

LETTER

A neutral sampling formula for multiple samples and an 'exact' test of neutrality

Rampal S. Etienne*
Community and Conservation
Ecology Group, University of
Groningen, PO Box 14, 9750 AA
Haren, The Netherlands
*Correspondence: E-mail:
r.s.etienne@rug.nl

Abstract

As the utility of the neutral theory of biodiversity is increasingly being recognized, there is also an increasing need for proper tools to evaluate the relative importance of neutral processes (dispersal limitation and stochasticity). One of the key features of neutral theory is its close link to data: sampling formulas, giving the probability of a data set conditional on a set of model parameters, have been developed for parameter estimation and model comparison. However, only single local samples can be handled with the currently available sampling formulas, whereas data are often available for many small spatially separated plots. Here, I present a sampling formula for multiple, spatially separated samples from the same metacommunity, which is a generalization of earlier sampling formulas. I also provide an algorithm to generate data sets with the model and I introduce a general test of neutrality that does not require an alternative model; this test compares the probability of the observed data (calculated using the new sampling formula) with the probability of model-generated data sets. I illustrate this with tree abundance data from three large Panamanian neotropical forest plots. When the test is performed with model parameters estimated from the three plots, the model cannot be rejected; however, when parameter estimates previously reported for BCI are used, the model is strongly rejected. This suggests that neutrality cannot explain the structure of the three Panamanian tree communities on the local (BCI) and regional (Panama Canal Zone) scale simultaneously. One should be aware, however, that aspects of the model other than neutrality may be responsible for its failure. I argue that the spatially implicit character of the model is a potential candidate.

Keywords

Biodiversity, dispersal limitation, Ewens sampling formula, likelihood, metacommunity, neutral model, niche differentiation, sequential construction, species abundance distribution.

Ecology Letters (2007) 10: 608–618

INTRODUCTION

Neutral theory in community ecology, strongly revived by Hubbell (2001) in his already seminal work, is starting to mature, in the same spirit as its ancestor in population genetics. Ecologists are increasingly accepting the merits of neutral theory (Alonso *et al.* 2006; Adler *et al.* 2007) as a null model, both in a philosophical and a technical sense. Philosophically, the assumption of ecological equivalence, i.e. differences between species do not matter for their abundance and diversity in ecological communities, is the most parsimonious description only to be replaced by a

more complex description when data convincingly tell us so. Technically, neutrality functions as a first approximation that may work well for some practical purposes in ecology (e.g. as an emergent property on the right spatial and temporal scale, Holt 2006) in the same way that the ideal gas law has proved useful in physics. However, this does not mean that neutrality itself is the generally accepted explanation for the structure of ecological communities. On the contrary, empirical evaluation of neutral theory has shown many limitations (McGill *et al.* 2006). Part of those limitations may be ascribed to other parts of the particular model formulation than neutrality *per se*. For example, the

effect of the assumed speciation mechanism (which has little to do with neutrality) on the shape of species-abundance distributions can be large (Etienne *et al.* 2007b) and a spatially explicit neutral model has been shown to give a much better description of species–area relationships than the original spatially implicit version (J. Rosindell & S.J. Cornell 2007). At the same time, the effect of the zero-sum assumption in the model seems to be nil (R.S. Etienne, D. Alonso, A.J. McKane, unpublished data). Hence, more model extensions and relaxation of other assumptions than neutrality are needed to explore the domain of applicability of the neutrality assumption.

Here, I extend Hubbell's local community model in such a way that it can be applied to multiple, spatially separated samples of species abundances simultaneously. Current analytical results in neutral theory apply only to a single, dispersal-limited, local sample that is linked to or embedded in (Alonso *et al.* 2006) a metacommunity dictated by a speciation-extinction balance (Volkov *et al.* 2003; Etienne 2005). Often, however, data consist of samples from several discontinuous plots. Lumping these together to produce a single species abundance data set assumes that all individuals in the total area sampled have the same probability to colonize an empty site. In a single sample from a contiguous plot this assumption is approximately satisfied if the plot is not too large relative to the dispersal distance, but for several samples from spatially distinct plots such a hand-waving argument has little credibility, and thus lumping is not appropriate. This problem has been recognized and a preliminary solution has been offered (Jabot, F. Chave, J. & Etienne, R.S., unpublished data; Munoz *et al.* 2007), but here I derive a full analytical sampling formula for this situation. It includes the single sample formula (Etienne 2005) as a special case. In addition, I present a new statistical test of the neutral model (given a data set) that does not require an alternative (niche-based) model. This is an important step as it avoids discussion of validity of the alternative model in empirical evaluations. I illustrate the new formula and test by applying them to three neotropical forest samples across a rainfall gradient. The results support the (already known) fact that the three communities are not ecologically equivalent. However, I also discuss other interpretations of the test results.

THE SAMPLING FORMULA FOR MULTIPLE SAMPLES

Suppose that there are N samples from N different local communities, each of which contains J_i individuals ($i = 1 \dots N$), summing to a total of J individuals in all samples together and a total of S different species. The N samples sizes can be summarized by the vector $\vec{J} = (J_1, \dots, J_N)$. The species found in these samples are indicated by an arbitrary order $k = 1 \dots S$ and the data set of all species

abundances \vec{D} can be written as a vector of vectors $\vec{D} = (\vec{D}_1, \dots, \vec{D}_N) = [(n_{11}, \dots, n_{1S}), \dots, (n_{N1}, \dots, n_{NS})]$ where n_{ik} represents the number of individuals of species k in sample i . The new sampling formula that I will derive is an expression for the probability of observing such a data set under the neutral model. The derivation follows the genealogical approach (Etienne & Olff 2004b; Etienne 2005). Instead of the assumption that all individuals regardless of spatial location have the same probability to colonize an empty site (which would allow lumping the samples together and implies treating all individuals as members of the same large local community), I assume that only individuals in the same plot can colonize an empty site with equal probability; immigrants from elsewhere can colonize this site, but these immigrants originate directly from the metacommunity, i.e. the regional species pool that is in speciation-extinction balance, and not from one of the other plots (see Fig. 1). In other words, the plots should be sufficiently far apart for one plot to have a negligible contribution to immigration into another plot, compared with the contribution of the whole metacommunity. Nevertheless, the plots do depend on one another because they share a common source pool of immigrants. This set-up bears some similarity to the scattering and collecting phases in population genetics (Wakeley 1999).

