neutral model is not a true metacommunity.

The neutral model has been criticized because it predicts species that are impossibly old (Nee 2005). However, the old age of species does not seem to be much affected by the mode of speciation; even the species' initial abundance does not matter much: translating a result from population genetics (Kimura

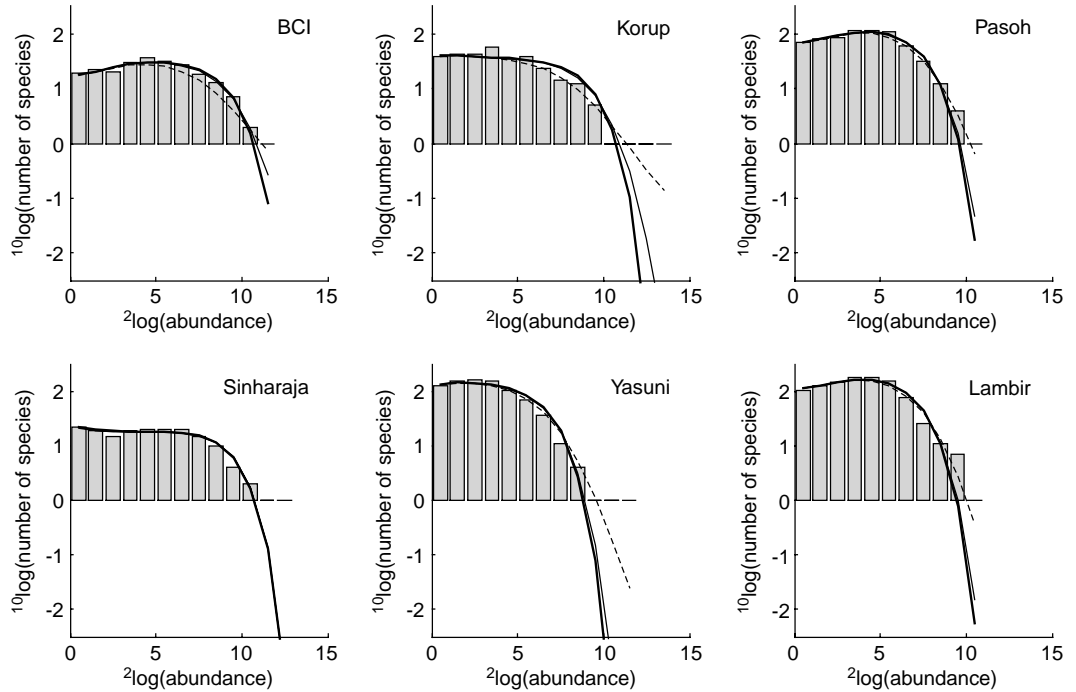


Fig. 4. Observed (bars) and expected (curves) number of species, according to DLM_s, the neutral model with saturating speciation rate, for six data sets (15–20 in Table 1 and 2), and three values of the saturation parameter x_s : $x_s = \infty$ (thick solid curve), $x_s = 0.1$ (thin solid curve), $x_s = 0.01$ (dotted curve). The binning method and scaling is the same as in Fig. 3. The case $x_s = \infty$ is identical to DLM₁. Observe that the curves are indistinguishable for the Sinharaja data set. This is due to the fact that this particular data set has its maximum likelihood at extreme dispersal limitation, in which case the contribution of the metacommunity to the local abundance pattern is very small. There is an almost equally likely parameter set with much less dispersal limitation (which is a common phenomenon, Etienne et al. 2006a). In that case the curves are distinguishable.

and Ohta 1973) leads to an expected age (expressed in generations) of

$$\begin{aligned}
 A(x, x_0) &= -2J_M \left[1 + \frac{1-x_0}{x_0} \log(1-x_0) + \frac{x}{1-x} \log(x) \right] \\
 &\approx -2J_M \left[\frac{x}{1-x} \log(x) \right] \quad \text{for } x_0 \ll x
 \end{aligned}
 \tag{11}$$

for a species currently at relative abundance x with initial relative abundance x_0 . Strictly, this result is only valid for M_1 and for $\theta \rightarrow 0$, but we expect that the assumption of ecological drift in a metacommunity of constant size has a much greater impact on the age of species than the mode of speciation, the amount of dispersal limitation or the value of θ . The zero-sum assumption (constant metacommunity size) only allows species to grow in abundance at the cost of the abundance of other species, but in a non-zero-sum metacommunity species can grow as rapidly as allowed by their reproductive number and hence reach an appreciable abundance in much less time. Yet, the equilibrium sampling formulas of zero-sum and non-zero-sum communities are identical, at least for the standard Hubbell model M_1 (Etienne et al. 2006b).

The ecological drift assumption makes demographic stochasticity the primary driver of species dynamics, whereas in macro-evolutionary time environmental stochasticity and catastrophes play a much more important role and will shorten species ages substantially. However, when averaged over evolutionary time, these processes may well be neutral, leading to equilibrium distributions that are similar to the ones presented here. As an analogy from physics, imagine two initially separated gases that are allowed to mix. With only diffusion it will take eons to reach equilibrium, but with some stirring, this time is enormously shortened, yet the equilibrium will be similar. In sum, our results, which are based on equilibrium sampling formulas, are a priori incompatible with realistic species ages.

Even if a realistic speciation mode will eventually force us to reject neutrality, the neutral model has merits that should be retained (Alonso et al. 2006). The model has brought Ockham's razor back to community ecology and should continue to inspire the construction of new neutral and non-neutral models. The sampling formulas it has generated (Etienne 2005) give excellent descriptions of species-abundance data in terms of interpretable parameters that can be calculated from,

and compared between data sets (Etienne and Olff 2005). Furthermore, it has stimulated work on stochastic community models in a field where deterministic models were the norm. Finally, it has provided a new perspective on how to model dispersal in community ecology and has recognized it as a sampling phenomenon (Etienne and Alonso 2005). This study is the first example of dispersal-limited sampling applied to a metacommunity that is described by a different process than in the standard Hubbell (2001) model.

Our results stress a more general point linking evolutionary biology and community ecology (McPeck and Miller 1996): the nature of the speciation process greatly affects community structure. For example, speciation rates strongly reflect the mating system (Arnqvist et al. 2000), which may therefore play a more crucial role for community structure than typically acknowledged in ecological models based on asexual reproduction. Also, genetic diversity may be strongly related to species diversity (Vellend 2003, 2005); a simple (individual-based, spatial) neutral model of genetic drift leading to reproductive isolation, and hence speciation, combined with ecological drift, may already provide important new insights. We hope that our paper will stimulate research on speciation in ecological communities, and in particular the relationship between speciation rate, species abundances and species diversity, where alternative hypotheses are tested against the observed structure of real ecological communities.

Acknowledgements – We thank Drew Allen for the tree community data files and references to them, and Nigel Pitman for the data set of Manu N.P. We thank Jérôme Chave, John Harte and Brian Maurer for helpful comments.

