

# A dispersal-limited sampling theory for species and alleles

RAMPAL S. ETIENNE<sup>1,\*</sup> & DAVID ALONSO<sup>2</sup>

<sup>1</sup>Community and Conservation Ecology Group, University of Groningen, PO Box 14, 9750 AA Haren, The Netherlands.

<sup>2</sup>Ecology and Evolutionary Biology, University of Michigan, 830 North University Av, Ann Arbor MI 48109-1048, USA.

\*Email for correspondence: r.s.etienne@rug.nl

**Keywords:** biodiversity, community, neutral model, Ewens sampling formula, random sampling, binomial sampling, hypergeometric sampling, dispersal-limited sampling

**Running Head:** A dispersal-limited sampling theory

**Words in abstract:** 222

**Words in main text:** approx. 4500

**Number of references:** 44

## Abstract

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

## Introduction

14 The importance of dispersal in ecology has long been recognized (*e.g.* Grinnell 1922, MacArthur & Wilson 1967,  
15 Levins & Culver 1971, Brown & Kodric-Brown 1977, Hanski 1983, Tilman 1994, Loreau & Mouquet 1999).  
16 Yet, seldom has a more vigorous (quantitative) case been made than by Hubbell (1997, 2001) who presented a  
17 comprehensible suite of stochastic neutral models of community structure based on the fundamental processes of  
18 speciation, extinction and dispersal. In the most often cited model of these, the local community consists of  $J$   
19 individuals of different species whose off-spring compete for sites that are left open after an individual dies. They  
20 do not only compete with one another, but they also compete with immigrants from outside the local community:  
21 there is a probability  $m$  that an open site is colonized by an immigrant. If  $m < 1$  the local community is called  
22 dispersal-limited. With probability  $1 - m$ , the open site is colonized by off-spring of a local individual. Each  
23 individual in the local community, regardless of species, has an equal chance of colonizing the open site (the neu-  
24 trality assumption). Each open site is immediately recolonized so community size remains constant (the zero-sum

1 assumption). The immigrants come from a regional species pool (the metacommunity, Hubbell 2001) that is in a  
2 stochastic balance between speciation and extinction. This balance is characterized by the parameter  $\theta$ , a compos-  
3 ite of the speciation rate  $\nu$  and metacommunity size  $J_M$ . Speciation in this model occurs by “point mutation” (in  
4 other models Hubbell uses “random fission” speciation which is a first step towards modelling allopatric specia-  
5 tion). This model resembles the continent-island infinite alleles model with Moran-like reproduction in population  
6 genetics (Wright 1931, Moran 1962, Ewens 1972); the difference with Moran reproduction is that the individual  
7 that dies does not produce any offspring that could replace it. We note that the terminology “continent-island” is  
8 only historical; the theory also applies to a local sample from a continuous landscape.

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

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

25 The forwards-in-time perspective uses a master equation approach with a Markovian description of states and  
26 transitions (McKane *et al.* 2000, Volkov *et al.* 2003, Vallade & Houchmandzadeh 2003, McKane *et al.* 2004,  
27 Alonso & McKane 2004). This has resulted in exact analytical expressions and various approximations for the  
28 *expected number of species with a certain abundance* in a sample of  $J$  individuals from a dispersal-limited local  
29 community: if  $n$  is the abundance, then  $E[S_n|\theta, m, J]$  denotes the expected number of species with this abundance

1 in this sample. Vallade & Houchmandzadeh (2003) and subsequent studies used the shorthand notation of  $\langle \phi_n \rangle$  or  
2  $S(n)$  for this expectation, but we employ the longer notation to emphasize that this is an expectation that follows  
3 from the model in contrast to the actually observed number of species with abundance  $n$ , which we will denote  
4 by  $\Phi_n$  as in Etienne (2005). The expected number of species with a certain abundance is the classical approach  
5 to study commonness and rarity in community ecology and also a very useful tool in exploring the behavior of  
6 community models. However, it cannot be used to obtain accurate estimates of the model parameters.

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

15 Because both lines of research work on the same model and have provided exact analytical results, they must  
16 somehow be related, but until now the common framework has not been made explicit. In this paper, after pre-  
17 senting the basic results of the two lines of research, we build such a framework. Its most important property is  
18 the sampling nature of the theory and the role that dispersal plays in it. We introduce new distributions, called the  
19 dispersal-limited binomial and dispersal-limited hypergeometric distributions by which the results of both lines of  
20 research arise naturally. As a result we find that the expression for  $E[S_n|\theta, m, J]$  for finite metacommunity size, as  
21 reported by Vallade & Houchmandzadeh (2003) is incorrect. An important consequence is that it is not possible to  
22 estimate metacommunity size and hence the speciation rate from species-abundance data, as was suggested based  
23 on this formula (Alonso & McKane 2004, p. 904). Next, we link the two lines of research by expressing results of  
24 one line of research in terms of the other and vice versa, by making use of the concept of subsamples. Most of our  
25 results are summarized in Table I. We end with a discussion of our results that tries to open new doors to further  
26 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

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

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

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

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

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

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

6 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:

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

9 In Pochhammer notation, (1) becomes even more compact:

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

10 Note that  $J_M$  does not enter equations (1) and (2), except by its role in  $\theta$ . Below, we make this more explicit.

## Dispersal limitation

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

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

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

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

14 and  $I$  is a transformed immigration parameter,

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

1 The parameter  $I$  is called  $\mu$  in Vallade & Houchmandzadeh 2003 and  $\gamma$  in Alonso & McKane 2004, while  $Ix$  is  
 2 called  $\lambda$  in Volkov *et al.* 2003.  $I$  is related to the immigration probability  $m$  and local community size  $J$  as the  
 3 fundamental biodiversity number  $\theta$  is related to the speciation probability  $\nu$  and metacommunity size  $J_M$  (Vallade  
 4 & Houchmandzadeh 2003, Alonso & McKane 2004, Etienne 2005),

$$\theta := \frac{\nu}{1-\nu} (J_M - 1) \quad (9)$$

5 In analogy to  $\theta$ , we will call  $I$  the **fundamental dispersal number**.

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

$$\star E[S_n|\theta, m, J_M, J] = \binom{J}{n} \sum_{j=1}^{J_M} \frac{\binom{I \frac{j}{J_M}}{n} \left( I \left( 1 - \frac{j}{J_M} \right) \right)^{J-n}}{(I)_J} E[S_j|\theta, J_M] \quad \star \quad (10)$$

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

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

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

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

$$K(\vec{D}, A) := \sum_{\{a_1, \dots, a_S | \sum_{i=1}^S a_i = A\}} \prod_{i=1}^S \frac{\bar{s}(n_i, a_i) \bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} \quad (12)$$

13 In appendix A we show that (11) can also be written in integral notation

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

14 where

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

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

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

# The sampling nature of the neutral theory

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

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

$$P_{\text{hyp}}[n|j, J_1, J_2] := \frac{\binom{j}{n} \binom{J_1-j}{J_2-n}}{\binom{J_1}{J_2}} \quad (15)$$

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

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

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

19 Invariance under sampling then means

$$E[S_n|\theta, m, J_2] = \sum_{j=n}^{J_1} P_{\text{hyp}}[n|j, J_1, J_2] E[S_j|\theta, m, J_1] \quad (17a)$$

