Backbones of Evolutionary History Test Biodiversity Theory for Microbes: Supplementary Online Material

James P. O'Dwyer [1,*], Steven W. Kembel[2], Thomas Sharpton[3]
1 Department of Plant Biology, University of Illinois, Urbana IL USA
2 Départment des sciences biologiques, Université du Québec à Montréal, Montreal, Quebec, Canada
3 Department of Microbiology and Department of Statistics, Oregon State University, Corvallis, OR, USA $*$ **E-mail: jodwyer@illinois.edu**

A Tree building and robustness

Quality controlled 16S rRNA sequences were downloaded from various public data repositories. We obtained human skin and gut (Roche 454) data from the HMP DACC (`http://www.hmpdacc.org/`) [1], marine data from the ICOMM MICROBIS web portal (`http://vamps.mbl.edu/portals/icomm/icomm.php/microbis/`) [2], and phyllosphere data (paired-end 150bp Illumina HiSeq reads) from [3]. For the HMP data, we analyzed 4 samples from four subjects' human stool and Anterior nares. For the phyllosphere data, we analyzed three samples, one each taken from Barro Colorado Island tree species Inga acuminata, Alseis blackiana, and Virola surinamensis. For the MICROBIS data, we analyzed a subset of samples selected as part of [4], including the Deep Arctic Ocean (DAO), Census Arctic Marine (CAM), Amundsun Sea Antarctica (ASA), and Active but Rare (ABR) datasets. In all cases, samples were chosen prior to analysis, without selection for particular phylogenetic tree structure.

We then aligned sequences from each repository to create multiple sequence alignments for phylogenetic reconstruction. For the HMP and MICROBIS samples, we used INFERNAL [5] in conjunction with a stochastic context free grammer (i.e., model) that was trained on a high-quality and curated alignment (i.e., reference alignment) produced by the Ribosomal Database Project [6], to align 16S reads to one another by way of the model, as in [7] (settings: –dna –hbanded –sub). We generated versions of these alignments that either included only the 16S reads or both 16S reads and reference sequences. Alignments were masked and filtered using the QIIME script filter_alignment.py QIIME (settings: -g 0.5 -s).

Quality controlled alignments were then subject to various phylogenetic reconstruction methods to evaluate their influence on the EAD/site frequency spectrum. First, we used RAxML [8] to generate a high-quality maximum likelihood tree as well as 100 bootstrapped trees. We also generated a de novo phylogeny using FastTree [9]. In addition, we repeated these analyses using alignments that included phylogenetically diverse and full-length reference sequences, a strategy that has proven useful in the phylogenetic analysis of short sequences [7, 10]. Lineages in the subsequent phylogenies that corresponded to reference sequences were pruned using the ape software package in R [11] to produce a phylogeny that relates 16S reads. The EAD was then independently calculated for each of these phylogenies and compared to assess

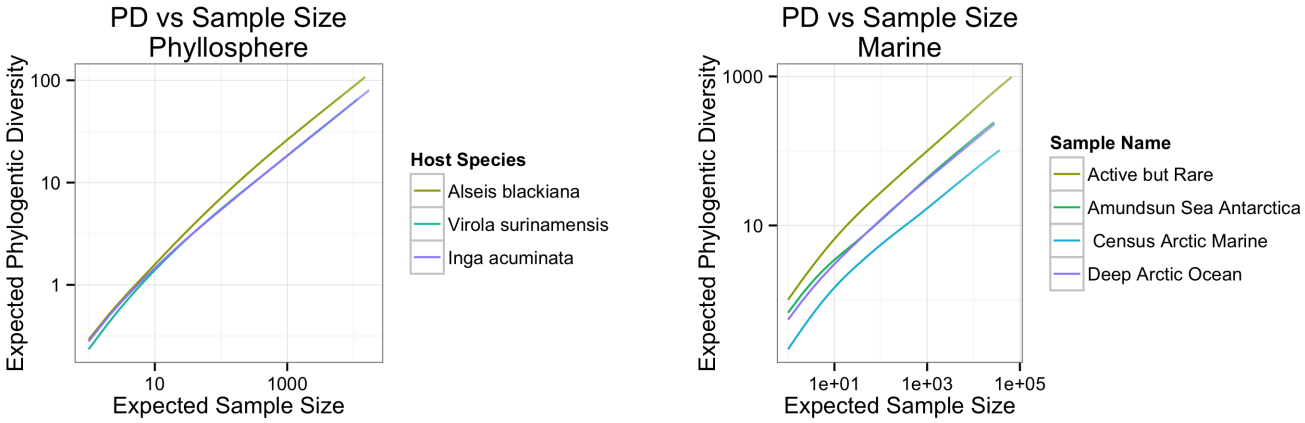

Figure 1. Empirical scaling of Phylogenetic Diversity (PD) with Sample size. PD increases with sample size approximately as a power law, for samples taken from phyllosphere communities in tropical forests and from marine environments at various latitudes and depths. This figure recapitulates the plots in Figure 1 of the main manuscript, but now with a legend describing which samples correspond to which line.

the impact of informatic strategy on EAD inference, with relatively minor changes observed across alignment methods (see SI Section D).

B Expected PD and Edge-Length Abundance Plots

In this section, we include expanded versions of Figures 1 and 3 of the main manuscript, with the primary addition being legends assigning each specific sample to each plotted line. These extended figures are labeled as SI Figure 1 and SI Figure 2, respectively.