Supplementary material

Derivation of the new sampling formulas (for M_0 , M_c and DLM_0), of the abundance distribution (for M_s and DLM_s) and details on the selection and analysis of and references to data sets in Table 1.

References

- Abrams, P. A. 2001. A world without competition. – *Nature* 412: 858–859.
- Alonso, D. et al. 2006. The merits of neutral theory. – *Trends Ecol. Evol.* 21: 451–457.
- Arnqvist, G. et al. 2000. Sexual conflict promotes speciation in insects. – *Proc. Natl Acad. Sci. USA* 97: 10460–10464.
- Bell, G. 2001. Neutral macroecology. – *Science* 293: 2413–2418.
- Brown, J. H. 2001. Toward a general theory of biodiversity. – *Evolution* 55: 2137–2138.
- Bush, G. L. et al. 1977. Rapid speciation and chromosomal evolution in mammals. – *Proc. Natl Acad. Sci. USA* 74: 3942–3946.
- Chave, J. 2004. Neutral theory and community ecology. – *Ecol. Lett.* 7: 241–253.
- Chave, J. and Leigh, E. H. 2002. A spatially explicit neutral model of β -diversity in tropical forests. – *Theor. Popul. Biol.* 62: 153–168.
- Chave, J. et al. 2006. Comparing models of species abundance. – *Nature* 441: E1.
- Clark, J. S. and McLachlan, J. S. 2003. Stability of forest biodiversity. – *Nature* 423: 635–638.
- Cohen, J. E. 1968. Alternate derivations of a species-abundance relation. – *Am. Nat.* 102: 165–172.
- Coyne, J. A. and Orr, H. A. 2004. *Speciation*. – Sinauer Associates.
- De Mazancourt, C. 2001. Consequences of community drift. – *Science* 239: 1772.
- Dieckmann, U. et al. (eds) 2004. *Adaptive speciation*. – Cambridge Univ. Press.
- Dornelas, M. et al. 2006. Coral reef diversity refutes the neutral theory of biodiversity. – *Nature* 440: 80–82.
- Emerson, B. C. and Kolm, N. 2005. Species diversity can drive speciation. – *Nature* 434: 1015–1017.
- Etienne, R. S. 2005. A new sampling formula for neutral biodiversity. – *Ecol. Lett.* 8: 253–260.
- Etienne, R. S. and Olff, H. 2005. Confronting different models of community structure to species-abundance data: a Bayesian model comparison. – *Ecol. Lett.* 8: 493–504.
- Etienne, R. S. and Alonso, D. 2005. A dispersal-limited sampling theory for species and alleles. – *Ecol. Lett.* 8: 1147–1156.
- Etienne, R. S. and Alonso, D. 2006. Neutral community theory: how stochasticity and dispersal-limitation can explain species coexistence. – *J. Stat. Phys.* In press. DOI: 10.1007/s10955-006-9163-2.
- Etienne, R. S. et al. 2006a. Comment on “Neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot”. – *Science* 311: 610b.
- Etienne, R. S. et al. 2006b. The zero-sum assumption in neutral biodiversity theory. Manuscript.
- Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. – *Theor. Popul. Biol.* 3: 87–112.
- Fargione, J. et al. 2003. Community assembly and invasion: an experimental test of neutral versus niche processes. – *Proc. Natl Acad. Sci. USA* 100: 8916–8920.
- Gavrilets, S. 2004. *Fitness landscapes and the origin of species*. – Princeton Univ. Press.
- Howard, D. J. and Berlocher, S. H. (eds) 1998. *Endless forms: species and speciation*. – Oxford Univ. Press.
- Hubbell, S. P. 1979. Tree dispersion, abundance and diversity in a tropical dry forest. – *Science* 203: 1299–1309.
- Hubbell, S. P. 2001. *The unified neutral theory of biodiversity and biogeography*. – Princeton Univ. Press.
- Hubbell, S. P. 2003. Modes of speciation and the lifespans of species under neutrality: a response to the comment by Robert E. Ricklefs. – *Oikos* 100: 193–199.
- Hubbell, S. P. and Lake, J. 2003. The neutral theory of biodiversity and biogeography, and beyond. – In: Blackburn, T. and Gaston, K. (eds) *Macroecology: patterns and processes*. Blackwell, pp. 45–63.
- Hu, X-S. et al. 2006. Neutral theory in macroecology and population genetics. – *Oikos* 113: 548–556.

- Jablonski, D. and Roy, K. 2003. Geographical range and speciation in fossil and living molluscs. – *Proc. R. Soc. Lond. B.* 270: 401–406.
- Jetschke, G. 2002. The unified neutral theory of biodiversity and biogeography. – *Ecology* 83: 1771–1772.
- Joyce, P. 1995. Robustness of the Ewens sampling formula. – *J. Appl. Probab.* 32: 609–622.
- Joyce, P. et al. 2003. When can one detect overdominant selection in the infinite-alleles model? – *Ann. Appl. Probab.* 13: 181–212.
- Karlin, S. and McGregor, J. 1972. Addendum to a paper of W. Ewens. – *Theor. Popul. Biol.* 3: 113–116.
- Kendall, R. G. 1948. On some modes of population growth giving rise to R.A. Fisher's logarithmic series distribution. – *Biometrika* 35: 6–15.
- Kimura, M. 1983. The neutral theory of molecular evolution. – Cambridge Univ. Press.
- Kimura, M. and Ohta, T. 1973. The age of a neutral mutant persisting in a finite population. – *Genetics* 75: 199–212.
- Losos, J. B. and Schluter, D. 2000. Analysis of an evolutionary species-area relationship. – *Nature* 408: 847–850.
- Makarieva, A. M. and Gorshkov, V. G. 2004. On the dependence of speciation rates on species abundance and characteristic population size. – *J. Biosci.* 29: 119–128.
- Magurran, A. E. and Henderson, P. A. 2003. Explaining the excess of rare species in natural species abundance distributions. – *Nature* 422: 714–716.
- Mayr, E. 1970. Populations, species and evolution. – Harvard Univ. Press.
- McGill, B. J. 2003a. A test of the unified neutral theory of biodiversity. – *Nature* 422: 881–885.
- McGill, B. J. 2003b. Strong and weak tests of macroecological theory. – *Oikos* 102: 679–685.
- McPeck, M. A. and Miller, T. E. 1996. Evolutionary biology and community ecology. – *Ecology* 77: 1319–1320.
- Moran, P. A. P. 1958. Random processes in genetics. – *Proc. Camb. Philos. Soc.* 54: 60–71.
- Nee, S. 2005. The neutral theory of biodiversity: do the numbers add up? – *Funct. Ecol.* 19: 173–176.
- Palmer, M. W. and Maurer, T. A. 1997. Does diversity beget diversity? A case study of crops and weeds. – *J. Veg. Sci.* 8: 235–240.
- Poorter, L. et al. 2005. Beyond the regeneration phase: differentiation of height-light trajectories among tropical tree species. – *J. Ecol.* 93: 256–267.
- Pueyo, S. 2006. Diversity: between neutrality and structure. – *Oikos* 112: 392–405.
- Purves, D. W. and Pacala, S. W. 2005. Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. – In: Burslem, D. et al. (eds), *Biotic interactions in the tropics*. Cambridge Univ. Press, pp. 107–138.
- Ricklefs, R. E. 2003. A comment on Hubbell's zero-sum ecological drift model. – *Oikos* 100: 185–192.
- Schluter, D. 2000. The ecology of adaptive radiation. – Oxford Univ. Press.
- Tavaré, S. and Ewens, W. J. 1997. Multivariate Ewens distribution. – In: Johnson, N. L. et al. (eds), *Discrete multivariate distributions*. Wiley, pp. 232–246.
- Vallade, M. and Houchmandzadeh, B. 2003. Analytical solution of a neutral model of biodiversity. – *Phys. Rev. E* 68: 061902.
- Vellend, M. 2003. Island biogeography of genes and species. – *Am. Nat.* 162: 358–365.
- Vellend, M. 2005. Species diversity and genetic diversity: parallel processes and correlated patterns. – *Am. Nat.* 166: 199–215.
- Van Doorn, G. S. et al. 2004. Sympatric speciation by sexual selection: a critical reevaluation. – *Am. Nat.* 163: 709–725.
- Volkov, I. et al. 2005. Density dependence explains tree species abundance and diversity in tropical forests. – *Nature* 438: 658–661.
- Watterson, 1974 and Watterson, G. A. 1974. and Wootton, J. T. 2005. Field parameterization and experimental test of the neutral theory of biodiversity. *Nature*. Vol. 4332005: 97–159309312.
- Wright, 1931 and Wright, S. 1931. and Wright, I. J. et al. 2004. The worldwide leaf economics spectrum. – *Nature* 428: 821–827

