The Rarefaction of Phylogenetic Diversity: Formulation, Extension and Application

David A. Nipperess

Abstract Like other measures of diversity, Phylogenetic Diversity (PD) increases monotonically and asymptotically with increasing sample size. This relationship can be described by a rarefaction curve tracing the expected PD for a given number of accumulation units. Accumulation units represent individual organisms, collections of organisms (e.g. sites), or even species (or equivalent), giving individual-based, sample-based and species-based curves respectively. The formulation for the exact analytical solution for the rarefaction of PD is given in an expanded form to demonstrate congruence with the classic formulation for the rarefaction of species richness. Rarefaction is commonly applied as a standardisation for diversity values derived from differing numbers of sampling units. However, the solution can be simply extended to create measures of phylogenetic evenness, phylogenetic beta-diversity and phylogenetic dispersion, derived from individual-based, sample-based and species-based curves respectively. This extension, termed ΔPD, is simply the initial slope of the rarefaction curve and is related to entropy measures such as PIE (Probability of Interspecific Encounter) and Gini-Simpson entropy. The application of rarefaction of PD to sample standardisation and measurement of phylogenetic evenness, phylogenetic beta-diversity and phylogenetic dispersion is demonstrated. Future prospects for PD rarefaction include the recognition of evolutionary hotspots (independent of species richness), the basis for ecological theory such as phylogeny-area relationships, and the prediction of unseen biodiversity.

Keywords Alpha diversity • Beta diversity • Evenness • Phylogeny • Sampling curves

D.A. Nipperess (✉)
Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia
e-mail: david.nipperess@mq.edu.au

R. Pellens, P. Grandcolas (eds.), *Biodiversity Conservation and Phylogenetic Systematics*, Topics in Biodiversity and Conservation 14,
DOI 10.1007/978-3-319-22461-9_10

Introduction

Phylogenetic Diversity (PD) is a simple, intuitive and effective measure of biodiversity. The PD of a set of taxa, represented as the tips of a phylogenetic tree, is the sum of the branch lengths connecting those taxa (Faith 1992). PD is a particularly flexible measure because it can be applied to any set of relationships among entities that can be reasonably portrayed as a tree. Thus, the tips do not, by necessity, need to represent species but could be higher taxa, Operational Taxonomic Units, Evolutionarily Significant Units, individual organisms or unique haplotypes. Further, the tree itself might not portray evolutionary relationships but instead be, for example, a cluster dendrogram portraying functional relationships among taxa (Petchey and Gaston 2002).

Since the original formulation by Faith (1992), PD has come to be not just a single measure equating to a phylogenetically weighted form of richness, but rather a general class of measures dealing with various aspects of alpha and beta-diversity (Faith 2013). The common feature of this class of measures is the summation of branch lengths rather than the counting of tips. By substituting branch segments (intervals between nodes on a phylogenetic tree) for species, and including a weighting for the length of that segment, it is possible to modify many of the classic measures of Species Diversity (SD) to a PD equivalent (Faith 2013). By this means, phylogenetically weighted measures of endemism (Faith et al. 2004; Rosauer et al. 2009), ecological resemblance (Ferrier et al. 2007; Nipperess et al. 2010), and entropy (Chao et al. 2010, and chapter "Phylogenetic Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers") have been developed, for example.

In its classic form, PD, like species richness, has the property of concavity (Lande 1996). That is, the addition of individuals or sets of individuals to a community can increase PD but never decrease it. Thus, just like species richness, PD increases monotonically with increasing sampling effort, creating a classic sampling curve that reaches an asymptote when all species (and branch segments) are represented (Fig. 1). Gotelli and Colwell (2001) recognise two general types of sampling curve, individuals-based and sample-based, that are distinguished by the units on the x-axis, representing either individual organisms or samples, respectively. Samples, in this context, are collections of individuals bounded in space and time, corresponding to the common ecological usage of the term. For PD, we can recognise a third type of sampling curve where the units on the x-axis are species or their equivalent (Fig. 1). Species, like samples, are also collections of individuals bounded, in this case, by some minimum degree of relatedness. Obviously, species-based sampling curves are meaningless when plotting species richness but have real value when plotting PD. For the purposes of generalisation, it is useful to be able to refer to these units (individuals, samples, species) with a single term. Chiarucci et al. (2008) used "accumulation units" to refer to individuals and samples. I extend this term to also include species as an additional unit of sampling effort in sampling curves. While these different units (individuals, samples, species) all measure

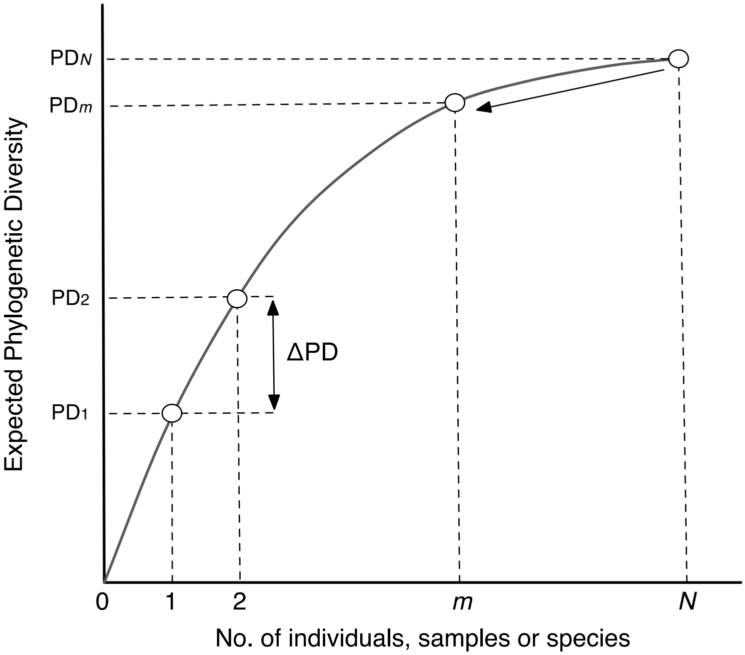

Fig. 1 Sampling curve showing the relationship between Phylogenetic Diversity (PD) and sampling depth. The level of sampling is measured in accumulation units of individuals, samples (collections of individuals) or species as required. PD_N is the Phylogenetic Diversity of the full set of N accumulation units. Rarefaction is the process (indicated by *unidirectional arrow*) of randomly subsampling (rarefying) the pool of N accumulation units to a subset of size m and calculating the expected PD of that subset (PD_m). ΔPD is the expected gain in PD between the first and second accumulation unit, and can be used as a measure of phylogenetic evenness, beta-diversity or dispersion, depending on the nature of the unit of accumulation

sampling effort in some sense, they are not equivalent and sampling curves derived from them must be interpreted differently in each case.

