A dispersal-limited sampling theory for species and alleles

Etienne, RS; Alonso, D

Published in:
Ecology Letters

DOI:
10.1111/j.1461-0248.2005.00817.x

Publication date:
2005

Ecology Letters, (2005) **8**: 1147–1156 doi: 10.1111/j.1461-0248.2005.00817.x

LETTER

A dispersal-limited sampling theory for species and alleles

Rampal S. Etienne[1]* and David Alonso[2]
[1]Community and Conservation Ecology Group, University of Groningen, PO Box 14, 9750 AA Haren, The Netherlands
[2]Ecology and Evolutionary Biology, University of Michigan, 830 North University Av, Ann Arbor, MI 48109-1048, USA
**Correspondence: E-mail: r.s.etienne@rug.nl*

Abstract

The importance of dispersal for biodiversity has long been recognized. However, it was never advertised as vigorously as Stephen Hubbell did in the context of his neutral community theory. After his book appeared in 2001, several scientists have sought and found analytical expressions for the effect of dispersal limitation on community composition, still in the neutral context. This has been done along two relatively independent lines of research that have a different mathematical approach and focus on different, yet related, types of results. Here, we study both types in a new framework that makes use of the sampling nature of the theory. We present sampling distributions that contain binomial or hypergeometric sampling on the one hand, and dispersal limitation on the other, and thus views dispersal limitation as ubiquitous as sampling effects. Further, we express the results of one line of research in terms of the other and vice versa, using the concept of subsamples. A consequence of our findings is that metacommunity size does not independently affect the outcome of neutral models in contrast to a previous assertion (*Ecol. Lett.*, 7, 2004, p. 904) based on an incorrect formula (*Phys. Rev. E*, 68, 2003, p. 061902, eqns 11–14). Our framework provides the basis for development of a dispersal-limited non-neutral community theory and applies in population genetics as well, where alleles and mutation play the roles of species and speciation respectively.

Keywords

Binomial sampling, biodiversity, community, dispersal-limited sampling, Ewens sampling formula, hypergeometric sampling, neutral model, random sampling.

Ecology Letters (2005) 8: 1147–1156

INTRODUCTION

The importance of dispersal in ecology has long been recognized (e.g. Grinnell 1922; MacArhur & Wilson 1967; Levins & Culver 1971; Brown & Kodric-Brown 1977; Hanski 1983; Tilman 1994; Loreau & Mouquet 1999). Yet, seldom has a more vigorous (quantitative) case been made than by Hubbell (1997, 2001) who presented a comprehensible suite of stochastic neutral models of community structure based on the fundamental processes of speciation, extinction and dispersal. In the most often cited model of these, the local community consists of J individuals of different species whose offspring compete for sites that are left open after an individual dies. They do not only compete with one another, but they also compete with immigrants from outside the local community: there is a probability m that an open site is colonized by an immigrant. If $m < 1$ the local community is called dispersal-limited. With probability $1 - m$, the open site is colonized by offspring of a local individual. Each individual in the local community, regardless of species, has an equal chance of colonizing the open site (the neutrality assumption). Each open site is immediately recolonized so community size remains constant (the zero-sum assumption). The immigrants come from a regional species pool (the metacommunity; Hubbell 2001) that is in a stochastic balance between speciation and extinction. This balance is characterized by the parameter θ, a composite of the speciation rate v and metacommunity size J_M. Speciation in this model occurs by 'point mutation' [in other models Hubbell (2001) uses 'random fission' speciation which is a first step towards modelling allopatric speciation]. This model resembles the continent-island infinite alleles model with Moran (1962)-like reproduction in population genetics (Wright 1931; Moran 1962; Ewens 1972); the difference with Moran (1962) reproduction is that the individual that dies does not produce any offspring that

could replace it. We note that the terminology 'continent-island' is only historical; the theory also applies to a local sample from a continuous landscape.

Hubbell's (2001) model has been heavily criticized, mostly because of its neutrality assumption. But even if this assumption turns out to be untenable, we should not reject the theory completely, as this would be throwing out the baby with the bath water. It is now realized that the neutral model is the appropriate null model with which other models containing more processes should be compared. Hubbell (2001) thus effectively introduced Ockham's razor to community ecology, i.e. the maxim that science should aim at finding the minimal set of processes that can satisfactorily explain observed phenomena. However, less attention has been given to the fact that Hubbell (2001) put dispersal at the top of this minimal set. In the present study, we argue that dispersal is just as ubiquitous as sampling effects and can even be framed in the same mathematical setting.

While Hubbell (2001) presented analytical results for his model without dispersal limitation ($m = 1$) because these were already known in population genetics (Ewens 1972; Karlin & McGregor 1972), he provided only simulation results for the biologically more interesting case with dispersal limitation ($m < 1$). This made it difficult to test accurately whether the neutral model can explain observed diversity patterns, such as the species-abundance distribution, better or worse than other community models (McGill 2003). Recently, however, analytical results for the case $m < 1$ have been found, along two distinct lines of research. These lines of research study the problem from the two perspectives that result from the duality of the theory (Etienne & Olff 2004b) with respect to time: forwards- and backwards-in-time.

The forwards-in-time perspective uses a master equation approach with a Markovian description of states and transitions (McKane *et al.* 2000, 2004; Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Alonso & McKane 2004). This has resulted in exact analytical expressions and various approximations for the 'expected number of species with a certain abundance' in a sample of J individuals from a dispersal-limited local community: if n is the abundance, then $E[S_n|\theta, m, J]$ denotes the expected number of species with this abundance in this sample. Vallade & Houchmandzadeh (2003) and subsequent studies used the shorthand notation of $\langle\phi_n\rangle$ or $S(n)$ for this expectation, but we employ the longer notation to emphasize that this is an expectation that follows from the model in contrast to the actually observed number of species with abundance n, which we will denote by Φ_n as in Etienne (2005). The expected number of species with a certain abundance is the classical approach to study commonness and rarity in community ecology and also a very useful tool in exploring the behaviour of community models. However, it cannot be used to obtain accurate estimates of the model parameters.

