The mean and variance of phylogenetic diversity under rarefaction

David A. Nipperess[1] and Frederick A. Matsen IV[2]*

[1]*Department of Biological Sciences, Faculty of Science, Macquarie University, NSW 2109, Australia; and* [2]*Computational Biology Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Mail stop: M1-B514, Seattle, WA 98109-1024, USA*

Summary

1. Phylogenetic diversity (PD) depends on sampling depth, which complicates the comparison of PD between samples of different depth. One approach to dealing with differing sample depth for a given diversity statistic is to rarefy, which means to take a random subset of a given size of the original sample. Exact analytical formulae for the mean and variance of species richness under rarefaction have existed for some time, but no such solution exists for PD.

2. We have derived exact formulae for the mean and variance of PD under rarefaction. We confirm that these formulae are correct by comparing exact solution mean and variance to that calculated by repeated random (Monte Carlo) subsampling of a data set of stem counts of woody shrubs of Toohey Forest, Queensland, Australia. We also demonstrate the application of the method using two examples: identifying hot spots of mammalian diversity in Australasian ecoregions and characterizing the human vaginal microbiome.

3. There is a very high degree of correspondence between the analytical and random subsampling methods for calculating mean and variance of PD under rarefaction, although the Monte Carlo method requires a large number of random draws to converge on the exact solution for the variance.

4. Rarefaction of mammalian PD of ecoregions in Australasia to a common standard of 25 species reveals very different rank orderings of ecoregions, indicating quite different hot spots of diversity than those obtained for unrarefied PD. The application of these methods to the vaginal microbiome shows that a classical score used to quantify bacterial vaginosis is correlated with the shape of the rarefaction curve.

5. The analytical formulae for the mean and variance of PD under rarefaction are both exact and more efficient than repeated subsampling. Rarefaction of PD allows for many applications where comparisons of samples of different depth are required.

Key-words: alpha diversity, phylogenetic diversity, rarefaction sampling depth

Introduction

Phylogenetic diversity (PD), the total branch length of a phylogenetic tree, has been extensively used as a measure of biodiversity. Originally conceived of as a method for prioritising regions for conservation (Faith 1992), PD has seen wider use in other applications such as biogeography (Davies & Buckley 2011), macroecology (Meynard *et al.* 2011) and microbial ecology (Lozupone & Knight 2008; Turnbaugh *et al.* 2008; Caporaso *et al.* 2012; Yu *et al.* 2012; Phillips *et al.* 2012). This increasing breadth of application can be attributed to a number of desirable properties including the following: (1) explicitly addressing the nonequivalence of species in their contribution to overall diversity, (2) acting as a surrogate for other aspects of diversity such as functional diversity (Cadotte *et al.* 2009, but see also Faith 1996), (3) incorporating informa-

tion on the evolutionary history of communities and biotas and (4) being robust to problems of species delineation because the relationships between populations and even individuals can be represented by relative branch lengths without the need to establish absolute species identity. Further, the original simple formulation of Faith (1992) has been built on to produce a broader 'PD calculus' measuring such aspects of diversity as phylogenetic endemism (Faith *et al.* 2004; Rosauer *et al.* 2009), evenness (Hill 1973; Allen *et al.* 2009) and resemblance (Ferrier *et al.* 2007; Lozupone & Knight 2008; Faith *et al.* 2009; Nipperess *et al.* 2010). For the purposes of this study, when referring to 'phylogenetic diversity' and 'PD', we refer explicitly to the definition of Faith (1992), where diversity is measured as the sum of branch lengths of a phylogenetic tree.

Phylogenetic diversity increases with increasing sampling effort just like many other measures of biodiversity. Thus, the comparison of the phylogenetic diversity of communities is not straightforward when sample sizes differ, as is common with real data sets. Unless data are standardized in some sense to

*Correspondence author. E-mail: matsen@fhcrc.org

account for differences in sample size or effort, the relative diversity of communities can be profoundly misinterpreted (Gotelli & Colwell 2001).

The established solution to the problem of interpreting diversity estimates with samples of varying size is rarefaction. The rarefaction of a given sample of size n to a level k is simply the uniform random choice of k of the n observations (typically without replacement). The observations are typically of either individual organisms or collections of organisms, giving either individual-based or sample-based rarefaction curves (Gotelli & Colwell 2001). To consider a given measure of diversity under rarefaction, the measure of diversity is simply applied to the rarefied sample. Researchers are typically interested in the expectation and variance of a measure of diversity under rarefaction.

Rarefaction curves can be used to understand the depth of sampling of a community compared with its total diversity. Additionally, rarefaction curves capture information about evenness (Olszewski 2004) and beta-diversity (Crist & Veech 2006), depending on whether observations are of individuals or collections. Rarefaction curves have been computed for phylogenetic diversity (Lozupone & Knight 2008; Turnbaugh *et al.* 2008; Caporaso *et al.* 2012; Yu *et al.* 2012). In each of these cases, rarefaction was not by counts of individual organisms or collections of such, but was instead based on counts of unique sequences or operational taxonomic units. Rarefaction by such units, including taxonomic species, makes sense in the context of phylogenetic diversity where it might not with other measures of biodiversity. In effect, with these examples, rarefaction is by the tips of the tree, and the resulting curve gives an indication of tree shape and distribution of sample observations among the tips of the tree.