Beside the units by which sampling effort is measured, Gotelli and Colwell (2001) distinguished between "accumulation curves" and "rarefaction curves", based on the process by which the sampling curve is calculated. An accumulation curve plots a single ordering of individuals or samples (or species) against a cumulatively calculated concave diversity measure. The jagged shape of the resulting curve is highly dependent on the, often arbitrary, order of the accumulation units. To resolve this problem, rarefaction curves instead plot the *expected* value of the diversity measure against the corresponding number of accumulation units. Rarefaction can be achieved using an algorithmic procedure of repeated random sub-sampling of the full set of accumulation units and calculating the mean diversity (Gotelli and Colwell 2001). However, Hurlbert (1971) and Simberloff (1972) showed that expected diversity can be calculated using an exact analytical solution, obviating the need for computer-intensive repeated sub-sampling. Initially, this solution was for individuals-based rarefaction curves, but it has since been shown that the same solution applies to sample-based rarefaction (Kobayashi 1974; Ugland et al. 2003; Mao et al. 2005; Chiarucci et al. 2008).

The original purpose of rarefaction was to allow the comparison of datasets with differing amounts of sampling effort (Sanders 1968). Assemblages can be compared “fairly” when rarefied to the same number of accumulation units (Gotelli and Colwell 2001). However, rarefaction has broader application than this single purpose. Depending of the unit of accumulation, the shape of the rarefaction curve provides information on ecological evenness (Olszewski 2004) and beta-diversity (Crist and Veech 2006). Rarefaction of species richness also forms the basis of estimators of species richness, including unseen species (Colwell and Coddington 1994). In the case of PD, species-based rarefaction curves also allow for a measure of phylogenetic dispersion (Webb et al. 2002), effectively the expected PD for some given number of species (Nipperess and Matsen 2013). A solution for the rarefaction of PD is therefore desirable as it will allow for these applications to be realised for phylogenetically explicit datasets.

Rarefaction of Phylogenetic Diversity, using an algorithmic solution of repeated sub-sampling, has now been done several times (see for example Lozupone and Knight 2008; Turnbaugh et al. 2009; Yu et al. 2012). However, an analytical solution for PD rarefaction, similar to that determined by Hurlbert (1971) for species richness, is preferable both because its results are exact (not dependent on the number of repeated subsamples) and substantially more computationally efficient. Nipperess and Matsen (2013) recently published just such a solution for both the mean and variance of PD under rarefaction. This solution is quite general, being applicable to rooted and unrooted trees, and even allowing partition of the tree into smaller components than the individual branch segments. As a result, the solution is given in a very generalised form and its relationship with classic rarefaction formula for species richness is not immediately clear.

In this chapter, I provide a detailed formulation for the exact analytical solution for expected (mean) Phylogenetic Diversity for a given amount of sampling effort. This formulation is for the specific but common case of a rooted phylogenetic tree where whole branch segments are selected under rarefaction. I use the same form of expression as used by Hurlbert (1971) to demonstrate the direct relationship between rarefaction of PD and rarefaction of species richness. I do not include a solution for variance of PD under rarefaction due to its complexity when given in this form and instead refer the reader to Nipperess and Matsen (2013). I extend this framework to show how the initial slope of the rarefaction curve (ΔPD) can be used as a flexible measure of phylogenetic evenness, phylogenetic beta-diversity or phylogenetic dispersion, depending on the unit of accumulation. I apply PD rarefaction and the derived ΔPD measure to real ecological datasets to demonstrate its usefulness in addressing ecological questions. Finally, I discuss some future directions for the extension and application of PD rarefaction.

Formulation

To begin, the classic rarefaction formula for species richness will be reviewed in order to demonstrate how it can be extended to the case of Phylogenetic Diversity. The expected species richness (S) for a given amount of sampling is simply the sum of probabilities (p) of each species occurring in a subset of m accumulation units (Eq. 1).

$$E\left[S\right]_m = \sum_i^S {}_m p_i \tag{1}$$

To solve Eq. 1, we need to determine the probability (p) of each species being selected by a random draw of m accumulation units from the total set of N units. Regardless of whether the accumulation unit is an individual or a sample, this probability is a function of the frequency (n) with which species i occurs across the set of N accumulation units (Chiarucci et al. 2008). Since N is a set of finite size, random draws from that set should be without replacement and thus p is defined by the hypergeometric distribution (Hurlbert 1971). Substituting into Eq. 1, the expected species richness is as follows (Eq. 2).

$$E\left[S\right]_m = \sum_i^S \left[1 - \frac{\begin{pmatrix} N - n_i \\ m \end{pmatrix}}{\begin{pmatrix} N \\ m \end{pmatrix}} \right] \tag{2}$$

The quantity within the square brackets in Eq. 2 corresponds to p in Eq. 1. Note that the expressions in curved brackets are binomial coefficients and not simple fractions, while the quantity subtracted from one within the square brackets is a fraction. The denominator in this fraction gives the number of distinct subsets of size m that can be drawn from the total set of N units. The numerator gives the number of distinct subsets of size m that do not contain species i. Equation 2 is the same as that originally proposed by Hurlbert (1971).

Phylogenetic Diversity is simply the sum of a set of branch lengths spanning a set of species (or, more generally, tips). So, for a set of S species, there is a corresponding set of T branch segments. Each branch segment (j) has a length (L) measured as sequence substitutions, millions of years, or some other biologically meaningful estimate of difference. Considering only rooted phylogenetic trees, PD is calculated as follows (Eq. 3).

$$PD = \sum_j^T L_j \tag{3}$$

In the original definition intended by Faith (1992), the PD of a subset of species is calculated by summing the branch lengths connecting that set of species to the root of the tree, even when the common ancestor of that subset is not the same as the root. In this definition, a subset containing a single species (or even a single individual) has a non-zero PD value, which in this case, would be the total path length from the tip to the root. This corresponds to the *rooted PD* value of Pardi and Goldman (2007). The alternative, called *unrooted PD* by Pardi and Goldman (2007), includes only the branch segments connecting a subset of species to their common ancestor, and thus a subset containing only a single species would have zero PD. The former definition, rooted PD, is adopted here because it allows for the straightforward formulation of a whole class of derived PD measures (Faith 2013), and because it is concordant with the original idea of PD acting as a surrogate for the feature diversity of a set (Faith 1992; Faith et al. 2009). Obviously, rooted PD requires a rooted phylogenetic tree, even if the choice of root is arbitrary (Nipperess and Matsen 2013).

