Supporting Information (Appendix E)

Rarefaction and Extrapolation: Making Fair Comparison of Abundance-sensitive Phylogenetic Diversity among Multiple Assemblages

T. C. HSIEH AND ANNE CHAO

Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan

Appendix E. Rarefaction and extrapolation analysis for bat data

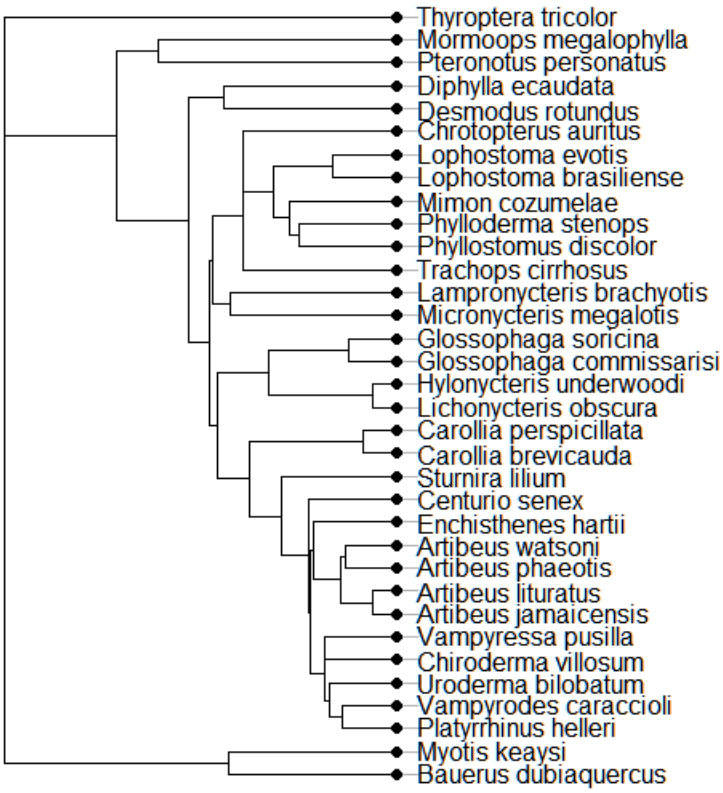
In the main text, we used bacterial data to illustrate our proposed sample-size- and coverage-based rarefaction/extrapolation (R/E) sampling curves for the phylogenetic diversity measure ${}^{q}PD$. Here we use phyllostomid bat data collected by Medellín et al. (2000) to further illustrate the method. These data were also analyzed by Allen et al. (2009). In this data set, 34 bat species were observed in four habitats (rainforest, cacao plantations, inactive oldfield plantations, and cornfield plantations) in the Selva Lacandona of Chiapas, Mexico. The original species abundance data for the four habitats are listed in Table E.1; the four habitats are referred to as rainforest, cacao, oldfield and cornfield, respectively. The rooted ultrametric phylogenetic tree of the 34 species was constructed by taking a sub-tree from the life tree on the earth provided in (Rosindell and Harmon 2012) by using the software OneZoom. The rooted tree is shown in Fig. E.1 in the Newick format. The tree depth is 60.4 Myr (the age of the root of the observed tree). Without loss of generality, we selected the root of all observed 34 species as our reference point in all our computation of estimates. Although the root of the observed taxa varies with sampling data, we can easily transform all our estimates to those for a new reference point; see Discussion in the main text and Appendix G for transformations. Because different investigation may require different reference points, we suggest reporting estimates as a function of the time or age of the reference point, as suggested in Chao et al. (2010).

For illustrative purposes and for display simplicity, we only considered three habitats by excluding the data from the oldfield habitat. The reference sample sizes for the rainforest, cacao and cornfield habitats are respectively (444, 699, 572), with the corresponding observed species richnesses (27, 21, 17). A summary of the data including the observed

species diversity (Hill numbers), the observed abundance-based phylogenetic diversity ${}^{q}PD$ for the order $q = 0, 1, 2$, and the estimated sample coverage of the three reference samples are shown in Table E.2.

TABLE E.1. The original species abundance/frequency data for four habitats of bat assemblages (Medellín et al. 2000).