Supplementary material

Appendix A. Derivation of the new sampling formulas (for M_0 , M_c and DLM_0)

Deterministic model

We first construct a deterministic continuous-time model of the expected number of species with abundance n which we denote by S_n . Let g_n and r_n be the rate of increase and decrease of the number of species with abundance n respectively. Assume that there is a ceiling $c \leq J_M$ to the abundance of a species, $g_n = r_{n+1} = 0$ for $n \geq c$. New species enter at a rate g_0 that may depend on the S_n . The dynamics of S_n are then described by (Alonso 2004, Etienne and Alonso 2006, Mckane, A.J., pers.comm.)

$$\frac{dS_1}{dt} = g_0 + r_2 S_2 - (r_1 + g_1) S_1 \quad (\text{A-1a})$$

$$\frac{dS_n}{dt} = g_{n-1} S_{n-1} + r_{n+1} S_{n+1} - (r_n + g_n) S_n \quad \text{for } 1 < n < c \quad (\text{A-1b})$$

$$\frac{dS_c}{dt} = g_{c-1} S_{c-1} - r_c S_c \quad (\text{A-1c})$$

It is fairly easy to see that the steady-state solution must be

$$S_1 = \frac{g_0}{r_1} \quad (\text{A-2a})$$

$$S_n = \frac{g_{n-1}}{r_n} S_{n-1} \quad \text{for } 2 < n \leq c \quad (\text{A-2b})$$

which can be written as

$$S_n = \frac{g_0}{r_1} \prod_{j=2}^n \frac{g_{j-1}}{r_j} = S_1 \frac{r_1}{r_n} \prod_{j=1}^{n-1} \frac{g_j}{r_j} \quad (\text{A-3})$$

with the convention that the product returns 1 if $j > n$.

In the continuous-time version of Hubbell's (2001) model with zero-sum dynamics (constant community size J_M), the parameters g_n and r_n are given by

$$g_n = \frac{J_M - n}{J_M} \beta n \quad (\text{A-4a})$$

$$r_n = \frac{n}{J_M} (\beta(J_M - n) + \hat{v}_1 J_M) \quad (\text{A-4b})$$

$$g_0 = \hat{v}_1 J_M \quad (\text{A-4c})$$

where β is the per capita birth rate and \hat{v}_1 is the per capita speciation rate (where the subscript 1 refers to the Hubbell model, see the main text). We see here that the overall speciation rate, g_0 , is proportional to metacommunity size J_M and the per species speciation rate is proportional to the abundance n . We can rescale the parameters g_n and r_n by introducing a scaling parameter K ,

$$K := \beta J_M + \hat{v}_1 J_M \quad (\text{A-5})$$

Rescaling time and the speciation rate as

$$\tau := Kt \quad (\text{A-6a})$$

$$v_1 := \frac{g_0}{K} = \frac{\hat{v}_1 J_M}{K} \quad (\text{A-6b})$$

we obtain the following dimensionless parameters,

$$g'_n = \frac{(J_M - n)n}{J_M^2} (1 - v_1) \quad (\text{A-7a})$$

$$r'_n = \frac{n}{J_M} \left(\frac{J_M - n}{J_M} + v_1 \frac{n}{J_M} \right) \quad (\text{A-7b})$$

$$g'_n = v_1 \quad (\text{A-7c})$$

These dimensionless rates correspond exactly to the transition probabilities in the Moran version of Hubbell's model in discrete-time. The Moran (1958, 1962) model in population genetics is identical to Hubbell's except that, in contrast to Hubbell's (2001) formulation, individuals that will die in the next time step are allowed to produce offspring before they die. For example, seeds may be produced before the mother plant dies. The rates (Eq. A-4) or the rescaled rates (Eq. A-7) lead to Hubbell's standard model with its well-known distribution of expected species abundances:

$$S_n(\theta, J_M) = \frac{\theta (J_M + 1 - n)_n}{n (J_M + \theta - n)_n} \quad (\text{A-8})$$

where $\theta = \theta_{\text{Moran}}$,

$$\theta_{\text{Moran}} := \frac{v_1}{1 - v_1} J_M \quad (\text{A-9})$$

For comparison, the original Hubbell model (where there is no reproduction before death) has

$$g'_n = \frac{(J_M - n)n}{J_M(J_M - 1)} (1 - v_1) \quad (\text{A-10a})$$

$$r'_n = \frac{n}{J_M} \left(\frac{J_M - n}{J_M - 1} + \frac{n - 1}{J_M - 1} v_1 \right) \quad (\text{A-10b})$$

$$g'_n = v_1 \quad (\text{A-10c})$$

and (Vallade and Houchmandzadeh 2003, Etienne 2005).

$$\theta_{\text{Hubbell}} = \frac{v_1}{1 - v_1} (J_M - 1) \quad (\text{A-11})$$

and these also lead to Eq. A-8. Hence, the only difference between the Moran version and the original version is the slight difference in θ which is negligible in practical cases where J_M is assumed to be very large; moreover J_M and v_1 cannot be independently estimated from species abundance data, only θ can be estimated. [A technical detail: Hubbell (2001) actually stated that $\theta = 2\hat{v}_1 J_M$; this is the θ for one of his models which has non-overlapping generations (Etienne and Alonso 2006), whereas the model cited mostly in the literature (Vallade and Houchmandzadeh 2003, Etienne 2005) has overlapping generations, as does the Moran model]. Formula A-7 is invariant under sampling, i.e. for a sample of size J , simply replace each J_M by J (but not the J_M that appears in θ):

$$S_n(\theta, J) = \frac{\theta (J + 1 - n)_n}{n (J + \theta - n)_n} \quad (\text{A-12})$$