We also show plots summarizing the fitted exponents

C Phylogenetic Sampling and Small Sample Sizes

Using binomial sampling, expected phylogenetic diversity (PD) is related to the expected number, n, of individual tips sampled from N total tips as [12]:

$$PD(n) = \sum_k S(k)(1 - (1 - n/N)^k), \tag{1}$$

where $S(k)$ is the Edge-length Abundance Distribution, described in the main text and in [12], and equivalent to the site-frequency spectrum in the population genetics Infinite Sites model.

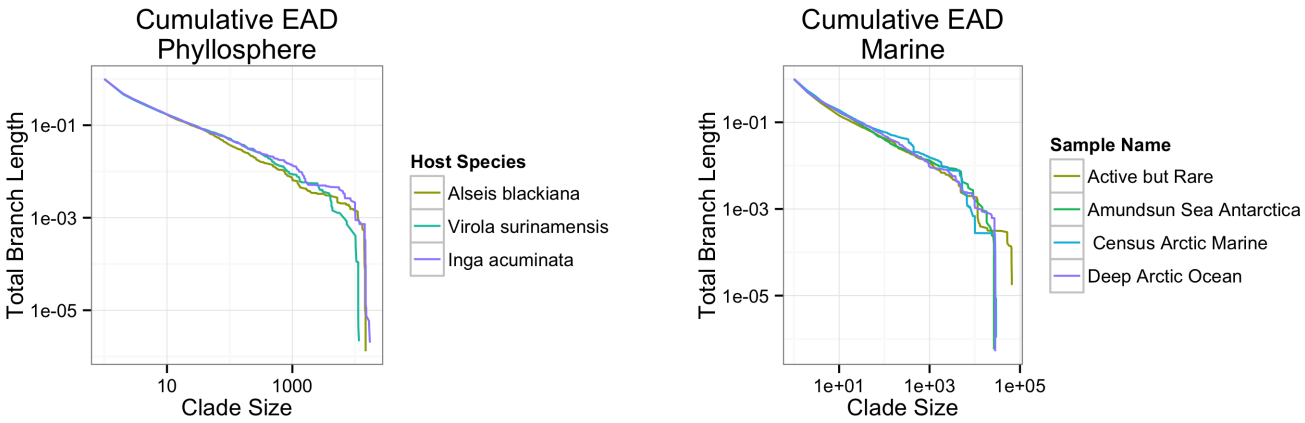

Figure 2. Empirical Edge-length abundance Distributions follow an approximately power law distribution. The two panels show the EAD from a set of sample trees, inferred using FastTree, taken from each of two habitats: phyllosphere and marine. These recapitulate the plots in Figure 3 of the main manuscript, but now with a legend describing which samples correspond to which line.

We can expand this expression in powers of n/N:

$$PD(n) = \frac{n}{N}\sum_k kS(k) - \left(\frac{n}{N}\right)^2 \sum_k \frac{k(k-1)}{2} S(k) + \ldots, \tag{2}$$

and for small enough n/N the scaling must be approximately linear—sometimes referred to as a 'sampling' phase [13]. The condition on sample size for this linearity to hold is that

$$\frac{n}{N} \ll \frac{\sum_k kS(k)}{\sum_k \frac{k(k-1)}{2} S(k)}. \tag{3}$$

D Fitting Power Law Exponents to the Edge-length Abundance Distribution

For each empirical tree, we computed the Edge-length Abundance Distribution, or EAD, as described in main text Fig. 2 and in [12]. We used a maximum likelihood method for fitting power law exponents to this distribution for each sample, due to its accuracy and performance advantage over other methods for fitting power law frequency distributions (see e.g. [14]).

One complication in fitting the EAD is that for each value of k, $S(k)$ is weighted by branch length. So it takes a real value, rather than an integer number of counts. To compute an effective number of counts for each value of k, we split up each branch length into discrete segments, with grain size chosen to match the smallest branch length between any two nodes

(or node and tip) in the tree. So, a long branch length with three tips downstream will correspond to a large number of 'counts' of three tips—we generate a number of counts proportional to the branch length.

Using this count data, we then apply the maximum likelihood method described in [15], which assumes the probability of a single observation having a number of tips, k as,

$$p(k) \propto k^{-\alpha} \tag{4}$$

and maximizes the likelihood corresponding to a large number of observations,

$$\mathcal{L}(\alpha|k_1, k_2...) \propto \prod_i k_i^{-\alpha}, \tag{5}$$

with respect to α. We plan to upload the R code for computing EAD exponents to the Picante package in the CRAN repository.

We used this approach to fit exponents in each of our samples, with results by habitat summarized in SI Figure 3. The exponent of the power law fit lies between $\alpha = 1.3$ and $\alpha = 1.7$. In SI Fig. 3, we also show that our results are robust to methodology, comparing trees inferred using a rapid tree inference algorithm (FastTree) to the range of exponents arising from a maximum-likelihood tree inference algorithm (RAxML). In our SI Section A we described additional systematic variation in the alignment and inference steps. We find the same or very similar scaling in all cases, providing evidence that this scaling does not arise due to the bias of one particular alignment or tree-builiding algorithm.

