EVE 161:
Microbial Phylogenomics

Class #9:
rRNA Ecology

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

Presenters

Tree of Life
Eukaryotes
Plants
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Green algae
Land plants
Red algae
Glaucophyte algae
Ciliates
Alveolates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Trypanosomes
Leishmanias
Core Jakobids
Parabasalids
Diplomonads
Oxymonads
Excavates
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Bacteria
OP11
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids
Mycoplasmas
Firmicutes

Scales of Diversity

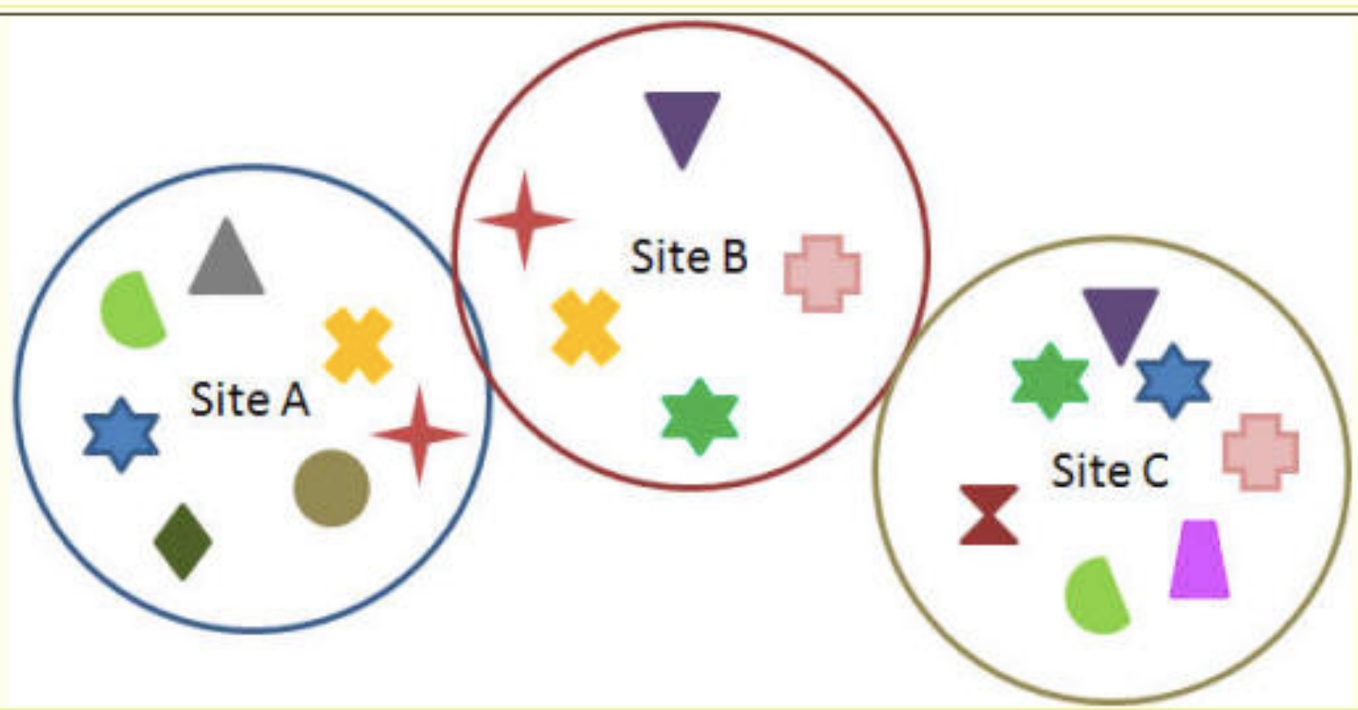
Biodiversity Can be Expressed at Several Scales

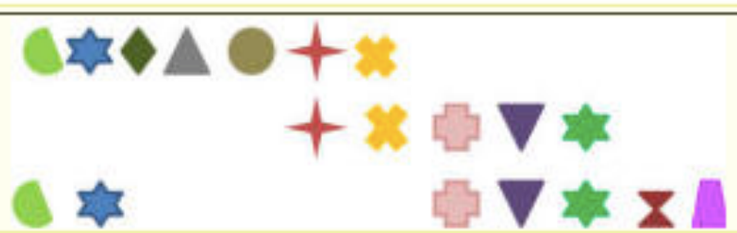
Biodiversity can be measured and monitored at several spatial scales.

Alpha Diversity = richness and evenness of individuals within a habitat unit. For example in the figure below, **Alpha Diversity** of Site A = 7 species, Site B = 5 species, Site C = 7 species.

Beta Diversity = expression of diversity between habitats. In the example below, the greatest **Beta Diversity** is observed between Site A and C with 10 species that differ between them and only 2 species in common.