The backwards-in-time perspective takes a genealogical, coalescent-type approach where community members are traced back to the ancestors that once immigrated into the community (Etienne & Olff 2004a,b; Etienne 2005). This line has resulted in an analytical expression for the 'joint multivariate probability of observing S species with abundances' $n_1, n_2,\ldots,n_S$ in a sample of J individuals from the local community. Let us denote this collection by $\vec{D}$, i.e. $\vec{D} = (n_1, n_2, \ldots, n_S)$. The joint multivariate probability is thus the likelihood $P[\vec{D}|\theta, m, J]$, which can be used in maximum likelihood estimation of model parameters from species-abundance data (Etienne 2005) or other methods based on the likelihood (Etienne & Olff 2005), but is less useful for studying the behaviour of the model.

Because both lines of research work on the same model and have provided exact analytical results, they must somehow be related, but until now the common framework has not been made explicit. In the present study, after presenting the basic results of the two lines of research, we build such a framework. Its most important property is the sampling nature of the theory and the role that dispersal plays in it. We introduce new distributions, called the dispersal-limited binomial and dispersal-limited hypergeometric distributions by which the results of both lines of research arise naturally. As a result we find that the expression for $E[S_n|\theta, m, J]$ for finite metacommunity size, as reported by Vallade & Houchmandzadeh (2003) is incorrect. An important consequence is that it is not possible to estimate metacommunity size and hence the speciation rate from species-abundance data, as was suggested based on this formula (Alonso & McKane 2004, p. 904). Next, we link the two lines of research by expressing results of one line of research in terms of the other and vice versa, by making use of the concept of subsamples. Most of our results are summarized in Table 1. We end with a discussion of our results that tries to open new doors to further development of neutral as well as non-neutral theories in community ecology and population genetics.

RESULTS OF THE TWO LINES OF RESEARCH

No dispersal limitation

Without dispersal limitation ($m = 1$), $E[S_n|\theta, J]$ is given by (Moran 1958, Watterson 1974 and Vallade & Houchmandzadeh 2003):

$$E[S_n|\theta,J] = \frac{\theta}{n}\frac{\Gamma(J+1)}{\Gamma(J+1-n)}\frac{\Gamma(J+\theta-n)}{\Gamma(J+\theta)} \qquad (1)$$

The multivariate probability distribution is given by the Ewens sampling formula (Ewens 1972)

Table 1 Overview of the analytical results for the species-abundance distribution of a local sample in neutral community theory

Quantity	$J_M \to \infty$	$J_M < \infty$
$m = 1$		
$E[S_n \mid \theta, J]$	$\int_0^1 P_{bin}[n \mid x, J]\Omega(x)dx$	$= \sum_{j=1}^{J_M} P_{hyp}[n \mid j, J_M, J]E[S_j \mid \theta, J_M]$
$P[\vec{D} \mid \theta, J]$	$\frac{\prod_{i=1}^{S}\int_0^1 P_{bin}[n_i \mid x, J_i]\Omega(x)dx}{\prod_{j=1}^{J}\Phi_j!}$	$= \frac{\prod_{i=1}^{S}\sum_{j=1}^{J_M} P_{hyp}[n_i \mid j, J_M, J]E[S_j \mid \theta, J_M]}{\prod_{j=1}^{J}\Phi_j!}$
$m < 1$		
$E[S_n \mid \theta, m, J_M, J]$	$\int_0^1 P_{bin}^{DL}[n \mid m, x, J]\Omega(x)dx$	$= \sum_{j=1}^{J_M} P_{hyp}^{DL}[n \mid m, j, J_M, J]E[S_j \mid \theta, J_M]$
$P[\vec{D} \mid \theta, m, J_M, J]$	$\frac{\prod_{i=1}^{S}\int_0^1 P_{bin}^{DL}[n_i \mid m, x, J_i]\widehat{\Omega}[x \mid \theta, m, \vec{D}_{i+1}]dx}{\prod_{j=1}^{J}\Phi_j!}$	$= \frac{\prod_{i=1}^{S}\int_0^1 P_{bin}^{DL}[n_i \mid m, x, J_i]\widehat{\Omega}[x \mid \theta, m, \vec{D}_{i+1}]dx}{\prod_{j=1}^{J}\Phi_j!}$

Let the entire metacommunity consist of J_M individuals and let the sample consist of J individuals of S different species with abundances $n_1, n_2, \ldots, n_S$. Let us denote this sample by $\vec{D}$, i.e. $\vec{D} = (n_1, n_2, \ldots, n_S)$; Φ_j is the number of species in the sample that have abundance j. The model parameters are the fundamental biodiversity number θ, which is a measure of the regional diversity, and the fundamental dispersal number I. The immigration probability m is a function of I, see eqn 8, $m = \frac{I}{I+J-1}$. The quantities $E[S_n \mid \theta, J]$ and $E[S_n \mid \theta, m, J_M, J]$ represent the expected number of species with abundance n in the cases without dispersal limitation ($I = \infty$, i.e. $m = 1$) and with dispersal limitation ($I < \infty$, i.e. $m < 1$) respectively, according to the neutral model. $\Omega(x)dx$, where $\Omega(x)$ is given by eqn 21, is the number of species with relative abundance between x and $x + dx$ in the metacommunity (regional species pool); $\widehat{\Omega}[x \mid \theta, m, \vec{D}_{i+1}]dx$ is a modified version of that, see eqn 39. The probabilities $P[\vec{D} \mid \theta, J]$ and $P[\vec{D} \mid \theta, m, J]$ represent the joint multivariate probability of observing S species with abundances $n_1, n_2, \ldots, n_S$ in a sample of J individuals, again for the cases without and with dispersal limitation respectively. $P_{bin}[n \mid x, J]$, $P_{hyp}[n \mid j, J_M, J]$, $P_{bin}^{DL}[n \mid m, x, J]$ and $P_{hyp}^{DL}[n \mid m, j, J_M, J]$ are the binomial, hypergeometric, dispersal-limited binomial and dispersal-limited hypergeometric distributions respectively, given in eqns 15, 20, 24 and 28. These four distributions are the distributions by which the expressions for the regional species-abundance distribution must be weighed to obtain the expressions for the local sample. The binomial distribution $P_{bin}[n \mid x, J]$ and the hypergeometric distribution $P_{hyp}[n \mid j, J_M, J]$ are the limits of the dispersal-limited hypergeometric distribution $P_{hyp}^{DL}[n \mid m, j, J_M, J]$ for $m \to 1$ in the cases $J_M \to \infty$ and $J_M < \infty$ respectively.

