

## LETTER

## Hubbell's fundamental biodiversity parameter and the Simpson diversity index

Fangliang He\* and Xin-Sheng Hu

Department of Renewable Resources, University of Alberta, 751 General Service Building, Edmonton, Alberta, T6G 2H1 Canada

\*Correspondence: E-mail: fhe@ualberta.ca

### Abstract

Central to Hubbell's neutral theory of biodiversity is a universal, dimensionless fundamental biodiversity parameter that is the product of community size and speciation rate. One of the most important discoveries of Hubbell's theory is that the species-abundance distribution and the species–area relationship of the neutral metacommunity is completely determined by this fundamental biodiversity parameter, although the diversity patterns of the local community are collectively determined by the biodiversity parameter and migration. Using the relative abundance of species and following the concept of heterozygosity of population genetics, here we developed an analytical relationship between this biodiversity parameter and the well-known Simpson diversity index. This relationship helps bridge the evolutionary aspect of biodiversity to the ecological and statistical aspect of the diversity. The relationship between these two parameters suggests that diversity patterns of the metacommunity can also be equally described by the Simpson index. This relationship provides an alternative approach to interpret and estimate the fundamental biodiversity parameter for the metacommunity.

### Keywords

Metacommunity, neutral theory, population genetics, species-abundance distribution, species–area curve, the Simpson index.

*Ecology Letters* (2005) 8: 386–390

### INTRODUCTION

Mutation, selection, migration and genetic drift are the primary evolutionary forces that affect the population dynamics of a species (Wright 1955). When the selection effect is negligible or much smaller than the effect of genetic drift, the distribution of an allele in a population displays a pure neutral process; otherwise selection may play a decisive role in determining the fate of the allele. Hubbell (1997, 2001) extends this theory of population genetics to investigate the species assembly rules in community ecology. Analogous to the neutral theory of population genetics, Hubbell's theory assumes that the distribution of species abundance in a community is predominated by the neutral drift of species abundances. In this case, no selection pressure is involved in the per capita probabilities of birth, death, migration and speciation for all the organisms in a community (Bell 2001; Hubbell 2001). As such, species abundance in a community bears no footprint of events such as species interactions or heterogeneous habitat conditions. While there may be a good chance for the neutral assumption to be violated in reality and indeed there

have been considerable criticisms, on various grounds, of the theory (Clark & McLachlan 2003; McGill 2003; Nee & Stone 2003), the neutral theory provides a tractable and testable mechanistic null hypothesis for understanding species assembly rules.

Central to Hubbell's neutral theory is the universal, dimensionless fundamental biodiversity parameter  $\Theta$  that is the product of the size of the community ( $J_M$ ) and speciation rate ( $v$ ), i.e.  $\Theta = 2J_M v$ . This fundamental biodiversity parameter completely determines the diversity patterns [the species-abundance distribution (SAD) and the species–area relationship (SAR)] of the metacommunity, while for the local community the SAD and SAR are determined by the biodiversity parameter together with immigration. It is worth noting that Hubbell's metacommunity model is spatially implicit and the complete determination of the SAD and SAR by  $\Theta$  requires species be distributed at random on landscapes. Here we established a functional relationship between Hubbell's biodiversity parameter and the Simpson diversity index for the metacommunity. Such a relationship is of significance for bridging the evolutionary aspect of biodiversity to the

ecological and statistical aspect of the diversity, thus helping to understand and interpret the neutral theory of biodiversity. It may also provide a new approach to estimate the fundamental biodiversity parameter of the metacommunity.