Gamma Diversity = landscape diversity or diversity of habitats within a landscape or region. In this example, the gamma diversity is 3 habitats with 12 species total diversity.



Site A = 7 Species Site B = 5 Species Site C = 7 Species		A vs B = 8 species B vs C = 4 species A vs C = 10 species

http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

rRNA Ecology Workflow

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2001, p. 4399–4406 Vol. 67, No. 10
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.10.4399–4406.2001

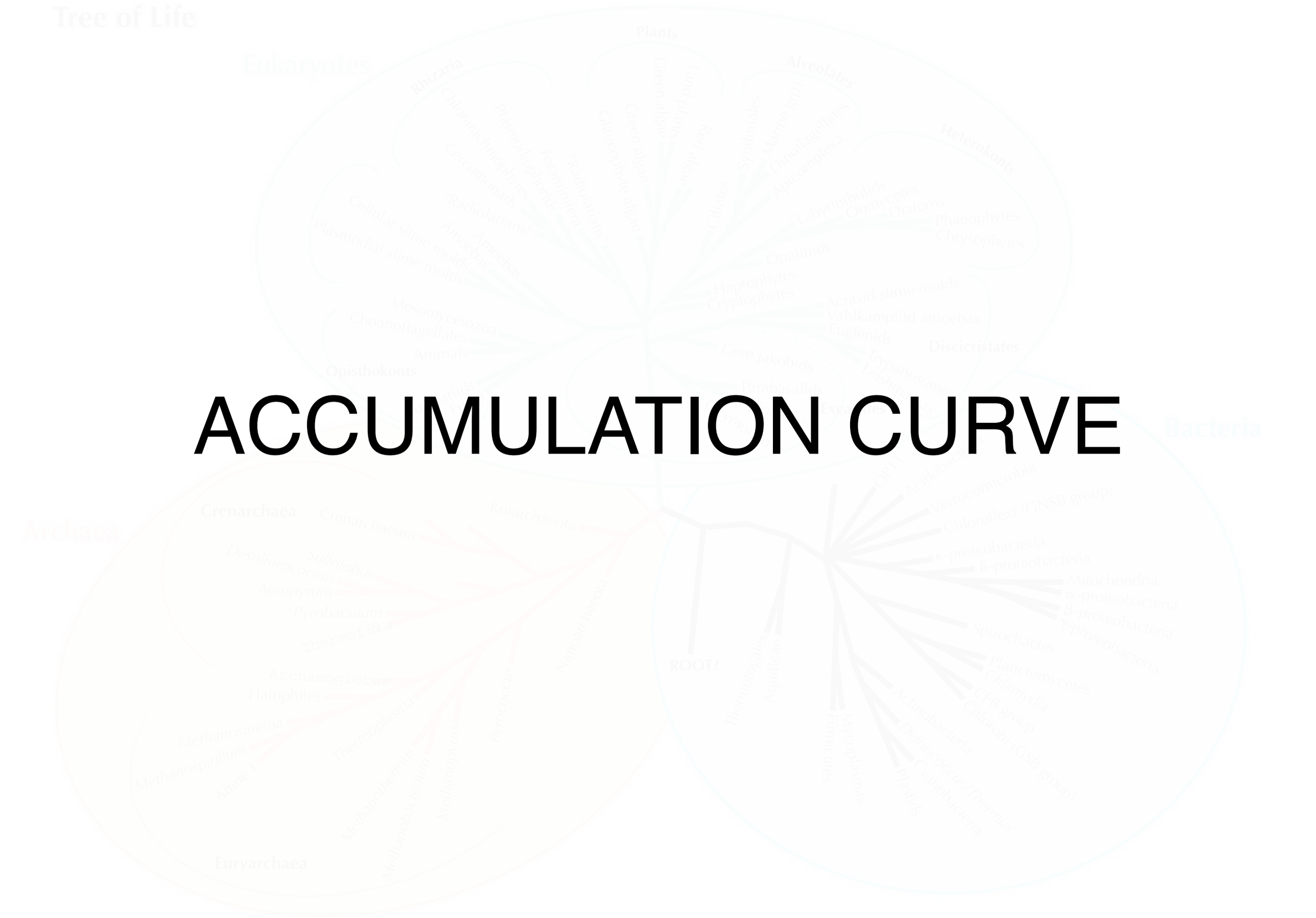
MINIREVIEW

Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity

JENNIFER B. HUGHES,* JESSICA J. HELLMANN,† TAYLOR H. RICKETTS, AND BRENDAN J. M. BOHANNAN

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020

ACCUMULATION CURVE



Accumulation curves

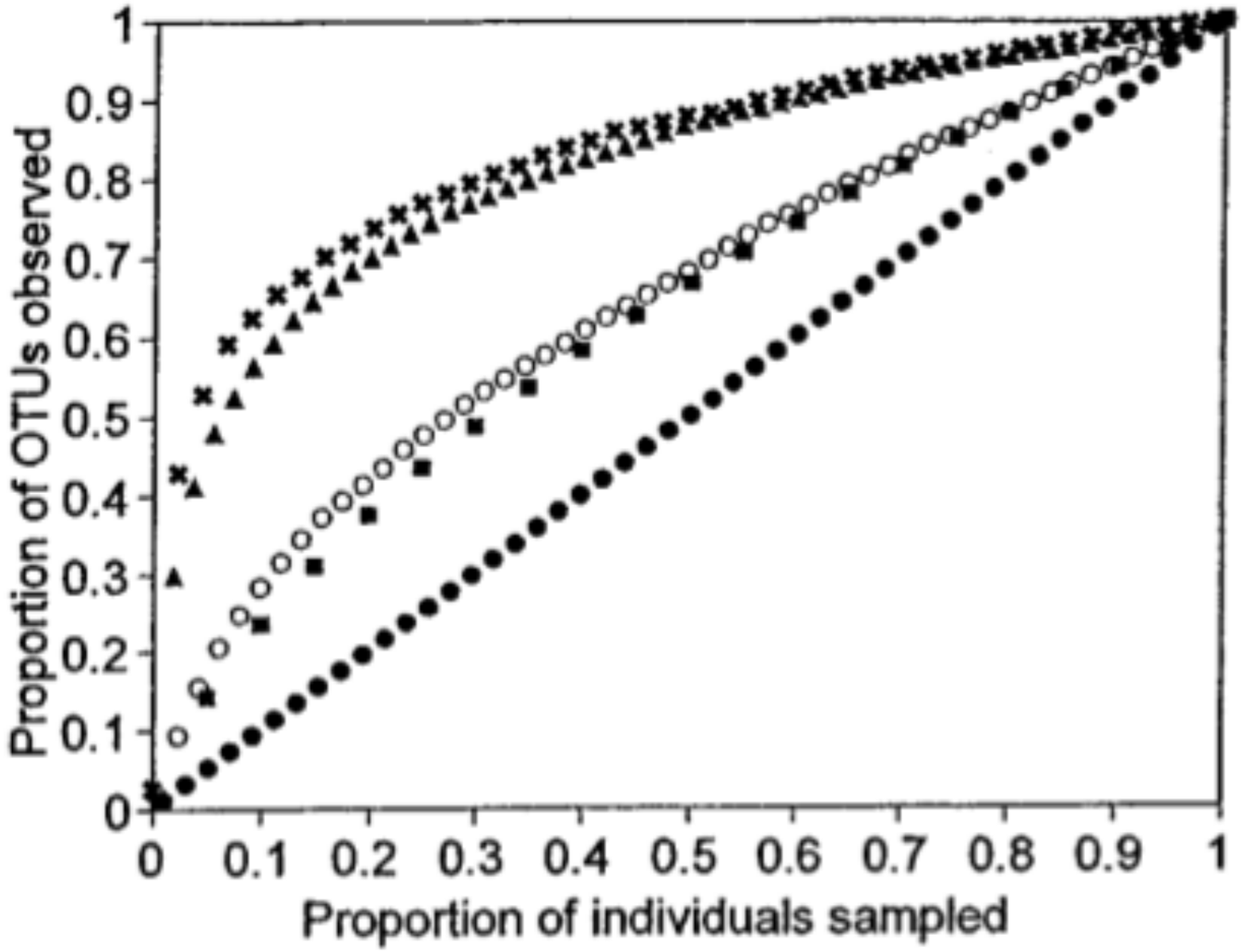
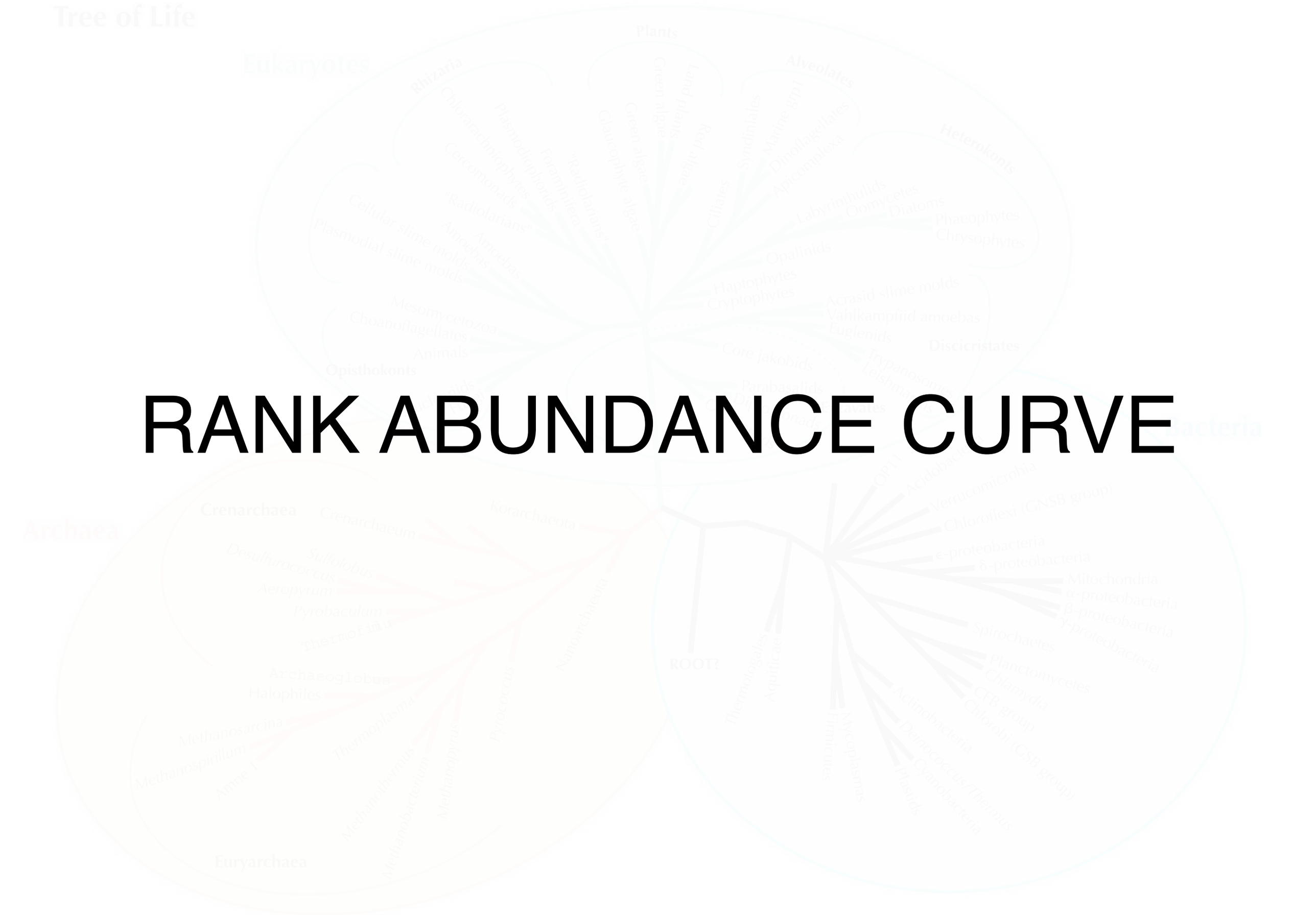