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

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

## No dispersal limitation

1 When there is no dispersal limitation, a local community is a simple sample from the metacommunity. We then  
 2 have (17a) with  $J_1 = J_M$  and  $J_2 = J$ ; hence

$$E[S_n|\theta, J] = \sum_{j=1}^{J_M} P_{\text{hyp}}[n|j, J_M, J] E[S_j|\theta, J_M] \quad (18)$$

3 For infinite metacommunity size  $J_M$  this can also be written as

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

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

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

5 and

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

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

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

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

## Dispersal limitation

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

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

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

2 where

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

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

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

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

15 For finite metacommunity size the analog of the dispersal-limited binomial distribution  $P_{\text{bin}}^{\text{DL}}$  will be called the  
 16 dispersal-limited hypergeometric distribution  $P_{\text{hyp}}^{\text{DL}}$ . Here we derive an expression for this distribution. We follow  
 17 the second line of research in tracing back individuals in a sample from the local community to their ancestors that  
 18 once immigrated into that local community (Etienne & Olf 2004). These ancestors represent a sample from the  
 19 metacommunity and thus obey all the formula we have presented for the case  $m = 1$ . We only need to establish  
 20 the link between the current sample and this sample of ancestors. Let the sample of ancestors contain  $A$  ancestors.  
 21 Its probability distribution is also governed by the Ewens sampling formula, with parameter  $I$  (Etienne & Olf  
 22 2004):

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

23 (See Wakeley 1998 for similar equation in population genetics). Let there be  $a$  ancestors of the species under  
 24 consideration. The probability of finding  $a$  ancestors of this species, given that there are  $j$  individuals of this

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

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

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

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

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

18 With the new distribution (28), we can write the analog of (23) for finite  $J_M$  (see also Table I):

$$E[S_n|\theta, m, J_M, J] = \sum_{j=1}^{J_M} P_{\text{hyp}}^{\text{DL}}[n|m, j, J_M, J] E[S_j|\theta, J_M] \quad (29)$$

19 When we compare this to the result of Vallade & Houchmandzadeh (2003) given in (10), we see that these expres-  
 20 sions are different in general, being only equal for infinite  $J_M$  for which we have (23). The expression of Vallade &  
 21 Houchmandzadeh (2003) given in (10) is incorrect, because it is not invariant under hypergeometric sampling. In  
 22 fact, it corresponds to an approximate discretization of the exact integral result (6) and only converges to (6) when  
 23  $J_M$  tends to infinity (see Appendix B). In Figure 1 we show that (10) converges to the exact result (6) when  $J_M$  is

1 large enough, but substantially deviates from it for lower values of  $J_M$ . As in the case without dispersal limitation,  
 2 the expressions (23) and (29) for infinite and finite metacommunity size  $J_M$  are identical, as we show in Appendix  
 3 C (see also Table I).

4 The dispersal-limited hypergeometric distribution can be generalized to

$$P_{\text{hyp}}^{\text{DL}}[\vec{D}_1 | m, \vec{D}_2, J_M, J] = \tag{30}$$

$$\frac{J!}{n_1! \dots n_S!} \sum_{A=1}^J \sum_{a_1=1}^{n_1} \dots \sum_{a_{S-1}=1}^{n_{S-1}} \left( \prod_{i=1}^{S-1} \bar{s}(n_i, a_i) \right) \bar{s} \left( J - \sum_{i=1}^{S-1} n_i, A - \sum_{i=1}^{S-1} a_i \right) \frac{I^A}{(I)_J} \frac{a_1! \dots a_S!}{A!} P_{\text{hyp}}[\vec{a} | \vec{j}, J_M, A]$$

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

## The subsample approach

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

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

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

16 where  $\Theta$  represents the model parameters ( $\theta$  or  $(\theta, m)$ ). Equation (31) just follows from the definition of condi-  
 17 tional probabilities.

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

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

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

$$J_i := J - \sum_{k=1}^{i-1} n_k \quad (33)$$

3 This definition implies, for instance, that  $J_1 = J$ ,  $J_S = n_S$  and  $J_{S+1} = 0$ . For later convenience, we define the  
 4 partial datasets  $\vec{D}_i$ :

$$\vec{D}_i = (n_i, \dots, n_S) \quad (34)$$

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

7 With definition (33), (31) becomes

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

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

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

9 and

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

10 with

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

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

$$\hat{\Omega}(x|\theta, m, \vec{D}_{i+1}) = \Omega(x) F(x|\theta, m, \vec{D}_{i+1}) \quad (39)$$

12 with  $\Omega(x)$  given by (21) and  $F(x|\theta, m, \vec{D}_{i+1})$  defined in equation (D-6) in Appendix D. Comparing (23) and  
 13 (38) we can interpret (38) as having an abundance distribution  $\Omega(x)$  that is modified by a factor that takes into  
 14 account the subsample  $\vec{D}_{i+1}$ . We further note that (36) and (37) are even simpler when species are labelled: then  
 15 there is only  $S!$  in the denominator.

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

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

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

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

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

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

## Discussion

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

8 For the first time in neutral community ecology, the main results of two lines of research -  $E[S_n|\theta, m, J]$ , the  
 9 expected number of species with abundance  $n$  in a sample of size  $J$ , and  $P[\vec{D}|\theta, m, J]$ , the joint multivariate prob-  
 10 ability of observing  $S$  species with abundances  $n_1, n_2, \dots, n_S$  in a sample of size  $J$  - have been presented together  
 11 and related to one another. In the case without dispersal limitation ( $m = 1$ ),  $P[\vec{D}|\theta, J]$  can even be expressed in  
 12 terms of  $E[S_{n_i}|\theta, J_i]$  using subsamples  $\vec{D}_i$ , whereas in the case with dispersal limitation, this expression must  
 13 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  
 14  $P[\vec{D}|\theta, m, J]$ . Although this has been derived in the mathematically identical theory in population genetics for  
 15 the case without dispersal limitation, the derivation for the case with dispersal limitation is given here for the first  
 16 time. Relating expected values to multivariate distributions is important because it is much easier to write and  
 17 solve for stationarity dynamical one-dimensional models involving expected values (McKane *et al.* 2000, Val-  
 18 lade & Houchmandzadeh 2003, McKane *et al.* 2004) than it is for their corresponding multivariate distributions.  
 19 However, we emphasize that precisely these exact multivariate sampling distributions taken as likelihood functions  
 20 are actually needed to perform maximum likelihood estimation of model parameters (Etienne 2005) and sound  
 21 statistical model comparisons (Etienne & Olf 2005).

22 Moreover, our sampling framework has enabled us to show that the sampling distributions are valid for a meta-  
 23 community of any size  $J_M$ . In other words, two samples of equal size from two metacommunities of different  
 24 sizes  $J_{M,1}$  and  $J_{M,2}$  are characterized by exactly the same sampling distributions, as long as both metacommuni-

1 ties are described by the same biodiversity number ( $\theta_1 = \theta_2$ ). This has not been emphasized in previous work.  
2 This is important for two reasons. First, an already existing expression  $E[S_n|\theta, m, J_M, J]$  when  $J_M$  is finite  
3 (Vallade & Houchmandzadeh 2003) turns out to be incorrect. Alonso & McKane (2004), assuming Vallade &  
4 Houchmandzadeh (2003) to be correct, suggested that species-abundance data can be used to estimate the meta-  
5 community size and hence the speciation rate  $\nu$  because  $\theta := \frac{\nu(J_M-1)}{1-\nu}$  (Vallade & Houchmandzadeh 2003, Alonso  
6 & McKane 2004, Etienne 2005). The independence of metacommunity size that we have shown in this paper,  
7 however, implies that this is not possible. Second, since metacommunity size does not matter, we can safely as-  
8 sume infinite metacommunity size which simplifies our formulae, because we can use binomial sampling instead  
9 of hypergeometric sampling. We want to stress, however, that it is invariance under hypergeometric sampling that  
10 provided the basis for our sampling theory.