The speciation probability per individual in Hubbell's model is a constant v_1 which leads to a speciation probability per species of approximately $v_1 n$ (and a speciation rate of exactly $\hat{v}_1 n$). Now we move to the new model with a speciation probability that has the same value v_0 for each species. This results in a speciation probability of $\frac{v_0}{n}$ per individual (to a very good approximation). We can substitute this for v_1 in Eq. A-7 to obtain the dimensionless rates

$$g'_n = \frac{J_M - n}{J_M} \frac{n}{J_M} \left(1 - \frac{v_0}{n} \right) \quad (\text{A-13a})$$

$$r'_n = \frac{n}{J_M} \left(\frac{J_M - n}{J_M} + \frac{n}{J_M} \frac{v_0}{n} \right) \quad (\text{A-13b})$$

$$g'_n = \frac{v_0 S}{J_M} \quad (\text{A-13c})$$

This leads to

$$\begin{aligned}
S_n &= \frac{S_1}{n} \frac{J_M - 1 + v_0}{J_M - n + v_0} \prod_{i=1}^{n-1} \frac{(J_M - i)(i - v_0)}{i(J_M - i + v_0)} \\
&= \frac{\Gamma(n - J_M)\Gamma(n - v_0)\Gamma(2 - v_0 - J_M)}{\Gamma(1 - J_M)\Gamma(1 - v_0)\Gamma(1 + n - v_0 - J_M)} \frac{S_1}{n!} \\
&= S_1 \frac{(1 - J_M)_{n-1} (1 - v_0)_{n-1}}{n!(2 - J_M - v_0)_{n-1}} \quad (\text{A-14})
\end{aligned}$$

We can calculate S_1 by using the fact that $J_M = \sum_{n=1}^{J_M} n S_n$ which yields

$$\begin{aligned}
S_n(v_0, J_M) &= \frac{J_M(v_0)_{J_M-1}}{(J_M - 1)!} \frac{(1 - J_M)_{n-1} (1 - v_0)_{n-1}}{n!(2 - J_M - v_0)_{n-1}} \\
&= \binom{J_M}{n} \frac{\Gamma(n - v_0)\Gamma(J_M - n + v_0)}{\Gamma(J_M)\Gamma(v_0)\Gamma(1 - v_0)} \quad (\text{A-15})
\end{aligned}$$

Summing over n yields

$$S(v_0, J_M) = \frac{(1 + v_0)_{J_M-1}}{(J_M - 1)!} = \frac{1}{v_0 B(v_0, J)} \quad (\text{A-16})$$

where $B(x,y)$ is the Beta function. For a sample from the metacommunity of size J we must apply the hypergeometric distribution and it can be shown that the formula for S_n (and hence for S) is invariant under sampling:

$$\begin{aligned}
S_n(v_0, J) &= \sum_{j=n}^{J_M} \frac{\binom{j}{n} \binom{J_M - j}{J - n}}{\binom{J_M}{J}} \frac{J_M(v_0)_{J_M-1}}{(J_M - 1)!} \\
&\quad \times \frac{(1 - J_M)_{j-1} (1 - v_0)_{j-1}}{j!(2 - J_M - v_0)_{j-1}} \\
&= \frac{J(v_0)_{J-1}}{(J - 1)!} \frac{(1 - J)_{n-1} (1 - v_0)_{n-1}}{n!(2 - J - v_0)_{n-1}} \\
&= \binom{J}{n} \frac{\Gamma(n - v_0)\Gamma(J - n + v_0)}{\Gamma(J)\Gamma(v_0)\Gamma(1 - v_0)} \quad (\text{A-17})
\end{aligned}$$

This is an important result, but it does not give us the required sampling formulas. We will presently turn to the derivation of these sampling formulas, using the full stochastic discrete-time model, but first we want to emphasize an important difference between the two modes of speciation. We saw that Hubbell's standard model has $g'_n = \frac{v_1 J_M}{J_M}$ and that the modified model has

$g'_n = \frac{v_0 S}{J_M}$. This suggests an alternative interpretation of the two models. In the first case the overall speciation rate is proportional to metacommunity size, whereas in the second case it is proportional to metacommunity species richness.

Stochastic model

We use an inductive approach, where we model the transition of one state of community structure (i.e. the exact abundances of each species) to another in discrete time. Hubbell's model is a Markov model (Markov chain), as the probability of transition of one state to another only depends on the state of the former state, and not on any previous states. For Markov models the stationary state (the probability distribution of all possible states in equilibrium) is given by the normalized eigenvector of the transition matrix corresponding to the dominant eigenvalue ($\lambda = 1$). Hence, if we can construct the transition matrix and determine its leading eigenvector, then we have the stationary state of the community, and when we take a sample from that, we have the sampling formula. The number of possible states increases very rapidly with community size, which makes construction of the transition matrix, let alone eigenvector calculations, unfeasible for even small communities size ($J_M > 10$). We therefore extract a general pattern for very small community size and then check it for larger values by other means (see below). For example, for $J_M = 4$, we have 5 possible states: (1,1,1,1), (1,1,2), (2,2), (1,3) and (4). The transition matrix T is given by

$$T = \begin{pmatrix} \frac{v_0}{1} + \frac{1}{4} \left(1 - \frac{v_0}{1}\right) & \frac{2}{4} \left(\frac{2v_0}{4 \cdot 1} + \frac{2v_0}{4 \cdot 2}\right) & 0 & 0 & 0 \\ \frac{3}{4} \left(1 - \frac{v_0}{1}\right) & \frac{2}{4} \left(\frac{2v_0}{4 \cdot 1} + \frac{2v_0}{4 \cdot 2} + \frac{1}{4} \left(1 - \frac{v_0}{1}\right)\right) + \frac{2}{4} \left(\frac{2}{4} \left(1 - \frac{v_0}{1}\right) + \frac{2}{4} \left(1 - \frac{v_0}{2}\right)\right) & \frac{2v_0}{4 \cdot 2} + \frac{2v_0}{4 \cdot 2} & \frac{3}{4} \left(\frac{1v_0}{4 \cdot 1} + \frac{3v_0}{4 \cdot 3}\right) & 0 \\ 0 & \frac{2}{4} \left(\frac{1}{4} \left(1 - \frac{v_0}{1}\right)\right) & \frac{2}{4} \left(1 - \frac{v_0}{2}\right) & \frac{3}{4} \left(\frac{1}{4} \left(1 - \frac{v_0}{1}\right)\right) & 0 \\ 0 & \frac{2}{4} \left(\frac{2}{4} \left(1 - \frac{v_0}{2}\right)\right) & \frac{2}{4} \left(1 - \frac{v_0}{2}\right) & \frac{3}{4} \left(\frac{3}{4} \left(1 - \frac{v_0}{3}\right)\right) + \frac{1}{4} \left(\frac{1}{4} + \frac{3v_0}{4 \cdot 3}\right) & \frac{v_0}{4} \\ 0 & 0 & 0 & \frac{13}{44} \left(1 - \frac{v_0}{3}\right) & 1 - \frac{v_0}{4} \end{pmatrix} \quad (\text{A-18})$$