FIG. 1. Accumulation curves for Michigan plants (✖; n = 1,783) (26), Costa Rican birds (▲; n = 5,007) (J. B. Hughes, unpublished data), human oral bacteria (○; n = 264) (33), Costa Rican moths (■; n = 4,538) (56), and East Amazonian soil bacteria (●; n = 98) (6). Curves are averaged over 100 simulations using the computer program EstimateS and are standardized for the number of individuals and species observed.

RANK ABUNDANCE CURVE



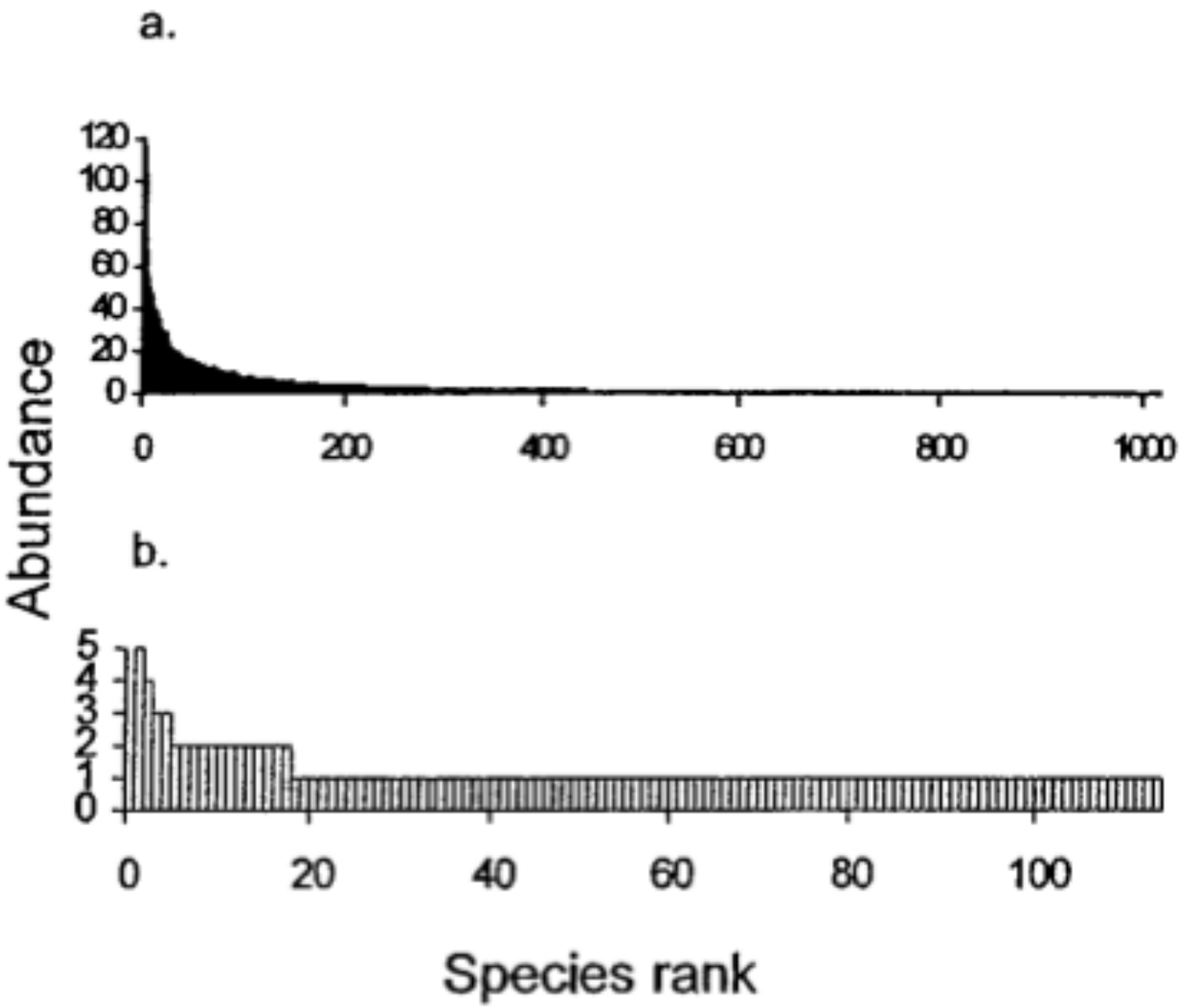


FIG. 2. Rank-abundance curves for (a) tropical moths (n = 4,538) (56) and (b) temperate soil bacteria (n = 137) (39). The two most abundant species of moths (396 and 173 individuals) are excluded from panel a to shorten the y axis.

Another way to compare how well communities have been sampled is to plot their rank-abundance curves. The species are ordered from most to least abundant on the *x* axis, and the abundance of each type observed is plotted on the *y* axis. The moth and soil bacteria communities exhibit a similar pattern (Fig.2), one that is typical of superdiverse communities such as tropical insects. A few species in the sample are abundant, but most are rare, producing the long right-hand tail on the rank-abundance curve.

RICHNESS ESTIMATORS

In contrast to rarefaction, richness estimators estimate the total richness of a community from a sample, and the estimates can then be compared across samples. These estimators fall into three main classes: extrapolation from accumulation curves, parametric estimators, and nonparametric estimators (14, 23, 47). To date, we have found only two studies that apply richness estimators to microbial data (33, 43).

CURVE EXTRAPOLATION

Most curve extrapolation methods use the observed accumulation curve to fit an assumed functional form that models the process of observing new species as sampling effort increases. The asymptote of this curve, or the species richness expected at infinite effort, is then estimated. These models include the Michaelis-Menten equation (13, 51) and the negative exponential function (61). The benefit of estimating diversity with such extrapolation methods is that once a species has been counted, it does not need to be counted again. Hence, a surveyor can focus effort on identifying new, generally rarer, species. The downside is that for diverse communities in which only a small fraction of species is detected, several curves often fit equally well but predict very different asymptotes (61). **This approach therefore requires data from relatively well sampled communities, so at present curve extrapolation methods do not seem promising for estimating microbial diversity in most natural environments.**

PARAMETRIC ESTIMATION

Parametric estimators are another class of estimation methods. These methods estimate the number of unobserved species in the community by **fitting sample data to models of relative species abundances**. These models include the lognormal (49) and Poisson lognormal (7). For instance, Pielou (48) derived an estimator that assumes species abundances are distributed lognormally; that is, if species are assigned to log abundance classes, the distribution of species among these classes is normal. By fitting sample data to the lognormal distribution, the parameters of the curve can be evaluated. Pielou's estimator uses these parameter values to estimate the number of species that remain unobserved and thereby estimate the total number of species in the community.

There are three main impediments to using parametric estimators for any community.