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

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

25 Neutral community theory as formulated by Hubbell (2001) can be seen as an extension of Ewens' (1972)  
26 theory into the ecological arena. This extension is far from trivial because Hubbell's main intuition is that, in  
27 addition to neutral (or ecological) drift, it is dispersal limitation that is the leading factor structuring ecological  
28 communities. All recent theoretical advances in neutral community theory based on Hubbell's (2001) formulation  
29 can now be translated back to population genetics to extend Ewens' work as "a dispersal-limited sampling theory

1 of selectively neutral alleles". With the dispersal-limited sampling distributions introduced in this work, we can  
2 not only examine whether a certain allelic polymorphism is maintained neutrally, but we can also easily estimate  
3 the amount of dispersal limitation (or degree of isolation) of the locality where this allelic polymorphism comes  
4 from. It also enables computation of the ages of alleles in dispersal-limited populations.

5 Concerning the evolutionary age of *species* (or, equivalently, species time-to-extinction), the neutral theory  
6 has been strongly criticized for yielding unrealistically old species (Lande *et al.* 2003, Nee 2005). However, this  
7 finding may depend more on other model assumptions than on the assumption of neutrality. For instance, Nee's  
8 (2005) estimates of species ages are based on Ewens' equilibrium model for fixed community size with  $\theta \rightarrow 0$  and  
9  $m = 1$ . Griffiths & Lessard (2005) recently presented a formula for any value of  $\theta$  that makes species ages already  
10 a few orders of magnitude smaller. Species ages might also be appreciably different if dispersal limitation is taken  
11 into account. Furthermore, non-equilibrium dynamics and fluctuations in community size may substantially affect  
12 effective community size and thereby the times scales of species origination. Also, even if species ages are better  
13 explained by non-neutral processes at evolutionary times scales, such as ecological succession (a process involving  
14 ecologically non-equivalent species interacting through non-neutral processes such as facilitation and hierarchical  
15 competition), the final mature community that we observe today may still be consistent with neutral dynamics. In  
16 sum, the use of species ages to falsify the neutral theory is rather premature.

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

25 We have expressed the local community as a sample from the larger regional metacommunity, a sample which  
26 may or may not be affected by dispersal limitation. In our expressions the metacommunity is purely regulated  
27 by speciation and extinction, and thus governed by the Ewens sampling formula, but this is not necessary. Our  
28 dispersal-limited hypergeometric distribution can also be applied to metacommunities that are structured according  
29 to other, even non-neutral, rules. Although at the local community level the dynamics are neutral, any differences

1 in species abundances due to (non-neutral) metacommunity structure propagate to this local level. This allows for  
2 a dispersal-limited sampling theory for non-neutral communities. A more exact but more challenging approach  
3 would be to replace the dispersal-limited hypergeometric distribution of equations (28) and (30) that assume local  
4 neutrality by a new dispersal-limited distribution that takes into account, at the local level, the same non-neutral  
5 factors controlling abundances in the metacommunity. This can potentially be done in essentially the same formal-  
6 ism we have presented here (possibly following suggestions in the population genetics literature (*e.g.* Wakeley &  
7 Takahashi 2004 and Slade & Wakeley 2005). Our expressions are however good approximations that are fully in  
8 line with the model assumptions on the time scale discussed above.

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

## Acknowledgements

17 We thank three anonymous reviewers, John Wakeley, Jérôme Chave and Han Olf for very constructive comments.  
18 D.A. thanks the support of the James S. McDonnell Foundation through a Centennial Fellowship to Mercedes  
19 Pascual.

## Supplementary material

20 The following material is available from <http://www.blackwellpublishing.com/products/journals/suppmat/ELE/???>.

21 **Appendix S1.pdf.** A pdf-file containing the following four appendices:

22 Appendix A. Derivation of equation (13).

23 Appendix B. The relation of the approximation (10) to the exact result (6).

- 1 Appendix C. Proof of the equality of (23) and (29).
- 2 Appendix D. Derivation of (36) and (37).
- 3 Appendix E. Derivation of (1) and (6) from (2) and (11).
- 4 Appendix F. A historical note on the origins of the binomial and hypergeometric distributions.

## References

- 5 Alonso, D. & A.J. McKane (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecology Letters* 7: 901-910.
- 6 Brown, J.H. & A. Kodric-Brown (1977). Turnover rate in insular biogeography: effect of immigration on extinction. *Ecology* 58: 445-449.
- 7
- 8 Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* 30: 101-110.
- 9
- 10 Dewdney, A.K. (1998). A general theory of the sampling process with applications to the "veil line". *Theoretical Population Biology* 54: 294-302.
- 11
- 12 Dewdney, A.K. (2000). A dynamical model of communities and a new species-abundance distribution. *Biological Bulletin* 35: 152-165.
- 13
- 14 Diserud, O.H. & S. Engen (2000). A general and dynamic species abundance model, embracing the lognormal and the gamma models. *American Naturalist* 155: 497-511.
- 15
- 16 Engen, S. & R. Lande (1996a). Population dynamic models generating the lognormal species abundance distribution. *Mathematical Biosciences* 132: 169-183.
- 17
- 18 Engen, S. & R. Lande (1996b). Population dynamic models generating the species abundance distributions of the Gamma type. *Journal of theoretical Biology* 178: 325-331.
- 19
- 20 Etienne, R.S. (2005). A new sampling formula for neutral biodiversity. *Ecology Letters* 8: 253-260.
- 21
- 22 Etienne, R.S. & H. Olf (2004a). How dispersal limitation shapes species - body size distributions in local communities. *American Naturalist* 163: 69-83.
- 23
- 24 Etienne, R.S. & H. Olf (2004b). A novel genealogical approach to neutral biodiversity theory. *Ecology Letters* 7: 170-175.
- 25
- 26 Etienne, R.S. & H. Olf (2005). Bayesian analysis of species-abundance data: assessing the relative importance of dispersal and niche-partitioning for the maintenance of biodiversity. *Ecology Letters* 8: 493-504.