$$P[\vec{D} \mid \theta, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \Phi_j!}\frac{\theta^S}{(\theta)_J} \tag{2}$$

where Φ_j is the observed number of species with abundance j, as we noted above, and $(\theta)_J$ is the Pochhammer symbol defined as

$$(\theta)_J = \prod_{i=1}^{J}(\theta + i - 1) = \frac{\Gamma(\theta + J)}{\Gamma(\theta)} = \sum_{j=1}^{J} \bar{s}(J, j)\theta^j \tag{3}$$

where $\Gamma(x)$ is the gamma function and $\bar{s}(j, k)$ is the so-called unsigned Stirling number of the first kind. We will frequently use the last two equalities in our formulas below. We also note that $\bar{s}(j, 1) = \Gamma(j) = (j - 1)!$. Below we will also frequently use the definition of the beta function:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} = \int_0^1 x^{a-1}(1 - x)^{b-1}dx \tag{4}$$

In Pochhammer notation, eqn 1 becomes even more compact:

$$E[S_n \mid \theta, J] = \frac{\theta}{n}\frac{(J + 1 - n)_n}{(J + \theta - n)_n} \tag{5}$$

Note that J_M does not enter eqns 1 and 2, except by its role in θ. Below, we make this more explicit.

Dispersal limitation

With dispersal limitation ($m < 1$) and metacommunity size J_M tending to infinity, $E[S_n \mid \theta, m, J]$ is given by Vallade & Houchmandzadeh (2003) and Alonso & McKane (2004):

$$E[S_n \mid \theta, m, J] = \frac{\theta}{(I)_J}\binom{J}{n}\int_0^1 (Ix)_n[I(1 - x)]_{J-n}\frac{(1 - x)^{\theta-1}}{x}dx \tag{6}$$

where, we used notation of Etienne (2005) for later comparison. Here, $\binom{J}{n}$ is the usual binomial coefficient,

$$\binom{J}{n} = \frac{J!}{n!(J - n)!} \tag{7}$$

and I is a transformed immigration parameter,

$$I = \frac{m}{1 - m}(J - 1) \tag{8}$$

The parameter I is called μ in Vallade & Houchmandzadeh (2003) and γ in Alonso & McKane (2004), while Ix is called λ in Volkov *et al.* (2003). I is related to the immigration probability m and local community size J as the fundamental biodiversity number θ is related to the speciation probability v and metacommunity size J_M (Vallade & Houchmandzadeh 2003; Alonso & McKane 2004; Etienne 2005),

$$\theta = \frac{v}{1-v}(J_M - 1) \tag{9}$$

In analogy to θ, we will call I the 'fundamental dispersal number'.

Vallade & Houchmandzadeh (2003) derived a different expression for $E[S_n|\theta, m, J_M, J]$ for finite metacommunity J_M:

$$* \; E(S_n|\theta, m, J_M, J] = \binom{J}{n} \sum_{j=1}^{J_M} \frac{\left(I\frac{j}{J_M}\right)_n \left[I\left(1-\frac{j}{J_M}\right)\right]_{J-n}}{(I)_J} \times E\left[S_j|\theta, J_M\right] * \tag{10}$$

We will show below that this expression is incorrect (hence the *), and that the expression for $E[S_n|\theta, m, J_M, J]$ for finite J_M is also given by eqn 6. This important finding that J_M only enters the formulae through θ, see eqn 9, will be discussed later.

The joint multivariate probability distribution for $m < 1$ is given by a new sampling formula (Etienne 2005)

$$P[\vec{D}|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^{J} K(\vec{D}, A) \frac{I^A}{(\theta)_A} \tag{11}$$

Here, the $K(\vec{D}, A)$ for $A = S, \ldots, J$ are coefficients fully determined by the data, being defined as

$$K(\vec{D}, A) = \sum_{\{a_1,\ldots,a_S | \sum_{i=1}^{S} a_i = A\}} \prod_{i=1}^{S} \frac{\overline{s}(n_i, a_i)\overline{s}(a_i, 1)}{\overline{s}(n_i, 1)} \tag{12}$$

In Appendix A (see Supplementary Material) we show that eqn 11 can also be written in integral notation

$$P[\vec{D}|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \Phi_j!} \frac{\theta^S}{(I)_J} \int_0^1 \cdots \int_0^1 \prod_{i=1}^{S} \left[(I_i x_i)_{n_i} \frac{(1-x_i)^{\theta-1}}{x_i} \right] dx_1 \ldots dx_S \tag{13}$$

where

$$I_i = I \prod_{k=1}^{i-1} (1 - x_k) \tag{14}$$

Equation 13 provides a way to avoid Stirling numbers in computing the multivariate probability, e.g. by Monte Carlo integration. This will, however, be very computationally intensive for a large number of species S.

We also note that eqns 2 and 11 must be multiplied by $\frac{\prod_{j=1}^{J} \Phi_j!}{S!}$ if the species are labelled in some way because their identity matters (Johnson *et al.* 1997, chapter 41).

THE SAMPLING NATURE OF THE NEUTRAL THEORY

The essential difference between the actual distribution of species abundances in the whole community and the observed abundance distribution in samples was already recognized by Fisher *et al.* (1943), and addressed by using Poisson random sampling (Pielou 1969; Bulmer 1974) and, more recently and in a fully exact way, by using hypergeometric random sampling (Dewdney 1998). In population genetics, it was immediately acknowledged that the Ewens sampling formula represents a theory where such sampling effects are fully taken into account (hence the name). However, it has not been emphasized enough in community ecology that this is also true for Hubbell's (2001) extension of the theory that includes dispersal limitation. In this section, we emphasize this by building a single sampling framework that contains the previous expressions that come from the two separate lines of research.

A particular property of our model formulation is the invariance of the formulae under hypergeometric sampling (drawing without replacement), i.e. if we take a subsample of size J_2 from a sample of size J_1 ($J_1 > J_2$), then the formulae for the subsample are identical to those for the sample when we simply substitute J_2 for J_1. The mathematical formulation is as follows. We first define the hypergeometric distribution as

$$P_{hyp}[n|j, J_1, J_2] = \frac{\binom{j}{n}\binom{J_1-j}{J_2-n}}{\binom{J_1}{J_2}} \tag{15}$$

which is the probability of sampling n individuals of a species in a subsample of size J_2 given that there are j individuals of this species in the sample of size J_1. More generally, given a sample of size J_1 that contains S_1 species with abundances $j_1, \ldots, j_{S_1}$, the probability of drawing a subsample of size J_2 with abundances $n_1, \ldots, n_{S_1}$ (some of which may equal 0) is given by

$$P_{hyp}[\vec{D}_2 \,|\, \vec{D}_1, J_1, J_2] = \frac{\prod_{i=1}^{S_1} \binom{j_i}{n_i}}{\binom{J_1}{J_2}} \tag{16}$$

where $\vec{D}_1 = (j_1, \ldots, j_{S_1})$ and $\vec{D}_2 = (n_1, \ldots, n_{S_1})$ with some of the n_i equalling 0 if $S_2 < S_1$.