As in the single-sample case, the metacommunity is governed by the fundamental biodiversity parameter θ (Hubbell 2001) and there is immigration to each local community. Hubbell (2001) described the probability of immigration to the single local community in his model with the parameter m . This parameter m can be related to the fundamental dispersal limitation parameter I by $m := \frac{I}{I+J-1}$ (Etienne & Alonso 2005) which shows that m depends on sample size. This becomes problematic in a sampling formula for multiple samples and instead the fundamental dispersal number I (or I_i for each sample i) should be used as a measure of dispersal limitation. Replacing m by I calls for a clear interpretation of I . This interpretation was first given in Etienne & Olff (2004a): I is the potential number of immigrants that compete with the local individuals for available sites. Thus, I does not depend on sample size, but rather on sample area, because the number of immigrants that compete with the local individuals decreases as the sample area increases and thus the average distance between immigrants and locals becomes larger. When I is assumed the same for all local samples, the sampling formula for multiple samples should therefore be used only on multiple samples taken from similarly sized and similarly shaped plots. Samples from plots of different sizes can be dealt with by (repeated) subsampling of the plots such that all subsamples are similarly sized. I will provide an example below.

The genealogical approach (Etienne & Olff 2004b) assigns immigrating ancestors to all individuals in a local

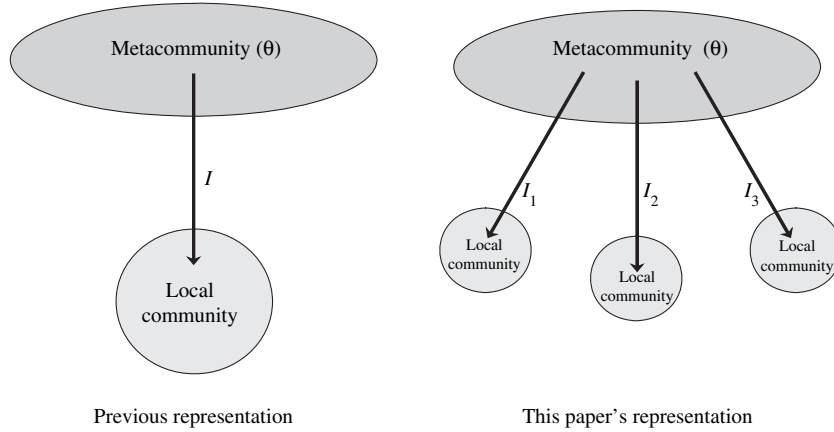


Figure 1 Schematic representation of the neutral model considered previously (left) and of the model considered in this article (right). On the left only one local sample is considered, or samples are lumped together into one big sample. On the left the spatial distinctiveness of the samples is taken into account. Immigrants come directly from the metacommunity; direct exchange between the local communities in the sample is assumed to be negligible. In both cases the metacommunity is completely described by the parameter θ , and immigration from the metacommunity into the local communities is described by the parameter I_i for sample i .

sample. The ancestors are a random sample from the metacommunity and the distribution of these ancestors over species are therefore described by the Ewens sampling formula with parameter θ . The distribution of the individuals over the ancestors is also described by a Ewens sampling distribution, but with parameter I_i (Etienne & Olf 2004b). Combining these Ewens sampling distributions (with parameter θ or parameter I_i) by summing over all the possible abundance vectors of the ancestors that are compatible with the local community data, one obtains the required probability $P[\vec{D}|\vec{J}, \theta, \vec{J}]$. Here I will only present this resulting sampling formula and I refer to Appendix S1 in the Supplementary Material for the complete derivation, because this is quite technical:

$$P_{bc}[\vec{D}|\vec{J}, \theta, \vec{J}] = F_{bc} \left(\prod_{i=1}^N \frac{J_i!}{(I_i)_{j_i} \prod_{k=1}^S n_{ik}!} \right) \sum_{\{a_{11}, \dots, a_{NS}\}} \left[\prod_{k=1}^S (a_k - 1)! \left(\prod_{i=1}^N \bar{s}(n_{ik}, a_{ik}) I_i^{A_i} \right) \right] \frac{\theta^S}{(\theta)_A} \quad (1)$$

I will elaborate on the indices b and c below. The a_{ik} represent the number of ancestors of species k in sample i . They are unknown, but they are integrated out by the summation in eqn (1) which is over all a_{ik} ($i = 1 \dots N$, $k = 1 \dots S$) with each a_{ik} taking all integer values between 1 and n_{ik} . For example, for two samples and two species, the summation $\sum_{a_{11}=1}^{n_{11}} \sum_{a_{12}=1}^{n_{12}} \sum_{a_{21}=1}^{n_{21}} \sum_{a_{22}=1}^{n_{22}}$ means $\sum_{\{a_{11}, a_{12}, a_{21}, a_{22}\}}$. The quantities A_i , a_k and A are functions of these a_{ik} : $A_i = \sum_k a_{ik}$ and $a_k = \sum_i a_{ik}$ and $A = \sum_i A_i = \sum_{i,k} a_{ik} = \sum_k a_k$. The quantity $(x)_y$ is the Pochhammer symbol defined as

$$(x)_y := \prod_{i=1}^y (x + i - 1) = \frac{\Gamma(x + y)}{\Gamma(x)} = \sum_{j=1}^y \bar{s}(y, j) x^j, \quad (2)$$

where $\Gamma(x)$ is the Gamma function and $\bar{s}(j, k)$ is the unsigned Stirling number of the first kind (Etienne 2005).