- 1 Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* 3: 87-112.
- 2 Fisher, R.A., A.S. Corbet & C.B. Williams (1943). The relation between the number of species and the number of  
3 individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42-58.
- 4 Griffiths, R.C. & S. Lessard (2005). Ewens' sampling formula and related formulae: Combinatorial proofs, exten-  
5 sions to variable population size and applications to ages of alleles. *Theoretical Population Biology*. In press.
- 6 Grinnell, J. (1922). On the role of the accidental. *Auk* 39: 373-380.
- 7 Hanski, I. (1983). Coexistence of competitors in patchy environment. *Ecology* 64: 493-500.
- 8 He, F.L. (2005). Deriving a neutral model of species abundance from fundamental mechanisms of population  
9 dynamics *Functional Ecology* 19: 187-193.
- 10 Hubbell, S.P. (1997). A unified theory of biogeography and relative species abundance and its application to tropi-  
11 cal rain forests and coral reefs. *Coral Reefs* 16: S9-S21.
- 12 Hubbell, S.P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton, NJ: Princeton Uni-  
13 versity Press.
- 14 Johnson, N.L., S. Kotz & N. Balakrishnan (1997). *Discrete multivariate distributions*. New York, NY: Wiley.
- 15 Karlin, S. & J. McGregor (1972). Addendum to a paper of W. Ewens. *Theoretical Population Biology* 3: 113-116.
- 16 Lande, R., S. Engen & B.-E. Saether (2003). *Stochastic population dynamics in ecology and conservation*. Oxford  
17 Series in Ecology and Evolution. Oxford, U.K.: Oxford University Press.
- 18 Levins, R. & D. Culver. 1971. Regional coexistence of species and competition between rare species. *Proceedings*  
19 *of the National Academy of Science of the USA* 68: 1246-1248.
- 20 Loreau, M. & N. Mouquet (1999). Immigration and the maintenance of local species diversity. *American Naturalist*  
21 154: 427-440.
- 22 MacArthur, R.H. & E.O. Wilson (1967). *Island biogeography*. Princeton, NJ: Princeton University Press.
- 23 McGill, B.J. (2003). A test of the unified neutral theory of biodiversity. *Nature* 422: 881-885.
- 24 McKane, A.J., D. Alonso & R.V. Solé (2000). A mean field stochastic theory for species rich assembled commu-  
25 nities. *Physical Review E* 62: 8466-8484.
- 26 McKane, A.J., D. Alonso & R.V. Solé (2004). Analytic solution of Hubbell's model of local community dynamics.  
27 *Theoretical Population Biology* 65: 67-73.
- 28 Moran, P.A.P. (1958). Random processes in genetics. *Proceedings of the Cambridge Philosophical Society* 54:  
29 60-71.

- 1 Moran, P.A.P. (1962). *Statistical processes of evolutionary theory*. Oxford, U.K.: Clarendon Press.
- 2 Nee, S. (2005). The neutral theory of biodiversity: do the numbers add up? *Functional Ecology* 19: 173-176.
- 3 Pielou, E.C. (1969). *An introduction to mathematical ecology*. New York, N.Y.: Wiley.
- 4 Slade, P.F. & J. Wakeley (2005). The structured ancestral selection graph and the many-demes limit. *Genetics* 169:  
5 1117–1131.
- 6 Solé, R.V., D. Alonso & A.J. McKane (2000). Scaling in a network model of multispecies communities. *Physica*  
7 A 286: 337–344.
- 8 Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology* 75: 2-16.
- 9 Vallade, M. & B. Houchmandzadeh (2003). Analytical solution of a neutral model of biodiversity. *Physical Review*  
10 E 68: 061902.
- 11 Volkov, I., J.R. Banavar, S.P. Hubbell & A. Maritan (2003). Neutral theory and relative species abundance in  
12 ecology. *Nature* 424: 1035-1037.
- 13 Wakeley, J. (1998). Segregating sites in Wright's island model. *Theoretical Population Biology* 53: 166-175.
- 14 Wakeley, J. (1999). Non-equilibrium migration in human history. *Genetics* 153: 1863–1871.
- 15 Wakeley, J. & N. Aliacar (2001). Gene genealogies in a metapopulation. *Genetics* 159: 893-905. Corrigendum in  
16 *Genetics* 160:1263 (2002).
- 17 Wakeley, J. & T. Takahashi (2004). The many-demes limit for selection and drift in a subdivided population,  
18 *Theoretical Population Biology* 66: 83–91.
- 19 Watterson, G.A. (1974). Models for the logarithmic species abundance distribution. *Theoretical Population Biology*  
20 6: 217-250.
- 21 Wright, S. (1931). Evolution in Mendelian populations. *Genetics* 16: 97-159.

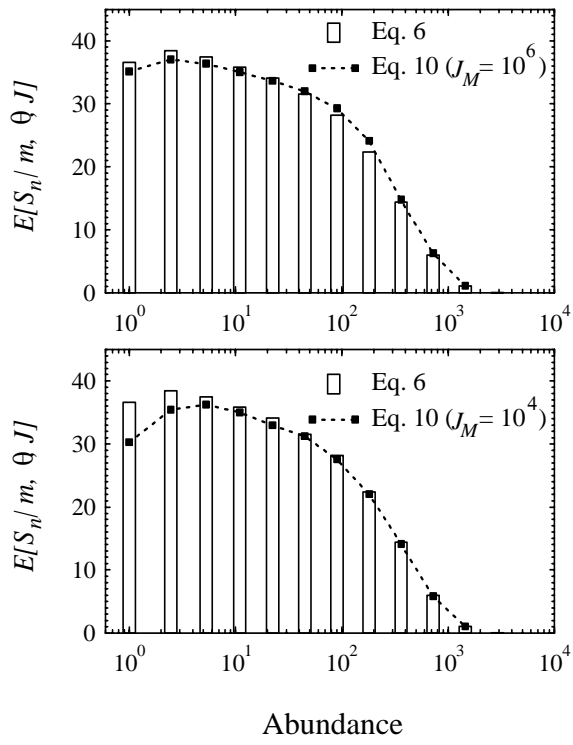
1 **Table I.** Overview of the analytical results for the species-abundance distribution of a local sample in neu-  
2 tral community theory. Let the entire metacommunity consist of  $J_M$  individuals and let the sample consist of  
3  $J$  individuals of  $S$  different species with abundances  $n_1, n_2, \dots, n_S$ . Let us denote this sample by  $\vec{D}$ , that is,  
4  $\vec{D} = (n_1, n_2, \dots, n_S)$ ;  $\Phi_j$  is the number of species in the sample that have abundance  $j$ . The model parameters are  
5 the fundamental biodiversity number  $\theta$ , which is a measure of the regional diversity, and the fundamental dispersal  
6 number  $I$ . The immigration probability  $m$  is a function of  $I$ , see (8),  $m = \frac{I}{I+J-1}$ . The quantities  $E[S_n|\theta, J]$   
7 and  $E[S_n|\theta, m, J_M, J]$  represent the expected number of species with abundance  $n$  in the cases without dispersal  
8 limitation ( $I = \infty$ , i.e.  $m = 1$ ) and with dispersal limitation ( $I < \infty$ , i.e.  $m < 1$ ) respectively, according to the  
9 neutral model.  $\Omega(x)dx$ , where  $\Omega(x)$  is given by (21), is the number of species with relative abundance between  $x$   
10 and  $x + dx$  in the metacommunity (regional species pool);  $\widehat{\Omega}[x|\theta, m, \vec{D}_{i+1}]dx$  is a modified version of that, see  
11 (39). The probabilities  $P[\vec{D}|\theta, J]$  and  $P[\vec{D}|\theta, m, J]$  represent the joint multivariate probability of observing  $S$   
12 species with abundances  $n_1, n_2, \dots, n_S$  in a sample of  $J$  individuals, again for the cases without and with dispersal  
13 limitation respectively.  $P_{\text{bin}}[n|x, J]$ ,  $P_{\text{hyp}}[n|j, J_M, J]$ ,  $P_{\text{bin}}^{\text{DL}}[n|m, x, J]$  and  $P_{\text{hyp}}^{\text{DL}}[n|m, j, J_M, J]$  are the binomial,  
14 hypergeometric, dispersal-limited binomial and dispersal-limited hypergeometric distributions respectively, given  
15 in (20), (15), (24) and (28). These four distributions are the distributions by which the expressions for the regional  
16 species-abundance distribution must be weighed to obtain the expressions for the local sample. The binomial dis-  
17 tribution  $P_{\text{bin}}[n|x, J]$  and the hypergeometric distribution  $P_{\text{hyp}}[n|j, J_M, J]$  are the limits of the dispersal-limited  
18 hypergeometric distribution  $P_{\text{hyp}}^{\text{DL}}[n|m, j, J_M, J]$  for  $m \rightarrow 1$  in the cases  $J_M \rightarrow \infty$  and  $J_M < \infty$  respectively.