E Coarse-Graining Cutoff Choice for Computing Heterogeneity of Branching Rates

In the main text, we showed the results of coarse-graining an empirical phyllosphere phylogeny, displayed in Fig. 5. In the left-hand panel of this figure, we show that as the tree is coarse-grained, there is an immediate reduction in the number of nodes as the coarse-graining scale is increased above the smallest segment of branch-length in the tree. There is then a plateau, as the tree structure remains relatively constant over several orders of magnitude. For increasing coarse-graining scales, the entire tree eventually collapses down, approaching a single node with a very large polytomy. We propose that choosing a coarse-graining to a point within this middle region reveals the heterogeneity in branching rates in the tree. To make a definite and consistent choice for this scale across multiple empirical trees, we explore a range of coarse-graining cutoff scales, S_i, evenly spaced on a logarithmic scale, and each of these cutoff scales corresponds to a number of remaining nodes in the tree, $\mathcal{N}_i$. We then choose the cutoff scale among these that minimizes

$$\frac{\log(\mathcal{N}_{i-1}) - \log(\mathcal{N}_i)}{\log(S_i) - \log(S_{i-1})}. \tag{6}$$

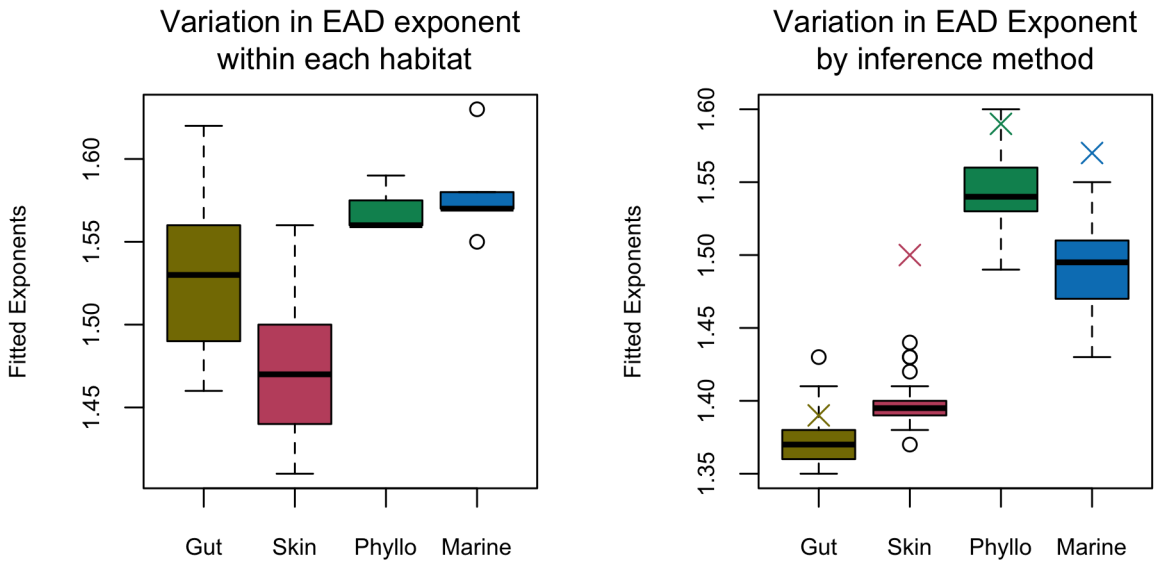

Figure 3. The left panel shows the range of exponents of the EAD across different habitat types and demonstrates that there is a narrow range of variation within each habitat type. The right panel explores variation in the method of tree inference: for each habitat type a single sample has been chosen, and a phylogenetic tree is inferred using FastTree. The EAD exponent from this single tree is represented by an X in the plot, and compared to the range of exponents for 100 bootstrapped trees inferred using RaXML and the same alignment method. The FastTree exponent in some habitats differs from the range of RaXML exponents, though does not alter the qualitative scaling behavior, providing evidence that the scaling of the EAD is not an artifact of one particular tree-building methodology.

Informally, our choice of cutoff is an approximation to minimizing $\frac{d\log(\mathcal{N})}{dlog(S)}$, for cutoff scale, S, and number of nodes $\mathcal{N}(S)$. In practice, this choice is somewhat arbitrary—it picks out a coarse-graining scale immediately following the first drop in nodes with changing cutoff, where the change in number of nodes with cutoff slows down. We could equally well pick a larger cutoff further along the plateau, and it could be that the full range of cutoff scales is informative about processes acting on differing scales.

In SI Figures 4–7 we show the coarse-graining according to this criterion for all of the datasets analyzed in this study, analogous to Figure 6 from the main manuscript. In SI Figure 8 we show that the qualitative conclusions of the number of nodes and distribution of branch length sizes does not depend strongly on the number of coarse-graining scales we choose—there are always distinct peaks in branch length size distribution, and therefore regions of stability where coarse-grained tree structure is relatively stable. Finally, in SI Figure 9 we show the same three panels for a neutral tree with changing community size through time—the best case scenario for a neutral model, where the EAD has qualitatively the same power law scaling as an empirical

tree. We see that the qualitative features of this neutral tree are quite different under coarse-graining, with no peaks in the distribution of branch length sizes, and no fat-tailed distribution of polytomy sizes throughout the coarse-grained tree.

F Model Descriptions

F.1 Kingman Coalescent with Time-Varying Rate

In the main text we introduced a coalescent model [16] to approximate neutral evolution but with changing community size, so that community size increases over time proportional to $N(t) \sim t^{\beta}$.The coalescent is a continuous-time Markov process, in which times between coalescent events are independent exponential random variables, and the rate of coalescence for any two of n individuals in a community of total size N is proportional to $\frac{n(n-1)}{2N}$. To approximate changing community size, we model the waiting time between two coalescent events with an exponential distribution, but we change the rate *following each coalescent event*, in proportion to $t^{-\beta}$. In effect, we approximate a smooth change in population size by small jumps in the coalescence rate following each event. In effect, times between coalescent events nearer to the tips of the tree are stretched out relative to the Kingman coalescent, while coalescent times near the root of the tree are shortened.