There are three main impediments to using parametric estimators for any community. First, data on relative species abundances are needed. For macroorganisms, often only the presence or absence of a species in a sample or quadrat is recorded. In contrast, data on relative OTU abundances of microbes are often collected (see discussion below about potential biases). Second, one has to make an assumption about the true abundance distribution of a community. Although most communities of macroorganisms seem to display a lognormal pattern of species abundance (17, 36, 66), there is still controversy as to which models fit best (24, 30). In the absence of a variety of large microbial data sets, it is not clear which, if any, of the proposed distribution models describe microbial communities. Finally, even if one of these models is a good approximation of relative abundances in microbial communities, parametric estimators require large data sets to evaluate the distribution parameters. The largest microbial data sets currently available include only a few hundred individuals.

NON-PARAMETRIC ESTIMATION

The final class of estimation methods, nonparametric estimators, is the most promising for microbial studies. **These estimators are adapted from mark-release-recapture (MRR) statistics for estimating the size of animal populations** (32, 59). Nonparametric estimators based on MRR methods consider the proportion of species that have been observed before ("recaptured") to those that are observed only once. In a very diverse community, the probability that a species will be observed more than once will be low, and most species will only be represented by one individual in a sample. In a depauperate community, the probability that a species will be observed more than once will be higher, and many species will be observed multiple times in a sample.

CHAO vs ACE


Tree of Life
Eukaryotes
Plants
Green algae
Land plants
Red algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Glaucophyte algae
Alveolates
Ciliates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Trypanosomes
Leishmania
Core Jakobids
Parabasalids
Diplomonads
Amoebas
Cellular slime molds
Plasmodial slime molds
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Nanoarchaeota
Thermoplasma
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquificales
Microsporidia
Cytophagales
Flavobacteria
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Aquifex
Deinococcus/Thermus
Cyanobacteria
Plastids

The Chao1 and abundance-based coverage estimators (ACE) use this MRR-like ratio to estimate richness by adding a correction factor to the observed number of species (9, 11). (For reviews of these and other nonparametric estimators, see Colwell and Coddington [14] and Chazdon et al. [12].)

Tree of Life

CHAO

$$S_{\mathrm{Chao1}} = S_{\mathrm{obs}} + \frac{n_1^2}{2n_2}$$

For instance, Chao1 estimates total species richness as where *S*obs is the number of observed species, *n*1 is the number of singletons (species captured once), and *n*2 is the number of doubletons (species captured twice) (9). Chao (9) noted that this index is particularly useful for data sets skewed toward the low-abundance classes, as is likely to be the case with microbes.

ACE


Tree of Life
Eukaryotes
Plants
Land plants
Red algae
Green algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Glaucophyte algae
Ciliates
Syndinians
Alveolates
Apicomplexa
Dinoflagellates
Ichthyosporeans
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Heterokonts
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Trypanosomes
Leishmanias
Discicristates
Core Jakobids
Parabasalids
Diplomonads
Retortamonads
Excavates
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
Mesomycetozoa
Choanoflagellates
Animals
Nucleariids
Fungi
Opisthokonts
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Korarchaeota
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquificales
Gram +
Actinobacteria
Firmicutes
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Deinococcus/Thermus
Cyanobacteria
Plastids

The ACE (10) incorporate data from all species with fewer than 10 individuals, rather than just singletons and doubletons. ACE estimates species richness as

$$S_{\mathrm{ACE}} = S_{\mathbf{abund}} + \frac{S_{\mathbf{rare}}}{C_{\mathrm{ACE}}} + \frac{F_1}{C_{\mathrm{ACE}}}\gamma^2_{\mathrm{ACE}}$$

where *S*rare is the number of rare samples (sampled abundances ≤10) and *S*abund is the number of abundant species (sampled abundances >10). Note that *S*rare + *S*abund equals the total number of species observed. *C*ACE = 1 – F1/Nrare estimates the sample coverage, where F1 is the number of species with *i* individuals and Finally

- http://www.ncbi.nlm.nih.gov/pmc/articles/instance/93182/equ/M4

$$\gamma^2_{\mathrm{ACE}} = \max\left[\frac{S_{\mathrm{rare}}\sum_{i=1}^{10} i(i-1)F_i}{C_{\mathrm{ACE}}\left(N_{\mathrm{rare}}\right)\left(N_{\mathrm{rare}}-1\right)} - 1, 0\right]$$

which estimates the coefficient of variation of the *Fi*'s (R. Colwell, User's Guide to EstimateS 5 [http://viceroy.eeb.uconn.edu/estimates]).

EVALUATING MEASURES

Both Chao1 and ACE underestimate true richness at low sample sizes. For example, the maximum value of *S*Chao1 is (*S*2obs + 1)/2 when one species in the sample is a doubleton and all others are singletons. Thus, *S*Chao1 will strongly correlate with sample size until *S*obs reaches at least the square root of twice the total richness (14).

BIAS

Bias describes the difference between the expected value of the estimator and the true, unknown richness of the community being sampled (in other words, whether the estimator consistently under- or overestimates the true richness).