For instance, element T(2,2) gives the probability to remain in state (1,1,2). This probability can be calculated as follows. In the model, one individual must die. This means that the state between two time units is (1,2), with probability $\frac{2}{4}$, or (1,1,1), also with probability $\frac{2}{4}$. In the first case the state (1,1,2) can be re-entered if there is speciation or if the dying individual produces offspring that replaces itself. The former event has probability $\frac{2v_0}{4} + \frac{2v_0}{4}$ whereas the latter event has probability $\frac{1}{4}(1 - \frac{v_0}{1})$. In the second case the state (1,1,2) can be regained if any of the individuals reproduces without speciation which has probability $\frac{2}{4}(1 - \frac{v_0}{1}) + \frac{2}{4}(1 - \frac{v_0}{2})$. Adding all this contribution results in T(2,2). The normalized eigenvector of T corresponding to eigenvalue $\lambda = 1$ is

$$\begin{pmatrix} v_0^3 \\ 2v_0^2(1 - v_0) \\ \frac{1}{2}v_0(1 - v_0)^2 \\ \frac{2}{3}v_0(1 - v_0)(2 - v_0) \\ \frac{1}{6}(3 - v_0)(2 - v_0)(1 - v_0) \end{pmatrix} \quad (\text{A-19})$$

For example, the fourth element, $\frac{2}{3}v_0(1 - v_0)(2 - v_0)$, gives the probability of state (1,3). Repeating this procedure for other small values of metacommunity size and computing the eigenvector corresponding to $\lambda = 1$ gives stationary distributions that can be generalized to

$$\begin{aligned} P_{M_0}[\vec{D}|v_0, J_M] &= \frac{J_M}{\prod_{i=1}^S n_i! \prod_{j=1}^J \Phi_j!} v_0^{S-1} (S-1)! \prod_{i=1}^S (1 - v_0)_{n_i-1} \\ &= \frac{J_M!}{\prod_{i=1}^S n_i! \prod_{j=1}^J \Phi_j!} v_0^{S-1} \frac{(S-1)!}{(J_M - 1)!} \\ &\quad \times \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \end{aligned} \quad (\text{A-20})$$

The second line is written in such a way that it has the same prefactor as the Ewens sampling formula. So far we have no formal proof of this result for arbitrary values of J_M . However, it can be verified numerically for any choice of J_M (within computational power). Also, it is fairly straightforward to show (and easy to check numerically) that this distribution is invariant under hypergeometric sampling (show that it is true for a sample of size $J = J_M - 1$ and by repeating this, it follows that it is true for any J). Hence Eq. A-20 also holds for samples of size J if we replace J_M by J . Furthermore,

it can be shown analytically that the expected number of species with abundance n that follows from the stochastic Eq. A-20, which we will denote by $E_{M_0}[S_n|v_0, J]$, is equal to S_n given in Eq. A-17, which was obtained from deterministic differential equations (hence the difference in notation). First note that (see also Etienne and Alonso 2005),

$$\begin{aligned} E_{M_0}[S_n|v_0, J] &= \sum_{\Phi_n=1}^J \sum_{\{\vec{D}|\Phi_n\}} \Phi_n \frac{J!}{\prod_{i=1}^S n_i! \prod_{j=1}^J \Phi_j!} v_0^{S-1} \frac{(S-1)!}{(J-1)!} \\ &\quad \times \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \\ &= \sum_{\Phi_n=1}^J \sum_{\{\vec{D}|\Phi_n\}} \frac{J!}{\prod_{i=1}^S n_i! \prod_{j=1}^J (\Phi_j - \delta_{jn})!} v_0^{S-1} \\ &\quad \times \frac{(S-1)!}{(J-1)!} \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \end{aligned} \quad (\text{A-21})$$

Defining $\Phi'_j = \Phi_j - \delta_{jn}$ where δ_{jn} is Kronecker's delta (which is equal to 1 if $j = n$ and 0 otherwise), we can rewrite this as a sum of probabilities of observing exactly $\Phi_n - 1$ species with abundance n in a subsample of size $J - n$,

$$\begin{aligned} E_{M_0}[S_n|v_0, J] &= \frac{v_0}{n} \frac{J!}{(J-n)!} \frac{(J-n-1)!}{(J-1)!} (S-1) \prod_{j=1}^{n-1} \left(1 - \frac{1}{j} v_0\right) \\ &\quad \times \sum_{\Phi_n=0}^{J-n} \sum_{\{\vec{D}|\Phi_n\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i! \prod_{j=1}^{J-n} \Phi'_j!} v_0^{S-1-1} \\ &\quad \times \frac{(S-1-1)!}{(J-n-1)!} \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \\ &= \frac{v_0}{n} \frac{J}{J-n} \prod_{j=1}^{n-1} \binom{j-v_0}{j} \sum_{\{\vec{D}\}} (S-1) \\ &\quad \times \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i! \prod_{j=1}^{J-n} \Phi'_j!} v_0^{S-1-1} \frac{(S-1-1)!}{(J-n-1)!} \\ &\quad \times \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) = \frac{v_0}{n} \frac{J}{J-n} \frac{(1-v_0)_{n-1}}{(n-1)!} \\ &\quad \times \sum_{\{\vec{D}\}} (S-1) \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i! \prod_{j=1}^{J-n} \Phi'_j!} v_0^{S-1-1} \\ &\quad \times \frac{(S-1-1)!}{(J-n-1)!} \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \end{aligned} \quad (\text{A-22})$$