The indices b and c , which can both take the values 0 or 1, indicate whether or not it is necessary to uniquely identify species and samples, that is, whether or not one puts specific labels (names) on each individual with respect to its species name and its sample location (Johnson *et al.* 1997; Etienne & Alonso 2005). When species (or samples) are not labelled, this means that the order of species (or samples) in the vector \vec{D} is considered irrelevant. For example, if the nature of each species is considered irrelevant (which is the standard assumption in the literature on species abundance distributions), then a data set $\vec{D} = [(a, b), (c, d)]$ is treated as indistinguishable from $\vec{D} = [(b, a), (d, c)]$. But note that $\vec{D} = [(b, at), (c, d)]$ is different, that is, species identity is kept track of; all samples must still have the same (but arbitrary) order of species. Similarly, if the nature of each site is considered irrelevant, then $\vec{D} = [(a, b), (c, d)]$ is treated as indistinguishable from $\vec{D} = [(c, d), (a, b)]$. Evidently, there are four possibilities: (a) both species and samples are labelled, (b) samples are labelled, but species are not; (c) species are labelled, but samples are not; and (d) neither species, nor samples are labelled. The pre-factor F_{bc} in (1) deals with these possibilities (see Appendix S1), with the indices b and c indicating whether species are labelled ($b = 1$) or not ($b = 0$) and whether samples are labelled ($c = 1$) or not ($c = 0$):

$$F_{11} = \frac{1}{S!}, \quad (3a)$$

$$F_{01} = \frac{1}{\prod_j \Phi_j!}, \tag{3b}$$

$$F_{10} = \left(\frac{N!}{\prod_{\vec{k}} \Psi_{\vec{k}}!} \right) \frac{1}{S!}, \tag{3c}$$

$$F_{00} = \frac{1}{\prod_l \Omega_l!} \left(\frac{N!}{\prod_{\vec{k}} \Psi_{\vec{k}}!} \right) \frac{1}{\prod_j \Phi_j!}. \tag{3d}$$

Here, Φ_j is the number of species that have abundance vector \vec{j} across the samples (where samples are not labelled) and $\Psi_{\vec{k}}$ is the number of samples that have the same species-abundance distribution \vec{k} (where species are not labelled). The quantity Ω_l covers the rare occasions where permuting species and samples at the same time would give the exactly same distribution data set. See Appendix S1 for details. Note that the F_{bc} are only important when comparing different data sets \vec{D} (see Johnson *et al.* 1997 and Slatkin 1996), but they do not matter when estimating the parameters from a data set \vec{D} using likelihood-based methods such as maximum likelihood or Bayesian methods. Also note that $c = 1$ in the case that the I_i are different, because different I_i already implies that we distinguish between different samples. If the I_i are identical, it is most natural to assume $c = 0$ (see also below). The value of c only matters in practice if there are samples with exactly the same size J_i .

If all the I_i are different, eqn (1) is extremely difficult to evaluate in practice for realistic sample sizes and species numbers because of the enormous amount of summations involved (over all the a_{ik}). However, if dispersal limitation is highly correlated between communities (for example because they are similar samples from the metacommunity), then we may assume that the I_i are all the same, $I_i = I$. The number of parameters reduces to two (θ and I) and eqn (1) can be further simplified to

$$P_{bc}[\vec{D}|\vec{I}, \theta, \vec{J}] = F_{bc} \left[\prod_{i=1}^N \frac{J_i!}{(I_i)_{J_i} \prod_{k=1}^S n_{ik}!} \right] \sum_A M(\vec{D}, A) \frac{I^A \theta^S}{(\theta)_A}, \tag{4a}$$

where

$$M(\vec{D}, A) := \sum_{\{a_{11}, \dots, a_{NS} \mid \sum_{i,k} a_{ik} = A\}} \prod_{k=1}^S \left[(a_k - 1)! \prod_{i=1}^N \bar{s}(n_{ik}, a_{ik}) \right]. \tag{4b}$$

Equation (4) can certainly be evaluated within reasonable time. Note that (4) reduces to the formula given in Etienne (2005) for $N = 1$ (the value of c is then irrelevant):

$$P_b[\vec{D}|\vec{I}, \theta, \vec{J}] = F_b \left[\frac{J!}{(I)_J \prod_{k=1}^S n_k!} \right] \sum_A K(\vec{D}, A) \frac{I^A \theta^S}{(\theta)_A} \tag{5a}$$

with

$$K(\vec{D}, A) := \sum_{\{a_1, \dots, a_S \mid \sum_k a_k = A\}} \prod_{k=1}^S (a_k - 1)! \bar{s}(n_k, a_k) \tag{5b}$$

thus (4) is a generalization of the formula in Etienne (2005).

'EXACT' TEST OF NEUTRALITY

The sampling formula (4) can be used in likelihood-based parameter estimation techniques, such as maximum likelihood estimation (MLE) or Bayesian estimation (see Etienne 2005 and Etienne & Olf 2005 for examples for the one-sample sampling formula), and model comparison and goodness-of-fit tests. Below I will give an example of parameter estimation for a real data set, but first I will show how the sampling formula functions in a novel goodness-of-fit test of the neutral model. This test is based on the sequential construction scheme associated with the new sampling formula.

Distributions (sampling formulas) in the Ewens family correspond to sequential construction schemes, also called urn schemes, that generate samples from the distribution. Hubbell (2001), borrowing from Ewens (1972), presented the scheme that corresponds to the Ewens Sampling Formula, to generate a species-abundance distribution when there is no dispersal limitation. These urns are known as Hoppe urns in population genetics, after Hoppe (1984, 1987). Likewise, Etienne (2005) presented the sequential construction scheme for a single sample from a dispersal limited community. For applications see for example, Alonso & McKane (2004) and Etienne & Olf (2005). Here I present the sequential construction scheme when there are several samples from different localities. This scheme forms the core of the neutrality test to be discussed below, and a pseudo-code for it is provided in Appendix S2 in the Supplementary Material. It generates the species-ancestry-abundance distribution, given the model parameters θ and I_i and sample sizes J_i . Each individual in each sample is given an ancestry label and a species label according to some specific rules. When all individuals have been labelled, the species-ancestry-abundance distribution can easily be computed by counting all individuals with the same ancestry label and species label. If one ignores the ancestry labels (i.e. effectively sums over the ancestries) and only looks at the species labels, one obtains the species-abundance distribution. As with the sequential construction scheme for a single sample (Etienne 2005), the new sequential construction scheme for multiple samples has a simple but important application in the computation of the Simpson (1949) diversity (Appendix S3 in the SM).