Invariance under sampling then means

$$E[S_n|\theta, m, J_2] = \sum_{j=n}^{J_1} P_{hyp}[n|j, J_1, J_2] E\left[S_j|\theta, m, J_1\right] \tag{17a}$$

$$P[\vec{D}_2|\theta, m, J_2] = \sum_{\{\vec{D}_1\}} P_{hyp}[\vec{D}_2 \,|\, \vec{D}_1, J_1, J_2] P[\vec{D}_1|\theta, m, J_1] \tag{17b}$$

where the sum in the second line is over all distinct data sets $\vec{D}_1$ that have size J_1.

No dispersal limitation

When there is no dispersal limitation, a local community is a simple sample from the metacommunity. Then we have eqn 17a with $J_1 = J_M$ and $J_2 = J$; hence

$$E[S_n|\theta,J] = \sum_{j=1}^{J_M} P_{hyp}[n|j,J_M,J]E\left[S_j|\theta,J_M\right] \tag{18}$$

For infinite metacommunity size J_M this can also be written as

$$E[S_n|\theta,J] = \int_0^1 P_{bin}[n|x,J]\Omega(x)dx \tag{19}$$

where $P_{bin}[n|x,J]$ is the binomial distribution (drawing with replacement),

$$P_{bin}[n|x,J] = \binom{J}{n} x^n(1-x)^{J-n} \tag{20}$$

and

$$\Omega(x) = \frac{\theta(1-x)^{\theta-1}}{x} \tag{21}$$

is the abundance distribution in the infinite metacommunity (Ewens 1972; Alonso & McKane 2004; see also Table 1). We remark that the binomial distribution is the limit of the hypergeometric distribution for infinite metacommunity size (in which case there is no difference between sampling with and without replacement).

Equations 18 and 19 are identical for finite J_M as well: they both lead to eqn 1, the former due to the sampling nature of the theory expressed in eqn 17a, the latter by recognizing the beta distribution in the integrand and writing factorials as gamma functions:

$$\begin{aligned} E[S_n|\theta,J] &= \binom{J}{n}\int_0^1 x^n(1-x)^{J-n}\frac{\theta(1-x)^{\theta-1}}{x}dx \\ &= \theta\frac{\Gamma(J+1)}{\Gamma(n+1)\Gamma(J-n+1)}\frac{\Gamma(n)\Gamma(\theta+J-n)}{\Gamma(\theta+J)} \\ &= \frac{\theta}{n}\frac{\Gamma(J+1)}{\Gamma(J-n+1)}\frac{\Gamma(\theta+J-n)}{\Gamma(\theta+J)} \end{aligned} \tag{22}$$

Dispersal limitation

With dispersal limitation, the local community is no longer a simple hypergeometric sample from the metacommunity. It is a dispersal-limited hypergeometric sample (which is dispersal-limited binomial for infinite J_M). We will derive an expression for the corresponding distribution.

We first consider a metacommunity of infinite size. Let us write eqn 6 as (see also Table 1)

$$E[S_n|\theta,m,J] = \int_0^1 P_{bin}^{DL}[n|m,x,J]\Omega(x)dx \tag{23}$$

where

$$P_{bin}^{DL}[n|m,x,J] = \binom{J}{n}\frac{(Ix)_n(I(1-x))_{J-n}}{(I)_J} \tag{24}$$

and $\Omega(x)$ is given by eqn 21. Equation 24 was first calculated in the context of a stochastic model of community dynamics based on the community matrix (McKane *et al.* 2000; Solé *et al.* 2000), and then applied to the context of neutral community ecology (Volkov *et al.* 2003; McKane *et al.* 2004). It also appears in a similar model in population genetics (Wakeley & Takahashi 2004). Mathematically, it is known as the negative hypergeometric distribution which is a special case of the Pólya-Eggenberger distribution which in turn is a special case of the unified hypergeometric distribution (Johnson *et al.* 1997, chapters 39 and 40). In eqn 23, $P_{bin}^{DL}[n|m, x, J]$ must be interpreted as the probability for a dispersal-limited species of relative abundance x in the metacommunity (with infinite size) to be represented by exactly n individuals in a sample of size J (McKane *et al.* 2004). Our notation of $P_{bin}^{DL}[n|m, x, J]$ refers to the fact that eqn 24 is the dispersal-limited binomial distribution; it becomes the binomial distribution (eqn 20) as $m \to 1$ (Alonso & McKane 2004). We can generalize eqn 24 to

$$P_{bin}^{DL}[\vec{D}_1|m,\vec{D}_2,J] = \frac{J!}{n_1!\ldots n_S!}\frac{\prod_{i=1}^{S}(I_i x_i)_{n_i}}{(I)_J} \tag{25}$$

where, I_i is given by eqn 14 and $\vec{D}_2$ is a vector of relative abundances x_i. This provides an alternative derivation of eqn 13; this is most easily done with the 'labelled-species' form of eqn 11.