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

## Figure captions

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

# Figures



1

2 Figure 1.

## Appendix A. Derivation of equation (13)

- 1 Here we derive the integral notation of (11). We first note that  $\bar{s}(n_i, a_i) = 0$  for  $a_i > n_i$ . Therefore, (11) can be  
 2 written as

$$\begin{aligned}
 P \left[ \vec{D} | \theta, m, J \right] &= \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(\theta)_A} = \\
 &= \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \left( \prod_{i=1}^S \sum_{a_i=1}^{n_i} \bar{s}(n_i, a_i) \frac{\bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} \right) \frac{I^A}{(\theta)_A} = \\
 &= \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{a_1=1}^{n_1} \dots \sum_{a_S=1}^{n_S} P_1 \frac{I^{A_1}}{(\theta)_{A_1}} \tag{A-1}
 \end{aligned}$$

- 3 where

$$A_i := \sum_{j=i}^S a_j \tag{A-2}$$

- 4 (so  $A_1 = A$ ) and

$$P_i := \prod_{j=i}^S \bar{s}(n_j, a_j) \frac{\bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} = \prod_{j=i}^S \bar{s}(n_j, a_j) \frac{\Gamma(a_j)}{\Gamma(n_j)} \tag{A-3}$$

- 5 We write the first summation of (A-1) in terms of  $P_2$ ,  $A_2$  and  $a_1$ :

$$\sum_{a_1=1}^{n_1} P_1 \frac{I^{A_1}}{(\theta)_{A_1}} = \sum_{a_1=1}^{n_1} P_1 \frac{I^{A_2+a_1}}{(\theta)_{A_2+a_1}} = P_2 \sum_{a_1=1}^{n_1} \bar{s}(n_1, a_1) \frac{\Gamma(a_1)}{\Gamma(n_1)} \frac{I^{A_2+a_1}}{(\theta)_{A_2+a_1}} \tag{A-4}$$

- 6 Expressing  $(\theta)_{A_2+a_1}$  in terms of Gamma functions and noting that  $\Gamma(n_1) = (n_1 - 1)!$ , we obtain after rearranging  
 7 terms

$$\sum_{a_1=1}^{n_1} P_1 \frac{I^{A_1}}{(\theta)_{A_1}} = P_2 I^{A_2} \frac{1}{(n_1 - 1)! \Gamma(\theta + A_2)} \sum_{a_1=1}^{n_1} \bar{s}(n_1, a_1) I^{a_1} \frac{\Gamma(a_1) \Gamma(\theta + A_2)}{\Gamma(\theta + A_2 + a_1)} \tag{A-5}$$

- 8 The last quotient is the Beta function that can be written in its integral form:

$$\sum_{a_1=1}^{n_1} P_1 \frac{I^{A_1}}{(\theta)_{A_1}} = P_2 I^{A_2} \frac{1}{(n_1 - 1)! \Gamma(\theta + A_2)} \sum_{a_1=1}^{n_1} \bar{s}(n_1, a_1) I^{a_1} \int_0^1 x_1^{a_1} (1 - x_1)^{A_2} \frac{(1 - x_1)^{\theta-1}}{x_1} dx_1 \tag{A-6}$$

- 9 Changing the order of summation and integration and using Pochhammer notation leads to:

$$\begin{aligned}
 \sum_{a_1=1}^{n_1} P_1 \frac{I^{A_1}}{(\theta)_{A_1}} &= \int_0^1 \frac{1}{(n_1 - 1)!} P_2 \frac{I^{A_2}}{(\theta)_{A_2}} (1 - x_1)^{A_2} \frac{(1 - x_1)^{\theta-1}}{x_1} \sum_{a_1=1}^{n_1} \bar{s}(n_1, a_1) (I x_1)^{a_1} dx_1 = \\
 &= \frac{1}{(n_1 - 1)!} \int_0^1 P_2 \frac{(I(1 - x_1))^{A_2} (1 - x_1)^{\theta-1}}{(\theta)_{A_2} x_1} (I x_1)_{n_1} dx_1 \tag{A-7}
 \end{aligned}$$

- 10 This means that the summation of  $P_1 \frac{I^{A_1}}{(\theta)_{A_1}}$  over  $a_1$  is written in terms of an integral over  $P_2 \frac{(I(1-x_1))^{A_2}}{(\theta)_{A_2}}$  and some  
 11 additional terms that do not depend on any  $a_i$ . More generally,

$$\sum_{a_i=1}^{n_i} P_i \frac{\left( I \prod_{k=1}^{i-1} (1 - x_k) \right)^{A_i}}{(\theta)_{A_i}} = \frac{1}{(n_i - 1)!} \int_0^1 P_{i+1} \frac{\left( I \prod_{k=1}^i (1 - x_k) \right)^{A_{i+1}}}{(\theta)_{A_{i+1}}} \left( I x_i \prod_{k=1}^{i-1} (1 - x_k) \right)_{n_i} \frac{(1 - x_i)^{\theta-1}}{x_i} dx_i \tag{A-8}$$

1 for  $i = 1 \dots S$ . The  $S$  summations in (A-1) require repeated application of (A-8), from  $i = 1$  to  $i = S$ , which  
 2 yields

$$P \left[ \vec{D} | \theta, m, J \right] = \frac{J!}{\prod_{i=1}^S n_i! \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \int_0^1 \dots \int_0^1 \prod_{i=1}^S \left( I x_i \prod_{k=1}^{i-1} (1 - x_k) \right)_{n_i} \frac{(1 - x_i)^{\theta-1}}{x_i} dx_1 \dots dx_S \quad (\text{A-9})$$

1 which is (13) with (14).

## Appendix B. The relation of the approximation (10) to the exact result (6)

A rigorous expansion of the approximate result (10) in Vallade & Houchmandzadeh (2003) in terms of  $(1/J_M)$  powers can be written after some algebra (Alonso & McKane, *unpublished*):

$$S(n) = S_0(n) + \mathcal{O}\left(\frac{1}{J_M}\right) \quad (\text{B-1})$$

where  $S(n)$  is given by (10) and  $S_0(n)$  is given by (6). This expansion confirms that Vallade & Houchmandzadeh's result (10) converges to (6) when  $J_M$  tends to infinity. The existence of this expansion suggested a quantitative path to independently estimate metacommunity sizes and biodiversity numbers (and hence also speciation rates) from species-abundance data. In the main text we have shown that expected sampling abundances do not depend on metacommunity size. Therefore, that initial hope of independent estimation must be abandoned. In addition, Vallade & Houchmandzadeh's approximation is not very useful anyway, because the integral in 6 can be evaluated much faster and more robustly than the sum in (10).