	Habitat			
Species	Rainforest	Cacao	Oldfield	Cornfield
Pteronotus_paraguanensis	20	4	4	0
Mormoops_megalophylla	0	0	1	0
Micronycteris_megalotis	0	2	1	0
Lampronycteris_brachyotis	1	0	0	0
Mimon_cozumelae	2	1	4	0
Phyllostomus_discolor	2	0	5	0
Phylloderma_stenops	3	0	0	0
Lophostoma_brasiliense	1	2	0	1
Lophostoma_evotis	0	2	0	0
Trachops_cirrhosus	2	0	0	0
Chrotopterus_auritus	5	0	0	0
Glossophaga_soricina	20	53	70	58
Glossophaga_commissarisi	15	20	33	27
Lichonycteris_obscura	0	0	0	1
Hylonycteris_underwoodi	0	0	0	2
Carollia_brevicauda	72	159	250	45
Carollia_perspicillata	49	100	92	60
Sturnira_lilium	56	137	102	234
Artibeus_lituratus	76	82	51	66
Artibeus_jamaicensis	73	96	35	30
Enchisthenes_hartii	1	0	0	0
Artibeus_phaeotis	6	4	5	3
Artibeus_watsoni	17	13	12	15
Platyrrhinus_helleri	6	10	17	18
Vampyressa_pusilla	1	3	0	0
Vampyrodes_caraccioli	1	0	0	0
Chiroderma_villosum	0	0	1	5
Uroderma_bilobatum	7	4	2	4
Centurio_senex	1	2	0	2
Desmodus_rotundus	4	3	0	1
Diphylla_ecaudata	1	0	1	0
Thyroptera_tricolor	1	1	0	0
Bauerus_dubiaquercus	1	0	2	0
Myotis_keaysi	0	1	2	0



((Thyroptera_tricolor:60.4,((Mormoops_megalophylla:36.7,Pteronotus_parnellii:36.7)Mormoopidae{Mustached_Bats_and_more}:6.5,((Diphylla_ecaudata:26.8,Desmodus_rotundus:26.8)3099:5.3,((((Chrotopterus_auritus:23.7,((Lophostoma_evotis:9.9,Lophostoma_brasiliense:9.9)3107:9.1,(Mimon_cozumelae:16.6,(Phylloderma_stenops:15,Phyllostomus_discolor:15)3113:1.6)3114:2.4)3115:4.7)3116:0,Trachops_cirrhosus:23.7)3117:4.7,(Lampronycteris_brachyotis:25.6,Micronycteris_megalotis:25.6)3126:2.8){Big-eared_Bats_and_more}:0.4,(((Glossophaga_soricina:7.4,Glossophaga_commissarisi:7.4)Glossophaga{Long-tongued_Bats[some]}:12.4,(Hylonycteris_underwoodi:3.7,Lichonycteris_obscura:3.7)3147:16.1)3154:7.8,((Carollia_perspicillata:5.2,Carollia_brevicauda:5.2)3158:17.6,(Sturnira_lilium:17.8,(Centurio_senex:13.7,((Enchisthenes_hartii:12.8,((Artibeus_watsoni:8,Artibeus_phaeotis:8)3185:0.6,(Artibeus_lituratus:3.8,Artibeus_jamaicensis:3.8)3186:4.8)Artibeus:4.2){Fruit-eating_Bats[some]}:0.7,((Vampyressa_pusilla:11.3,Chiroderma_villosum:11.3)3204:0,(Uroderma_bilobatum:10.4,(Vampyrodes_caraccioli:8.5,Platyrrhinus_helleri:8.5)3213:1.9)3214:0.9)3215:2.2)3216:0.2)3217:4.1)3218:5)3219:4.8)3220:1.2)3221:3.3)Phyllostomidae:11.1)3222:17.2)3225:0,(Myotis_keaysi:25.9,Bauerus_dubiaquercus:25.9)3668:34.5)3670;

FIGURE E.1. The phylogenetic tree for 34 phyllostomid bat species in the Newick format (Rosindell and Harmon 2012). The tree depth is 60.4 Myr.