For finite metacommunity size the analogue of the dispersal-limited binomial distribution P_{bin}^{DL} will be called the dispersal-limited hypergeometric distribution P_{hyp}^{DL}. Here, we derive an expression for this distribution. We follow the second line of research in tracing back individuals in a sample from the local community to their ancestors that once immigrated into that local community (Etienne & Olff 2004b). These ancestors represent a sample from the metacommunity and thus obey all the formula we have presented for the case $m = 1$. We only need to establish the link between the current sample and this sample of ancestors. Let the sample of ancestors contain A ancestors. Its probability distribution is also governed by the Ewens

sampling formula, with parameter I (Etienne & Olff 2004b; see Wakeley 1998 for similar equation in population genetics):

$$P[A|m(I),J] = \overline{s}(J,A)\frac{I^A}{(I)_J} \tag{26}$$

Let there be a ancestors of the species under consideration. The probability of finding a ancestors of this species, given that there are j individuals of this species in the metacommunity, is the hypergeometric distribution $P_{\text{hyp}}[a|j, J_M, A]$ of eqn 15. The probability that a ancestors have n descendants among the J individuals in our dispersal-limited sample is computed as follows. From combinatorics it is known that there are $\overline{s}(J, A)$ partitions of J individuals into A groups (each group containing at least one individual). For example, if $J = 4$ and $A = 3$, the possible partitions are (a, b, cd), (a, bc, d), (ab, c, d), (ac, b, d), (ad, b, c) and (a, bd, c). Likewise there are $\overline{s}(n, a)$ partitions of n individuals into a groups and $\overline{s}(J - n, A - a)$ partitions of the remaining $J - n$ individuals into $A - a$ groups. There are $\binom{J}{n}$ ways of choosing n out of J individuals. Likewise, there are $\binom{A}{a}$ ways of choosing a out of A ancestors. The probability $P[n|a, A, J]$ that n individuals in our local community sample descend from exactly a ancestors in our metacommunity sample is given by Wakeley (1999)

$$P[n|a,A,J] = \frac{\binom{J}{n}}{\binom{A}{a}}\frac{\overline{s}(n,a)\,\overline{s}\,(J-n,A-a)}{\overline{s}(J,A)} \tag{27}$$

The dispersal-limited hypergeometric distribution is therefore a sum of the product of the three probabilities given in eqns 15, 26 and 27 over all possible values of A and a:

$$\begin{aligned} P^{\text{DL}}_{\text{hyp}}[n|m,j,J_{\text{M}},J] &= \sum_{A=1}^{J}\sum_{a=1}^{n} P[n|a,A,J]P_{\text{hyp}}[a|j,J_{\text{M}},A] \\ P[A|m(I),J] &= \binom{J}{n}\sum_{A=1}^{J}\sum_{a=1}^{n}\overline{s}(n,a)\overline{s}(J-n,A-a) \\ &\quad\times\frac{I^A}{(I)_J}\frac{1}{\binom{A}{a}}P_{\text{hyp}}[a|j,J_{\text{M}},A] \end{aligned} \tag{28}$$

For $m \to 1$, I becomes infinite and only the term $A = J$ and $a = n$ contribute to the sum, so eqn 28 becomes $P_{\text{hyp}}[n|j, J_M, J]$, because $\overline{s}(n, n) = 1$. For $J_M \to \infty$, the hypergeometric distribution $P_{\text{hyp}}[a|j, J_M, A]$ becomes the binomial with parameter $x = j/J_M$ and the remaining sums in terms of Stirling numbers and powers of x can be written as Pochhammer symbols resulting in eqn 24. So, the new dispersal-limited hypergeometric distribution has the right limit behaviour. For any value of J_M, when m tends to 1, it tends to the random hypergeometric sampling distribution. When J_M tends to infinity, for any value of m, it tends to the dispersal-limited binomial distribution. With the new

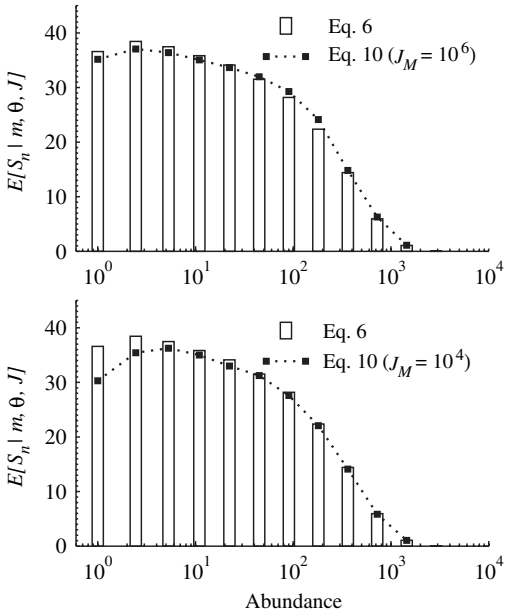

Figure 1 Example of the difference in expected number of species between the exact result (eqn 6) and the approximation (eqn 10) by Vallade & Houchmandzadeh (2003) for two different values of metacommunity size. The parameter values used are $\theta = 50$ and $m = 0.5$. Local community size is $J = 20\,000$. Particularly the diversity of species with low abundances is underestimated with eqn 10. The lower and upper boundaries of the abundance classes are such that abundance class i contains all abundances n for which $2^{i-1} \le n < 2^i$.

distribution (eqn 28), we can write the analogue of eqn 23 for finite J_M (see also Table 1):

$$E[S_n|\theta,m,J_{\text{M}},J] = \sum_{j=1}^{J_{\text{M}}} P^{\text{DL}}_{\text{hyp}}[n|m,j,J_{\text{M}},J]E\left[S_j|\theta,J_{\text{M}}\right] \tag{29}$$

When we compare this to the result of Vallade & Houchmandzadeh (2003) given in eqn 10, we see that these expressions are different in general, being only equal for infinite J_M for which we have eqn 23. The expression of Vallade & Houchmandzadeh (2003) given in eqn 10 is incorrect, because it is not invariant under hypergeometric sampling. In fact, it corresponds to an approximate discretization of the exact integral result (eqn 6) and only converges to eqn 6 when J_M tends to infinity (see Appendix B). In Fig. 1 we show that eqn 10 converges to the exact result (eqn 6) when J_M is large enough, but substantially deviates from it for lower values of J_M. As in the case without dispersal limitation, the expressions (eqns 23 and 29) for infinite and finite metacommunity size J_M are identical, as we shown in Appendix C (see also Table 1).

The dispersal-limited hypergeometric distribution can be generalized to

$$P_{\text{hyp}}^{\text{DL}}\left[\vec{D}_1\,|m,\vec{D}_2,J_{\text{M}},J\right]=\frac{J!}{n_1!\ldots n_S!}\sum_{\mathcal{A}=1}^{J}\sum_{a_1=1}^{n_1}\cdots\sum_{a_{S-1}=1}^{n_{S-1}}\left[\prod_{i=1}^{S-1}\overline{s}(n_i,a_i)\right]\overline{s}\left(J-\sum_{i=1}^{S-1}n_i,\mathcal{A}-\sum_{i=1}^{S-1}a_i\right)\times\frac{I^{\mathcal{A}}}{(I)_J}\frac{a_1!\ldots a_S!}{\mathcal{A}!}P_{\text{hyp}}[\vec{a}\,|\,\vec{j},J_{\text{M}},\mathcal{A}] \tag{30}$$

which leads to eqn 11 when applied to a sample from the metacommunity [which is governed by the ('labelled-species' form of the) Ewens sampling formula (eqn 2)]. While eqn 28 has a parallel expression in population genetics (Wakeley 1999), its generalization (eqn 30) is, to our knowledge, entirely new.