**METACOMMUNITY DIVERSITY MODELS**

Our derivation of the relationship between the fundamental biodiversity parameter and the Simpson index is based on the relative abundance of species which is equivalent to the allele frequency (a continuous variable), defined as the ratio of the abundance of individual species over the total abundance in a community. This treatment of species abundance (or allele frequency) as a relative ratio differs from Ewens' (1972) multivariate distribution that considers the combinatorial probability of different types of alleles sampled from a population. Following the footsteps of Hubbell (2001), under the neutral hypothesis the dynamics of a metacommunity is governed by ecological drift (equivalent to genetic drift) and speciation. The size of the metacommunity,  $J_M$ , is assumed to be a constant so that the ecological drift of the metacommunity is equal to  $1/J_M$  per time step. We further assume that the change in the relative abundance of each species (i.e. the ratio of the abundance of each species over  $J_M$ , denoted by  $x$ ) is very small within one step of time and hence is considered as a continuous stochastic process. Define  $\Phi(x)$  as the (absolute) number of species which have relative abundances of  $x$  in the metacommunity so that  $\Phi(x)dx$  is the expected number of species whose relative abundances fall within the range of  $(x, x + dx)$ . According to the classic theory of population genetics of the infinite allele model (Kimura & Crow 1964),  $\Phi(x)$  reads

$$\Phi(x) = \frac{\Theta(1-x)^{\Theta-1}}{x}, \quad 0 < x \leq 1 \tag{1}$$

where  $\Theta = 2J_M\nu$  is the fundamental biodiversity parameter, and  $\nu$  the speciation rate (Hubbell 2001). It is clear that  $\int_{1/J_M}^1 \Phi(x)dx$  is the total number of species in the metacommunity. The lower cut-off of the integration is defined as  $1/J_M$  because no species (alleles) can have abundance (individuals) less than 1. This is well recognized in population genetics (Wright 1969). Note Vallade & Houchmandzadeh (2003) recently independently obtained eqn 1 from the birth-death mutation neutral process for an infinite metacommunity.

$\Phi(x)$  of eqn 1 is the number of species, its probability density function (PDF)  $\phi(x)$  can simply be obtained by dividing eqn 1 by the total number of species  $\int_{1/J_M}^1 \Phi(x)dx$ :

$$\phi(x) = \frac{C\Theta(1-x)^{\Theta-1}}{x}, \tag{2}$$

where

$$C = \left[ \int_{1/J_M}^1 \Phi(x)dx \right]^{-1}$$

is a normalization factor to make eqn 2 a PDF. This PDF is also traditionally used to model allele frequency in molecular evolution (e.g. Wright 1969, p. 398), and later on it will be used to estimate the maximum likelihood estimate (MLE) of  $\Theta$ .

The SAD can then be expressed either in terms of eqns 1 or 2. According to eqn 1, the number of species whose relative abundances fall within interval  $(x_i, x_{i+1})$  is plainly

$$S_M(i) = \int_{x_i}^{x_{i+1}} \frac{\Theta(1-x)^{\Theta-1}}{x} dx,$$

while eqn 2 is the relative species-abundance distribution. These models describe a highly skewed SAD, with a large number of rare species similar to Fisher's logseries distribution (Alonso & McKane 2004).

Another important pattern developed by Caswell (1976) and Hubbell (1997, 2001) for the metacommunity is the SAR. Caswell and Hubbell's SAR was first derived from the infinite neutral allele model of Ewens (1972). Like the species-abundance pattern, Caswell and Hubbell's metacommunity SAR is also uniquely determined by the fundamental diversity parameter  $\Theta$ . This same SAR, defined as the average number of species in a sample of size  $n$  drawn at random from the metacommunity, can also be derived from eqn 1:

$$\bar{S}_n = \int_0^1 [1 - (1-x)^n] \Phi(x) dx = \sum_{i=0}^{n-1} \frac{\Theta}{\Theta + i}. \tag{3}$$