TABLE E.2. Data summary of phyllostomid bat data in three habitats

	Habitat		
Statistics	Rainforest	Cacao	Cornfield
Sample size n	444	699	572
Sample coverage	0.980	0.996	0.995
(i) Observed species diversity (Hill numbers)			
${}^{0}D_{obs}$	27	21	17
${}^{1}D_{obs}$	11.269	8.341	7.121
${}^{2}D_{obs}$	8.449	6.643	4.651
(ii) Observed phylogenetic diversity			
${}^{0}PD_{obs}$	609.9	492.7	294.9
${}^{1}PD_{obs}$	134.8	115.5	109.4
${}^{2}PD_{obs}$	91.0	86.3	84.3

Following Chao et al. (2014), we illustrate below how to construct the two proposed types of sampling curves (sample-size and coverage-based) for the abundance-based ${}^{q}PD$ measures. The sampling curves are then used to compare the phylogenetic diversity among habitats.

STEP (1): compare sample-size-based R/E curves up to a maximum size

First, we compare in Fig. E.2 the integrated sample-size-based R/E curves along with 95% confidence intervals based on 200 bootstrap replications (Chao et al. 2014) up to a maximum sample size of 1000. The estimated diversities for equally-large samples then can be compared across the three habitats with any specific sample size. All panels reveal a consistent phylogenetic diversity ordering pattern: rainforest > cacao > cornfield, but the three confidence intervals are not completely disjoint, implying at least one of the tests of significant difference in diversity of a fixed order q between any two habitats is not conclusive. Notice that for each habitat, the sampling curve for Faith's PD (${}^{0}PD$) increases with sample size, but the curves for q = 1 and 2 level off beyond the reference sample,

illustrating that higher-order ${}^{q}PD$ are increasingly dominated by the species/lineages with large node abundances (near the root) and are therefore less sensitive to sampling effects.

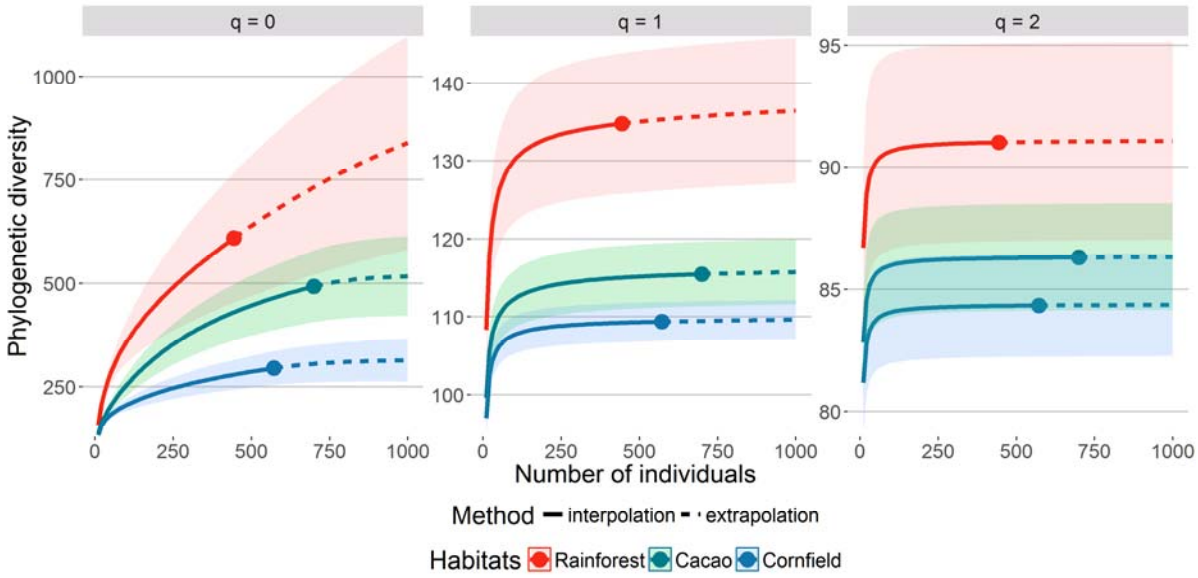


FIGURE E.2. Sample-size-based rarefaction (solid lines) and extrapolation (dotted lines) for abundance-based phylogenetic measure ${}^{q}PD$ of order $q = 0$ (left panel), $q = 1$ (middle panel), and $q = 2$ (right panel), up to the maximum sample size of 1000 individuals for phyllostomid bat species data in three habitats: rainforest, cacao, and cornfield. The reference point for all calculations was fixed to be 60.4 Myr (the depth of the observed tree). Reference samples are denoted by solid dots. The 95% confidence bands (shaded areas) are obtained by a bootstrap method based on 200 replications. Each curve is extrapolated up to the maximum size of 1000.