We now use the fact that

$$\begin{aligned} & \sum_{\{\vec{D}\}} (S-1) \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^{J-n} \Phi_j!} v_0^{S-1-1} \frac{(S-1-1)!}{(J-n-1)!} \\ & \times \prod_{i=1}^{S-1} \prod_{j=1}^{n_i-1} \left(1 - \frac{1}{j} v_0\right) \\ & = E_{M_0} [S|v_0, J-n] = \frac{(1+v_0)_{J-n-1}}{(J-n-1)!} \end{aligned} \quad (\text{A-23})$$

where the latter equality can be derived by summing over the $E[S_n|v_0, J]$. This equality can also be independently established (similar to the derivation for the original Hubbell model performed in Etienne and Alonso 2005). With Eq. A-23, Eq. A-22 becomes

$$\begin{aligned} E_{M_0} [S_n|v_0, J] &= \frac{v_0}{n!} \frac{J}{J-n} (1-v_0)_{n-1} \frac{(1+v_0)_{J-n-1}}{(J-n-1)!} \\ &= \frac{v_0}{n!} \frac{J}{(J-n)!} (1-v_0)_{n-1} (1+v_0)_{J-n-1} \\ &= \frac{v_0}{n!} \frac{J}{(J-n)!} \prod_{j=1}^{n-1} (1-v_0+j-1) \\ & \times \prod_{j=1}^{J-n-1} (1+v_0+j-1) \\ &= \frac{Jv_0}{n!(J-1)!} \prod_{j=1}^{n-1} (J-j) \prod_{j=1}^{n-1} (1-v_0+j-1) \\ & \times \prod_{j=2}^{J-1} (v_0+j-1) \prod_{j=J-n+1}^{J-1} \frac{1}{v_0+j-1} \\ &= \frac{J}{(J-1)!n!} \prod_{j=1}^{J-1} (v_0+j-1) \prod_{j=1}^{n-1} (j-J) \\ & \times \prod_{j=1}^{n-1} (1-v_0+j-1) \prod_{j=J-n+1}^{J-1} \frac{1}{1-v_0-j} \\ &= \frac{J}{(J-1)!n!} \prod_{j=1}^{J-1} (v_0+j-1) \prod_{j=1}^{n-1} (1-J+j-1) \\ & \times \prod_{j=1}^{n-1} (1-v_0+j-1) \prod_{j=1}^{n-1} \frac{1}{2-J-v_0+j-1} \\ &= \frac{J(v_0)_{J-1}}{(J-1)!} \frac{(1-J)_{n-1} (1-v_0)_{n-1}}{n!(2-J-v_0)_{n-1}} \end{aligned} \quad (\text{A-24})$$

which is identical to Eq. A-17.

By the same reasoning we find the expressions for M_c and again it can be shown that the resulting

sampling formula is also invariant under hypergeometric sampling (but a separate dependence on J_M remains, see Eq. 5) and that it leads to the expected number of species with abundance n given by Eq. A-3 with the appropriate substitutions for g_n and r_n . The latter expression can also be shown to be invariant under hypergeometric sampling.

Dispersal limitation

To obtain a dispersal-limited version of the sampling formula (4) we apply the dispersal-limited sampling formula given by Etienne and Alonso (2005):

$$\begin{aligned} P[\vec{D}|I, \Theta, J] &= \frac{1}{\prod_{j=1}^J \Phi_j!} \frac{J!}{n_1! \dots n_S!} \\ & \times \sum_{A=1}^J \sum_{\{a_1, \dots, a_S | \sum_{k=1}^S a_k = A\}} \left(\prod_{i=1}^S \bar{s}(n_i, a_i) \right) \\ & \times \frac{I^A}{(I)_J} \frac{a_1! \dots a_S!}{A!} P[a_1, \dots, a_S | \Theta, A] \end{aligned} \quad (\text{A-25})$$

where I is the fundamental dispersal number (Etienne and Alonso 2005), Θ represents the model parameters of the non-dispersal-limited model and $P[a_1, \dots, a_S | \Theta, A]$ is the sampling formula for a sample of size A from the metacommunity. If we substitute Eq. 2 for $P[a_1, \dots, a_S | \Theta, A]$, we arrive at the Etienne sampling formula (Etienne 2005, Alonso et al. 2006).

$$\begin{aligned} P_{DLM_1} [\vec{D}|\theta, I, J] &= \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(\theta)_A} \\ &= P_{M_1} [\vec{D}|\theta, J] \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(I)_J} \frac{(\theta)_J}{(\theta)_A} \end{aligned} \quad (\text{A-26})$$

with $K(\vec{D}, A)$ given by

$$K(\vec{D}, A) = \sum_{\{a_1, \dots, a_S | \sum_{i=1}^S a_i = A\}} \prod_{i=1}^S \bar{s}(n_i, a_i) \frac{(a_i - 1)!}{(n_i - 1)!} \quad (\text{A-27})$$

Here $\bar{s}(x, y)$ denotes the unsigned Stirling number of the first kind (see Etienne 2005). If, instead of Eq. 2 we substitute Eq. 4 for $P[a_1, \dots, a_S | \Theta, A]$, we obtain

$$\begin{aligned} P_{DLM_0} [\vec{D}|v_0, I, J] &= P_{M_0} [\vec{D}|v_0, J] \sum_{A=S}^J L(\vec{D}, v_0, A) \frac{I^A}{(I)_J} \frac{(J-1)!}{(A-1)!} \end{aligned} \quad (\text{A-28})$$

where $L(\vec{D}, v_0, A)$ is given by

$$L(\vec{D}, v_0, A) = \sum_{\{a_1, \dots, a_s \mid \sum_{i=1}^s a_i = A\}} \prod_{i=1}^s \bar{s}(n_i, a_i) \frac{(1 - v_0)^{a_i - 1}}{(1 - v_0)^{n_i - 1}} \quad (\text{A-29})$$

Likewise we can calculate $E_{\text{DLM}_0}[S_n | v_0, I, J]$. From Etienne and Alonso (2005) we have

$$\begin{aligned} E[S_n | I, \Theta, J] &= \binom{J}{n} \sum_{A=1}^J \sum_{a=1}^n \bar{s}(n, a) \bar{s}(J - n, A - a) \frac{I^A}{(I)_J} \\ &\times \frac{1}{\binom{A}{a}} E[S_a | \Theta, A] \end{aligned} \quad (\text{A-30})$$

This yields, substituting Eq. A-17 for $E[S_a | \Theta, A]$,

$$\begin{aligned} E_{\text{DLM}_0}[S_n | v_0, I, J] &= \binom{J}{n} \sum_{A=1}^J \sum_{a=1}^n \bar{s}(n, a) \bar{s}(J - n, A - a) \frac{I^A}{(I)_J} \frac{1}{\binom{A}{a}} \\ &\times \binom{A}{a} \frac{\Gamma(a - v_0) \Gamma(A - a + v_0)}{\Gamma(A) \Gamma(v_0) \Gamma(1 - v_0)} \\ &= \binom{J}{n} \sum_{a_1=1}^n \sum_{a_2=1}^{J-n} \bar{s}(n, a_1) \bar{s}(J - n, a_2) \frac{I^{a_1 + a_2}}{(I)_J} \\ &\times \frac{B(a_1 - v_0, a_2 + v_0)}{\Gamma(v_0) \Gamma(1 - v_0)} \\ &= \frac{1}{\Gamma(v_0) \Gamma(1 - v_0)} \binom{J}{n} \sum_{a_1=1}^n \sum_{a_2=1}^{J-n} \bar{s}(n, a_1) \bar{s}(J - n, a_2) \\ &\times \frac{I^{a_1 + a_2}}{(I)_J} \int_0^1 x^{a_1 - v_0 - 1} (1 - x)^{a_2 + v_0 - 1} dx \\ &= \frac{1}{(I)_J \Gamma(v_0) \Gamma(1 - v_0)} \\ &\times \binom{J}{n} \int_0^1 \left(x^{-v_0 - 1} (1 - x)^{v_0 - 1} \sum_{a_1=1}^n \bar{s}(n, a_1) (Ix)^{a_1} \right. \\ &\times \left. \sum_{a_2=1}^{J-n} \bar{s}(J - n, a_2) (I(1 - x))^{a_2} \right) dx \\ &= \frac{1}{\Gamma(v_0) \Gamma(1 - v_0)} \binom{J}{n} \int_0^1 \frac{(Ix)_n (I(1 - x))_{J-n}}{(I)_J} \\ &\times \frac{(1 - x)^{v_0 - 1}}{x^{v_0 + 1}} dx \end{aligned} \quad (\text{A-31})$$