The complete proof for the expansion given in (B-1) will be given elsewhere, but an easy argument to show to what extent (10) and (6) converge to each other as  $J_M$  increases goes as follows. The sum in (10) corresponds to an approximate discretization of the exact integral result given by (6). Let us write (6) as

$$S_0(n) = \int_0^1 G(x) dx \quad (\text{B-2})$$

where  $G(x) = P_{\text{bin}}^{\text{DL}}[n|m, x, J] \Omega(x)$ . We divide the interval  $(0, 1)$  in  $J_M$  points to obtain

$$S_0(n) \approx \sum_{j=1}^{J_M} G(x_j) \Delta x \quad (\text{B-3})$$

where  $\Delta x = 1/J_M$ ,  $x_j = j \Delta x$  and

$$G(x_j) = P_{\text{bin}}^{\text{DL}}\left[n|m, \frac{j}{J_M}, J\right] \Omega\left(\frac{j}{J_M}\right) \quad (\text{B-4})$$

Hence  $S_0(n)$  can be approximated by the following sum:

$$S_0(n) \approx \sum_{j=1}^{J_M} P_{\text{bin}}^{\text{DL}}\left[n|m, \frac{j}{J_M}, J\right] \frac{\theta}{j} \left(1 - \frac{j}{J_M}\right)^{\theta-1} \quad (\text{B-5})$$

Compare this to Vallade & Houchmandzadeh's (2003) formula, given by (10) but repeated here for easier comparison:

$$S(n) = \sum_{j=1}^{J_M} P_{\text{bin}}^{\text{DL}}\left[n|m, \frac{j}{J_M}, J\right] E[S_j|\theta, J_M] \quad (\text{B-6})$$

2 Because each term  $E[S_j|\theta, J_M]$  can be approximated by (Alonso & McKane 2004)

$$E[S_j|\theta, J_M] = \frac{\theta}{j} \left(1 - \frac{j}{J_M}\right)^{\theta-1} + \mathcal{O}\left(\frac{1}{J_M^2}\right), \quad (\text{B-7})$$

3 we conclude that (10), apart from vanishingly small terms, corresponds to the discretization (B-5) in  $J_M$  points,

1 which by definition converges to the integral in the limit of an infinite number of points.

## Appendix C. Proof of the equality of (23) and (29)

2 Here we show that equations (23) and (29) are identical:

$$\begin{aligned}
E[S_n|\theta, m, J] &= \frac{\theta}{(I)_J} \binom{J}{n} \int_0^1 (Ix)_n (I(1-x))_{J-n} \frac{(1-x)^{\theta-1}}{x} dx = \\
&= \frac{\theta}{(I)_J} \binom{J}{n} \sum_{j=1}^n \sum_{k=1}^{J-n} \bar{s}(n, j) \bar{s}(J-n, k) I^{j+k} \frac{\Gamma(j) \Gamma(k+\theta)}{\Gamma(j+k+\theta)} = \\
&= \frac{\theta}{(I)_J} \binom{J}{n} \sum_{A=1}^J \sum_{a=1}^n \bar{s}(n, a) \bar{s}(J-n, A-a) I^A \frac{\Gamma(a) \Gamma(A-a+\theta)}{\Gamma(A+\theta)} = \\
&= \binom{J}{n} \sum_{A=1}^J \sum_{a=1}^n \bar{s}(n, a) \bar{s}(J-n, A-a) \frac{I^A}{(I)_J} \frac{\Gamma(A+1-a)}{\Gamma(A+1)} a \Gamma(a) \frac{\theta}{a} \frac{\Gamma(A+1)}{\Gamma(A+1-a)} \frac{\Gamma(A+\theta-a)}{\Gamma(A+\theta)} = \\
&= \binom{J}{n} \sum_{A=1}^J \sum_{a=1}^n \bar{s}(n, a) \bar{s}(J-n, A-a) \frac{I^A}{(I)_J} \frac{1}{\binom{A}{a}} E[S_a|\theta, A] = \\
&= \sum_{j=1}^{J_M} \sum_{A=1}^J \sum_{a=1}^n \frac{\binom{J}{n}}{\binom{A}{a}} \bar{s}(n, a) \bar{s}(J-n, A-a) \frac{I^A}{(I)_J} P_{\text{hyp}}[a|j, J_M, A] E[S_j|\theta, J_M] = \\
&= \sum_{j=1}^{J_M} P_{\text{hyp}}^{\text{DL}}[n|m, j, J_M, J] E[S_j|\theta, J_M] \tag{C-1}
\end{aligned}$$

3 where in the first line we have used the polynomial form of the Pochhammer symbol (3) and the definition of the

1 beta function (4).

## Appendix D. Derivation of (36) and (37)

### No dispersal limitation

2 Without dispersal limitation, each probability  $P[n_i|\theta, J_i]$  in (35) can be written as

$$\begin{aligned} P[n_i|\theta, J_i] &= \frac{P[n_i, \dots, n_S|\theta, J_i]}{P[n_{i+1}, \dots, n_S|\theta, J_{i+1}]} = \frac{\frac{J_i!}{\prod_{k=i}^S n_k \prod_{j=1}^{J_i} \Phi_j!} \frac{\theta^{S-(i-1)}}{(\theta)^{J_i}}}{\frac{J_{i+1}!}{\prod_{k=i+1}^S n_k \prod_{j=1}^{J_{i+1}} (\Phi_j - \delta_{jn})!} \frac{\theta^{S-i}}{(\theta)^{J_{i+1}}}} = \\ &= \frac{1}{\Phi_{n_i}} \frac{\theta}{n_i} \frac{J_i!}{(J_i - n_i)!} \frac{(\theta)^{J_i - n_i}}{(\theta)^{J_i}} = \frac{E[S(n_i)|\theta, J_i]}{\Phi_{n_i}} \end{aligned} \quad (\text{D-1})$$

3 Here,  $\delta_{jn}$  is Kronecker's delta which is equal to 1 if  $j = n$  and 0 otherwise. Substituting this in (35) leads to (36).

4 One can easily check that substituting (1) in (36) leads back to the multivariate probability distribution (2).

### Dispersal limitation

5 When dispersal is limited,  $P[n_i|\theta, m, J_i]$  in (35) is given by

$$\begin{aligned} P[n_i|\theta, m, J_i] &= \frac{P[n_i, \dots, n_S|\theta, J_i]}{P[n_{i+1}, \dots, n_S|\theta, J_{i+1}]} = \\ &= \frac{\frac{J_i!}{\prod_{k=1}^{S_i} n_k \prod_{j=1}^{J_i} \Phi_j!} \frac{\theta^{S-(i-1)}}{(I)^{J_i}} \sum_{A=S-(i-1)}^{J_i} K(\vec{D}_i, A) \frac{I^A}{(\theta)_A}}{\frac{J_{i+1}!}{\prod_{k=1}^{S_{i+1}} n_k \prod_{j=1}^{J_{i+1}} (\Phi_j - \delta_{jn})!} \frac{\theta^{S-i}}{(I)^{J_{i+1}}} \sum_{A=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, A) \frac{I^A}{(\theta)_A}} \end{aligned} \quad (\text{D-2})$$

6 The key step is now to realize that  $K(\vec{D}_i, A)$  is a combination of terms involving  $\frac{\bar{s}(n_i, j)\bar{s}(j, 1)}{\bar{s}(n_i, 1)}$  and  $K(\vec{D}_{i+1}, k)$

7 where  $k = A - j$  (Etienne 2005). Equation (D-2) then simplifies to, also cancelling equal terms in numerator and

8 denominator,

$$P[n_i|\theta, m, J_i] = \frac{1}{\Phi_{n_i}} \frac{\theta}{n_i} \frac{J_i!}{(J_i - n_i)!} \frac{(I)^{J_i - n_i}}{(I)^{J_i}} \frac{\sum_{j=1}^{n_i} \frac{\bar{s}(n_i, j)\bar{s}(j, 1)}{\bar{s}(n_i, 1)} I^j \sum_{k=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, k) \frac{I^k}{(\theta)_{j+k}}}{\sum_{A=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, A) \frac{I^A}{(\theta)_A}} \quad (\text{D-3})$$