One way to obtain summary statistics such as expectation and variance under rarefaction is to compute these statistics on samples drawn using a Monte Carlo procedure, that is, calculate the desired statistics on a collection of random draws. On the other hand, there are closed-form solutions for the mean of many measures of biodiversity under rarefaction. For example, an analytical solution is well known for species diversity, can be calculated for rarefaction by individuals and samples and is much more efficient than re-sampling (Hurlbert 1971; Ugland *et al.* 2003; Chiarucci *et al.* 2008). However, we are not aware of such a formula for any phylogenetic diversity metrics.

In this study, we establish analytical formulae for the mean and variance of phylogenetic diversity under rarefaction. We develop these formulas in the setting of a phylogenetic tree with 'marks', which are a simple generalization allowing multiplicity of observations and arbitrary positions of observations along the tree.

Materials and methods

There are two different notions of the induced phylogenetic diversity (PD) of a subset K of the leaves L of a tree T; these notions have been called *unrooted* and *rooted* PD (Pardi & Goldman 2007). Unrooted PD is the total branch length of the smallest unrooted subtree contained in T that has all of the leaves in

K. Rooted PD is the total branch length of the smallest rooted tree containing the original root of T as well as the selected leaves K. The rooted definition was that originally intended by Faith (1992): see Faith (2006) for a historical discussion. These two need not be the same: for example, any K consisting of a single element will have zero unrooted phylogenetic diversity, but nonzero rooted phylogenetic diversity. It is important to make a distinction between rooted and unrooted PD vs. rooted and unrooted trees. In our formulation, we are effectively treating all trees as rooted, by assigning an arbitrary root if necessary, and that unrooted and rooted PD refers specifically to the forced inclusion (or not) of a special root.

The two definitions of PD are useful in different domains of application. For example, for conservation applications, keeping a single species has significant value, thus it makes sense to have nonzero PD for a single species. On the other hand, when comparing the level of ecological diversity between environments, it may not make sense to keep the root, in which case the diversity between the members of a set of size one is zero.

We will derive formulae for both definitions of PD. However, the description of the variance of unrooted PD will be deferred to the Appendix.

Formulae for rarefaction of phylogenetic diversity can be easily and productively generalized from the notion of a tree to the notion of a *tree with marks*, which allows more flexibility in abundance weighting and attachment locations. We define a *tree with marks* as a tree along with a collection of special points on the tree (marked with stars in Fig. 1), which may be present with multiplicity. The induced subtree of a collection of marks on a phylogenetic tree is the smallest connected set that contains all of those marks. The phylogenetic diversity of a (sub)tree is the total branch length of the tree.

In this setting, marks represent observations. Thus, if a certain leaf taxon t is observed x times, x marks are put at t. However, it is just easy to generalize to the setting where marks appear on the interior of tree edges. The motivation for working in terms of marks is that it
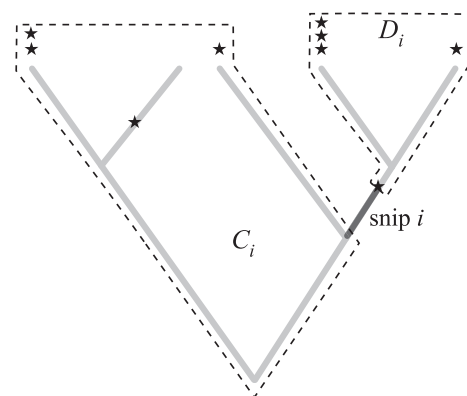
Fig. 1. A hypothetical phylogenetic tree illustrating key concepts in the formulation of the rarefaction of phylogenetic diversity. The tree is populated with marks (indicated by stars) that represent observations of particular points on the tree in a sample. Marks might commonly be placed only at the leaves (tips) of the tree, but allowing marks to occur anywhere provides for more flexible applications. Multiple marks indicate multiple observations: for example, several individuals of a species. The tree can then be broken up into snips, which are the edge segments between marks and/or internal nodes. For each snip i, there are two sets of marks, C_i and D_i, which name the set of marks that are on the proximal (towards the root) side of i vs. those on the distal (towards the leaves) side of i.

formalizes the notion of observation count and affords some extra flexibility for location of observations. In particular, microbial ecologists often census a given community by high-throughput sequencing, and it is not practical to build a phylogenetic tree on all of the sequences thus created. For this reason, sometimes scientists either map sequences to trees using similarity search plus a most recent common ancestor strategy, as in the work of Huson *et al.* (2007), or 'place' the sequences into the tree using a phylogenetic criterion (Berger *et al.* 2011; Matsen *et al.* 2010). The attachment point of a mapping of a sequence into a tree is then considered as a mark.