Given this definition, the rarefaction of PD involves finding the expected (average) sum of branch lengths (including the path to the root) for all possible distinct subsets of m accumulation units (Fig. 2). This is achieved by extending the classic rarefaction formula through a substitution of species for branch segments in a phylogenetic tree. Since PD is simply the sum of branch lengths, then the expected PD must also be the sum of branch lengths, each weighted by the probability (q) of its occurrence in a subset of size m (O'Dwyer et al. 2012). So, for a rooted phylogenetic tree represented as a set of T branch segments, the expected PD is given as follows (Eq. 4).

$$E\left[PD\right]_m = \sum_j^T L_j \times {}_m q_j \tag{4}$$

The probability of each branch segment occurring in a subset is again a function of the frequency with which it occurs among accumulation units. The frequency of occurrence of a particular branch segment (o) depends on the frequency of occurrence of species that are descendent from that branch segment. Let x be a binary value indicating whether species i is (1) or is not (0) a descendant of branch segment j. Multiplying x by n and summing across all species will give the total number of occurrences of branch segment j among N accumulation units (Eq. 5).

$$o_j = \sum_i^S \left(n_i \times x_{ij}\right) \tag{5}$$

Thus, by summing across branches instead of species, substituting branch occurrence for species occurrence, and including a branch length weighting, we are able to adapt the classic rarefaction formula for species richness for the purposes of calculating expected Phylogenetic Diversity (Eq. 6). Note this solution is equivalent to that of Nipperess and Matsen (2013) but is expressed in an expanded form for the specific case of calculating rooted PD. Equation 6 is very similar to the solution for

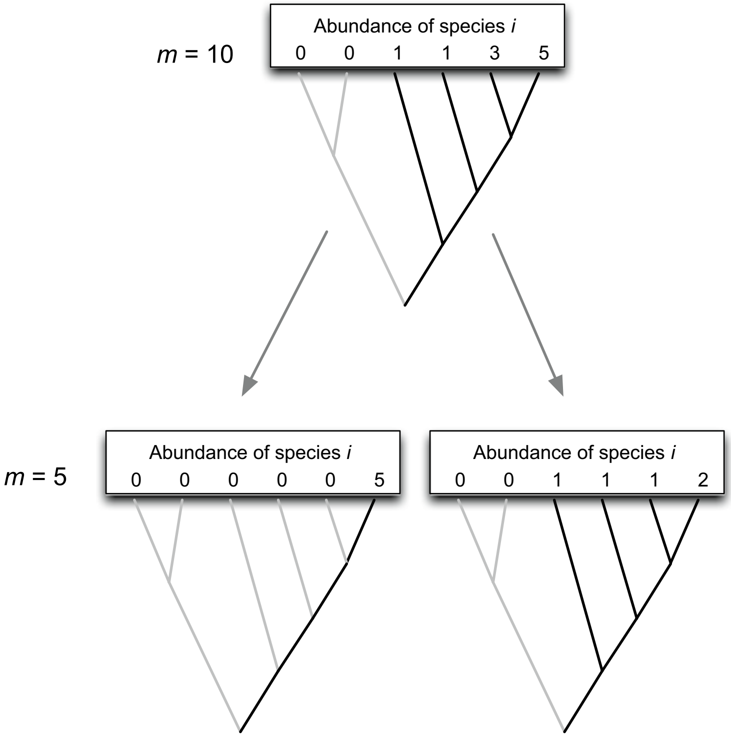

Fig. 2 An illustration of the process of rarefying Phylogenetic Diversity (PD) by units of individuals. An initial sample of ten individuals ($m=10$) distributed among four tips (species) is rarefied to a subset of five individuals ($m=5$) by a process of random sampling without replacement. For the rarefied samples, 2 of the 252 possible subsets are shown. The expected PD under rarefaction is the average sum of branch lengths represented by each of these distinct subsets. The branch lengths summed to calculate PD are *black* while those not represented (and thus not summed) are *grey*. Note that the rooted definition of PD is used where the path length to the root is always included, even in the case where only a single tip is represented

expected PD of Faith (2013) but differs in that random draws are without replacement following the hypergeometric distribution.

$$E[PD]_m = \sum_j^T \left[L_j \times \left(1 - \frac{\binom{N-o_j}{m}}{\binom{N}{m}} \right) \right] \tag{6}$$

Finally, it is now possible to calculate the expected PD for a given number of *species*. A species, in this context, is simply a collection of individuals in much the same way as a sample is a collection of individuals, and the same equations apply.

Under these circumstances, o_j is equal to the sum of x_{ij} (over all species) as n_i will always equal 1, and N is equal to S. Substituting into Eq. 6 gives the following formula for rarefaction by species (Eq. 7).

$$E\left[PD\right]_m = \sum_{j}^{T}\left[L_j \times \left(1 - \frac{\binom{S - \sum x_{ij}}{m}}{\binom{S}{m}}\right)\right] \tag{7}$$

Extension

It has previously been recognised (Lande 1996; Olszewski 2004) that there is a relationship between individuals-based rarefaction curves and measures of evenness. Specifically, the initial slope of the individuals-based curve for species richness is equal to the PIE (Probability of Interspecific Encounter) index of Hurlbert (1971). The initial slope of the rarefaction curve is the difference between the expected species richness for two individuals ($m=2$) and the expected species richness for one individual ($m=1$), and is the probability that the second individual will be a different species from the first (Olszewski 2004). The PIE index is directly related to the Gini-Simpson index – the probability that two individuals selected at random will be different species. The difference between these two indices is in the form of random sampling – Gini-Simpson samples with replacement (thus assuming infinite population size) while PIE, just like rarefaction, samples without replacement. Following Olszewski (2004), PIE can be expressed as the following (Eq. 8) where $E[S_1]$ and $E[S_2]$ refer to the expected species richness of one and two randomly drawn individuals respectively. Note that $E[S_1]$ always equals one in this case.

$$PIE = E[S_2] - E[S_1] \tag{8}$$

When considering a sample-based curve, it is clear that the initial slope is related to the beta-diversity of the set of samples from which the curve is calculated. In this case, the difference between $E[S_1]$ and $E[S_2]$ is the expected number of species in the second sample that are not found in the first. Thus, the PIE index can be used to measure beta-diversity if applied to sample-based rarefaction. This interpretation is directly related to the additive partitioning of species diversity into alpha and beta components where alpha-diversity is the mean (expected) richness of a single sample and beta-diversity is the gain in species richness from a single sample to a larger set of samples and can be read directly from a rarefaction curve (Crist and Veech 2006).