The discovery that the metacommunity SAD and SAR are fully determined by the fundamental biodiversity parameter  $\Theta$  is conceptually very fundamental, which links community diversity patterns to the evolutionary process of speciation. While in population genetics  $\Theta$  can be estimated from the coalescent theory (Hudson 1992), in ecology it is challenging how it may be estimated from species-abundance data (but see Alonso & McKane 2004; Etienne & Olff 2004). In the following we show that the estimation of  $\Theta$  may become possible if a species relative abundance is considered equivalent to an allele frequency, as Caswell (1976) and Hubbell (2001) have argued. It is well established in population genetics that the fundamental parameter  $\Theta$  is a function of the probability that two randomly drawn individuals belong to the same allele (or homozygosity) (Kimura 1983). The complement of this probability is called

heterozygosity, equivalent to the well-known Simpson diversity index in ecology (i.e. the probability of two randomly sampled individuals belonging to different species). The Simpson index is expressed as  $D = 1 - \sum_{i=1}^S x_i^2$ , where  $x_i$  is the relative abundance of the  $i$ th species. Obviously, the relationship between  $D$  and  $\Theta$  can be defined by

$$D = 1 - \int_0^1 x^2 \Phi(x) dx,$$

a well known function in population genetics (Kimura 1983; Hartl & Clark 1997). Substituting eqn 1 to this formula leads to a simple form,

$$D = \frac{\Theta}{1 + \Theta}$$

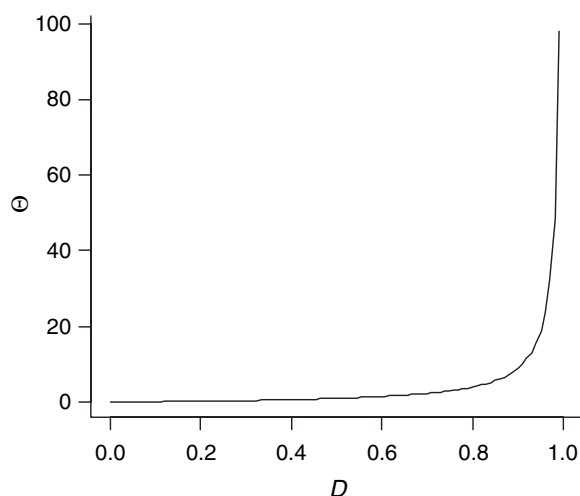
or

$$\Theta = \frac{D}{1 - D}. \quad (4)$$

This relationship as shown in Fig. 1 makes it possible to interpret and estimate the fundamental biodiversity parameter in terms of the Simpson index.  $D$  involves sampling and is subject to the drift in species abundance, it has variance (Stewart 1976):

$$V(D) = \frac{2\Theta}{(1 + \Theta)^2(2 + \Theta)(3 + \Theta)}.$$

Based on this variance and eqn 4, the variance of  $\Theta$  can be derived as:



**Figure 1** Relationship between Hubbell's fundamental biodiversity parameter  $\Theta$  and the Simpson diversity index  $D$ . Note when  $D$  is near 1,  $\Theta$  is a sharp function of  $D$ .

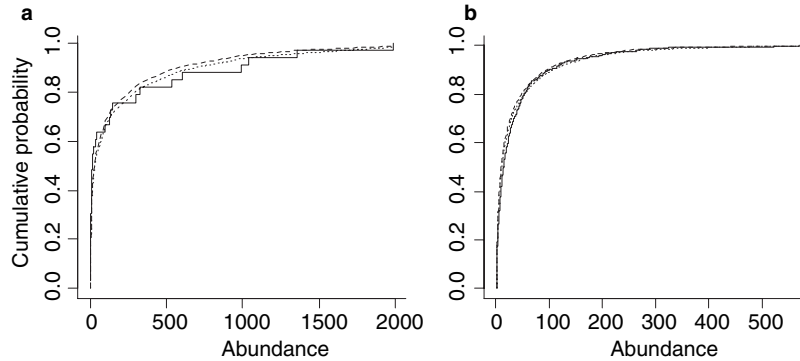
$$V(\Theta) = \frac{2\Theta(1 + \Theta)^2}{(2 + \Theta)(3 + \Theta)}. \quad (5)$$