The *unrooted phylogenetic diversity* of a tree with marks is the total branch length of the tree induced by those marks, that is, the total branch length of the smallest connected subset of the tree containing the marks. The *rooted phylogenetic diversity* of a rooted tree with marks is equal to the unrooted phylogenetic diversity of the tree with the given marks along with a mark added at the root; in this case, the path from the root to the selected leaves is always included in the PD calculation. These are simple generalizations of the corresponding definitions for leaf observations.

The following sections will be concerned with rarefying the collection of marks and computing phylogenetic diversities of the corresponding induced subtrees. We will use *proximal* to indicate the direction towards the root and *distal* to mean the opposite. If T is unrooted, we will still use these terms for descriptive purposes; in this case, an arbitrary root can be permanently assigned.

We fix a nonempty collection M of n marks on a tree T and some number $1 \leq k < n$ of marks to sample for our rarefaction. Again, marks can be present multiple times in a collection, enabling the expression of multiplicity of observation of a taxon or sequence.

Definition 1: *Define an edge snip* to be a maximal segment of an edge with no marks or internal nodes.

Say there are s snips on the tree with marks and that they are indexed by i. Let ℓ_i be their length for $1 \leq i \leq s$. Let C_i be the set of marks that are proximal to snip i and D_i be the set of marks that are distal to snip i (Fig. 1).

Definition 2: *For every $1 \leq i \leq s$,* let X_i^r be the random variable that is equal to one if there is at least one mark on the distal side of snip i after rarefaction and zero otherwise. Let X_i^u be the random variable that is equal to one if there is at least one mark on each side of snip i after rarefaction and zero otherwise.

The following two statements are true for $X \in \{X^r, X^u\}$ with the corresponding $Y \in \{Y^r, Y^u\}$. The phylogenetic diversity Y after rarefaction can be expressed as the random variable

$$Y = \sum_i \ell_i X_i. \qquad \text{eqn 1}$$

because the length of a snip i contributes to the PD exactly when the corresponding $X_i = 1$.

Thus, $\mathbb{E}(Y) = \sum_i \ell_i \mathbb{E}(X_i)$, and

$$\text{Var}(Y) = \sum_{i,j} \ell_i \ell_j \text{Cov}(X_i, X_j). \qquad \text{eqn 2}$$

To calculate expectations and covariances of the X_i, the following definition will be useful. Fix an $R \subset M$. Let $q_k(R)$ be the probability that nothing in R is selected in a uniform sample of size k from M without

replacement. Recalling that $n = |M|$, note that (from the hypergeometric distribution):

$$q_k(R) = \begin{cases} \binom{n - |R|}{k} / \binom{n}{k} & \text{when } n - |R| \geq k \\ 0 & \text{otherwise} \end{cases}$$

with the convention that $\binom{x}{0} = 1$ for all $x \in \mathbb{N}$.

Note that the $q_k(R)$ can be calculated for successive k by observing that

$$q_{k+1}(R) = \frac{n - |R| - k}{n - k} q_k(R).$$

Because the q_k only depend on the size of R, a computer implementation only needs to calculate the $q_k(R)$ once for any R of a given size; the $q_k(R)$ notation was chosen for convenience.

ROOTED PHYLOGENETIC DIVERSITY

As described above, rooted phylogenetic diversity does PD calculation while always including the root. By equations (1) and (2), all that is needed is the mean and the covariance matrix of the X_i^r's. Note that X_i^r is zero unless at least one element of D_i is sampled, in which case it is one. Thus,

$$\mathbb{E}[X_i^r] = 1 - q_k(D_i). \qquad \text{eqn 3}$$

$X_i^r X_j^r$ is zero unless the rarefaction samples at least one element of both D_i and D_j, in which case it is one. The probability that one or both of these are empty under rarefaction is $q_k(D_i) + q_k(D_j) - q_k(D_i \cup D_j)$, thus

$$\mathbb{E}[X_i^r X_j^r] = 1 - q_k(D_i) - q_k(D_j) + q_k(D_i \cup D_j).$$

By equation (3),

$$\mathbb{E}[X_i^r] \mathbb{E}[X_j^r] = 1 - q_k(D_i) - q_k(D_j) + q_k(D_i) q_k(D_j),$$

thus,

$$\text{Cov}(X_i^r, X_j^r) = q_k(D_i \cup D_j) - q_k(D_i) q_k(D_j).$$

In summary,

$$\mathbb{E}[Y^r] = \sum_i \ell_i [1 - q_k(D_i)]$$

$$\text{Var}[Y^r] = \sum_{i,j} \ell_i \ell_j [q_k(D_i \cup D_j) - q_k(D_i) q_k(D_j)].$$

This solution can be seen to be a generalization of the analytical formulae for the mean and variance of expected species richness under rarefaction (Hurlbert 1971; Heck Jr *et al.*, 1975) as follows. Consider the special case of a 'star' tree with all tips sharing a single common ancestor, where all marks are located at the tips of the tree (with the exception of one mark placed at the root) and where all branch lengths (and thus all snips) have a length of one. Under these particular circumstances, the species richness and phylogenetic diversity of the collection of marks are equal and the formulae for mean and variance of expected phylogenetic diversity simplify to their equivalents for species richness.