To test for bias, one needs to know the true richness to compare against the sample estimates. As yet, this comparison is impossible for microbes, because no communities have been exhaustively sampled. The bias of richness estimators has only been tested in a few natural communities in which the exact abundance of every species in an area is known (12, 14, 15, 26, 47).

PRECISION

Precision describes the variation of the estimates from all possible samples that can be taken from the population

In contrast, precision is a relatively simple property to assess. With multiple samples (or one large sample) from a microbial community, the variance of microbial richness estimates can be calculated and compared. Moreover, most ecological questions require only comparisons of relative diversity. For these questions, an estimator that is consistent with repeated sampling (is precise) is often more useful than one that on average correctly predicts true richness (has the lowest bias). Thus, if we use diversity measures for relative comparisons, we avoid the problem of not being able to measure bias. (This assumes that the bias of an estimator does not differ so radically among communities that it disrupts the relative order of the estimates. In the absence of alternative evidence, this initial assumption seems appropriate.)

Chao (8) derives a closed-form solution for the variance of *S*Chao1:

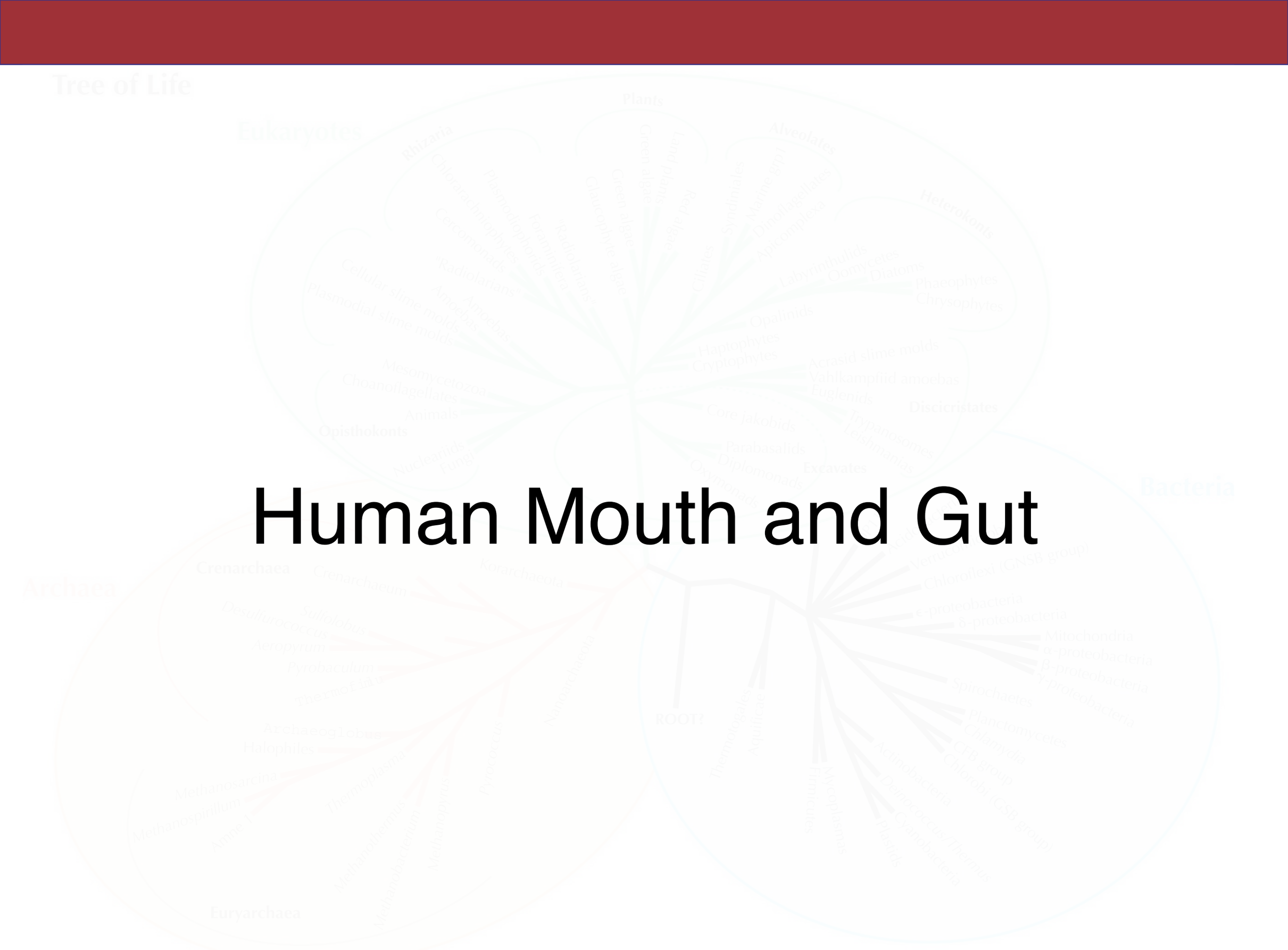
$$Var(S_{\text{Chao1}}) = n_2\left(\frac{m^4}{4} + m^3 + \frac{m^2}{2}\right), \text{where } m = \frac{n_1}{n_2}$$

This formula estimates the precision of Chao1; that is, it estimates the variance of richness estimates that one expects from multiple samples. A closed-form solution of variance for the ACE has not yet been derived.