We now turn to test this relationship for two sets of empirical data. The first data set is the abundances of filamentous fungi in the phylloplane of the grass *Lolium perenne*. The data are from Thomas & Shattock (1986) and is analysed by Magurran (1988). There are 33 species with total of 7861 individuals. It is already known that the data follows the logseries distribution ( $\chi^2$  test,  $P = 0.70$ ; Magurran 1988, page 135). We fitted our metacommunity model (eqn 2) to the data. The MLE of  $\Theta$  is 4.89 with asymptotic standard error = 2.19 (which is 2.50 if calculated from eqn 5) and the species-abundance data are described well by the metacommunity model [Kolmogorov–Smirnov (KS) test with  $P = 0.413$ ; see Fig. 2a]. The independent estimate of  $\Theta$  based on eqn 4 using the Simpson index ( $D = 0.8578$ ) is 6.03 which falls within the one standard error interval of the MLE  $\Theta$  (i.e.  $4.89 \pm 2.19$ ). We also tested if the same metacommunity model using the  $\Theta (=6.03)$  derived from the Simpson index would describe the observed data well. The result in Fig. 2a shows that this is indeed the case (KS test with  $P = 0.435$ ).

The second data set is the abundances of 678 tree species (26 554 trees in total) from the Pasoh Forest Reserve, Malaysia (He *et al.* 1997). We know that the Pasoh is a local community that can be adequately described by the lognormal-like multinomial zero-sum model (Hubbell 2001). As expected, our metacommunity model (eqn 2) fails to fit the Pasoh data (KS test with  $P < 0.0001$ ; Fig. 2b). The MLE  $\Theta$  for PDF (2) is 145.41 with asymptotic standard error = 13.52 (which is 16.88 if calculated from eqn 5). The independent estimate of  $\Theta$  from the Simpson index ( $D = 0.99407$ ) is 167.56 which falls outside the one standard error interval of MLE  $\Theta$  ( $145.41 \pm 13.52$ ) but within the two standard error interval. The metacommunity model using the  $\Theta (=167.56)$  estimated from the Simpson also fails to describe the observed Pasoh data (KS test with  $P < 0.0001$ ).

## DISCUSSION

Since the first application of the neutral allele model of population genetics to the study of relative species abundance (Caswell 1976; Hubbell 1979) and the recent synthetic treatment of the subject (Hubbell 2001), the neutral theory of biodiversity has become one of the most studied subjects in ecology (Bell 2001; Chave & Leigh 2002; Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Alonso & McKane 2004; Etienne & Olff 2004; Chave 2004; Mckane *et al.* 2004; He 2005). This progress has advanced our understanding of how neutral ecological and evolutionary processes would contribute to structuring and



**Figure 2** Cumulative probability functions for (a) the fungus species abundance data and (b) the Pasoh tree species-abundance data. The step (solid) curves in each figure are the observed cumulative probabilities, and the dotted curves are the predictions of metacommunity model (eqn 2) that was directly parameterized using the maximum likelihood method to the respective data ( $\hat{\Theta} = 4.89$  for the fungus data, 145.41 for the Pasoh data). The dashed curves are the predictions of eqn (2) using the  $\Theta$  derived from the Simpson index ( $\hat{\Theta} = 6.03$  for the fungus data, 167.56 for the Pasoh data). For the fungus data both dotted and dashed curves describe very well the species-abundance pattern [Kolmogorov–Smirnov (KS) test with  $P = 0.413$  for the dotted curve, 0.435 for the dashed curve]. Neither curve fits the Pasoh data well (KS test with  $P < 0.0001$ ; note the sample size, i.e. the number of species, is much larger for the Pasoh data).

maintaining species assemblages at multiple spatial scales from local to regional.