It follows that we can also define measures of phylogenetic evenness and phylogenetic beta-diversity using the initial slope of the PD rarefaction curve, where the units of accumulation are either individuals or samples respectively (Fig. 1). In either case, the initial slope is the expected gain in PD (ΔPD) when adding a second accumulation unit to the first. Further, because PD rarefaction curves can also meaningfully use species as accumulation units, we can extend this idea to include a measure of phylogenetic dispersion where the gain in PD is the expected branch length in the lineage (path from tip to root) of a second randomly selected species that is not shared with the first. Thus, we can define a general measure (ΔPD) for phylogenetic evenness, phylogenetic beta-diversity or phylogenetic dispersion, depending on the accumulation units chosen (Eq. 9, see also Fig. 1). ΔPD is very similar to the ΔPDq measure of Faith (2013) although in that case, probabilities are not derived from the hypergeometric distribution. Further, ΔPDq is specifically applied to the problem of estimating loss of PD from extinction – a problem that is mathematically similar to rarefaction.

$$\Delta PD = E[PD_2] - E[PD_1] \tag{9}$$

If branch lengths are measured as millions of years between branching events, then ΔPD is measured in units that make intuitive sense and allows for direct comparison across trees and systems. Alternatively, one could standardise the measure by dividing by its theoretical maximum. ΔPD will be maximum when all individuals, species or samples represent wholly distinct lineages with no shared branch lengths. For an ultrametric tree, the lineage length (path from tip to root) is invariant across species and is equal to the depth of the tree. When rarefaction is by units of individuals or species, $E[PD_1]$ is the lineage length. When rarefaction is by units of samples, $E[PD_1]$ will equal the average PD of a sample and will be equal to ΔPD in the extreme case where each sample shares no branch length with any other sample. Thus, whether referring to units of individuals, species or samples, $E[PD_1]$ represents the theoretical maximum of ΔPD and can be used to standardise the measure as follows.

$$\Delta PD_{standard} = \frac{\Delta PD}{\Delta PD_{\max}} = \frac{E[PD_2] - E[PD_1]}{E[PD_1]} \tag{10}$$

Application

The following is a demonstration of the application of PD rarefaction, and the derived ΔPD statistics, to real ecological datasets. These applications are not intended to provide definitive answers to ecologically important questions but are, rather, simple demonstrations of how PD rarefaction can allow new analyses to be undertaken and, hopefully, new insights gained.

In all these applications, I have used published data on mammals. This is principally for convenience as mammals (Bininda-Emonds et al. 2007) and birds (Jetz et al. 2012) are the only major taxonomic groups for which comprehensive species-level supertrees are available. I have used an updated version of the mammal supertree of Bininda-Emonds et al. (2007) published as supplementary material by Fritz et al. (2009). In this supertree, all branch lengths are measured in units of time (millions of years between branching events), allowing for a straight-forward interpretation of PD as cumulative evolutionary history (Proches et al. 2006).

All analyses were conducted using the statistical software, *R* version 2.15.2 (R Core Team 2012). Phylogenetic information was processed using the *ape* package in R (Paradis et al. 2004). PD rarefaction analyses used the *phylodiv*, *phylocurve* and *phylorare* functions, written by the author and available from: http://davidnipperess.blogspot.com.au.

Standardisation of Sampling

The most commonly used application for rarefaction is standardisation to allow comparisons to be made between datasets with differing amounts of sampling effort. Standardisation can be achieved by rarefying all datasets back to a common (typically the minimum) number of accumulation units (Sanders 1968; Gotelli and Colwell 2001).

Law et al. (1998) surveyed bats in ten State Forests of the south-west slopes region of New South Wales, Australia. Survey methods were a combination of ultrasonic detectors, harp-traps, mist-nets and trip-lines. For the purposes of this demonstration, only data from the harp-traps will be used. A harp-trap is a rectangular frame, stringed vertically with nylon line, placed so as to intercept the flight path of low-flying bats (Tidemann and Woodside 1978). A bat striking the nylon lines of the trap will tumble down into a collecting bag at the bottom.

Sampling effort among State Forests was variable with between 8 and 30 trap-nights. Comparison of bat diversity between State Forests is therefore confounded by variation in sampling effort, as can be seen when plotting separate PD rarefaction curves for each State Forest (Fig. 3). To correct for variation in trapping effort, expected PD for each State Forest was calculated for the common value of 15 individuals, which was the minimum number recovered from a State Forest (Fig. 3). While rarefying to eight trap-nights (samples) would also be an appropriate method of standardisation, data on the bat species caught per trap-night were not available in Law et al. (1998). Standardising for sample effort changed the rank order of the sites for Phylogenetic Diversity (Table 1). A test of the rank correlation between the standardised and non-standardised PD values was relatively high but non-significant (Spearman's correlation coefficient, rho = 0.57, p = 0.084). Therefore, what one concludes about the relative bat diversity (and perhaps conservation importance) among these sites is dependent upon whether or not sampling effort is taken into account.

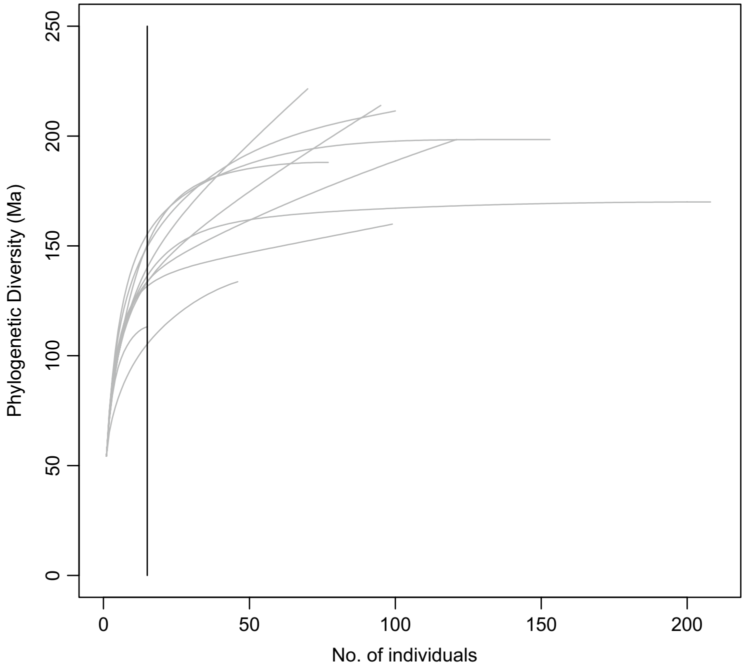

Fig. 3 An example of standardisation of Phylogenetic Diversity (PD) by rarefaction. Data are abundances of bats caught in harp-traps in State Forests of the south-west slopes region of New South Wales, Australia. See Law et al. (1998) for a description of the data. Plotting separate individuals-based curves (*grey lines*) for each site shows considerable variation in sampling effort, with the raw value of PD being dependent on the number of trapped individuals. To allow for comparison between sites, PD is rarefied to an expected value for 15 individuals for all sites (indicated by *black vertical line*)

Table 1 Comparison of diversity measures for bat assemblages for ten state forests of the south-west slopes region of New South Wales, Australia

State forest	Individuals (N)	Species richness (S)	Phylogenetic diversity (PD_N)	Standardised phylogenetic diversity (PD_{15})
Bago	99	6	159	132
Maragle	208	8	170	136
Buccleuch	100	7	211	150
Bungongo	70	6	221	140
Woomargama	121	7	198	133
Carabost	153	8	198	155
Murraguldrie	95	6	214	133
Ellerslie	46	4	134	105
Tumblong	77	7	188	151
Minjary	15	3	113	113

Original data was taken from Law et al. (1998). Phylogenetic Diversity is measured in units of millions of years

Phylogenetic Evenness

The extension of PD rarefaction to ΔPD allows for the measurement of phylogenetic evenness, which is essentially a measure of the distribution of individuals among branches in a phylogenetic tree (Webb and Pitman 2002). A phylogenetically even community is one where the most evolutionarily distinct species are also the most abundant. Because ΔPD will increase with both increasing phylogenetic evenness and phylogenetic diversity, it is more correctly a measure of entropy (Jost 2006), directly comparable to the PIE and Gini-Simpson indices. It has a particularly close relationship with the quadratic entropy measure of Rao (1982). Rao's quadratic entropy measures the average distance between individuals in an assemblage. When that distance is measured as patristic distance (path length on a phylogenetic tree), ΔPD will be approximately half of Rao's quadratic entropy. ΔPD is also similar in intent, but not in form, to the phylogenetic entropy index of Allen et al. (2009).