F.2 Λ-coalescent

The Λ coalescent [17, 18] is also a continuous-time Markov process, in which times between coalescent events are independent exponential random variables, but where now there are multiple distinct types of coalescent event, each with a different rate. The rates are determined by a coalescent parameter, which we called γ in the text (but is often termed α in the coalescent literature—we changed this parameter name to avoid confusion with other variables). For a fixed total community size, we can choose units where the Kingman coalescent rate is $\frac{n(n-1)}{2}$. If the coalescent parameter γ is between 1 and 2, then this model is also called the β-coalescent (a subfamily of Λ-coalescent models), where the rate for j lineages to merge is

$$\lambda_j = \binom{n}{j} \frac{B(j-\gamma, n-j+\gamma)}{B(2-\gamma, \gamma)}. \tag{7}$$

For sufficiently large j, this is well-approximated by $\lambda_j \propto j^{-(\gamma+1)}$.

We simulated all coalescent trees using functions drawn from the *Ape* R package, and plan to upload these functions to the *Picante* package.

References

1. Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, C.C. Baker, V. DiFrancesco, T.K. Howcraft, R.W. Karp, R.D. Lunsford, C.R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A.R. Little, H. Peavy, C. Pontzer, M. Portnoy, M.H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

2. Linda Amaral-Zettler, Luis Felipe Artigas, John Baross, Loka Bharathi, Antje Boetius, Dorairajasingam Chandramohan, Gerhard Herndl, Kazuhiro Kogure, Phillip Neal, Carlos Pedrós-Alió, A. Ramette, S. Schouten, L. Stal, A. Thessen, J. Leeuw, and M. Sogin. A global census of marine microbes. In *Life in the Worlds Oceans: Diversity, Distribution and Abundance*, pages 223–245. Blackwell Publishing Ltd, 2010.

3. Steven W Kembel, Timothy K O'Connor, Holly K Arnold, Stephen P Hubbell, S Joseph Wright, and Jessica L Green. Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest. *Proceedings of the National Academy of Sciences*, 111(38):13715–13720, 2014.

4. J Ladau, TJ Sharpton, SW Kembel, JP O'Dwyer, JL Green, JA Eisen, and KS Pollard. Global marine bacterial diversity peaks at high latitudes in winter. *ISME Journal*, 7:1669–1677, 2013.

5. Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

6. Bonnie L Maidak, Gary J Olsen, Niels Larsen, Ross Overbeek, Michael J McCaughey, and Carl R Woese. The ribosomal database project (rdp). *Nucleic acids research*, 24(1):82–85, 1996.

7. TJ Sharpton, J Ladau, SW Kembel, JP O'Dwyer, JL Green, JA Eisen, and KS Pollard. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data . *PLoS Computational Biology*, 7:e1001061, 2011.

8. Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

9. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

10. Samantha J Riesenfeld and Katherine S Pollard. Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis. *BMC genomics*, 14(1):419, 2013.

11. Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

12. JP O'Dwyer, SW Kembel, and JL Green. Phylogenetic Diversity Theory Sheds Light on the Structure of Complex Microbial Communities. *PLoS Computational Biology*, page 8(12): e1002832, 2012.

13. JL Green and JB Plotkin. A statistical theory for sampling species abundances. *Ecology Letters*, 10:1037–45, 2007.

14. Ethan P White, Brian J Enquist, and Jessica L Green. On estimating the exponent of power-law frequency distributions. *Ecology*, 89(4):905–912, 2008.

15. Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

16. JFC Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

17. J Pitman. Coalescents with multiple collisions. *The Annals of Probability*, 27:1870–1902, 1999.

18. N Berestycki. Recent Progress in Coalescent Theory. *Ensaios Matematicos*, 16:1–193, 2009.

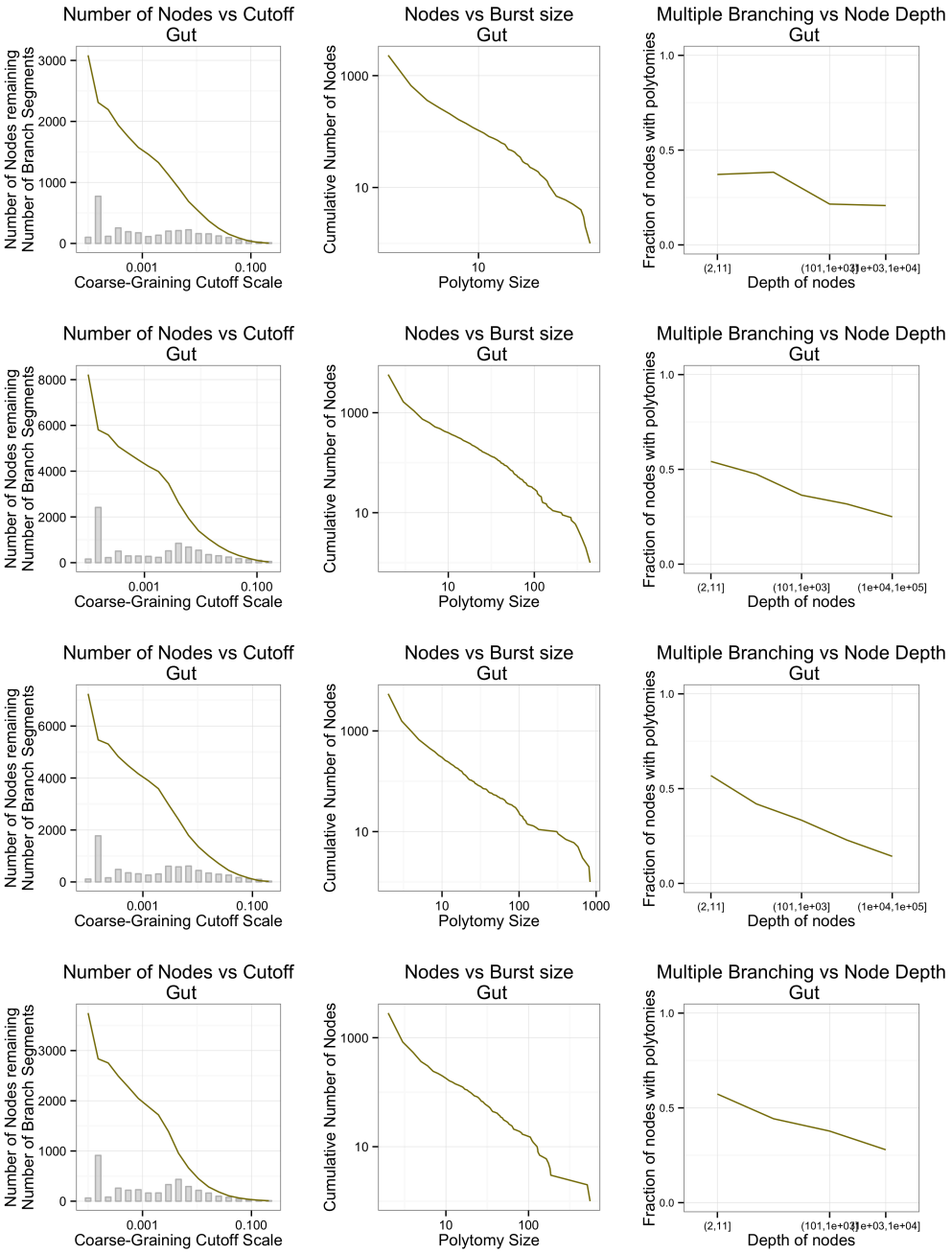

Figure 4. Coarse-graining empirical gut (stool) phylogenies. The left-hand panels show that the number of internal nodes in each tree changes as we change the coarse-graining cutoff, where the cutoff scale is in units of total tree depth, from root to furthest tip. We next choose a specific cutoff scale, immediately at the start of this range of stability (we quantify our criterion for choosing a specific cut-off scale in our SI methods section), though our results would be similar for any cutoff chosen somewhere within this range. The central panel plots the corresponding coarse-grained tree, and shows the cumulative distribution of the number of internal nodes as a function of polytomy size. Finally, the right-hand panel shows the fraction of nodes with multiple branches (i.e. with polytomy size ≥ 3) as a function of the total number of tips downstream of the node.