UNROOTED PHYLOGENETIC DIVERSITY

Assume as above that we are sampling $k > 0$ marks for our rarefaction. It is not possible for the rarefaction samples from both C_i and D_i to be empty. Thus, these two events are mutually exclusive, and

$$\mathbb{E}[X_i^u] = 1 - q_k(C_i) - q_k(D_i). \qquad \text{eqn 4}$$

Then, by equations (1) and (4),

$$\mathbb{E}(Y^u) = \sum_i \ell_i[1 - q_k(C_i) - q_k(D_i)]. \qquad \text{eqn 5}$$

The variance of the unrooted case is deferred to the Appendix.

EXAMPLE APPLICATIONS

We demonstrate our method for calculating the mean and variance of phylogenetic diversity under rarefaction using three examples. In the first, we compare the rarefaction curve generated by Monte Carlo randomization to that calculated by the exact analytical solution. The data are counts of stems of all woody shrubs in forty plots in Toohey Forest, Queensland, Australia. Within each plot, all plant stems above 0·3 m and below 3·0 m were counted; this figure was used as an index of abundance. All shrubs were identified with species, and a composite phylogeny was compiled from multiple published trees; see (Nipperess *et al.* 2010) for a more detailed description of the data. Stem counts were summed across all plots to produce a single value per species before rarefaction by individual stems. Of the total of 582 stems, rarefied values were calculated for every multiple of 10 stems from 10 to 580. For the Monte Carlo procedure, mean and variance of phylogenetic diversity were calculated from 2000 random subsamples of size k from the pool of 582 stems.

Our second example demonstrates rarefaction of phylogenetic diversity by units of species. Phylogenetic diversity of extant mammals was calculated for each terrestrial ecoregion on the Australian continental shelf (that is, Australia along with Tasmania, New Guinea and offshore islands). Terrestrial ecoregions are biogeographical units representing distinct species assemblages (Olson *et al.* 2001). Species lists of mammals for each ecoregion were sourced from the WildFinder database maintained by the World Wildlife Fund (http://www.worldwild-life.org/science/wildfinder/). Evolutionary relationships were sourced from a species-level supertree of the world's mammals (Bininda-Emonds *et al.* 2007). Because of the strong correlation between species richness and phylogenetic diversity, rarefaction allows for the comparison of ecoregions with the effect of spatial variation in species richness removed. To do this, the expected phylogenetic diversity for a subset of 25 mammalian species was calculated for each ecoregion. The value of 25 was chosen because it was the minimum species richness for this set of ecoregions.

Our third example comes from the human microbiome. We reanalyse a pyrosequencing data set describing bacterial communities from women with bacterial vaginosis (Srinivasan *et al.* 2012). Bacterial vaginosis (BV) has previously been shown to be associated with increased microbial community diversity (Fredricks *et al.* 2005). For this study, swabs were taken from 242 women from the Public Health, Seattle and King County Sexually Transmitted Diseases Clinic between September 2006 and June 2010, of which 220 samples resulted in enough material to analyse (data available as Sequence Read Archive submission SRA051298). Vaginal fluid for each specimen was also evaluated according to Nugent score, which provides a diagnostic score for BV that ranges from 0 (BV-negative) to 10 (BV-positive) based on the presence and absence of bacterial morphotypes as viewed under a microscope (Nugent *et al.* 1991). Selection of reference sequences and sequence preprocessing were performed using the methods described by Srinivasan *et al.* (2012). Of 452 358 reads passed quality filtering, with a median of 1779 reads per sample (range: 523–2366). For this application, we investigated the shape of the rarefaction curves with respect to resampling.

Results

There was a very high degree of correspondence between the analytical and Monte Carlo methods for the expected value and variance of phylogenetic diversity of the Toohey Forest data set under rarefaction (Figs 2 and 3; corresponding results for unrooted PD are not shown). In this application, the Monte Carlo estimate of the PD variance does not converge quickly to the exact value, as can be seen from the deviations of the points (generated from 2000 Monte Carlo samples) from the curve in Fig. 3. Such slow convergence provides further motivation for an exact formula.

Correcting phylogenetic diversity for the number of species present made a substantial difference to the ranking of terrestrial ecoregions in terms of their diversity (Fig. 4). With unrarefied phylogenetic diversity, the three highest ranked ecoregions (south-eastern Papuan rainforests, Southern New Guinea lowland rainforests and central range montane rainforests) are found in New Guinea. However, when variation in species richness is taken into account by rarefaction, two of the three highest ranked ecoregions (Australian Alps montane grasslands and Naracoorte woodlands) were in temperate Australia. Thus, the rarefied version demonstrates high phylogenetic diversity for this data set relative to the number of species present for those regions.