Low ecological evenness may be an indicator of disturbance where a small number of species are favoured. If those favoured species are also closely related, due to sharing a trait that allows exploitation of disturbance events, we can expect a reduction in phylogenetic evenness (Helmus et al. 2010). Medellin et al. (2000) surveyed the bat assemblages along a disturbance gradient in the Selva Lacandona, Chiapas, Mexico. The disturbance gradient consisted of four habitats, which, in order of disturbance, were cornfield, oldfield, cacao plantation and forest. Bats were sampled using mist nets and each habitat in the disturbance gradient was sampled using the same effort, thus making possible the comparison of habitats without the need for rarefaction. Medellin et al. (2000) found a trend of decreasing species richness and species evenness with increasing disturbance, and this trend is also reflected in the phylogenetic diversity and evenness of the assemblages (Table 2, Fig. 4).

The trend in phylogenetic evenness may simply be reflecting the abundance distribution among species. To determine the phylogenetic contribution to phylogenetic evenness, ΔPD was divided by the PIE index (Table 2). Since PIE is the probability that the second randomly selected individual is a different species to the

Table 2 Comparison of diversity measures for bat assemblages from four habitats along a disturbance gradient in the Selva Lacandona, Chiapas, Mexico

Habitat	Individuals	Species richness	PIE	Phylogenetic diversity	Phylogenetic evenness (ΔPD)	Phylogenetic component (ΔPD/PIE)
Cornfield	572	17	0.786	295	17.2	21.8
Oldfield	690	20	0.809	469	18.1	22.4
Cacao	699	21	0.851	493	18.2	21.3
Forest	444	27	0.884	609	20.4	23.0

Original data taken from Medellin et al. (2000). PIE refers to the Probability of Interspecific Encounter (Hurlbert 1971). Phylogenetic Diversity and phylogenetic evenness are measured in units of millions of years

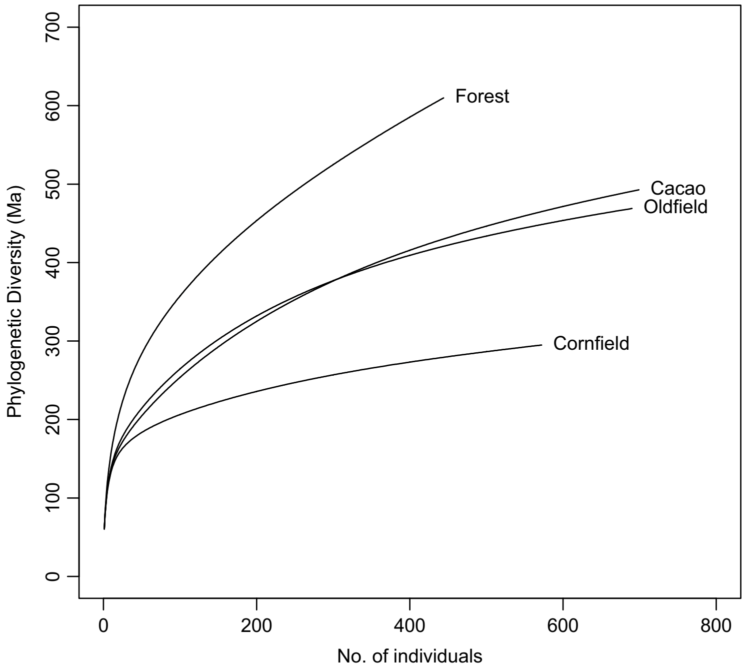

Fig. 4 Individuals-based PD rarefaction curves for bat assemblages from four habitats along a disturbance gradient in the Selva Lacandona, Chiapas, Mexico.(See Medellin et al. (2000) for a description of the data. Phylogenetic evenness (ΔPD) values are highest in the least disturbed habitat (Forest) and lowest in the most disturbed habitat (Cornfield)

first, we can divide ΔPD by PIE to get the expected branch length of that species (conditional on the second individual being a different species). This value is related to phylogenetic dispersion (ΔPD from a species-based rarefaction curve) but differs due to the conditional probability structure, and effectively measures the pure phylogenetic contribution to ΔPD independent of the abundance distributions among species. We see, in this case, that the phylogenetic component generally decreases with increasing disturbance (Cacao being the exception), supporting the notion that disturbance favours more closely related species.

Phylogenetic Beta-Diversity

Phylogenetic beta-diversity is effectively the turnover of branch lengths between samples in space and/or time. Like its species-level equivalent, phylogenetic beta-diversity can be measured on a pair-wise basis (Lozupone and Knight 2005; Bryant et al. 2008; Nipperess et al. 2010) or as a single value for a set of samples (Anderson et al. 2010). Rarefaction of PD provides a means for deriving a single value of beta-diversity for a set of samples of any size via the ΔPD measure, which is a phylogenetic analogue of the additive partitioning approach of Crist and Veech (2006).

Morton et al. (1994) compiled data on small mammal assemblages for 245 sites in arid Australia. I calculated beta-diversity for two regions from this dataset – Tanami desert and Uluru-Kata Tjuta National Park, Northern Territory. These regions had a similar number of sites (Table 3) covering a roughly similarly sized area but differed in the number of vegetation types. The Tanami sites were all spinifex grassland while the Uluru sites comprised a mix of spinifex grassland, acacia shrubland and woodland (Morton et al. 1994). It might be expected therefore that the Uluru sites will show higher beta-diversity due to the diversity of habitats represented. In addition to ΔPD, I used the additive partitioning method to calculate species-level beta-diversity as the difference between total species richness of all sites in a region and the mean species richness of a single site (Lande 1996; Crist and Veech 2006).

Contrary to expectations, the Tanami desert sites showed greater species beta-diversity and phylogenetic beta-diversity despite the lack of variation in vegetation type (Table 3). This pattern is driven by the much higher site-level (alpha) species richness in Uluru-Kata Tjuta National Park (Table 3, Fig. 5) without a concomitant increase in overall (gamma) species richness, resulting in a high degree of species overlap. Given the overlap in species among Uluru sites, it appears that most small mammals are not specialised for particular vegetation types.