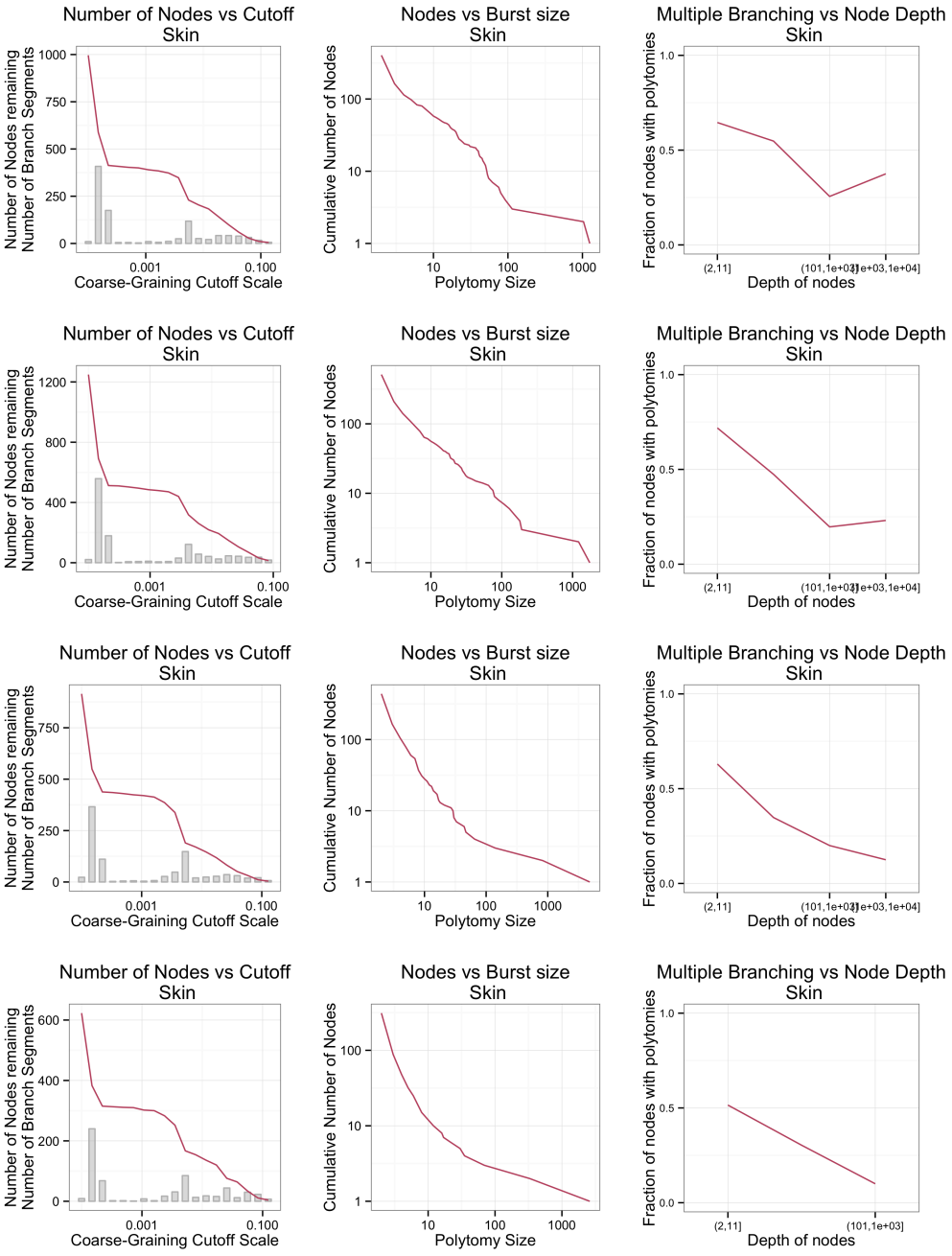

Figure 5. Coarse-graining empirical phylogenies sampled from exterior of the nostrils. The left-hand panels show that the number of internal nodes in each tree changes as we change the coarse-graining cutoff, where the cutoff scale is in units of total tree depth, from root to furthest tip. We next choose a specific cutoff scale, and the central panels plot the corresponding coarse-grained tree, and shows the cumulative distribution of the number of internal nodes as a function of polytomy size. Finally, the right-hand panels show the fraction of nodes with multiple branches (i.e. with polytomy size ≥ 3) as a function of the total number of tips downstream of the node.

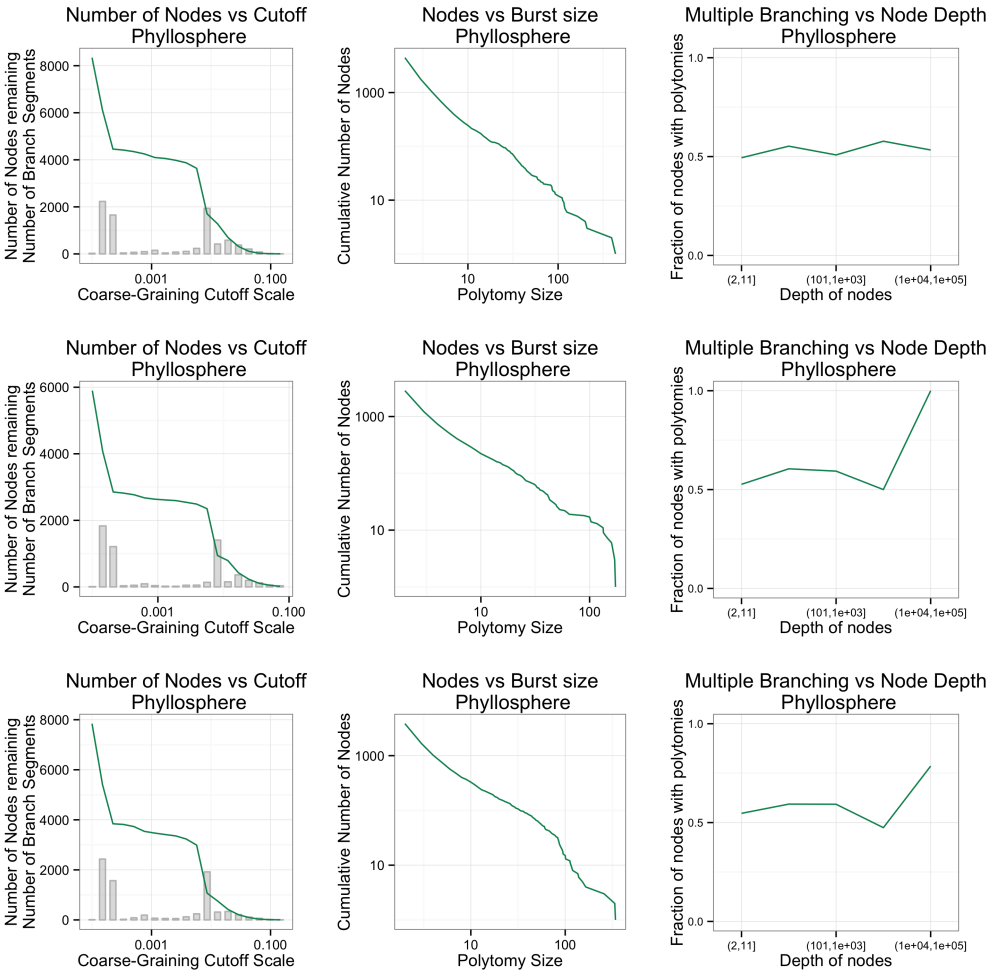

Figure 6. Coarse-graining empirical phylogenies sampled from leaves of tree species Inga acuminata, Alseis blackiana, and Virola surinamensis. (The top three panels also appear as Figure 6 in the main manuscript.) The left-hand panels show that the number of internal nodes in each tree changes as we change the coarse-graining cutoff, where the cutoff scale is in units of total tree depth, from root to furthest tip. We next choose a specific cutoff scale, and the central panels plot the corresponding coarse-grained tree, and shows the cumulative distribution of the number of internal nodes as a function of polytomy size. Finally, the right-hand panels show the fraction of nodes with multiple branches (i.e. with polytomy size ≥ 3) as a function of the total number of tips downstream of the node.