STEP (2): construct a sample completeness curve to link sample-size- and coverage-based R/E curves

The sample completeness curve (Fig. E.3) depicts how sample completeness (measured by sample coverage) increases with sample size with 95% confidence intervals for each habitat up to the maximum size of 1000. The estimated coverage values for the reference sample in the three habitats (rainforests, cacao plantations, cornfields) are respectively (0.980, 0.996, 0.995); see Table E.2. Fig. E.3 shows for any fixed size smaller than 1000 that the sample

completeness for the three habitats rainforest habitat for q = 0, 1 and 2 exhibits an opposite ordering to that for diversity comparison: cornfield > cacao > rainforest, although all 95% intervals are not completely separable. The sample coverage curves provide a bridge between sample-size- and coverage-based sampling accumulation curves.

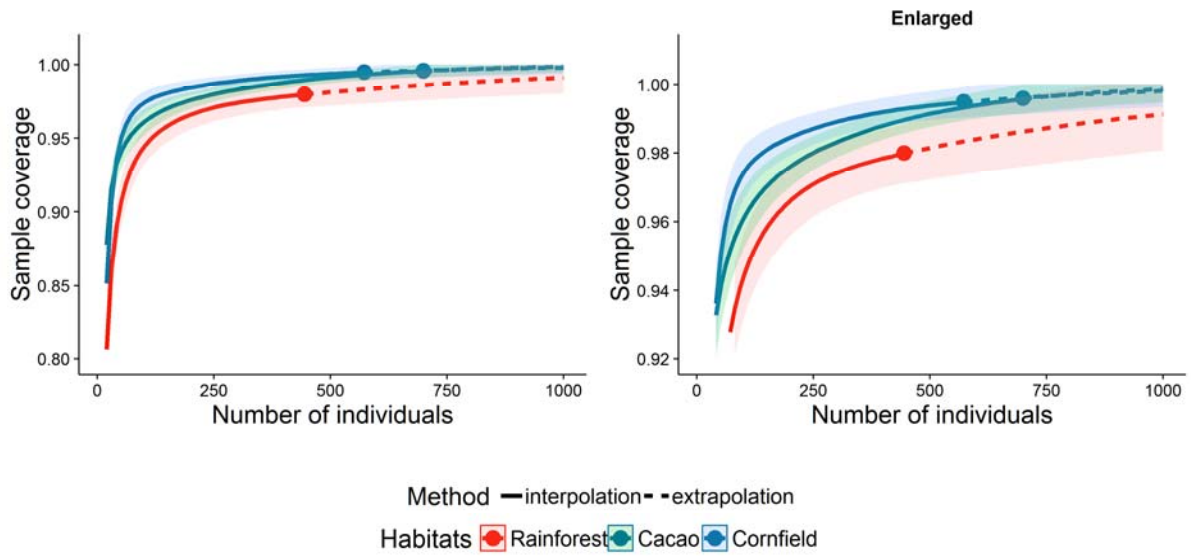


FIGURE E.3. Plot of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of sample size for phyllostomid bat species data in three habitats: rainforest, cacao, and cornfield. Reference samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. Each curve is extrapolated up to the maximum sample size of 1000.

STEP (3): compare coverage-based R/E curves up to a maximum coverage

Based on the sample completeness curve (Fig. E.3), the sample coverage estimates for the three habitats (rainforest, cacao, cornfield), when sample size is increased to 1000, change from (0.980, 0.996, 0.995) to (0.991, 0.999, 0.998). In Fig. E.4, we present the coverage-based R/E sampling curves with 95% confidence intervals for q = 0, 1 and 2 up to the coverage that corresponds to the size of 1000. Because the increase in coverage for the extrapolation is small, and the estimated diversity for q =1 and 2 hardly change beyond the reference samples, the extrapolation parts in Fig. E.4 are nearly invisible for these two orders of q. Similar phenomenon also occurs for q = 0 in the cacao and cornfield habitats. Nevertheless, all the three panels show the same diversity ordering as that in the sample-size-based R/E curves: rainforest > Cacao > cornfield. Note that for Faith's PD (^0PD),

none of the three confidence bands intersect when sample coverage is sufficiently high. This implies that data are sufficient to show significant difference in diversity between any two habitats for a fixed standardized coverage value in the restricted range. However, for q = 1 and 2, the confidence intervals for the Cacao and cornfield habitats overlap.