The significance of eqn 4 is that it not only establishes a conceptual connection between the evolutionary aspect of diversity and the ecological and statistical aspect of the diversity but also provides a potentially useful method for estimating  $\Theta$  from the readily derived Simpson diversity  $D$  for the metacommunity. The two examples and the results shown in Fig. 2 demonstrate the utility of the  $\Theta$ – $D$  relationship. In both cases the  $\Theta$  estimated from the Simpson index is similar to the MLE  $\Theta$  estimated directly from the species-abundance model. Even in the case of the Pasoh forest where the metacommunity model (2) does not fit the data well and  $D$  is near 1 ( $=0.99407$ ), the  $\Theta$ – $D$  relationship still gives a  $\Theta$  close to the MLE  $\Theta$ . These results are supported by the fact that the predicted species-abundance relationships using the MLE  $\Theta$  and the  $D$ -derived  $\Theta$  are not significantly different; the  $P$ -values of the KS test for the difference between the dotted curve and dashed curve in Fig. 2 are 0.896 and 0.988 for the fungus and Pasoh data respectively.

It is important to emphasize that because eqn 4 is derived for the neutral metacommunity defined by model (1), it should only be applied to such a community. Although we are almost certain that a community is not a neutral metacommunity if it fails to be fitted by metacommunity models (e.g. the Pasoh forest), a high goodness-of-fit does not mean a community is a neutral metacommunity (e.g. the fungus data). No practical method is available to determine if a community is a true neutral metacommunity. In the case of the Pasoh forest where neither  $\Theta$  predicts the species-abundance data well (Fig. 2b), no further interpretation of  $\Theta$  should be implied.

Like any other diversity indices, the Simpson index is also sensitive to sample size (increase with sample size) although its overall performance against several criteria is superior to many other indices such as the Shannon information entropy (Peet 1974; May 1975; Routledge 1979; see Magurran 1988 for a review). A sufficiently large sample is needed for a reliable estimation of  $D$ . The  $\Theta$ – $D$  relationship may also provide an avenue to estimate  $D$ , we may flip eqn 4 around to estimate an unbiased Simpson index from  $\Theta$  so that to take the advantage of the unbiased property of the MLE of  $\Theta$ .

Since the development of Hubbell’s zero-sum multinomial model, several metacommunity models have been derived (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Etienne & Olf 2004). Although all these models describe a similar logseries-like SAD and contain Hubbell’s fundamental parameter, they differ in mathematical forms and  $\Theta$  is entered into each of the models differently. Naturally, they should have  $\Theta$ – $D$  forms differing from eqn 4. For instance, for the logseries model (Volkov *et al.* 2003),

$$f(n) = \frac{\Theta x^n}{n}$$

for  $n = 1, 2, \dots$ ,  $\Theta$ – $D$  takes a relationship as

$$D = 1 - \frac{J_M + \Theta}{J_M \Theta}.$$

This result enforces the emphasis that a  $\Theta$ – $D$  relationship only holds for the metacommunity model from which it is derived. Note the metacommunity studied here has identical  $\Theta$  to Hubbell’s model as evidenced by its definition ( $\Theta = 2J_M v$ ) and the species area eqn 3.

Although  $\Theta$  and  $D$  have a unique relationship for a given metacommunity ( $J_M$  is known for a given community), the index alone is not sufficient to determine the fundamental diversity parameter of the local community; additional information on migration is needed. The relationship among Hubbell's parameter, Simpson's index and migration for the local community remains to be developed. Another interesting index is the analytical  $\beta$ -diversity given by Chave & Leigh (2002), which is the probability that two individuals separated by a certain distance belongs to the same species. This is virtually a spatial version of the Simpson index, suggesting that a spatially explicit metacommunity undergoing neutral speciation can also be characterized by the Simpson index. Obviously, in this case, this index is not related to  $\Theta$  in such a simple way as it is given in eqn 4.

As a conclusion, because biodiversity indices such as the Simpson's are well and widely received by ecologists and have very practical interpretations, the relationship between Hubbell's diversity parameter and the Simpson index is expected to stimulate and enhance the understanding of the neutral theory in general and the biodiversity parameter in particular.