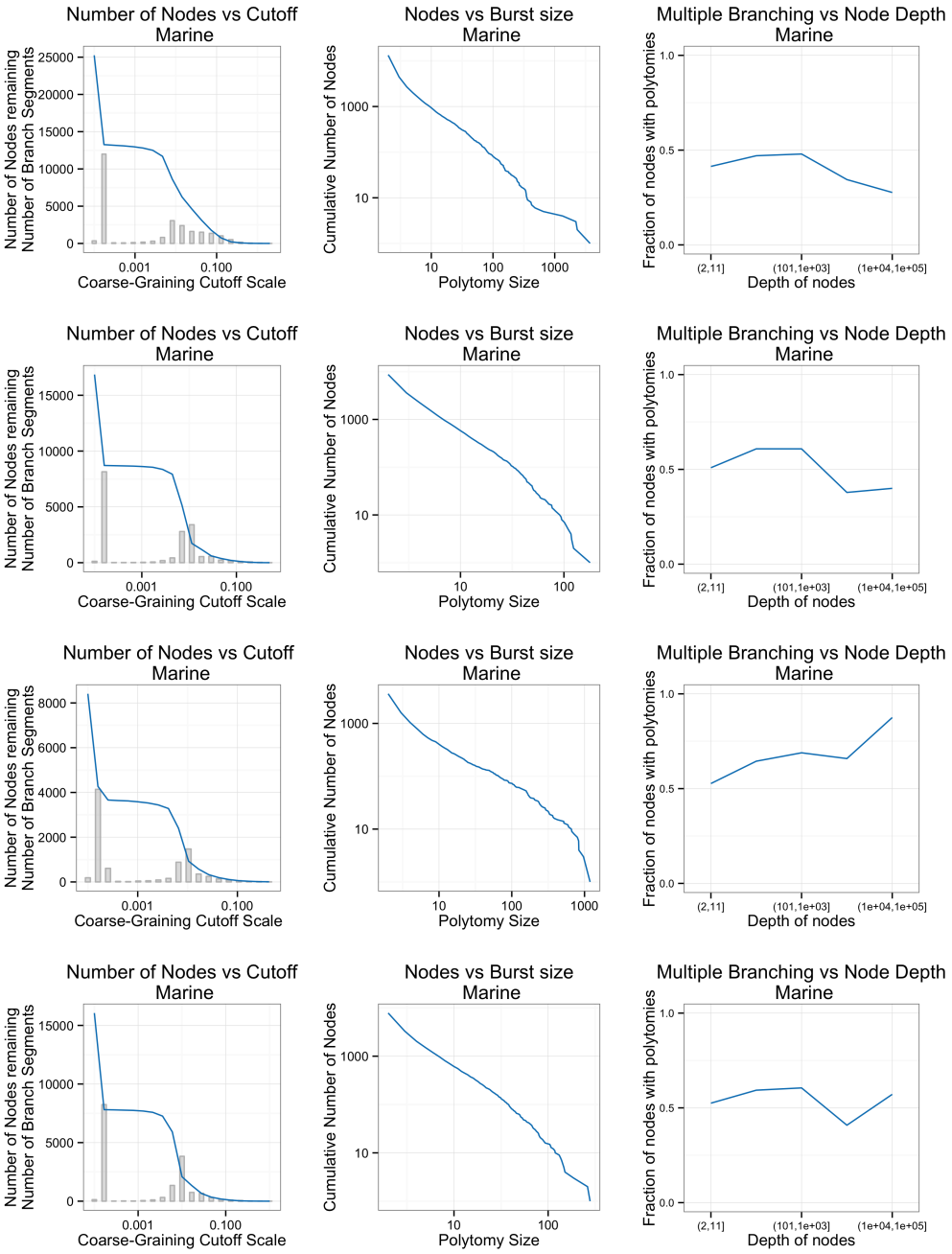

Figure 7. Coarse-graining empirical phylogenies sampled from MICROBIS marine samples Deep Arctic Ocean (DAO), Census Arctic Marine (CAM), Amundsun Sea Antarctica (ASA), and Active but Rare (ABR). The left-hand panels show that the number of internal nodes in each tree changes as we change the coarse-graining cutoff, where the cutoff scale is in units of total tree depth, from root to furthest tip. We next choose a specific cutoff scale, and the central panels plot the corresponding coarse-grained tree, and shows the cumulative distribution of the number of internal nodes as a function of polytomy size. Finally, the right-hand panels show the fraction of nodes with multiple branches (i.e. with polytomy size ≥ 3) as a function of the total number of tips downstream of the node.