The subsample approach

In this section, we relate the expected number of species, eqns 1 and 6, to the corresponding multivariate probability distributions, eqns 2 and 11. First, we examine whether eqns 2 and 11 can be expressed in terms of eqns 1 and 6, respectively, for the observed values $n_1,\ldots,n_S$. This does not only show the link between the two types of expressions (from two lines of research), but it has practical importance as well, because the expected number of species with a particular abundance is usually easier to obtain (using the master equation approach) than the multivariate probability distribution.

We need the concept of subsamples. First, we note that $P[\vec{D}\,|\Theta,J]=P[n_1,\ldots,n_S|\Theta,J]$ can, like every multivariate probability, be written as

$$P[\vec{D}\,|\Theta,J]=P[n_1,\ldots,n_S|\Theta,J]=P[n_1|\Theta,J]P[n_2|n_1,\Theta,J]\ldots P[n_S|n_1,\ldots,n_{S-1},\Theta,J] \tag{31}$$

where Θ represents the model parameters [θ or (θ, m)]. Equation 31 just follows from the definition of conditional probabilities.

The first term in eqn 31, $P\,[n_1\,|\,\Theta,J]$, is the probability of a species in a sample of size J to have exactly abundance n_1. The second term in eqn 31, $P\,[n_2\,|\,n_1,\Theta,J]$, is the probability of a species in sample size of size J to have exactly abundance n_2 given that another species in the sample has abundance n_1. This probability is equivalent to the probability of a species in sample of size $J-n_1$ to have exactly abundance n_2. It can therefore be expressed as

$$P[n_2|n_1,\Theta,J]=P[n_2|\Theta,J-n_1] \tag{32}$$

We call the sample size $J-n_1$ the effective sample size for species 2. More generally, we can define the effective sample size J_i for species i as

$$J_i=J-\sum_{k=1}^{i-1}n_k \tag{33}$$

This definition implies, for instance that $J_1=J, J_S=n_S$ and $J_{S+1}=0$. For later convenience, we define the partial data sets

$$\vec{D}_i=(n_i,\ldots,n_S) \tag{34}$$

entailing $\vec{D}_1=\vec{D}$ and $\vec{D}_S=n_S$. We further define Φ_{n_i} as the number of species with abundance n_i in the subsample $\vec{D}_i$.

With the definitions in eqn 33, eqn 31 becomes

$$P[\vec{D}\,|\Theta,J]=\prod_{i=1}^{S}P[n_i|\Theta,J_i] \tag{35}$$

In Appendix D we show that this leads to the following expressions (see also Table 1):

$$P[\vec{D}\,|\theta,J]=\frac{\prod_{i=1}^{S}E[S_{n_i}|\theta,J_i]}{\prod_{j=1}^{J}\Phi_j!} \tag{36}$$

and

$$P[\vec{D}\,|\theta,m,J]=\frac{\prod_{i=1}^{S}\widehat{E}[S_{n_i}|\theta,m,J_i]}{\prod_{j=1}^{J}\Phi_j!} \tag{37}$$

with

$$\widehat{E}[S_{n_i}|\theta,m,J_i]=\int_0^1 P_{\text{bin}}^{\text{DL}}[n_i|m,x,J_i]\widehat{\Omega}(x|\theta,m,\vec{D}_{i+1})dx \tag{38}$$

where $P_{\text{bin}}^{\text{DL}}[n_i|m,x,J_i]$ is defined in eqn 24 and $\widehat{\Omega}(x|\theta,m,\vec{D}_{i+1})$ is defined by

$$\widehat{\Omega}(x|\theta,m,\vec{D}_{i+1})=\Omega(x)F\left(x|\theta,m,\vec{D}_{i+1}\right) \tag{39}$$

with $\Omega(x)$ given eqn 21 and $F(x|\theta,m,\vec{D}_{i+1})$ defined in equation (D-7) in Appendix D. Comparing eqns 23 and 38 we can interpret eqn 38 as having an abundance distribution $\Omega(x)$ that is modified by a factor that takes into account the subsample $\vec{D}_{i+1}$. We further note that eqns 36 and 37 are even simpler when species are labelled: then there is only $S!$ in the denominator.

We also note that eqns 1 and 6 can be derived from the multivariate probability distributions (eqns 2 and 11) using the equality

$$E[S_n|\Theta,J]=\sum_{\Phi_n=0}^{J}\Phi_n P[\Phi_n|\Theta,J] \tag{40}$$

where $P[\Phi_n\,|\,\theta,J]$ is the probability that exactly Φ_n species with abundance n are observed. This is a sum over all possible data sets that have Φ_n species with abundance n:

$$E[S_n|\Theta,J] = \sum_{\Phi_n=0}^{J} \Phi_n \sum_{\left\{\vec{D}|\Phi_n\right\}} P[\vec{D}|\Theta,J] \quad (41)$$

In Appendix E we show that with help of the subsample concept this indeed leads to eqns 1 and 6.

Watterson (1974) already provided alternative derivations for the mathematically identical model in population genetics when $m = 1$. However, no such derivations have been given for the case with dispersal limitation.

DISCUSSION

We have presented previously obtained results of neutral community theory in a general framework where the dispersal-limited sampling nature of the theory plays a central role. We have summarized our results in Table 1.