#### ACKNOWLEDGEMENTS

This work was inspired from many discussions with Stephen Hubbell. The comments of Hubbell and Luis Borda De Agua greatly improved this article. We are grateful to Jonathan Chase and four referees for their very constructive reviews. This work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the Alberta Ingenuity Fund, and the Faculty of Agriculture, Forestry, and Home Economics, the University of Alberta.

#### REFERENCES

- Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.
- Bell, G. (2001). Neutral macroecology. *Science*, 293, 2413–2418.
- Caswell, H. (1976). Community structure: a neutral model analysis. *Ecol. Monogr.*, 46, 327–354.
- Chave, J. (2004). Neutral theory and community ecology. *Ecol. Lett.*, 7, 241–253.
- Chave, J. & Leigh, E.G. (2002). A spatially explicit neutral models of  $\beta$ -diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153–168.
- Clark, J.S. & McLachlan, J.S. (2003). Stability for forest biodiversity. *Nature*, 423, 635–638.
- Etienne, R.S. & Olf, H. (2004). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.*, 7, 170–175.
- Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3, 87–112.
- Hartl, D.L. & Clark, A.G. (1997). *Principles of Population Genetics*, 3rd edn. Sinauer Associates, Sunderland, MA.
- He, F. (2005). Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct. Ecol.*, 19, in press.
- He, F., Legendre, P. & LaFrankie, J.V. (1997). Distribution patterns of tree species in a Malaysian tropical rain forest. *J. Veget. Sci.*, 8, 105–114.
- Hubbell, S.P. (1979). Tree dispersion, abundance and diversity in a tropical dry forest. *Science*, 203, 1299–1309.
- 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.
- Hudson, R.R. (1992). Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.*, 7, 1–44.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, London.
- Kimura, M. & Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49, 725–738.
- Magurran, A.E. (1988). *Ecological Diversity and its Measurement*. Princeton University Press, Princeton, NJ.
- May, R.M. (1975). Patterns of species abundance and diversity. In: *Ecology and Evolution of Communities* (eds Cody, M.L. & Diamond, M.J.). Harvard University Press, Cambridge, MA, pp. 81–120.
- McGill, B.J. (2003). A test of the unified neutral theory of biodiversity. *Nature*, 422, 881–885.
- Mckane, A.J., Alonso, D. & Solé, R. (2004). Analytic solution of Hubbell's model of local community dynamics. *Theor. Popul. Biol.*, 65, 67–73.
- Nee, S. & Stone, G. (2003). The end of the beginning for neutral theory. *Trends Ecol. Evol.*, 18, 433–434.
- Peet, R.K. (1974). The measurement of species diversity. *Ann. Rev. Ecol. System.*, 5, 285–307.
- Routledge, R.D. (1979). Diversity indices: which ones are admissible. *J. Theor. Biol.*, 76, 503–515.
- Stewart, F.M. (1976). Variability in the amount of heterozygosity maintained by neutral mutations. *Theor. Popul. Biol.*, 9, 188–201.
- Thomas, M.R. & Shattock, R.C. (1986). Filamentous fungal associations in the phylloplane of *Lolium perenne*. *Trans. Brit. Mycol. Soc.*, 87, 255–268.
- Vallade, M. & Houchmandzadeh, B. (2003). Analytical solution of a neutral model of biodiversity. *Phys. Rev. E*, 68, 061902–061905.
- Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.
- Wright, S. (1955). Classification of the factors of evolution. *Cold Spring Harbor Symposium on Quantitative Biology*, 20, 16–24.
- Wright, S. (1969). *Evolution and the Genetics of Populations. The Theory of Gene Frequencies*, Vol. 2. The University of Chicago Press, Chicago, IL.

Editor, Jonathan Chase

Manuscript received 7 October 2004

First decision made 30 October 2004

Second decision made 3 January 2005

Manuscript accepted 7 January 2005