Even more importantly, the sequential construction scheme, in combination with the sampling formula and parameter estimation, gives way to an exact test of neutrality

without the need for an alternative (niche-based) model. The sequential construction scheme in Etienne (2005) also allows such a test for a single sample in a similar fashion. The test is as follows. First, model parameters are estimated from the data using MLE. For this set of model parameters (θ, \vec{I}) and sample size vector \vec{J} one can generate any number (let us call this C) of artificial data sets \vec{D}_s ($s=1\dots C$) using the sequential construction scheme. The sampling formula (1) gives the probability $P_{0c}[\vec{D}_s|\vec{I},\theta,\vec{J}]$ of any data set \vec{D} under the neutral model (b must be 0, see Slatkin 1996). If one generates a large number of artificial data sets \vec{D}_s , then one can construct a frequency distribution of these probabilities $P_{0c}[\vec{D}_s|\vec{I},\theta,\vec{J}]$, and compare the value obtained for the real data set \vec{D} to the frequency distribution of the values for the artificial data sets \vec{D}_s . If the probability of the real data set is significantly smaller than the bulk of the artificial data sets, it is unlikely that neutral processes are responsible for the observed community structure, because most neutral communities have a larger probability. If the probability is similar to the values of the artificial data sets, then the observed species abundance distribution is consistent with neutrality, and neutrality cannot be rejected on these grounds. Instead of using parameters estimated from the data to generate the artificial data sets, it is preferable to use independent parameter estimates, but these are often not available.

The neutrality test proposed here is similar to the test proposed by Slatkin (1994, 1996 for the Ewens distribution in that it compares the probability (the likelihood) of the realized configuration with the probabilities of the artificial configurations, but it has a crucial difference in that it is not parameter-free. It is a mixture of a Monte Carlo significance test (Barnard 1963; Bartlett 1963) and the parametric bootstrap (Efron & Tibshirani 1993) which both make inferences based on artificially generated data sets. Testing for goodness-of-fit using percentiles of the distribution of a test statistic (here the likelihood) over the artificial data sets stems from Monte Carlo significance testing. The test is called 'exact' in the sense that the type I error can be precisely specified (Marriott 1979); the type II error can be minimized by increasing C (Marriott 1979). Using estimated parameters to generate these is inherited from the parametric bootstrap, but the parametric bootstrap proceeds by estimating the parameters for the artificial data sets to obtain an error estimate for the parameters of the real data set, whereas the method proposed here computes the likelihoods of the artificial data sets to obtain an estimate of goodness-of-fit of the model to the real data.

EXAMPLE

I applied the sampling formula and the neutrality test to a data set consisting of three Panamanian forest plots (Condit

et al. 2002), see Fig. 2. I chose this data set mainly because the data set is publicly available, allowing anyone to check or compare with my results. Nevertheless, the data set is ecologically also very interesting. The three plots, Sherman (5.96 ha of which 5 ha is in the data file), Barro Colorado Island (50 ha) and Cocoli (4 ha) lie along a precipitation gradient (3030, 2616 and 1950 mm year⁻¹ respectively, Condit *et al.* 2004) and, although only tens of kilometres apart, have relatively few species in common (Condit *et al.* 2004). This may be explained (Bunker & Carson 2005) by clear habitat affinities (Bazzaz 1998; Clark *et al.* 1998) or by extreme dispersal limitation (*sensu* Hubbell 2001) or both. The new sampling formula and neutrality test may shed new light on these alternatives.

Parameter estimation

I used maximum likelihood estimation (MLE) to obtain parameters for each of the plots separately and combinations of them (lumped and not lumped). The results are shown in Table 1. A first remarkable result is that the new multiple-sample sampling formula (4) does not seem to show multiple likelihood optima, in contrast to the one-sample formula (5) (Etienne *et al.* 2006). I have plotted the loglikelihood surface for further inspection (Fig. 3a) and indeed it shows only one peak. This result implies that adding spatial information by using several subsamples can rule out one of the maxima, and thus increases the information contained in species abundance data. However, a second remarkable result is that this single likelihood

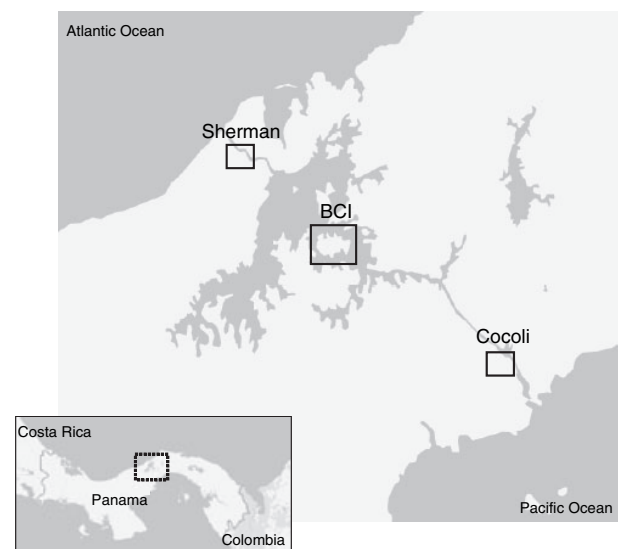


Figure 2 Map showing the location of the three Panamanian forest plots. They lie along a precipitation gradient from the Atlantic (wet) to the Pacific (dry).

Table 1 Maximum likelihood estimates of the model parameters and the corresponding maximum loglikelihoods for the three Panamanian forest plots separately and combined. If there are two local maxima in the loglikelihood surface (Etienne *et al.* 2006), the second maximum is also listed. Often this maximum is only marginally lower than the global maximum. The 10 reduced data sets result from combining Sherman and Cocoli with one of 10 different 5 ha subsamples of the BCI data set. The neutrality test is performed for all data sets consisting of (sub)samples from all three plots using the parameters estimated from the three samples (which yield the P -value p_{MLE}) or using the previously reported parameters for BCI only (which yield p_{BCI}).