9 We can now use the identities of the Stirling numbers of the first kind and the definition of the Beta function (4) to

1 find

$$\begin{aligned} P[n_i|\theta, m, J_i] &= \frac{1}{\Phi_{n_i}} \theta \frac{J_i!}{n_i! (J_i - n_i)!} \frac{(I)^{J_i - n_i}}{(I)^{J_i}} \frac{\sum_{j=1}^{n_i} \bar{s}(n_i, j) I^j \sum_{k=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, k) \frac{I^k}{(\theta)_k} \frac{\Gamma(j)\Gamma(\theta+k)}{\Gamma(\theta+j+k)}}{\sum_{A=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, A) \frac{I^A}{(\theta)_A}} = \\ &= \frac{1}{\Phi_{n_i}} \theta \binom{J_i}{n_i} \frac{(I)^{J_i - n_i}}{(I)^{J_i}} \frac{\sum_{j=1}^{n_i} \bar{s}(n_i, j) I^j \sum_{k=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, k) \frac{I^k}{(\theta)_k} \int_0^1 x^{j-1} (1-x)^{k+\theta-1} dx}{\sum_{A=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, A) \frac{I^A}{(\theta)_A}} = \\ &= \frac{1}{\Phi_{n_i}} \frac{\theta}{(I)^{J_i}} \binom{J_i}{n_i} (I)^{J_{i+1}} \frac{\int_0^1 \sum_{j=1}^{n_i} \bar{s}(n_i, j) (Ix)^j \sum_{k=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, k) \frac{(I(1-x))^k (1-x)^{\theta-1}}{(\theta)_k} dx}{\sum_{A=S-i}^{J_{i+1}} K(\vec{D}_{i+1}, A) \frac{I^A}{(\theta)_A}} \end{aligned} \quad (\text{D-4})$$

2 Using (3), multiplying the integrand by  $\frac{(I(1-x))_{J_{i+1}}}{(I(1-x))_{J_{i+1}}}$  and rearranging terms then leads to

$$P [n_i | \theta, m, J_i] = \frac{1}{\Phi_{n_i}} \frac{\theta}{(I)_{J_i}} \binom{J_i}{n_i} \int_0^1 (Ix)_{n_i} (I(1-x))_{J_i - n_i} \frac{(1-x)^{\theta-1}}{x} \frac{\sum_{k=S-i}^{J_{i+1}} \frac{K(\vec{D}_{i+1}, k)}{(\theta)_k} \frac{(I(1-x))^k}{(I(1-x))_{J_{i+1}}}}{\sum_{A=S-i}^{J_{i+1}} \frac{K(\vec{D}_{i+1}, A)}{(\theta)_A} \frac{I^A}{(I)_{J_{i+1}}}} dx \quad (\text{D-5})$$

3 Defining

$$F \left( x | \theta, m, \vec{D}_{i+1} \right) := \frac{\sum_{k=S-i}^{J_{i+1}} \frac{K(\vec{D}_{i+1}, k)}{(\theta)_k} \frac{(I(1-x))^k}{(I(1-x))_{J_{i+1}}}}{\sum_{A=S-i}^{J_{i+1}} \frac{K(\vec{D}_{i+1}, A)}{(\theta)_A} \frac{I^A}{(I)_{J_{i+1}}}} \quad (\text{D-6})$$

1 and using (39) and (21), we obtain our end result (37).

## Appendix E. Derivation of (1) and (6) from (2) and (11)

### No dispersal limitation

2 Equation (1) can be derived from (2) as follows. From (41) we get

$$\begin{aligned}
 E[S_n|\theta, J] &= \sum_{\Phi_n=1}^J \sum_{\{\vec{D}|\Phi_n\}} \Phi_n \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(\theta)_J} = \\
 &= \sum_{\Phi_n=1}^J \sum_{\{\vec{D}|\Phi_n\}} \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J (\Phi_j - \delta_{jn})!} \frac{\theta^S}{(\theta)_J} \tag{E-1}
 \end{aligned}$$

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

$$\begin{aligned}
 E[S_n|\theta, J] &= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{(\theta)_{J-n}}{(\theta)_J} \sum_{\Phi'_n=0}^{J-n} \sum_{\{\vec{D}|\Phi'_n\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^{J-n} \Phi'_j!} \frac{\theta^{S-1}}{(\theta)_{J-n}} = \\
 &= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{(\theta)_{J-n}}{(\theta)_J} \sum_{\{\vec{D}\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^{J-n} \Phi'_j!} \frac{\theta^{S-1}}{(\theta)_{J-n}} \tag{E-2}
 \end{aligned}$$

6 The sum on the right hand side is the sum of the probabilities of over all possible datasets with sample size  $J - n$ ,  
 7 which, evidently, equals unity, and after expressing Pochhammer symbols as quotients of gamma functions, we  
 8 obtain (1).

9 Also, the expected number of species (of any abundance) can be calculated as follows:

$$\begin{aligned}
 E[S|\theta, J] &= \sum_{S=1}^J SP[S|\theta, J] = \sum_{S=1}^J S \bar{s}(J, S) \frac{\theta^S}{(\theta)_J} = \\
 &= \theta \sum_{S=1}^J \bar{s}(J, S) \frac{S \theta^{S-1}}{(\theta)_J} = \theta \sum_{S=1}^J \bar{s}(J, S) \frac{1}{(\theta)_J} \frac{d}{d\theta} \theta^S = \\
 &= \theta \frac{d}{d\theta} \left( \sum_{S=1}^J \bar{s}(J, S) \frac{\theta^S}{(\theta)_J} \right) - \theta \sum_{S=1}^J \bar{s}(J, S) \theta^S \frac{d}{d\theta} \frac{1}{(\theta)_J} = -(\theta)_J \frac{d}{d\theta} \frac{1}{(\theta)_J} = \\
 &= \theta (\Psi(\theta + J) - \Psi(\theta)) = \sum_{i=1}^J \frac{\theta}{\theta + i - 1} \tag{E-3}
 \end{aligned}$$

10 where  $\Psi(x)$  is the digamma function or psi function,  $\Psi(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{1}{\Gamma(x)} \frac{d}{dx} \Gamma(x)$  and we have used the  
 11 identity  $\sum_{S=1}^J \bar{s}(J, S) \frac{\theta^S}{(\theta)_J} = 1$ .