Phylogenetic Dispersion

Phylogenetic dispersion is a measure of the average phylogenetic distance among species (or tips) (Webb et al. 2002) and is in effect a measure of tree shape (Davies and Buckley 2012). ΔPD provides a simple, intuitive measure of dispersion as the expected gain in PD of adding a second randomly selected species to the first. It can also be seen as a means of correcting for variation in species richness among samples, as it is well known that PD increases with species richness (Rodrigues and Gaston 2002).

Table 3 Comparison of diversity measures for small mammal assemblages of sites in the Tanami Desert and Uluru-Kata Tjuta National Park, Northern Territory, Australia

Region	No. of sites	Species richness (alpha)	Species richness (gamma)	Species beta diversity (additive)	Phylogenetic beta diversity (ΔPD)
Tanami	15	3.13	14	10.87	59.92
Uluru	13	6.54	13	6.46	22.54

Species beta diversity is calculated as the difference between the total species richness of a region (gamma) and the mean site-level species richness (alpha)

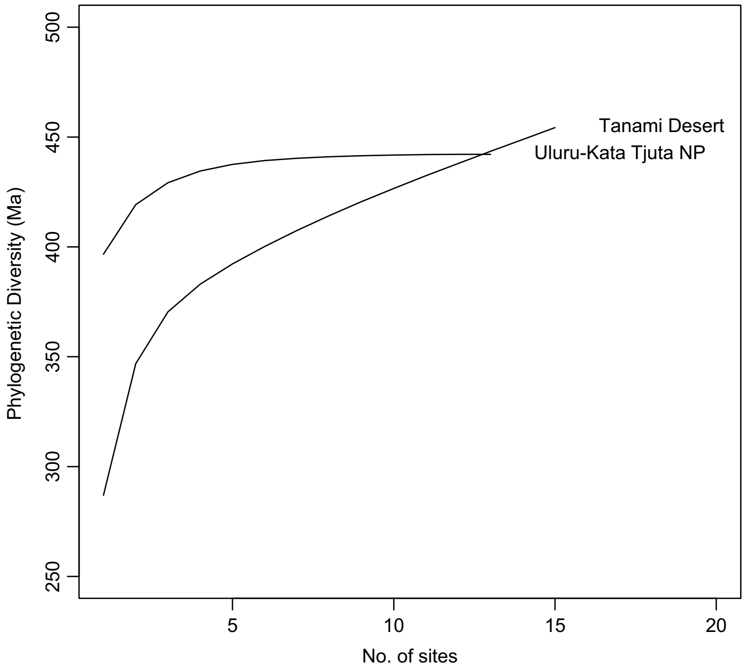

Fig. 5 Sample-based rarefaction curves for small mammal assemblages of sites in the Tanami Desert and Uluru-Kata Tjuta National Park, Northern Territory, Australia. See Morton et al. (1994) for a description of the data. Phylogenetic beta diversity (ΔPD) is higher among the Tanami sites than the Uluru sites

I generated PD rarefaction curves and ΔPD values for the mammal faunas of 71 of the 79 terrestrial ecoregions recognised by Olson et al. (2001) as constituting the Australasian biogeographic realm. Data were sourced from the wildfinder database (http://worldwildlife.org/pages/wildfinder) of the World Wildlife Fund. Eight ecoregions were excluded from the analysis because they had less than two species and thus a ΔPD value could not be calculated.

The ecoregions show huge variation in species richness and, as expected, Phylogenetic Diversity is highly dependent on species richness (Fig. 6). Tropical ecoregions (such as the central range Montane rainforests, New Guinea) have high species richness and high Phylogenetic Diversity (Fig. 6, Table 4). When considering phylogenetic dispersion, however, other ecoregions show unusually high or low values given their species richness (Table 4). The ecoregion with the lowest ΔPD is the New Caledonia dry forests. Because of its isolation, this fauna consists exclusively of bats and thus all the species are relatively closely related. The ecoregion with the highest ΔPD was the Mount Lofty woodlands of South Australia, reflecting relatively high numbers of marsupial species compared to the more tropically distributed bats and rodents.

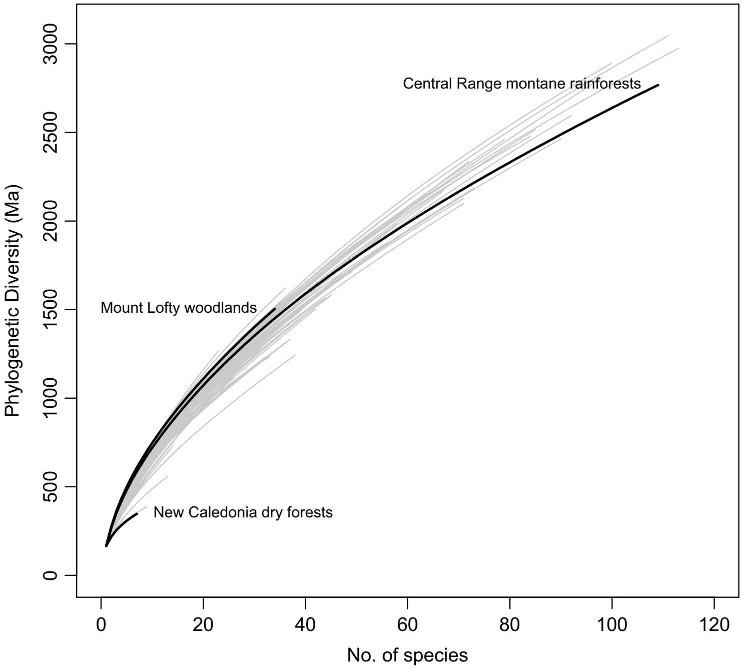

Fig. 6 Species-based rarefaction curves for mammal assemblages of terrestrial ecoregions of the Australasian biogeographic realm Ecoregions are as defined by Olson et al. (2001). Data are sourced from the wildfinder database (http://worldwildlife.org/pages/wildfinder). Three ecoregions are highlighted, as having minimum (New Caledonia dry forests), maximum (Mount Lofty woodlands) or median (Central Range montane rainforests) values of phylogenetic dispersion (ΔPD)

Table 4 Comparison of diversity measures for mammal assemblages of selected ecoregions of the Australasian biogeographic realm

Ecoregion	Species richness	Phylogenetic diversity (Ma)	Phylogenetic dispersion (ΔPD)
New Caledonia dry forests	7	347	51.3
Central range montane rainforests	109	2768	103.6
Mount Lofty woodlands	34	1504	110.7

Future Directions

As demonstrated here, rarefaction of PD has a straightforward application in standardising PD across samples so that they can be compared directly. Further, depending on the accumulation unit, the rarefaction formula can be extended to the calculation of metrics of phylogenetic evenness, phylogenetic beta-diversity and phylogenetic dispersion. However, the application of the PD rarefaction formula and its extension to other metrics is still very much in its infancy. Here I will outline some future directions for PD rarefaction.