Data set	Samples size and species richness		Maximum likelihood parameter estimation						Neutrality test	
	\vec{J}	\vec{S}	$\hat{\theta}_I$	\hat{I}_I	Loglik _I	$\hat{\theta}_{II}$	\hat{I}_{II}	Loglik _{II}	p_{MLE}	p_{BCI}
<i>Single samples</i>										
Sherman	2860	125	29.77	2558	-112.46	264.9	36.31	-112.77	Not performed	
BCI	21457	225	47.67	2211	-308.73	242.0	63.54	-312.55	Not performed	
Cocoli	1079	99	26.37	∞	-70.23	∞	26.37	-70.23	Not performed	
<i>Multiple samples lumped</i>										
Sherman + BCI	24317	273	59.04	2483	-332.36	287.7	79.30	-337.37	Not performed	
Sherman + Cocoli	3939	209	47.09	18109	-131.513	∞	46.94	-131.515	Not performed	
BCI + Cocoli	22 536	266	47.28	12651	-316.31	539.4	57.42	-316.17	Not performed	
Sherman + BCI + Cocoli	25 396	312	59.53	7499	-339.70	508.0	73.54	-340.79	Not performed	
<i>Multiple samples combined</i>										
Sherman + BCI	(2860, 21457)	(125, 225)	215.4	55.10	-756.89	-	-	-	Not performed	
Sherman + Cocoli	(2860, 1079)	(125, 99)	789.3	28.85	-236.81	-	-	-	Not performed	
BCI + Cocoli	(21 457, 1079)	(99, 225)	273.7	48.47	-607.02	-	-	-	Not performed	
Sherman + BCI + Cocoli	(2860, 21457, 1079)	(125, 225, 99)	259.3	44.24	-1091.8	-	-	-	0.065	< 0.001
Sherman + BCI ₁ + Cocoli	(2860, 2359, 1079)	(125, 152, 99)	270.5	39.18	-679.87	-	-	-	0.392	< 0.001
Sherman + BCI ₂ + Cocoli	(2860, 2151, 1079)	(125, 150, 99)	273.9	39.21	-668.84	-	-	-	0.283	0.002
Sherman + BCI ₃ + Cocoli	(2860, 2076, 1079)	(125, 162, 99)	280.0	41.18	-673.74	-	-	-	0.401	< 0.001
Sherman + BCI ₄ + Cocoli	(2860, 2027, 1079)	(125, 171, 99)	282.2	42.63	-680.40	-	-	-	0.132	0.004
Sherman + BCI ₅ + Cocoli	(2860, 2000, 1079)	(125, 166, 99)	290.8	41.71	-679.28	-	-	-	0.262	< 0.001
Sherman + BCI ₆ + Cocoli	(2860, 2050, 1079)	(125, 153, 99)	297.3	39.13	-654.40	-	-	-	0.379	< 0.001
Sherman + BCI ₇ + Cocoli	(2860, 2364, 1079)	(125, 147, 99)	298.6	37.27	-652.12	-	-	-	0.435	< 0.001
Sherman + BCI ₈ + Cocoli	(2860, 2225, 1079)	(125, 138, 99)	296.5	36.32	-640.46	-	-	-	0.082	< 0.001
Sherman + BCI ₉ + Cocoli	(2860, 2076, 1079)	(125, 145, 99)	300.4	37.65	-647.22	-	-	-	0.173	< 0.001
Sherman + BCI ₁₀ + Cocoli	(2860, 2129, 1079)	(125, 157, 99)	271.5	40.47	-688.08	-	-	-	0.255	< 0.001

maximum occurs for high θ and low I , indicating high regional diversity and strong dispersal limitation, whereas the one-sample likelihood maxima and the likelihood maximum for the lumped data all occur for low θ and high I , representing low regional diversity and little dispersal limitation. A third interesting result concerns the bias in parameter estimates. For the Ewens distribution it is known that the estimate for θ is biased (Johnson *et al.* 1997), although this bias decreases with sample size. For my previous sampling formula, there also seems to be considerable bias in MLE parameters (Jabot, F., Chave, J. & Etienne, R.S., unpublished data). In contrast, Fig. 3b shows the MLE parameter combinations for 1000 artificial data sets generated with the sequential construction scheme (see below) and these combinations spread nicely around the values with which they were generated, suggesting very little bias. Apparently, using multiple samples reduces the bias substantially. More generally, where the single samples from Sherman and Cocoli are, by themselves, too small to yield

reliable parameter estimates (particularly Cocoli, see Table 1), they form an informative data set when combined with (a subset of) BCI.

As noted above, when I is assumed the same for all local samples, the sampling formula for multiple samples should be used preferably on multiple samples taken from similarly sized and similarly shaped plots. The Sherman-BCI-Cocoli example clearly violates this condition. For this reason, I also computed parameter estimates for the combination of the Sherman and Cocoli plots with a 5 ha subsample of BCI, and averaged over all 10 combinations (one for each of 10 subsamples from BCI). The estimates of θ are all higher and the estimates of I are all lower than those for the full data set (Table 1).

Neutrality test

I also applied the 'exact' neutrality test to the Panamanian tree data set described above. I chose two sets of

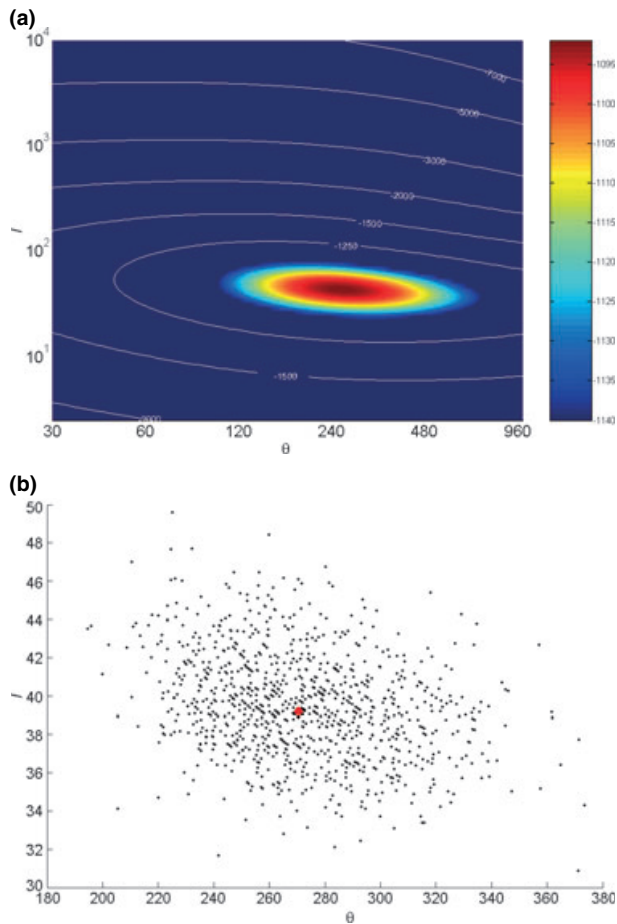


Figure 3 Characteristics of the sampling formula related to parameter estimation. (a) Loglikelihood surface for the three plots combined (full data set). Values below -1140 are indicated with contour lines, values above -1140 are indicated with colours. (b) Parameter estimates of each of 1000 simulated data sets generated with $\theta = 270.5$ and $I = 39.18$, the MLE parameters for subset 1 (see Table 1), and sample sizes \bar{j} equal to those of subset 1. The cloud of points provides insight in the error in the MLE parameter values of subset 1. The nice spread around the values with which the data were simulated (indicated in red) implies that the bias in the MLE parameter values is small.