and summing over n gives

$$\begin{aligned} E_{\text{DLM}_0}[S | v_0, I, J] &= \frac{1}{\Gamma(v_0) \Gamma(1 - v_0)} \int_0^1 \left(1 - \frac{I(1 - x)_J}{(I)_J} \right) \\ &\times \frac{(1 - x)^{v_0 - 1}}{x^{v_0 + 1}} dx \end{aligned} \quad (\text{A-32})$$

where we have used Vandermonde's formula,

$$(x + y)_N = \sum_{n=0}^N \binom{N}{n} (x)_n (y)_{N-n} \quad (\text{A-33})$$

It may be interesting to note that

$$\frac{1}{\Gamma(v_0) \Gamma(1 - v_0)} = \frac{\sin \pi v_0}{\pi} \quad (\text{A-34})$$

References

- Alonso, D. 2004. The stochastic nature of ecological interactions. Communities, metapopulations, and epidemics. – PhD thesis. Tech. Univ. Catalonia, Spain.
- Etienne, R.S. and Alonso, D. 2005. A dispersal-limited sampling theory for species and alleles. – *Ecol. Lett.* 8: 1147–1156.
- Etienne, R.S. and Alonso, D. 2006. Neutral community theory: how stochasticity and dispersal-limitation can explain species coexistence. – *J. Stat. Phys.* In press. DOI: 10.1007/s10955-006-9163-2.
- Hubbell, S.P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.
- Moran, P.A.P. 1958. Random processes in genetics. – *Proc. Camb. Philos. Soc.* 54: 60–71.
- Moran, P.A.P. 1962. Statistical processes of evolutionary theory. – Clarendon Press.
- Vallade, M. and Houchmandzadeh, B. 2003. Analytical solution of a neutral model of biodiversity. – *Phys. Rev. E* 68: 061902.

Appendix B. Derivation of the abundance curve of M_s and DLM_s

In the discussion we suggested that a saturating speciation-abundance relationship would be a more realistic model. Although a tractable sampling formula is not available, we can derive the abundance curve for this model, M_s and its dispersal-limited version, DLM_s , that is expressions for $E[S_n | \Theta, J]$ and $E[S_n | I, \Theta, J]$ in the limit of $J_M \rightarrow \infty$ (where $v_1 \rightarrow 0$ such that θ becomes a finite number). For this we use again Eq. A-3 with the following rates

$$g'_s = \frac{J_M - n}{J_M} \frac{n}{J_M} \left(1 - \frac{x_s v_1}{x_s + \frac{n}{J_M}} \right) \quad (\text{B-1a})$$

$$r'_n = \frac{n}{J_M} \left(\frac{J_M - n}{J_M} + \frac{n}{J_M} \frac{x_s v_1}{x_s + \frac{n}{J_M}} \right) \quad (\text{B-1b})$$

$$g'_0 = \sum_{i=1}^{J_M} \frac{x_s v_1}{x_s + \frac{i}{J_M}} \frac{i S_i}{J_M} \quad (\text{B-1c})$$

where x_s is a characteristic relative abundance at which the speciation-abundance curve starts to saturate (x_s may exceed unity as it is not a real relative abundance). For $x_s \rightarrow \infty$ we retrieve the standard Hubbell model, M_1 . For small x_s we obtain something similar to M_0 . With these rates Eq. A-3 becomes

$$S_1 = \frac{g'_0}{r'_1} \quad (\text{B-2a})$$

$$S_n = S_1 \frac{r'_1}{r'_n} \prod_{j=1}^{n-1} \frac{g'_j}{r'_j} = S_1 \frac{(J_M - 1)(1 + \gamma) + \theta}{n(J_M - n)(1 + \gamma n) + \theta} \times \prod_{j=2}^{n-1} \frac{(J_M - j)(1 + \gamma j)}{(J_M - j)(1 + \gamma j) + \theta} \quad (\text{B-2b})$$

where

$$\gamma = \frac{1}{J_M x_s (1 - v_1)} \quad (\text{B-3})$$

We will compute S_1 by using the fact that $J_M = \sum_{n=1}^{J_M} n S_n$ but only after we take the limit for $J_M \rightarrow \infty$. After some straightforward algebra we obtain

$$S_n = \frac{S_1}{n} \frac{\Gamma(J_M)}{\Gamma(J_M - n + 1)} \frac{\Gamma(\frac{1}{\gamma} + n)}{\Gamma(\frac{1}{\gamma} + 1)} \frac{\Gamma(-y + 2)}{\Gamma(-y + n + 1)} \times \frac{\Gamma(J_M - n - y - \frac{1}{\gamma})}{\Gamma(J_M - 1 - y - \frac{1}{\gamma})} \quad (\text{B-4})$$

where

$$y = \frac{1}{2} \left(J_M - \frac{1}{\gamma} \right) - \frac{1}{2} \sqrt{\left(J_M + \frac{1}{\gamma} \right)^2 + \frac{4\theta}{\gamma}} \quad (\text{B-5})$$

The asymptotic behavior of the Gamma function is known as Stirling's formula,

$$\Gamma(z) \approx e^{-z} z^{z-\frac{1}{2}} \sqrt{2\pi} \quad \text{for large } z \quad (\text{B-6})$$

and the asymptotic behavior of y is given by

$$y \approx -J_M x_s - \theta \frac{x_s}{1 + x_s} \quad \text{for large } J_M \text{ and small } v_1 \quad (\text{B-7})$$

When we let J_M go to infinity, we need the continuous version of S_n which we call $\Omega(x)dx$; this is the number of species with relative abundance between x and $x+dx$. Because $dx = \frac{1}{J_M}$ and $n = xJ_M$ we have, using B-6 and B-7,

$$\begin{aligned} \Omega(x) &= \lim_{J_M \rightarrow \infty} J_M S_{xJ_M} \\ &= S_1 \frac{(1-x)^{\theta \frac{x_s}{1+x_s} - 1}}{x} \left(\frac{x_s}{x+x_s} \right)^{\theta \frac{x_s}{1+x_s} + 1} \end{aligned} \quad (\text{B-8})$$

(see also Vallade and Houchmandzadeh 2003). The constant S_1 can now readily be determined by

$$\int_0^1 x \Omega(x) dx = 1 \quad (\text{B-9})$$

and this gives $S_1 = \theta$, so

$$\Omega(x) = \frac{\theta}{x} (1-x)^{\theta \frac{x_s}{1+x_s} - 1} \left(\frac{x_s}{x+x_s} \right)^{\theta \frac{x_s}{1+x_s} + 1} \quad (\text{B-10})$$