The rarefaction curves for the vaginal samples show a connection between the Nugent score of the sample and the shape of the curve (Fig. 5). The rarefaction curves for low Nugent score samples tend to start low and stay low. The high Nugent score samples typically start higher than low Nugent score samples and stay high.

Discussion

We have presented exact formulae for the mean and variance of rooted and unrooted phylogenetic diversity under rarefaction. This solution gives results that are indistinguishable from those given by Monte Carlo randomization. The analytical method is preferred both because its results are exact and can be more efficient than sampling.

Rarefaction of phylogenetic diversity is seeing growing use in a variety of biological disciplines and we highlight two specific applications here. Rarefaction of phylogenetic diversity by units of species allows for the assessment of phylogenetic diversity independent of species richness. Removing the influence of species richness can allow for the fairer comparison of the evolutionary history of fauna and flora. While it is possible to make this correction by taking the residuals from a regression between species richness and PD (Forest *et al.* 2007; Davies *et al.* 2011), the expected PD for a given species richness has also been determined by repeated subsampling of a species pool (Davies *et al.* 2006; Forest *et al.* 2007; Morlon *et al.* 2011; Yu *et al.* 2012). This latter method describes the relationship between phylogenetic diversity and species richness as a rarefaction curve. Our example of the mammalian faunas of the Australian continental shelf shows that such a correction can now be implemented with an exact analytical
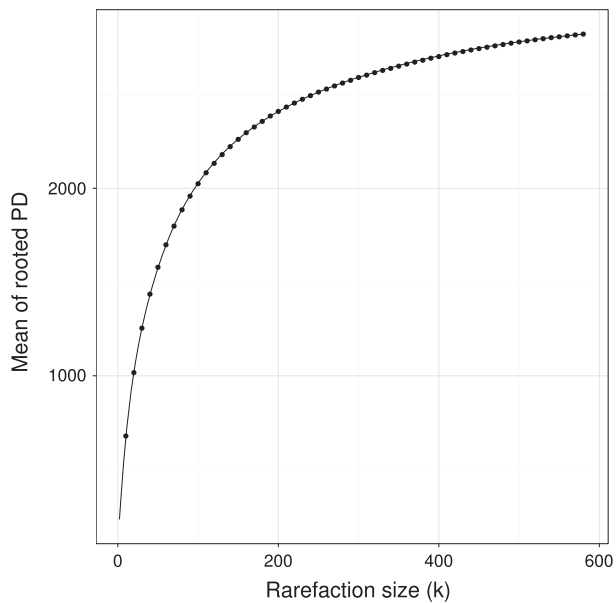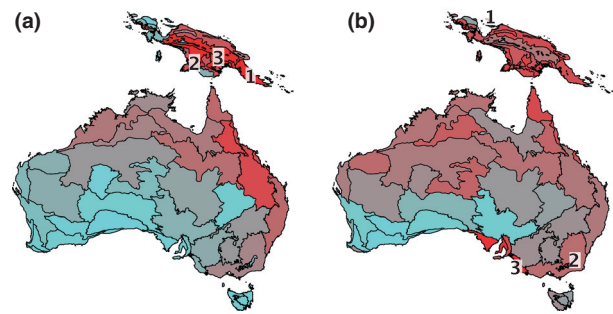
Fig. 2. Comparison of analytical value (curve) with Monte Carlo calculation with 2000 samples (points) for the mean of rooted PD under rarefaction.
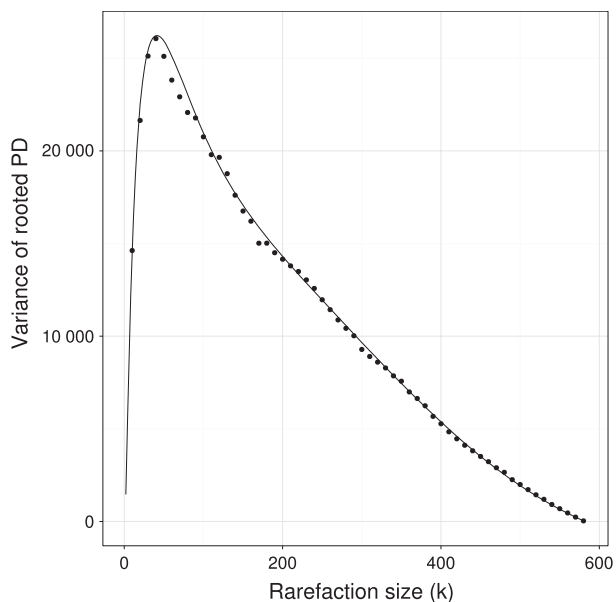

Fig. 3. Comparison of analytical value (curve) with Monte Carlo calculation with 2000 samples (points) for the variance of rooted PD under rarefaction.