For the first time in neutral community ecology, the main results of two lines of research – $E[S_n|\theta, m, J]$, the expected number of species with abundance n in a sample of size J, and $P[\vec{D}|\theta, m, J]$, the joint multivariate probability of observing S species with abundances $n_1, n_2,\ldots,n_S$ in a sample of size J – have been presented together and related to one another. In the case without dispersal limitation ($m = 1$), $P[\vec{D}|\theta, J]$ can even be expressed in terms of $E[S_{n_i}|\theta, J]$ using subsamples D_i, whereas in the case with dispersal limitation, this expression must be somewhat modified, but has a similar form. Also, we have derived $E[S_n|\theta, m, J]$ and $E[S|\theta, m, J]$ from $P[\vec{D}|\theta, m, J]$. Although this has been derived in the mathematically identical theory in population genetics for the case without dispersal limitation, the derivation for the case with dispersal limitation is given here for the first time. Relating expected values to multivariate distributions is important because it is much easier to write and solve for stationarity dynamical one-dimensional models involving expected values (McKane *et al.* 2000, 2004; Vallade & Houchmandzadeh 2003) than it is for their corresponding multivariate distributions. However, we emphasize that precisely these exact multivariate sampling distributions taken as likelihood functions are actually needed to perform maximum likelihood estimation of model parameters (Etienne 2005) and sound statistical model comparisons (Etienne & Olff 2005).

Moreover, our sampling framework has enabled us to show that the sampling distributions are valid for a metacommunity of any size J_M. In other words, two samples of equal size from two metacommunities of different sizes $J_{M,1}$ and $J_{M,2}$ are characterized by exactly the same sampling distributions, as long as both metacommunities are described by the same biodiversity number ($\theta_1 = \theta_2$). This has not been emphasized in previous work. This is important for two reasons. First, an already existing expression $E[S_n|\theta, m, J_M, J]$ when J_M is finite (Vallade & Houchmandzadeh 2003) turns out to be incorrect. Alonso & McKane (2004), assuming Vallade & Houchmandzadeh (2003) to be correct, suggested that species-abundance data can be used to estimate the metacommunity size and hence the speciation rate v because $\theta := \frac{v(J_M-1)}{1-v}$ (Vallade & Houchmandzadeh 2003; Alonso & McKane 2004; Etienne 2005). The independence of metacommunity size that we have shown in the present study, however, implies that this is not possible. Second, as metacommunity size does not matter, we can safely assume infinite metacommunity size, which simplifies our formulae, because we can use binomial sampling instead of hypergeometric sampling. We want to stress, however that it is invariance under hypergeometric sampling that provided the basis for our sampling theory.

Thus, mathematically, our formulas are valid for any J_M. Nevertheless, we need to remember the model assumption of separation of spatiotemporal scales: a local scale with immigration as the source of new species vs. a regional metacommunity scale with speciation as the source of new species. We cannot, therefore, choose any size J_M we want; we need to require that $J_M \gg J$. This assumption allows us to safely ignore speciation at the local level, and to assume that local dynamics are much faster than regional dynamics, so the metacommunity composition does not change appreciably when the ancestors are sampled (which occurs at different instances). The assumption $J_M \gg J$ is biologically very realistic, because, within our framework, J is the sample size that is in practice much lower than the metacommunity size.

We already noted that sampling effects have been recognized since Fisher *et al.* (1943). However, other stochastic models of communities do not (fully) take this into account (Volkov *et al.* 2003; He 2005), or impose Poisson sampling afterwards (Engen & Lande 1996a,b, Dewdney 2000; Diserud & Engen 2000). This makes comparison of different models difficult, even in the latter case, because the expressions may be conditioned differently. Some (implicitly) assume the number of sampled species S and others assume the number of sampled individuals J, as do our formulas. For a correct comparison, we need to condition on both (Etienne & Olff 2005).

Neutral community theory as formulated by Hubbell (2001) can be seen as an extension of Ewens' (1972) theory into the ecological arena. This extension is far from trivial because Hubbell's (2001) main intuition is that, in addition to neutral (or ecological) drift, it is dispersal limitation that is the leading factor structuring ecological communities. All recent theoretical advances in neutral community theory based on Hubbell's (2001) formulation can now be translated back to population genetics to extend Ewens' (1972) work as 'a dispersal-limited sampling theory of selectively neutral alleles'. With the dispersal-limited sampling distributions introduced in this work, we can not only examine whether a certain allelic polymorphism is maintained neutrally, but we can also easily

estimate the amount of dispersal limitation (or degree of isolation) of the locality where this allelic polymorphism comes from. It also enables computation of the ages of alleles in dispersal-limited populations.

Concerning the evolutionary age of species (or, equivalently, species time-to-extinction), the neutral theory has been strongly criticized for yielding unrealistically old species (Lande *et al.* 2003; Nee 2005). However, this finding may depend more on other model assumptions than on the assumption of neutrality. For instance, Nee's (2005) estimates of species ages are based on Ewens' (1972) equilibrium model for fixed community size with $\theta \rightarrow 0$ and $m = 1$. Griffiths & Lessard (2005) recently presented a formula for any value of θ that makes species ages already a few orders of magnitude smaller. Species ages might also be appreciably different if dispersal limitation is taken into account. Furthermore, non-equilibrium dynamics and fluctuations in community size may substantially affect effective community size and thereby the time scales of species origination. Also, even if species ages are better explained by non-neutral processes at evolutionary time scales, such as ecological succession (a process involving ecologically non-equivalent species interacting through non-neutral processes such as facilitation and hierarchical competition), the final mature community that we observe today may still be consistent with neutral dynamics. In sum, the use of species ages to falsify the neutral theory is rather premature.

A stronger test of neutrality than the goodness-of-fit of a single species-abundance distribution is a test whether two local communities that are both dispersal-limited hypergeometric samples from the same metacommunity, but are separated by a known distance have the (dis)similarity in their species-abundance distributions that one would expect from neutrality. We believe that our sampling framework is able to provide such a test in principle. As the distance between the local communities obviously matters, a spatially explicit model seems to be unavoidable, but perhaps the spatially implicit model with appropriately chosen parameters may be used as a proxy that captures the essence. In any case, this is a difficult task mathematically, but one that merits further study. Ideas in population genetics involving 'isolation by distance' (e.g. Wakeley & Aliacar 2001) may provide fruitful starting points.

We have expressed the local community as a sample from the larger regional metacommunity, a sample which may or may not be affected by dispersal limitation. In our expressions the metacommunity is purely regulated by speciation and extinction, and thus governed by the Ewens sampling formula, but this is not necessary. Our dispersal-limited hypergeometric distribution can also be applied to metacommunities that are structured according to other, even non-neutral, rules. Although at the local community level the dynamics is neutral, any differences in species abundances because of (non-neutral) metacommunity structure propagate to this local level. This allows for a dispersal-limited sampling theory for non-neutral communities. A more exact but more challenging approach would be to replace the dispersal-limited hypergeometric distribution of eqns 28 and 30 that assume local neutrality by a new dispersal-limited distribution that takes into account, at the local level, the same non-neutral factors controlling abundances in the metacommunity. This can potentially be done in essentially the same formalism we have presented here (possibly following suggestions in the population genetics literature (e.g. Wakeley & Takahashi 2004; Slade & Wakeley 2005). Our expressions are however, good approximations that are fully in line with the model assumptions on the time scale discussed above.