parameters to generate the artificial data with: first I chose the MLE parameters for all the samples together (but not lumped), and second, I chose the parameters estimated previously (and reported again in Table 1) for BCI only. For each parameter set I generated 1000 artificial data sets, each consisting of three subsamples with sample sizes set equal to those for the real data. Table 1 lists the P -value of the test, i.e. the fraction of the artificial data sets that have a lower loglikelihood than the real data set. Although the P -values are not significant for the data sets generated with the MLE parameters optimized for the full data set, they are very significant for the data sets generated with the

parameter values optimized for BCI only. Thus, in the former case, the model cannot be rejected, whereas in the latter case, the model is strongly rejected. This means that if the neutral model is believed to apply to the scale of BCI, it cannot apply (with the same parameters) to the scale of the Panama Canal Zone. However, if, from the outset, the neutral model is hypothesized to apply to the scale of the Panama Canal Zone, the data used cannot reject this null hypothesis.

For further analysis, I recorded a second test statistic, the Sørensen distance or Bray–Curtis dissimilarity index d_{ij} (Sørensen 1948; Bray & Curtis 1957), defined as

$$d_{ij} = \frac{\sum_{k=1}^S |n_{ik} - n_{jk}|}{\sum_{k=1}^S n_{ik} + n_{jk}}. \quad (6)$$

The results are as follows. The Sørensen distance for the artificial data is always lower than the observed when the artificial data sets are generated with the ML estimates for BCI only, no matter what pair of samples one is looking at. This may not be surprising, because the model was rejected for these parameter values. However, even for the MLE parameters for all three samples a consistent pattern arises regardless of what subsample is taken from BCI: Cocoli and Sherman are significantly less similar than expected under neutrality, but BCI and Sherman are substantially (consistently but not always significantly) more similar than expected under the neutral model. This pattern may be expected given the distances (both Euclidean and in terms of precipitation) between the plots, but it is surprising that it is significant even though neutrality cannot be rejected on the basis of the loglikelihood in this case.

DISCUSSION

My previous sampling formula (Etienne 2005) was a generalization of the Ewens sampling formula (Ewens 1972; Hubbell 2001) for a single, local sample for dispersal-limited communities. In this article, I have provided a generalization of this previous sampling formula for multiple local samples from the same metacommunity. Existing data sets often contain multiple samples that are by themselves too small to contain sufficient information for reliable parameter estimates or model comparisons, and the sampling formula of this article allows the information of all these samples to be used simultaneously, thereby greatly increasing the number of data sets the neutral theory can be tested on.

Above I have already identified a possible limitation of the new sampling formula: evaluation is extremely difficult in practice except when the I_i are assumed equal. This assumption in turn requires that plot sizes must be similar.

The latter restriction can be overcome by taking subsamples such that plot sizes are equalized. There is no loss of data if this is repeated for all different combinations of subsamples. Still, the equality of the I_i may be considered a limitation. Indeed, situations where different samples are thought to have different degrees of dispersal limitation cannot be dealt with. However, it may be doubted whether such situations are conceptually consistent with the current model because the model assumes a homogeneous metacommunity. Furthermore, for parameter estimation it is often undesirable to allow the I_i to differ, because the model would quickly be overparametrized and because potentially there are a multitude of local likelihood maxima in a high dimensional parameter space. In the example, the number of parameters would be still relatively low (4), but in many practical cases the number of plots and hence the number of parameters would be much higher. In fact, the sampling formula for multiple samples was designed to estimate θ and I for data sets with many (e.g. 50) small plots each of which would by itself be too small to allow for reliable parameter estimation, but together constitute a substantial data set. Thus, the sampling formula can be seen as a better alternative to naively lumping data: the additional information from many other small plots should be used to estimate θ and I more accurately, a benefit that would be lost if all I_i were allowed to differ. Having said this, I invite the mathematically and programming oriented reader to find an algorithm that can evaluate the sampling formula within reasonable time. It may be reassuring to note that for the previous, single-sample, formula (Etienne 2005), vastly faster algorithms were developed than I originally envisaged (e.g. Tetame by Chave & Jabot, <http://www.edb.ups-tlse.fr/equipe1/chave/tetame.htm>). For an impression of the possible values of the I_i one can use the two-stage approach of Munoz *et al.* 2007, but it should be noted that this approach does not provide simultaneous estimation of θ and I_i (or I) and can therefore point to an incorrect global likelihood maximum.

The previous paragraph touches upon a question of more general importance: when is the new sampling formula for multiple samples applicable? I already pointed out that plot sizes should be similar (from the outset or after subsampling). Because each sample is considered homogeneous, the length scale of each plot should not exceed the typical dispersal distance. I also noted that the samples should be sufficiently far apart for immigration from one sample to the other to be negligible (see also Fig. 1), that is, the plots should be separated by distances longer than the typical dispersal distance. At the same time, the samples should be close enough to belong to the same metacommunity, described by the Ewens distribution. A substantial environmental gradient that is known to strongly influence community composition should be avoided. Hence, all

plots should be separated by distances less than a characteristic distance of environmental change (the Panamanian example actually violates this condition). As I argue below, the new sampling formula is most suitable for samples from different islands that receive immigrants from a mainland without dispersal limitation.

In this article, I have provided a neutrality test, one that can be used without the need of sampling formulas of alternative models. Previously, the performance of neutral models in fitting species abundance distributions has been compared with the performance of alternative models, particularly the lognormal model (McGill 2003; Volkov *et al.* 2003; Etienne & Olff 2005). Such comparisons are hard to interpret because these alternative models – that are usually meant to describe niche differentiation – are often very different in nature, being a statistical description rather than a mechanistic one, or being static rather than dynamic, or they contain some form of neutrality by being symmetric rather than asymmetric. The lognormal model also suffers from this. It can be argued to be mechanistic, as it is a sequential breakage model where breakage occurs independently of fragment size (Bulmer 1974; Sugihara 1980), but sequential breakage models are static (and can be interpreted as symmetric). Another argument often invoked for a mechanistic underpinning of the lognormal is that the abundance of a species is governed by many more-or-less independent factors that interact multiplicatively rather than additively (May 1975) and that the Central Limit Theorem then ensures a lognormal distribution. However, it can be argued that this still requires species to be symmetric: the independent factors must affect abundance for each species in a similar way. Such hard-to-interpret comparisons are no longer necessary to test the neutral theory, as the test presented here is an internal test where goodness-of-fit is determined relative to the goodness-of-fit of data that are generated by the model itself. Nevertheless, progress should be still made in the development of dynamical, mechanistic niche-based models, because rejection of neutrality does not give us a clue about what type of niche differentiation is responsible for the rejection.