## Dispersal limitation

2 When dispersal is limited, with (41) we can derive (6) from (11):

$$\begin{aligned}
E[S_n|\theta, m, J] &= \sum_{\Phi_n=0}^S \Phi_n \sum_{\{\vec{D}|\Phi_n\}} P[\vec{D}|\theta, m, J] = \\
&= \sum_{\Phi_n=1}^S \sum_{\{\vec{D}|\Phi_n\}} \Phi_n \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(\theta)_A} = \\
&= \sum_{\Phi_n=1}^S \sum_{\{\vec{D}|\Phi_n\}} \frac{J!}{\prod_{i=1}^S n_i \prod_{j=1}^J (\Phi_j - \delta_{jn})!} \frac{\theta^S}{(I)_J} \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(\theta)_A} \quad (\text{E-4})
\end{aligned}$$

3 As in the case without dispersal limitation, we define  $\Phi'_j := \Phi_j - \delta_{jn}$  and we work towards a sum of probabilities  
4 of observing exactly  $\Phi_n - 1$  species with abundance  $n$  in a sample of size  $J - n$  (the corresponding dataset is  
5 called  $\vec{D}'$ ),

$$\begin{aligned}
E[S_n|\theta, m, J] &= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \sum_{\Phi'_n=0}^{S-1} \sum_{\{\vec{D}'|\Phi'_n\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi'_j!} \theta^{S-1} \sum_{A=S}^J K(\vec{D}, A) \frac{I^A}{(\theta)_A} = \\
&= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \sum_{\Phi'_n=0}^{S-1} \sum_{\{\vec{D}'|\Phi'_n\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi'_j!} \theta^{S-1} \sum_{j=1}^n \frac{\bar{s}(n, j) \bar{s}(j, 1)}{\bar{s}(n, 1)} I^j \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{I^k}{(\theta)_{j+k}} = \\
&= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \sum_{\{\vec{D}'\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi'_j!} \theta^{S-1} \sum_{j=1}^n \frac{\bar{s}(n, j) \bar{s}(j, 1)}{\bar{s}(n, 1)} I^j \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{I^k}{(\theta)_k} \frac{(\theta)_k}{(\theta)_{j+k}} = \\
&= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \sum_{\{\vec{D}'\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi'_j!} \theta^{S-1} \sum_{j=1}^n \frac{\bar{s}(n, j)}{\bar{s}(n, 1)} I^j \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{I^k}{(\theta)_k} \frac{\Gamma(j) \Gamma(\theta + k)}{\Gamma(\theta + j + k)} \quad (\text{E-5})
\end{aligned}$$

6 The term with Gamma functions can be written as a Beta function, and the Beta function can be written in its  
7 integral notation:

$$\begin{aligned}
E[S_n|\theta, m, J] &= \frac{\theta}{n} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \times \\
&\times \sum_{\{\vec{D}'\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi'_j!} \theta^{S-1} \sum_{j=1}^n \frac{\bar{s}(n, j)}{\bar{s}(n, 1)} I^j \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{I^k}{(\theta)_k} \int_0^1 x^{j-1} (1-x)^{k+\theta-1} dx \quad (\text{E-6})
\end{aligned}$$

2 Changing the order of integration and summation, and using (3), we obtain

$$\begin{aligned}
E[S_n|\theta, m, J] &= \frac{\theta}{n!} \frac{J!}{(J-n)!} \frac{1}{(I)_J} \times \\
&\times \int_0^1 \sum_{\{\vec{D}'\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi_j!} \theta^{S-1} \sum_{j=1}^n \bar{s}(n, j) (Ix)^j \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{(I(1-x))^k}{(\theta)_k} \frac{(1-x)^{\theta-1}}{x} dx \\
&= \theta \binom{J}{n} \frac{1}{(I)_J} \times \\
&\times \int_0^1 (Ix)_n (I(1-x))_{J-n} \frac{(1-x)^{\theta-1}}{x} \sum_{\{\vec{D}'\}} \frac{(J-n)!}{\prod_{i=1}^{S-1} n_i \prod_{j=1}^J \Phi_j!} \frac{\theta^{S-1}}{(I(1-x))_{J-n}} \sum_{k=S-1}^{J-n} K(\vec{D}', k) \frac{(I(1-x))^k}{(\theta)_k} dx \\
&= \theta \binom{J}{n} \frac{1}{(I)_J} \int_0^1 (Ix)_n (I(1-x))_{J-n} \frac{(1-x)^{\theta-1}}{x} dx \tag{E-7}
\end{aligned}$$

3 where in the last line we have used the fact that the sum of the probabilities of all possible datasets equals unity.

1 Also, the expected number of species (of any abundance) can be calculated as follows:

$$\begin{aligned}
E[S|\theta, m, J] &= \sum_{S=1}^J SP[S|\theta, m, J] = \sum_{S=1}^J S \sum_{A=S}^J \bar{s}(J, A) \frac{I^A}{(I)_J} \bar{s}(A, S) \frac{\theta^S}{(\theta)_A} = \\
&= \sum_{A=1}^J \bar{s}(J, A) \frac{I^A}{(I)_J} \sum_{S=1}^A S \bar{s}(A, S) \frac{\theta^S}{(\theta)_A} = \\
&= \sum_{A=1}^J \bar{s}(J, A) \frac{I^A}{(I)_J} \sum_{i=1}^A \frac{\theta}{\theta+i-1} = \\
&= \sum_{i=1}^J \frac{\theta}{\theta+i-1} \sum_{A=i}^J \bar{s}(J, A) \frac{I^A}{(I)_J} = \\
&= \sum_{i=0}^{J-1} \frac{\theta}{(\theta+i)(I)_J} \sum_{A=i+1}^J \bar{s}(J, A) I^A \tag{E-8}
\end{aligned}$$

## Appendix F. A historical note on the origins of the binomial and hypergeometric distributions

2 The first occurrence of “binomial distribution” is found in Yule (1911) on p. 305: “The binomial distribution only  
3 becomes approximately normal when  $n$  is large, and this limitation must be remembered in applying the table  
4 to cases in which the distribution is strictly binomial”. Fisher (1925) adopted the distribution (part III, section  
5 18). Although, the name for the distribution relatively new, the distribution itself has been studied since Bernoulli  
6 (1713, part 1.)

7 The earliest use of “hypergeometric distribution” appears in the title of Gonin (1936). Again, the name is  
8 relatively recent but the distribution itself is already suggested in Problem IV of Huygens (1657, p. 12). He and  
9 several other mathematicians solved the problem, whereas Bernoulli and de Moivre gave solutions for the general  
10 case (Hald 2003). At the end of the 19th. century Pearson (1899) wrote a paper in which he considered fitting the  
11 distribution (given by the “hypergeometrical series”) to data.

12 This brief historical note is based on <http://members.aol.com/jeff570/mathword.html>.

### References

- 13 Bernoulli, J. (1713). *Ars Conjectandi*. Basel, Switzerland. English translation available at [http://cerebro.xu.edu/  
14 math/Sources/JakobBernoulli/ars\\_sung/ars\\_sung.html](http://cerebro.xu.edu/math/Sources/JakobBernoulli/ars_sung/ars_sung.html).
- 15 Fisher, R.A. (1925). *Statistical methods for research workers*. London, U.K.: Oliver & Boyd. Available at [http://  
16 psychclassics.yorku.ca/Fisher/Methods/index.htm](http://psychclassics.yorku.ca/Fisher/Methods/index.htm)
- 17 Gonin, H.T. (1936). The use of factorial moments in the treatment of the hypergeometric distribution and in tests  
18 for regression. *Philosophical Magazine* 7: 215-226.
- 19 Hald (2003). A history of probability and statistics and their applications before 1750. New York, N.Y.: Wiley-  
20 Interscience.
- 21 Huygens, C. (1657). *De ratiociniis in ludo aleae*. Reprint of an English translation available at [http://www.leidenuniv.  
22 nl/fsw/verduin/stathist/huygens/huyg1714p.pdf](http://www.leidenuniv.nl/fsw/verduin/stathist/huygens/huyg1714p.pdf)
- 23 Pearson, K. (1899). On certain properties of the hypergeometrical series, and on the fitting of such series to obser-  
24 vation polygons in the theory of chance. *Philosophical Magazine* 47: 236-246.
- 529 Yule, G.U. (1911). *An introduction to the theory of statistics*. London, U.K.: Charles Griffin & Co. Ltd.