The picture that emerges is thus: species and niche assembly originate through evolutionary time shaping species abundances on the regional, long temporal scale. The very spatially extended nature of ecological systems involves dispersal limitation on the local and short temporal scale. So, if a particular locality is sampled, we will always have some degree of dispersal limitation in addition to other factors determining species abundances at the metacommunity level. The current challenge is to develop a dynamic community theory that can quantify the relative importance of dispersal limitation vs. other, neutral or non-neutral, factors determining species abundances through evolutionary time. We strongly believe that our dispersal-limited sampling theory provides the basis for such a unifying theoretical framework.

ACKNOWLEDGEMENTS

Authors thank three anonymous referees, John Wakeley, Jérôme Chave and Han Olff for very constructive comments. D.A. thanks the support of the James S. McDonnell Foundation through a Centennial Fellowship to Mercedes Pascual.

REFERENCES

Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.

Brown, J.H. & Kodric-Brown, A. (1977). Turnover rate in insular biogeography: effect of immigration on extinction. *Ecology*, 58, 445–449.

Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30, 101–110.

Dewdney, A.K. (1998). A general theory of the sampling process with applications to the 'veil line'. *Theor. Popul. Biol.*, 54, 294–302.

Dewdney, A.K. (2000). A dynamical model of communities and a new species-abundance distribution. *Biol. Bull.*, 35, 152–165.

Diserud, O.H. & Engen, S. (2000). A general and dynamic species abundance model, embracing the lognormal and the gamma models. *Am. Nat.*, 155, 497–511.

Engen, S. & Lande, R. (1996a). Population dynamic models generating the lognormal species abundance distribution. *Math. Biosci.*, 132, 169–183.

Engen, S. & Lande, R. (1996b). Population dynamic models generating the species abundance distributions of the Gamma type. *J. Theor. Biol.*, 178, 325–331.

Etienne, R.S. (2005). A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8, 253–260.

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). Bayesian analysis of species-abundance data: assessing the relative importance of dispersal and niche-partitioning for the maintenance of biodiversity. *Ecol. Lett.*, 8, 493–504.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3, 87–112.

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, 12, 42–58.

Griffiths, R.C. & Lessard, S. (2005). Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Popul. Biol.* (in press).

Grinnell, J. (1922). On the role of the accidental. *Auk*, 39, 373–380.

Hanski, I. (1983). Coexistence of competitors in patchy environment. *Ecology*, 64, 493–500.

He, F.L. (2005). Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct. Ecol.*, 19, 187–193.

Hubbell, S.P. (1997). A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, 16, S9–S21.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ, USA.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York, NY, USA.

Karlin, S. & McGregor, J. (1972). Addendum to a paper of W. Ewens. *Theor. Popul. Biol.*, 3, 113–116.

Lande, R., Engen, S. & Saether, B.-E. (2003). *Stochastic Population Dynamics in Ecology and Conservation. Oxford Series in Ecology and Evolution*. Oxford University Press, Oxford, UK.

Levins, R. & Culver, D. (1971). Regional coexistence of species and competition between rare species. *Proc Natl Acad Sci U S A*, 68, 1246–1248.

Loreau, M. & Mouquet, N. (1999). Immigration and the maintenance of local species diversity. *Am. Nat.*, 154, 427–440.

MacArhur, R.H. & Wilson, E.O. (1967). *Island Biogeography*. Princeton University Press, Princeton, NJ, USA.

McGill, B.J. (2003). A test of the unified neutral theory of biodiversity. *Nature*, 422, 881–885.

McKane, A.J., Alonso, D. & Solé, R.V. (2000). A mean field stochastic theory for species rich assembled communities. *Phys. Rev. E*, 62, 8466–8484.

McKane, A.J., Alonso, D. & Solé, R.V. (2004). Analytic solution of Hubbell's model of local community dynamics. *Theor. Popul. Biol.*, 65, 67–73.

Moran, P.A.P. (1958). Random processes in genetics. *Proc Camb Philol Soc*, 54, 60–71.

Moran, P.A.P. (1962). *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford, UK.

Nee, S. (2005). The neutral theory of biodiversity: do the numbers add up? *Funct. Ecol.*, 19, 173–176.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*. Wiley, New York, NY, USA.

Slade, P.F. & Wakeley, J. (2005). The structured ancestral selection graph and the many-demes limit. *Genetics*, 169, 1117–1131.

Solé, R.V., Alonso, D. & McKane, A.J. (2000). Scaling in a network model of multispecies communities. *Physica A*, 286, 337–344.

Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology*, 75, 2–16.

Vallade, M. & Houchmandzadeh, B. (2003). Analytical solution of a neutral model of biodiversity. *Phys. Rev. E*, 68, 061902.

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. (1998). Segregating sites in Wright's island model. *Theor. Popul. Biol.*, 53, 166–175.

Wakeley, J. (1999). Non-equilibrium migration in human history. *Genetics*, 153, 1863–1871.

Wakeley, J. & Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, 159, 893–905; Corrigendum in *Genetics* **160**, 1263 (2001).

Wakeley, J. & Takahashi, T. (2004). The many-demes limit for selection and drift in a subdivided population. *Theor. Popul. Biol.*, 66, 83–91.

Watterson, G.A. (1974). Models for the logarithmic species abundance distribution. *Theor. Popul. Biol.*, 6, 217–250.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article from http://www.Blackwell-Synergy.com:

Appendix A Derivation of eqn 13.
Appendix B The relation of the approximation (eqn 10) to the exact result (eqn 6).
Appendix C Proof of the equality of eqns 23 and 29.
Appendix D Derivation of eqns 36 and 37.
Appendix E Derivation of eqns 1 and 6 from eqns 2 and 11.
Appendix F A historical note on the origins of the binomial and hypergeometric distributions.

Editor, Jerome Chave
Manuscript received 11 May 2005
First decision made 20 June 2005
Second decision made 11 July 2005
Manuscript accepted 12 July 2005