Fig. 4. Phylogenetic diversity of mammal faunas for terrestrial ecoregions on the Australian continental shelf. Phylogenetic diversity is calculated for (a) all species present and (b) as an expected value after rarefaction to 25 species. Ecoregions are coloured light blue for low values to dark red for high values. The three highest ranked ecoregions in each case are indicated by number.
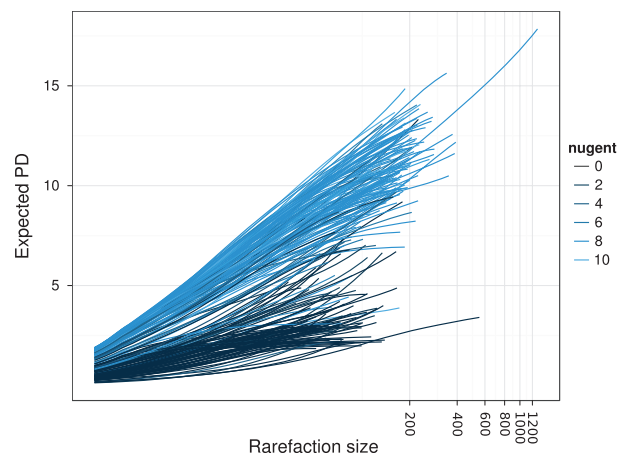
Fig. 5. Rarefaction curve of samples from Srinivasan *et al.* (2012). The Nugent score is a diagnostic score for bacterial vaginosis, with 0 being 'normal' and 10 being classified as BV.

solution rather than repeated subsampling. Further, as previously found by Forest *et al.* (2007) for the Cape Flora of South Africa, correction for the number of species makes a substantial difference to the rank order of phylogenetic diversity of sites.

The rarefaction curves for the vaginal samples give interesting information about the distribution of phylotypes in the vaginal microbiome. Some of this information recapitulates prior knowledge. For example, samples with low Nugent score are typically dominated by a handful of bacterial species in the

Lactobacillus genus. These rarefied curves start low and stay low. If there are also other distantly related organisms present, but in low abundance, the curve can start low and then curve up to a high level. The high Nugent score samples, which tend to start high and increase rapidly, indicate that there are a considerable number of taxa spread across the tree that appear in the samples with nontrivial count.

Software implementing the exact analytical solution for rarefaction of phylogenetic diversity is already available. The *phylorare* and *phylocurve* functions are implemented in the R statistical environment (R Development Core Team 2010). These functions calculate mean rooted phylogenetic diversity and can be used to standardize a set of samples to a particular level of sampling effort (*phylorare*) or generate a rarefaction curve for units of individuals, collections or species (*phylocurve*). These functions can be downloaded from http://david-nipperess.blogspot.com.au/. The pplacer suite of programs (http://matsen.fhcrc.org/pplacer/) is a collection of programs for 'phylogenetic placement' and associated analyses. The guppy software is the main binary to perform downstream analysis of collections of placements. It calculates Faith's

phylogenetic diversity as well as a number of other phylogenetic diversity measures, including the abundance-weighted $^q\bar{D}$ of Chao *et al.* (2010), a new one-parameter family of PD metrics (manuscript under preparation), phylogenetic entropy (Allen *et al.* 2009) and phylogenetic quadratic entropy (Rao 1982). It also calculates PD rarefaction curves with exact formulae as shown here, as well as those for phylogenetic quadratic entropy.

The work presented in this study relates to and extends previous work in similar areas. Faller *et al.* (2008) derived a central limit theorem for phylogenetic diversity under a model of random extinction. In doing so, they also derived the mean and variance of phylogenetic diversity under this model. This model is different than the setting of rarefaction in that the random variable signalling extinction is independent between species, which is not true for rarefaction to a given size considered here.

O'Dwyer *et al.* (2012) have also independently calculated a mean and variance under sampling, but with a different focus: they consider the distributions that might be achieved through a variety of sampling schemes from the 'metacommunity tree' of all extant lineages. They derive the expressions for the mean and variance of phylogenetic diversity that we use as a starting point for our proofs and then apply them to their various sampling distributions, using an approximation to bound the variance above. They consider the binomial, Poisson and negative binomial distributions, but do not consider the hypergeometric distribution as done here, which corresponds to the case of sampling without replacement. They do not derive exact expressions for the variance, nor do they consider unrooted PD or our more general setting.

Although we would like to extend the mean and variance formulas for PD under rarefaction to variants of PD, doing so may not be simple. For example, it would be interesting to investigate the mean and variance of $^q\bar{D}$, the abundance-weighted PD of Chao *et al.* (2010), under rarefaction. For $q = 0$, $^q\bar{D}$ is PD divided by T; for $q = 1$, $^q\bar{D}$ is $\exp(H_p/T)$; for $q = 2$, $^q\bar{D}$ is $1/(1-Q/T)$, where T is the maximum height of the phylogenetic tree, H_p is phylogenetic entropy (Allen *et al.* 2009), and Q is quadratic entropy (Rao 1982; Warwick *et al.*, 1995). Because these $q = 1$ and $q = 2$ cases are nonlinear functions of other abundance-weighted PD measures, the derivation of their mean and/or variance may be challenging.