Finally, we can use the following expressions for a random and a dispersal limited sample of size J (Etienne and Alonso 2005),

$$E_{M_S} [S_n | x_s, \theta, J] = \binom{J}{n} \int_0^1 x^n (1-x)^{J-n} \Omega(x) dx \quad (\text{B-11a})$$

$$\begin{aligned} E_{DLM_S} [S_n | I, x_s, \theta, J] \\ = \frac{1}{(I)_J} \binom{J}{n} \int_0^1 (Ix)_n (I(1-x))_{J-n} \Omega(x) dx \end{aligned} \quad (\text{B-11b})$$

References

- Etienne, R.S. and Alonso, D. 2005. A dispersal-limited sampling theory for species and alleles. – *Ecol. Lett.* 8: 1147–1156.
 Vallade, M. and Houchmandzadeh, B. 2003. Analytical solution of a neutral model of biodiversity. – *Phys. Rev. E* 68: 061902.

Appendix C. Data set selection and analysis

Data files and literature references for the first 13 the data sets of Table 1 were kindly provided by Andrew P. Allen, who searched for all the standard terms “permanent plots”, “stand plots”, “forest inventory”, etc. in SciSearch, JSTOR, and the Latin American electronic journal portal SciELO (www.scielo.org) to find references, and then checked bibliographies to find other references. This resulting in some 120 data sets of tree communities. From these only those with sample sizes larger than 5 000 were selected, as smaller sample sizes yield too unreliable parameter estimates. Seven large tropical tree data sets were added, one (Manu N.P.) provided by Nigel Pitman and six from a recent paper by Volkov et al. 2005, to extend the range of diversity values. In addition to Nigel Pitman, we thank John Terborgh and Percy Núñez for the Manu N.P. data set. Some of the data sets in Volkov et al. 2005 were also

found by Andrew Allen; in those cases we used the data file provided by Volkov et al. 2005.

The Akaike weight of model i in Table 1 and 2 is calculated as

$$w_i = \frac{e^{-\frac{\Delta_i}{2}}}{\sum_i e^{-\frac{\Delta_i}{2}}} \quad (\text{C-1})$$

where Δ_i is defined by the difference between the model's Akaike information criterion (AIC) and the AIC of the best model (i.e. the model with the lowest AIC):

$$\Delta_i = \text{AIC}_i - \text{AIC}_{\min} \quad (\text{C-2a})$$

$$\text{AIC}_i = -2\log(L_i) + 2K_i \quad (\text{C-2b})$$

where K_i is the number of parameters (degrees of freedom) of model i ($K=2$ in our dispersal limited models) and L_i is the likelihood of model i (in our dispersal limited models they are given by Eq. 7 and Eq. 8). Inserting Eq. C-2 in Eq. C-1 gives

$$w_i = \frac{e^{\log(L_i) - \log(L_{\min}) - K_i + K_{\min}}}{\sum_i e^{\log(L_i) - \log(L_{\min}) - K_i + K_{\min}}} \quad (\text{C-3})$$

For M_c we used 2 degrees of freedom. This is a conservative estimate. There are at least 2 degrees of freedom, because apart from θ there is one additional independent parameter, v_0 , in M_c . But there are actually more degrees of freedom, because the parameters v_1 and J_M that are contained in a single parameter θ in M_1 decouple in M_c . The number of degrees of freedom is, however, not 3, because v_1 and J_M remain highly correlated in M_c , so it is somewhere between 2 and 3. For the data sets we studied, the correlation between the estimated v_1 and J_M is indeed very high: choosing different initial values in the numerical optimization algorithm yields different values, but the θ -value determined by v_1 and J_M remains the same.

References corresponding to Table 1

Ayyappan, N. and Parthasarathy, N. 1999. Biodiversity inventory of trees in a large-scale permanent plot of tropical evergreen forest at Varagalaiar, Anamalais, Western Ghats, India. – *Biodivers. Conserv.* 8: 1533–1554.

- Beard, J. S. 1946. The Mora forests of Trinidad, British West-Indies. – *J. Ecol.* 33: 173–192.
- Bernadzki, E. et al. 1998. Compositional dynamics of natural forests in the Bialowieza National Park, northeastern Poland. – *J. Veg. Sci.* 9: 229–238.
- Bhat, D. M. et al. 2000. Forest dynamics in tropical rain forests of Uttara Kannada district in Western Ghats, India. – *Curr. Sci.* 79: 975–985.
- Dyrness, T. and Acker, S. 1999. Reference stands in or near the H.J. Andrews Experimental Forest. Permanent plots of the Pacific Northwest. Report Number 2. <http://www.fsl.orst.edu/lter/pubs/webdocs/reports/permpplot/hja.cfm?topnav=55>.
- Dyrness, T. and Acker, S. 2000. Permanent plots in the Mount Rainier National Park. Permanent Plots of the Pacific Northwest. Report Number 4. <http://www.fsl.orst.edu/lter/pubs/webdocs/reports/permpplot/rainier.cfm?topnav=55>.
- Eggeling, W. J. 1947. Observations on the ecology of the Budongo Rain Forest, Uganda. – *J. Ecol.* 34: 20–87.
- Fain, J. J. et al. 1994. 50 years of change in an upland forest in south-central New-York-General patterns. – *Bull. Torrey Bot. Club* 121: 130–139.
- Harmon, M. E. et al. 1998. Permanent plots surrounding the Wind River Canopy Crane. Permanent plots of the Pacific Northwest. Report Number 1. <http://www.fsl.orst.edu/lter/pubs/permpplot.htm>.
- Harrison, E. A. et al. 1989. Community dynamics and topographic controls on forest pattern in Shenandoah National Park, Virginia. – *Bull. Torrey Bot. Club* 116: 1–14.
- Nunes, Y. R. F. et al. 2003. Variations in tree community physiognomy, diversity, and species guild composition of a fragment of tropical semideciduous forest in Lavras, south-eastern Brazil. – *Acta Bot. Brasil.* 17: 213–229.
- Schulze, M. D. and Whitacre, D. F. 1999. A classification and ordination of the tree community of Tikal National Park, Petén, Guatemala. – *Bull. Fla. Mus. Nat. Hist.* 41: 169–297.
- Whittaker, R. H. et al. 1979. The Hubbard Brook ecosystem study: Forest nutrient cycling and element behavior. – *Ecology* 60: 203–220.
- Pitman, N. C. A. et al. 2001. Dominance and distribution of tree species in upper Amazonian terra firme forests. – *Ecology* 82: 2101–2117. (Manu N.P., Peru)
- Volkov, I. et al. 2005. Density dependence explains tree species abundance and diversity in tropical forests. – *Nature* 438: 658–661. (15. BCI, 16. Korup, 17. Pasoh, 18. Sinharaja, 19. Yasuni, 20. Lambir).