It cannot be emphasized enough that failure to reject neutrality does not imply acceptance: pattern does not equal process (Cohen 1968). This is true for any statistical test of any (null) model – not just neutral models. In the context of neutral theory it has indeed been explicitly shown that non-neutral processes can generate patterns consistent with neutrality (Purves & Pacala 2005; Walker 2007) and such patterns would fail to reject neutral theory. Nevertheless, successful rejection gives us information on the presence of deviations from the null hypothesis. Whether the neutral model is a true null model is still debated (Gotelli & McGill 2006), but it is now increasingly accepted as a parsimonious description of community structure containing factors that

are always present, such as sampling effects, dispersal (which can also be seen as a sampling effect, see Etienne & Alonso 2005) and stochasticity. Neutral theory is therefore useful in checking whether other factors play a major role as well.

For illustrative purposes, I have applied the new sampling formula to a data set consisting of three local samples from a neotropical forest metacommunity, one of which (BCI) has served frequently as an example data set in articles on neutral theory. The other two (Sherman and Cocoli) are, by themselves, too small to yield reliable parameter estimates (particularly Cocoli, see Table 1), but form an informative data set when combined with BCI. The parameter estimates obtained for the three plots combined are very different from those obtained for each of them alone, and especially from those for BCI which have been so often reported in the literature and which were believed to be reasonable. This by itself already suggests that neutrality and/or dispersal limitation is not a sufficient explanation for the structure of these tree communities at the metacommunity scale; niche differentiation (the obvious adaptation to the precipitation gradient) plays a major role at this scale. This is confirmed by the neutrality test using the parameter estimates obtained from BCI alone: the observed data occur in the tail of the neutral loglikelihood distribution meaning that it is highly unlikely that they are generated by purely neutral (dispersal-limited) processes. However, the neutrality test based on the parameter estimates from all three plots does not provide a significant confirmation (Table 1). Ideally, one should use truly independent parameter values to perform the neutrality test, and then the test is completely impartial. By using parameter values estimated from the data themselves – which is often the only available option – one introduces some circularity. This can be most easily understood in the case of the Ewens distribution (no dispersal limitation). Because then the observed number of species S is a sufficient statistic for θ , any data set generated with this θ will have an expected number of species of S and are therefore bound to be similar in this respect to the real data (this is why Slatkin conditioned his test on S), reducing the power of the neutrality test to reject it. Hence, the neutrality test can be considered conservative in this case. That is, it will sometimes yield false negatives but seldom false positives: if neutrality is not rejected, this does not necessarily mean that the observed pattern is generated by neutral processes, but if neutrality is rejected, then it is with much confidence.

Strictly speaking, as I already mentioned in the introduction, the test proposed here can only reject the particular model formulation under consideration (and the corresponding parameter estimates are then rendered meaningless), not neutrality itself. Hence, although it is tempting to discard neutral theory on the basis of the results of this

article, one must consider other characteristics of the model that may be responsible for rejection. I have considered the model's point mutation assumption and its zero-sum assumption elsewhere (Etienne *et al.* 2007; R.S. Etienne, D. Alonso, A.J. McKane, unpublished data). The consequences of the spatially implicit character of the model, however, have not been thoroughly explored. The sampling formula introduced in this article is probably the most complete sampling formula that is possible within the spatially implicit model framework using two spatial and temporal scales in agreement with the principle of ecological hierarchy (Allen & Starr 1982), and thus enables, for the first time, an exploration of these consequences. The analysis of the similarity index exposes the fact that dispersal limitation in a continuous landscape is a spatial phenomenon that a spatially implicit model has severe problems dealing with. Indeed, the model seems inconsistent in that it assumes no dispersal limitation in the metacommunity, whereas the local communities embedded in it are dispersal-limited. Perhaps, this can be remedied by assuming the dispersal-limited sampling formula (Etienne 2005) rather than the Ewens sampling formula (Ewens 1972) for the metacommunity. The mechanistic basis for this needs is not immediately obvious, however, and at any rate the resulting sampling formula would become very complicated and have an additional dispersal limitation parameter for the metacommunity scale which should be related to the parameter at the local scale in some unknown way. Alternatively, if there are independent data on metacommunity abundances, one can refrain from using a model distribution for the metacommunity altogether, but this still ignores spatial structure in a continuous landscape. Thus, the model (and these two variations) seems most appropriate for a mainland-island system where inter-island dispersal and local speciation can be neglected and the mainland's species abundance distribution is well described.

What is needed in a continuous landscape is a spatially explicit model that is similar in all other aspects. Although some work on spatially explicit neutral models has been done (Durrett & Levin 1996; Chave & Leigh 2002; Rosindell & Cornell 2007), this field is still underdeveloped. Particularly, it will be extremely demanding mathematically if not impossible to construct spatially explicit sampling formulas. The benefit of incorporating spatially explicit information will probably be offset by the loss of the ability to use the full abundance vector (\vec{D}) rather than summary statistics (diversity indices). This loss will be limited if informative summary statistics can be found. The fact that the most obvious summary statistic species richness (S) is a sufficient statistic for θ in the Ewens sampling formula (Johnson *et al.* 1997) gives some confidence that such a quest may be quite successful. Nevertheless, the complexity of the spatially explicit models will certainly call for an investigation to what

degree a spatially implicit model can be used as an approximation of spatially explicit models (see e.g. Vallade & Houchmandzadeh 2006). I expect that the sampling formula presented here will provide a valuable tool in this investigation.

ACKNOWLEDGEMENTS

I thank Emile Apol, Franck Jabot, Brian McGill, François Munoz, Han Olff, James Rosindell, Kevin Gross and two anonymous referees for stimulating discussions and/or helpful comments.