In summary, our proposed sample-size- and coverage-based rarefaction and extrapolation methods efficiently use all available data to make more robust and meaningful comparison of phylogenetic diversity among assemblages for a wide range of sample sizes and sample completeness.

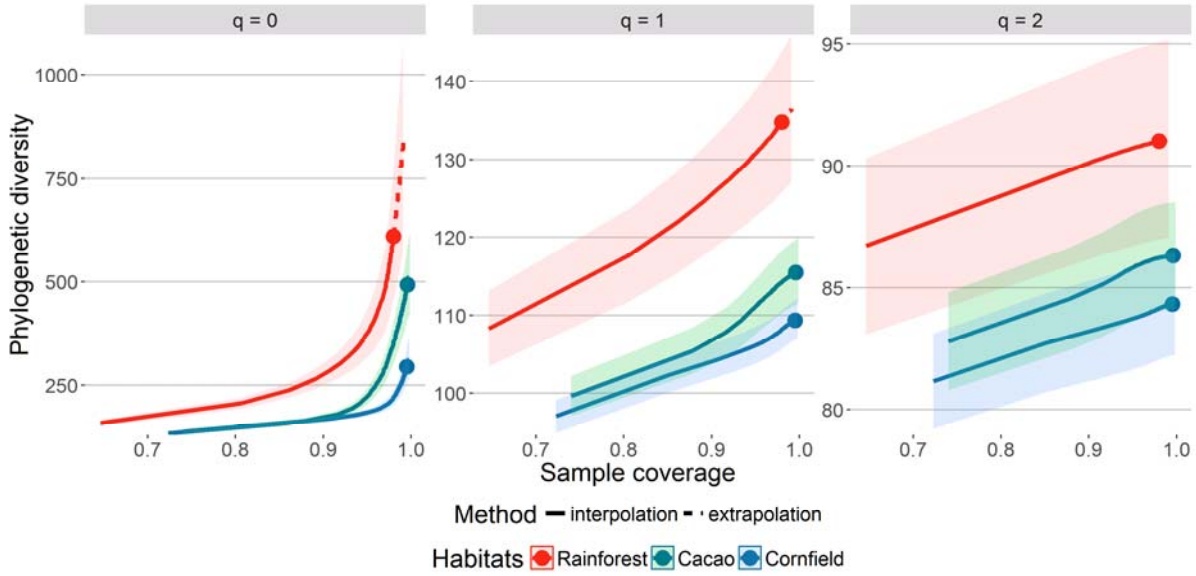


FIGURE E.4. Comparison of coverage-based rarefaction (solid lines) and extrapolation (dotted lines) for abundance-based phylogenetic measure ${}^{q}PD$ for order q = 0 (left panel), q = 1 (middle panel), and q = 2 (right panel), up to the maximum sample size of 1000 individuals for phyllostomid bat species data in three habitats: rainforest, cacao, and cornfield. The reference point for all calculations was fixed to be 60.4 Myr (the depth of the observed tree). Reference samples are denoted by solid dots. The 95% confidence bands (shaded areas) are obtained by a bootstrap method based on 200 replications. Each curve is extrapolated up to the sample coverage corresponding to the maximum sample size of 1000, i.e., 99.1% for rainforests, 99.9% for cacao habitat, and 99.8% for cornfield habitat.

References

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

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.

Chao A., Gotelli N.J., Hsieh T.C., Sander E.L., Ma K.H., Colwell R.K., Ellison A.M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs, 84:45-67.

Medellín R.A., Equihua, M., Amin, M.A. 2000. Bat diversity and abundance as indicators of disturbance in Neotropical rainforests. Conservation Biology, 14:1666-1675.

Rosindell J., Harmon, L.J. 2012. OneZoom: a fractal explorer for the tree of life. PLoS Biol, 10:e1001406.