Future work will include sensitivity of the PD rarefaction curve to tree shape and the distribution of individuals among species. It would also be interesting to investigate extensions of the present work to the 'coverage-based' framework recently proposed by Chao & Jost (2012), as well as an extension to 'unconditional variance' formulation of Colwell *et al.* (2012).

Acknowledgements

References

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

Berger, S.A., Krompass, D. & Stamatakis, A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, **60**, 291–302.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R., Beck, R.M.D., Grenyer, R., Price, S., Vos, R., Gittleman, J.L. & Purvis, A. (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Cadotte, M., Cavender-Bares, J., Tilman, D. & Oakley, T.H. (2009) Using phylogenetic, functional and trait diversity to understand patterns of plant community of productivity. *PLOS ONE*, **4**, e5695.

Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R. & Gilbert, J.A. (2012) The Western English Channel contains a persistent microbial seed bank. *ISME Journal*, **6**, 1089–1093.

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

Chao, A., Chiu, C.H. & Jost, L. (2010) Phylogenetic diversity measures based on hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **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 Ecology*, **9**, 121–123.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.

Crist, T.O. & Veech, J.A. (2006) Additive partitioning of rarefaction curves and species–area relationships: unifying α-, β-and γ-diversity with sample size and habitat area. *Ecology Letters*, **9**, 923–932.

Davies, R.G., Orme, C.D.L., Webster, A.J. Jones, K.E., Blackburn, T.M. & Gaston, K.J. (2006) Environmental predictors of global parrot (aves: Psittaciformes) species richness and phylogenetic diversity. *Global Ecology and Biogeography*, **16**, 220–233.

Davies, T.J & Buckley, L.B. (2011) Phylogenetic diversity as a window into the evolutionary and biogeographic histories of present-day richness gradients for mammals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **366**, 2414–2425.

Davies, T.J., Buckley, L.B., Davies, T.J. & Buckley, L.B. (2011) Phylogenetic diversity as a window into the evolutionary and biogeographic histories of present-day richness gradients for mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2414–2425.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

Faith, D.P. (1996) Conservation priorities and phylogenetic pattern. *Conservation Biology*, **10**, 1286–1289.

Faith, D.P. (2006) The role of the phylogenetic diversity measure, PD, in bioinformatics: Getting the definition right. *Evolutionary Bioinformatics Online*, **2**, 277–283.

Faith, D.P., Reid, C.A.M. & Hunter, J. (2004) Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment. *Conservation Biology*, **18**, 255–261.

Faith, D.P., Lozupone, C.A., Nipperess, D.A. & 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. *International Journal of Molecular Sciences*, **10**, 4723–4741.

Faller, B., Pardi, F. & Steel, M. (2008) Distribution of phylogenetic diversity under random extinction. *Journal of Theoretical Biology*, **251**, 286–296.

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. *Diversity and Distributions*, **13**, 252–264.

Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., der Bank, M., Reeves, G. et al. (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, **445**, 757–760.

Fredricks, D.N., Fiedler, T.L. & Marrazzo, J.M. (2005) Molecular identification of bacteria associated with bacterial vaginosis. *New England Journal of Medicine*, **353**, 1899–1911.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Heck Jr, K.L., van Belle, G. & Simberloff, D. (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, **56**, 1459–1461.

Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.

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

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Lozupone, C.A. & Knight, R. (2008) Species divergence and the measurement of microbial diversity. *FEMS Microbiology Reviews*, **32**, 557–578.

Matsen, F.A., Kodner, R.B. & Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

Meynard, C.N., Devictor, V., Mouillot, D., Thuiller, W., Jiguet, F. & Mouquet, N. (2011) Beyond taxonomic diversity patterns: how do α, β and γ components of bird functional and phylogenetic diversity respond to environmental gradients across France? *Global Ecology and Biogeography*, **20**, 893–903.

Morlon, H., Schwilk, D.W., Bryant, J.A., Marquet, P.A., Rebelo, A.G. Tauss, C., Bohannan, B.J.M. & Green, J.L. (2011) Spatial patterns of phylogenetic diversity. *Ecology Letters*, **14**, 141–149.

Nipperess, D.A., Faith, D.P. & Barton, K. (2010) Resemblance in phylogenetic diversity among ecological assemblages. *Journal of Vegetation Science*, **21**, 809–820.

Nugent, R.P., Krohn, M.A. & Hillier, S.L. (1991) Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology*, **29**, 297–301.

O'Dwyer, J.P., Kembel, S. W. & Green, J.L. (2012) Phylogenetic diversity theory sheds light on the structure of microbial communities. *PLOS Computational Biology* **8**, e1002832.