REFERENCES

- Adler, P.B., HilleRisLambers, J. & Levine, J.M. (2007). A niche for neutrality. *Ecol. Lett.*, 10, 95–104.
- Allen, T.F.H. & Starr, T.B. (1982). *Hierarchy: Perspectives for Ecological Complexity*. University of Chicago Press, Chicago, IL.
- Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.
- Alonso, D., Etienne, R.S. & McKane, A.J. (2006). The merits of neutral theory. *Trends Ecol. Evol.*, 21, 451–457.
- Barnard, G.A. (1963). Discussion of 'the spectral analysis of point processes'. *J. R. Stat. Soc. B*, 25, 294.
- Bartlett, M.S. (1963). The spectral analysis of point processes. *J. R. Stat. Soc. B*, 25, 264–296.
- Bazzaz, F.A. (1998). Tropical forests in a future climate: changes in biological diversity and impact on the global carbon cycle. *Clim. Change*, 39, 317–336.
- Bray, J.R. & Curtis, J.T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, 27, 325–349.
- Bulmer, M.G. (1974). On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, 30, 101–110.
- Bunker, D.E. & Carson, W.P. (2005). Drought stress and tropical forest woody seedlings: effect on community structure and composition. *J. Ecol.*, 93, 794–806.
- Chave, J. & Leigh, E.G. (2002). A spatially explicit neutral model of β -diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153–168.
- Clark, D.B., Clark, D.A. & Read, J.M. (1998). Edaphic variation and the mesoscale distribution of tree species in a neotropical rain forest. *J. Ecol.*, 86, 101–112.
- Cohen, J.E. (1968). Alternate derivations of a species-abundance relation. *Am. Nat.*, 102, 165–172.
- Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B. *et al.* (2002). Beta-diversity in tropical forest trees. *Science*, 295, 666–669.
- Condit, R., Aguilar, S., Hernandez, A., Perez, R., Lao, S., Angehr, G. *et al.* (2004). Tropical forest dynamics across a rainfall gradient and the impact of an el niño dry season. *J. Trop. Ecol.*, 20, 51–72.
- Durrett, R. & Levin, S. (1996). Spatial models for species-area curves. *J. Theor. Biol.*, 179, 119–127.
- Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* 57. Chapman & Hall, New York, NY.
- Etienne, R.S. (2005). A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8, 253–260.
- Etienne, R.S. & Alonso, D. (2005). A dispersal-limited sampling theory for species and alleles. *Ecol. Lett.*, 8, 1147–1156.
- Etienne, R.S. & Olff, H. (2004a). How dispersal limitation shapes species – body size distributions in local communities. *Am. Nat.*, 163, 69–83.
- Etienne, R.S. & Olff, H. (2004b). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.*, 7, 170–175.
- Etienne, R.S. & 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., Latimer, A.M., Silander, J.A. & Cowling, R.M. (2006). Comment on 'neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot'. *Science*, 311, 610b.
- Etienne, R.S., Apol, M.E.F., Olff, H. & Weissing, F.J. (2007). Modes of speciation and the neutral theory of biodiversity. *Oikos*, 116, 241–258.
- Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3, 87–112.
- Gotelli, N.J. & McGill, B.J. (2006). Null versus neutral models: what's the difference? *Ecography*, 29, 793–800.
- Holt, R.D. (2006). Emergent neutrality. *Trends Ecol. Evol.*, 21, 531–533.
- Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, 20, 91–94.
- Hoppe, F. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.*, 25, 123–159.
- Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.
- Johnson, N.L., Kotz, S. & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York, NY.
- Marriott, F.H.C. (1979). Barnard's monte carlo tests: how many simulations? *Appl. Stat.*, 28, 75–77.
- May, R.M. (1975). Patterns of species abundance and diversity. In: *Ecology and Evolution* (eds Cody, M. L. & Diamond, J.M.). Harvard University Press, Cambridge, MA, pp. 81–120.
- McGill, B.J. (2003). Strong and weak tests of macroecological theory. *Oikos*, 102, 679–685.
- McGill, B.J., Maurer, B.A. & Weiser, M.D. (2006). Empirical evaluation of neutral theory. *Ecology*, 87, 1411–1426.
- Munoz, F., Couteron, P., Ramesh, B.R. & Etienne, R.S. (2007). Inferring parameters of neutral communities: from one single large to several small samples. *Ecology* (in press).
- Purves, D.W. & Pacala, S.W. (2005). Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. In: *Biotic Interactions in the Tropics* (eds Burslem, D., Pinard, M. & Hartley, S.). Cambridge University Press, Cambridge, UK, pp. 107–138.
- Rosindell, J. & Cornell, S.J. (2007). Species-area relationships from a spatially explicit neutral model in an infinite landscape. *Ecol. L.*, 10, 586–595.
- Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Det Kongelige Danske Videnskaberne Selskab, Biologiske Skrifter*, 5, 1–34.
- Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163, 688.
- Slatkin, M. (1994). An exact test for neutrality based on the ewens sampling distribution. *Genet. Res.*, 64, 71–74.

- Slatkin, M. (1996). A correction to the exact test based on the ewens sampling distribution. *Genet. Res.*, 68, 259–260.
- Sugihara, G. (1980). Minimal community structure: an explanation of species-abundance patterns. *Am. Nat.*, 116, 770–787.
- Vallade, M. & Houchmandzadeh, B. (2006). Species abundance distribution and population dynamics in a two-community model of neutral ecology. *Phys. Rev. E.*, 74, 051914.
- Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.
- Wakeley, J. (1999). Non-equilibrium migration in human history. *Genetics*, 153, 1863–1871.
- Walker, S.C. (2007). When and why do non-neutral metacommunities appear neutral? *Theor. Popul. Biol.*, 71, 318–331.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article

Appendix.pdf A pdf-file consisting of three appendices:

- **Appendix S1** Derivation of the sampling formula for multiple samples.
- **Appendix S2** Sequential construction scheme for generating species-(ancestry-)abundance distributions.
- **Appendix S3** Simpson's diversity.

urn2.gp. The source code of the sequential construction scheme.

ML.zip. A zipfile containing a file with source code to compute the maximum likelihood parameter estimates for a given data set and loglikelihood values for plotting a loglikelihood surface.

The material is available as a part of the online article from: <http://www.blackwell-synergy.com/doi/full/10.1111/j.1461-0248.2007.01052.x>

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Editor, Kevin Gross

Manuscript received 11 January 2007

First decision made 14 February 2007

Second decision made 28 March 2007

Manuscript accepted 11 April 2007