Rarefaction by units of species allows for the comparison of locations while controlling for variation in species richness. This can easily be done by either rarefying all locations to a given number of species (Nipperess and Matsen 2013) or via ΔPD as demonstrated here. This kind of correction has previously been done by including species richness as an explanatory variable in a statistical model and taking the residuals (Davies et al. 2008) or by comparison to a null model derived by repeated subsampling (Davies et al. 2007). The latter method is often used as a statistical test of phylogenetic dispersion (also known as phylogenetic structure) where random draws are taken from a species pool, representing a null community assembly process (Webb 2000). Such methods are no longer necessary as the exact relationship between species richness and PD is described by the rarefaction curve (Nipperess and Matsen 2013). Further, the exact analytical solution is computationally efficient, allowing for practical application to very large datasets.

By removing the effect of species richness, we can identify "evolutionary hotspots" with higher than expected phylogenetic diversity (Davies et al. 2008; Nipperess and Matsen 2013) on a regional or global scale. We can then use the standardised PD values (called relative PD by Davies et al. 2007) to explore the environmental, ecological and historical processes that lead to the observed patterns of high or low phylogenetic dispersion (Kooyman et al. 2013). Ultimately, we may be able to develop the theory to predict these patterns (Davies et al. 2007), in a similar vein to what has been done for species richness (Arrhenius 1921; MacArthur and Wilson 1963; Rosindell et al. 2011). For example, the relationship of species richness with area is well known but the phylogeny-area relationship has only recently begun to be explored (Morlon et al. 2011). Rarefaction curves have an obvious connection to species-area curves (Olszewski 2004) and thus the development of PD rarefaction may well improve understanding of the phylogeny-area relationship. In particular, species-based rarefaction of PD allows for the separation of species diversity effects from those purely explained by phylogeny.

It is possible to predict how much Phylogenetic Diversity is yet to be sampled from the observed rarefaction curve. Rarefaction is the basis of several species diversity estimators, which attempt to calculate total diversity (including unseen species) for a set of individuals or samples by effectively extending the curve beyond the observed sampling depth (Colwell and Coddington 1994). It follows that a useful extension of PD rarefaction would be a PD estimator that predicts unseen branch length, given the observed rate of accumulation of PD. It is important to note that PD rarefaction calculates the expected branch length gained by adding additional accumulation units but does not predict where on the tree these branches will come from. Similarly, a biodiversity estimator based on PD rarefaction may be able to predict the amount of PD not yet sampled but would not be able to predict where these unseen branches would be added to an existing tree. This would be, nevertheless, an exciting development.

It has recently been proposed that the standardisation of samples for species diversity should not be done by rarefaction to the same size (i.e. no. of individuals), but rather by sample completeness (Alroy 2010; Jost 2010; Chao and Jost 2012). Completeness, when measured by a statistic known as coverage (Good 1953), is the proportion of individuals in a community that are represented by species in a sample from that community (Chao and Jost 2012). When samples differ in their coverage, they should be standardised to equal coverage before a "fair" comparison can be made. Much like expected species richness, the coverage of a sample can be estimated from the sample size and the distribution of individuals among the species in the sample (Chao and Jost 2012). Given that standardisation by sample completeness has been shown to yield a less biased comparison of species richness between communities (Chao and Jost 2012), it would be desirable to have a similar method of standardisation for PD. Since rarefaction of coverage is mathematically related to rarefaction of sample size, the recent work on estimating PD from sample size will no doubt form the basis from which estimated PD for sample coverage will be developed.

Finally, a general issue when considering any PD measure is uncertainty regarding the length of branches and the topology (branching pattern) of the tree. All PD measures (including those presented here) assume that the branch lengths and their arrangement in the tree are perfectly known. This is obviously an abstraction, although PD can be surprisingly robust to this source of variation (Swenson 2009). One solution to this dilemma is to calculate PD, including rarefied PD, for a large number of possible trees and report the mean and confidence limits. The output from a Bayesian phylogenetic analysis is a large number of trees, each with their own topology and corresponding branch lengths (see for example Jetz et al. 2012) and so lends itself well to this approach. However, when the possible trees number in the thousands and tens of thousands, this is obviously computationally intensive. An analytical solution, directly incorporating uncertainty into the calculation, would therefore be desirable. This is not an easy extension of the PD rarefaction solution because both variation in branch length and topology (affecting the probability of encountering internal branches) would need to be taken into account. It is worth remembering that phylogenetic relationships are not the only source of uncertainty when investigating real ecological communities – neither the abundance, nor even the presence (occupancy), of species are necessarily known with precision.

Conclusion

The formulation for the rarefaction of Phylogenetic Diversity (PD) is given in expanded form to show its simplicity and its connection to the classic formula for the rarefaction of species richness (Hurlbert 1971; Simberloff 1972). The method is exact and efficient and should be preferred over the algorithmic (Monte Carlo) solution involving repeated random sub-sampling. Further, the extension to the calculation of ΔPD provides a flexible and general framework for the measurement of biodiversity as phylogenetic evenness, phylogenetic beta-diversity or phylogenetic dispersion. The applications of PD rarefaction and ΔPD presented here are hopefully useful in improving understanding of the importance of rarefaction in ecology and in guiding future applications of the method. There are, I believe, exciting prospects for PD rarefaction in the future, including as a general method for standardising PD by removing variation with species richness, and for predicting unseen (i.e. un-sampled) PD. The recent availability of comprehensive phylogenies (Bininda-Emonds et al. 2007; Jetz et al. 2012) and rich data on species occurrences (Flemons et al. 2007), coupled with analytical advances such as PD rarefaction, allows us to better understand the distribution of Phylogenetic Diversity on the surface of the Earth and the processes giving rise to that distribution. This is valuable for its own sake but will also inform efforts to conserve as much of the Tree of Life as possible in the face of future extinctions (Rosauer and Mooers 2013).

Acknowledgements I would like to thank Frederick Matsen (Fred Hutchinson Cancer Research Center), Daniel Faith (Australian Museum), Peter Wilson (Macquarie University), Anne Chao (National Tsing Hua University) and Robert Colwell (University of Connecticut) for their input regarding the rarefaction of PD and what it all might mean. Daniel Miranda-Esquivel (Universidad Industrial de Santander) and Pedro Cardoso (University of Helsinki) provided helpful feedback on an earlier version of this chapter. This research was supported by the Australian Research Council (DP0665761 and DP1095200).