Olson, D.M., Dinerstein, E., Wikramanayake, E., Burgess, N., Powell, G., Underwood, E., D'Amico, J., Itoua, I., Strand, H., Morrison, J., Loucks, C., Allnutt, T., Ricketts, T.H., Kura, Y., Lamoreux, J., Wettengel, W., Hedao, P. & Kassem, K. (2001) Terrestrial ecoregions of the world: a new map of life on Earth. *BioScience*, **51**, 933–938.

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

Pardi, F. & Goldman, N. (2007) Resource-aware taxon selection for maximizing phylogenetic diversity. *Systematic Biology*, **56**, 431–444.

Phillips, C.D., Phelan, G., Dowd, S.E., MCdonough, M.M., Ferguson, A.W., Delton Hanson, J., Siles, L., Ordóñez-garza, N., San Francisco, M. & Baker, R.J. (2012) Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular Ecology*, **21**, 2617–2627.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject.org/.

Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43.

Rosauer, D., Laffan, S.W., Crisp, M.D., Donnellan, S.C. & Cook, L.G. (2009) Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology*, **18**, 4061–4072.

Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R. & Marrazzo, J.M., *et al.* (2012) Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLOS ONE*, **7**, e37818.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A. & Affourtit, J.P. et al. (2008) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Ugland, K.I., Gray, J.S. & Ellingsen, K.E. (2003) The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, **72**, 888–897.

Warwick, R.M. & Clarke, K.R., *et al.* (1995) New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–305.

Yu, D.W., Ji, Y., Emerson, B.C. & Wang, X. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Appendix

Here we will calculate the variance of the unrooted phylogenetic diversity Y^u via equation (2). First note that $(X_i^u)^2 = X_i^u$ because X_i^u only takes the values 1 and 0. Thus, if $i = j$, then $\mathrm{Cov}(X_i^u, X_j^u) = \mathbb{E}[X_i^u] - \mathbb{E}[X_i^u]^2$, which can be calculated using equation (4).

Now assume $i \neq j$. Instead of expressing the variance in terms of functions of C_i's and D_i's, we will express them in terms of the following quantities defined for a pair of edges i and j. If i is proximal to j, let $S_{i,j}$ be D_i (which is the <u>Same</u> side of i as j), $O_{i,j}$ be C_i (which is on the <u>Other</u> side of i from j), $S_{j,i}$ be C_j and $O_{j,i}$ be D_j. If j is proximal to i, the roles of i and j are reversed. If the path from i to j traverses the root, then let $S_{i,j}$ be C_i, $O_{i,j}$ be D_i, $S_{j,i}$ be C_j and $O_{j,i}$ be D_j. As $O_{i,j}$ and $S_{i,j}$ are just C_i and D_i in some order, we can think of X_i^u being the random variable equal to one if the rarefaction sample from both $O_{i,j}$ and $S_{i,j}$ is nonempty and zero otherwise; the corresponding definition is true for X_j^u. By these definitions, a key point is that $O_{i,j} \subset S_{j,i}$ and $O_{j,i} \subset S_{i,j}$.

To calculate $\mathrm{Cov}(X_i^u, X_j^u) = \mathbb{E}[X_i^u X_j^u] - \mathbb{E}[X_i^u]\mathbb{E}[X_j^u]$, note that $X_i^u X_j^u$ is zero unless the rarefaction samples at least one element of both $O_{i,j}$ and $O_{j,i}$, in which case it is one. The probability that one or both of these are empty under rarefaction is $q_k(O_{i,j}) + q_k(O_{j,i}) - q_k(O_{i,j} \cup O_{j,i})$, thus (for $i \neq j$)

$$\mathbb{E}[X_i^u X_j^u] = 1 - q_k(O_{i,j}) - q_k(O_{j,i}) + q_k(O_{i,j} \cup O_{j,i}).$$

By (4),

$$\begin{aligned}
\mathbb{E}[X_i^u]\mathbb{E}[X_j^u] &= [1 - q_k(O_{i,j}) - q_k(S_{i,j})][1 - q_k(O_{j,i}) - q_k(S_{j,i})] \\
&= 1 - q_k(O_{i,j}) - q_k(O_{j,i}) + q_k(O_{i,j})q_k(O_{j,i}) \\
&\quad - [1 - q_k(O_{i,j})]q_k(S_{j,i}) - q_k(S_{i,j})[1 - q_k(O_{j,i})] \\
&\quad + q_k(S_{i,j})q_k(S_{j,i}).
\end{aligned}$$

Thus, (again, for $i \neq j$),

$$\begin{aligned}
\mathrm{Cov}(X_i^u, X_j^u) &= q_k(O_{i,j} \cup O_{j,i}) - q_k(O_{i,j})q_k(O_{j,i}) \\
&\quad + [1 - q_k(O_{i,j})]q_k(S_{j,i}) + q_k(S_{i,j})[1 - q_k(O_{j,i})] \\
&\quad - q_k(S_{i,j})q_k(S_{j,i}).
\end{aligned}$$

These expressions can then be substituted back into equation (2) to obtain an expression for the variance of phylogenetic diversity.