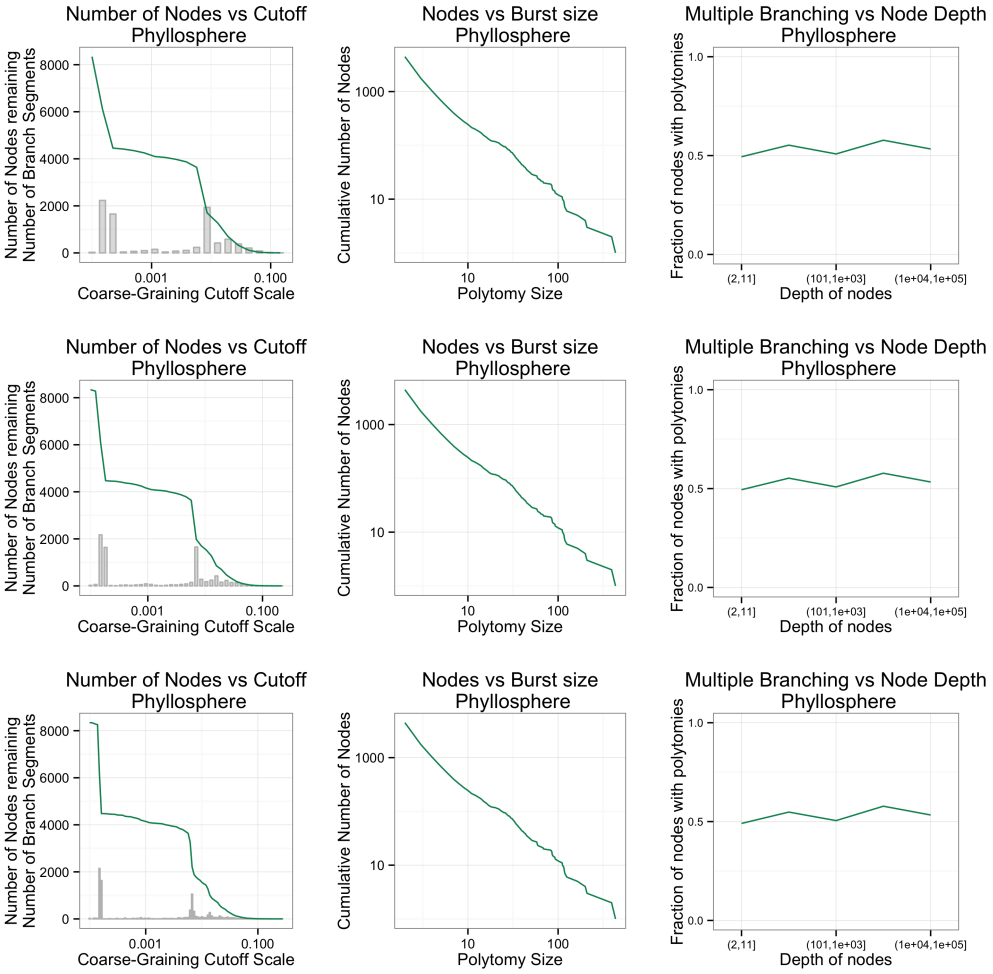

Figure 8. Coarse-graining an empirical phyllosphere phylogeny, sampled from leaves of angiosperm species Inga acuminata, but now showing multiple coarse-graining scales in the left-hand panel. The qualitative feature of a separation of scales is there at multiple choices of 'grain unit', and does not significantly change our results in the other panels.

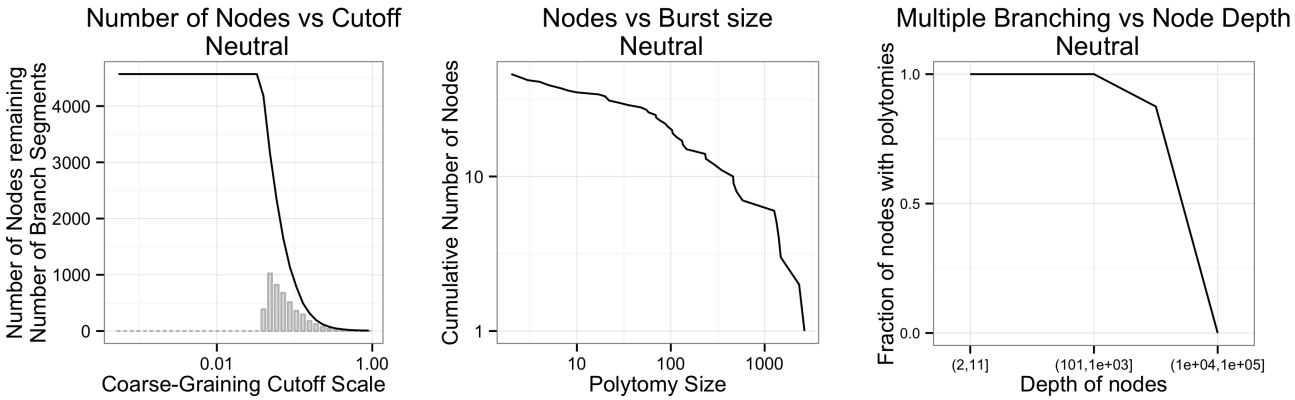

Figure 9. Coarse-graining a neutral tree with varying community size through time. We plotted some of the central results for this kind of tree in Figure 7 of the main manuscript. Here we now extend this description of the differences between neutral trees and empirical trees, by displaying the same three plots we've shown for empirical trees in Figure 6 of the main manuscript, and in SI Figures 3–7 above. The left-hand panel shows that the number of internal nodes in each tree changes as we change the coarse-graining cutoff, where the cutoff scale is in units of total tree depth, from root to furthest tip. We see no separation of scales (i.e. no distinct peaks in the distribution of branch length sizes). For a cutoff scale corresponding to the same number of remaining nodes as our empirical coarse-grained tree, we find in the central panel that the number of internal nodes as a function of polytomy size does not display a power law scaling. Finally, the right-hand panel shows the fraction of nodes with multiple branches (i.e. with polytomy size ≥ 3) as a function of the total number of tips downstream of the node, and we see that bursts of branching, such as they are in this tree, are not distributed evenly throughout.