References

Allen B, Kon M, Bar-Yam Y (2009) A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. Am Nat 174:236–243

Alroy J (2010) The shifting balance of diversity among major marine animal groups. Science 329:1191–1194

Anderson MJ, Crist TO, Chase JM et al (2010) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol Lett 14:19–28

Arrhenius O (1921) Species and area. J Ecol 9:95–99

Bininda-Emonds ORP, Cardillo M, Jones KE et al (2007) The delayed rise of present-day mammals. Nature 446:507–512

Bryant JA, Lamanna C, Morlon H et al (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proc Natl Acad Sci 105:11505–11511

Chao A, Jost L (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533–2547

Chao A, Chiu CH, Jost L (2010) Phylogenetic diversity measures based on Hill numbers. Phil Trans R Soc B 365:3599–3609

Chiarucci A, Bacaro G, Rocchini D, Fattorini L (2008) Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. Community Ecol 9:121–123

Colwell RK, Coddington J (1994) Estimating terrestrial biodiversity through extrapolation. Phil Trans R Soc B 345:101–118

Crist TO, Veech JA (2006) Additive partitioning of rarefaction curves and species-area relationships: unifying alpha, beta and gamma-diversity with sample size and habitat area. Ecol Lett 9:923–932

Davies TJ, Buckley LB (2012) Exploring the phylogenetic history of mammal species richness. Glob Ecol Biogeogr 21:1096–1105

Davies RG, Orme CDL, Webster AJ et al (2007) Environmental predictors of global parrot (Aves: Psittaciformes) species richness and phylogenetic diversity. Glob Ecol Biogeogr 16:220–233

Davies TJ, Fritz SA, Grenyer R et al (2008) Phylogenetic trees and the future of mammalian biodiversity. Proc Natl Acad Sci 105(Suppl 1):11556–11563

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61:1–10

Faith DP (2013) Biodiversity and evolutionary history: useful extensions of the PD phylogenetic diversity assessment framework. Ann N Y Acad Sci 1289:69–89

Faith DP, Reid CAM, Hunter J (2004) Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment. Conserv Biol 18:255–261

Faith DP, Lozupone CA, Nipperess DA, Knight R (2009) The cladistic basis for the phylogenetic diversity (PD) measure links evolutionary features to environmental gradients and supports broad applications of microbial ecology's 'Phylogenetic Beta Diversity' framework. Int J Mol Sci 10:4723–4741

Ferrier S, Manion G, Elith J, Richardson K (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Divers Distrib 13:252–264

Flemons P, Guralnick R, Krieger J et al (2007) A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). Ecol Informa 2:49–60

Fritz SA, Bininda-Emonds ORP, Purvis A (2009) Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecol Lett 12:538–549

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 4:379–391

Helmus MR, Keller W, Paterson MJ et al (2010) Communities contain closely related species during ecosystem disturbance. Ecol Lett 13:162–174

Hurlbert S (1971) The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577–586

Jetz W, Thomas GH, Joy JB et al (2012) The global diversity of birds in space and time. Nature 491:444–448

Jost L (2006) Entropy and diversity. Oikos 113:363–375

Jost L (2010) The relation between evenness and diversity. Diversity 2:207–232

Kobayashi S (1974) The species-area relation I. A model for discrete sampling. Res Popul Ecol 15:223–237

Kooyman RM, Rossetto M, Sauquet H, Laffan SW (2013) Landscape patterns in rainforest phylogenetic signal: isolated islands of refugia or structured continental distributions? PLoS One 8:e80685

Lande R (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos 76:5–13

Law B, Anderson J, Chidel M (1998) A bat survey in State Forests on the south-west slopes region of New South Wales with suggestions of improvements for future surveys. Aust Zool 30:467–479

Lozupone CA, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235

Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev 32:557–578

MacArthur RH, Wilson EO (1963) An equilibrium theory of insular zoogeography. Evolution 17:373–387

Mao C, Colwell RK, Chang J (2005) Estimating the species accumulation curve using mixtures. Biometrics 61:433–441

Medellin RA, Equihua M, Amin M (2000) Bat diversity and abundance as indicators of disturbance in neotropical rainforests. Conserv Biol 14:1666–1675

Morlon H, Schwilk DW, Bryant JA et al (2011) Spatial patterns of phylogenetic diversity. Ecol Lett 14:141–149

Morton SR, Brown JH, Kelt DA, Reid JRW (1994) Comparisons of community structure among small mammals of North American and Australian deserts. Aust J Zool 42:501–525

Nipperess DA, Matsen FA IV (2013) The mean and variance of phylogenetic diversity under rarefaction. Methods Ecol Evol 4:566–572

Nipperess DA, Faith DP, Barton K (2010) Resemblance in phylogenetic diversity among ecological assemblages. J Veg Sci 21:809–820

O'Dwyer JP, Kembel SW, Green JL (2012) Phylogenetic diversity theory sheds light on the structure of microbial communities. PLoS Comput Biol 8:e1002832

Olson D, Dinerstein E, Wikramanayake E et al (2001) Terrestrial ecoregions of the world: a new map of life on Earth. Bioscience 51:933–938

Olszewski TD (2004) A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. Oikos 104:377–387

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pardi F, Goldman N (2007) Resource-aware taxon selection for maximizing phylogenetic diversity. Syst Biol 56:431–444

Petchey OL, Gaston KJ (2002) Functional diversity(FD), species richness and community composition. Ecol Lett 5:402–411

Proches S, Wilson J, Cowling R (2006) How much evolutionary history in a 10 x 10 plot? Proc R Soc B 273:1143–1148

R Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21:24–43

Rodrigues ASL, Gaston KJ (2002) Maximising phylogenetic diversity in the selection of networks of conservation areas. Biol Conserv 105:103–111

Rosauer DF, Mooers AO (2013) Nurturing the use of evolutionary diversity in nature conservation. Trends Ecol Evol 28:322–323

Rosauer D, Laffan SW, Crisp MD et al (2009) Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. Mol Ecol 18:4061–4072

Rosindell J, Hubbell SP, Etienne RS (2011) The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol Evol 26:340–348

Sanders HL (1968) Marine benthic diversity: a comparative study. Am Nat 102:243–282

Simberloff D (1972) Properties of the rarefaction diversity measurement. Am Nat 106:414–418

Swenson NG (2009) Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities. PLoS One 4:e4390

Tidemann CR, Woodside DP (1978) A collapsible bat-trap and a comparison of results obtained with the trap and with mist-nets. Wildl Res 5:355–362

Turnbaugh PJ, Hamady M, Yatsunenko T et al (2009) A core gut microbiome in obese and lean twins. Nature 457:480–484

Ugland K, Gray JS, Ellingsen KE (2003) The species-accumulation curve and estimation of species richness. J Anim Ecol 72:888–897

Webb CO (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am Nat 156:145–155

Webb CO, Pitman N (2002) Phylogenetic balance and ecological evenness. Syst Biol 51:898–907

Webb CO, Ackerly DD, McPeek M, Donoghue MJ (2002) Phylogenies and community ecology. Ann Rev Ecol Syst 33:475–505

Yu DW, Ji Y, Emerson BC, Wang X (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623