Tree of Life

FOUR DATA SETS

Human Mouth and Gut



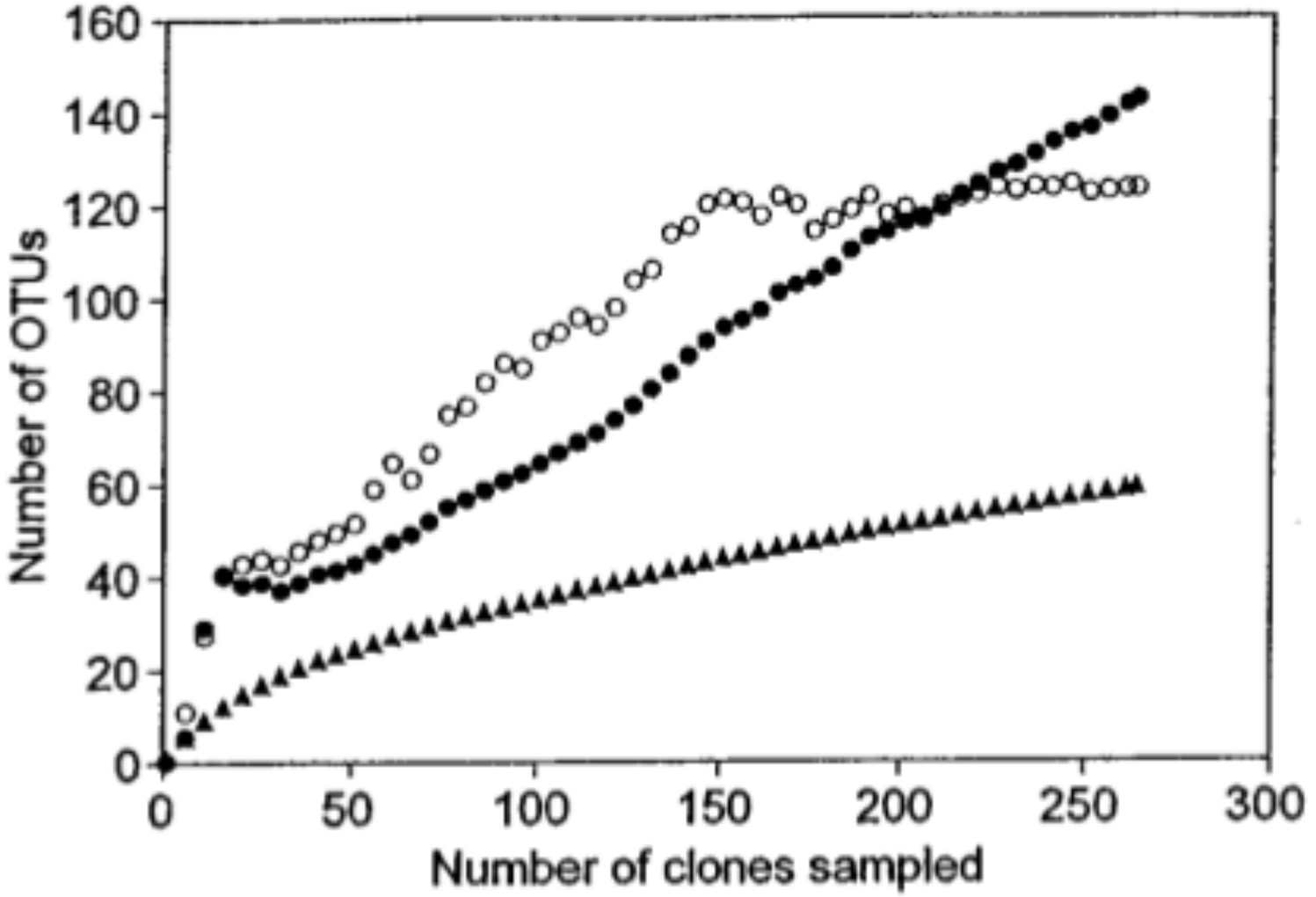


FIG. 3. Observed and estimated OTU richness of bacteria in a human mouth (33) versus sample size. The number of OTUs observed for a given sample size, or the accumulation curve, is averaged over 50 simulations (○). Estimated OTU richness is plotted for Chao1 (●) and ACE (▲) estimators.

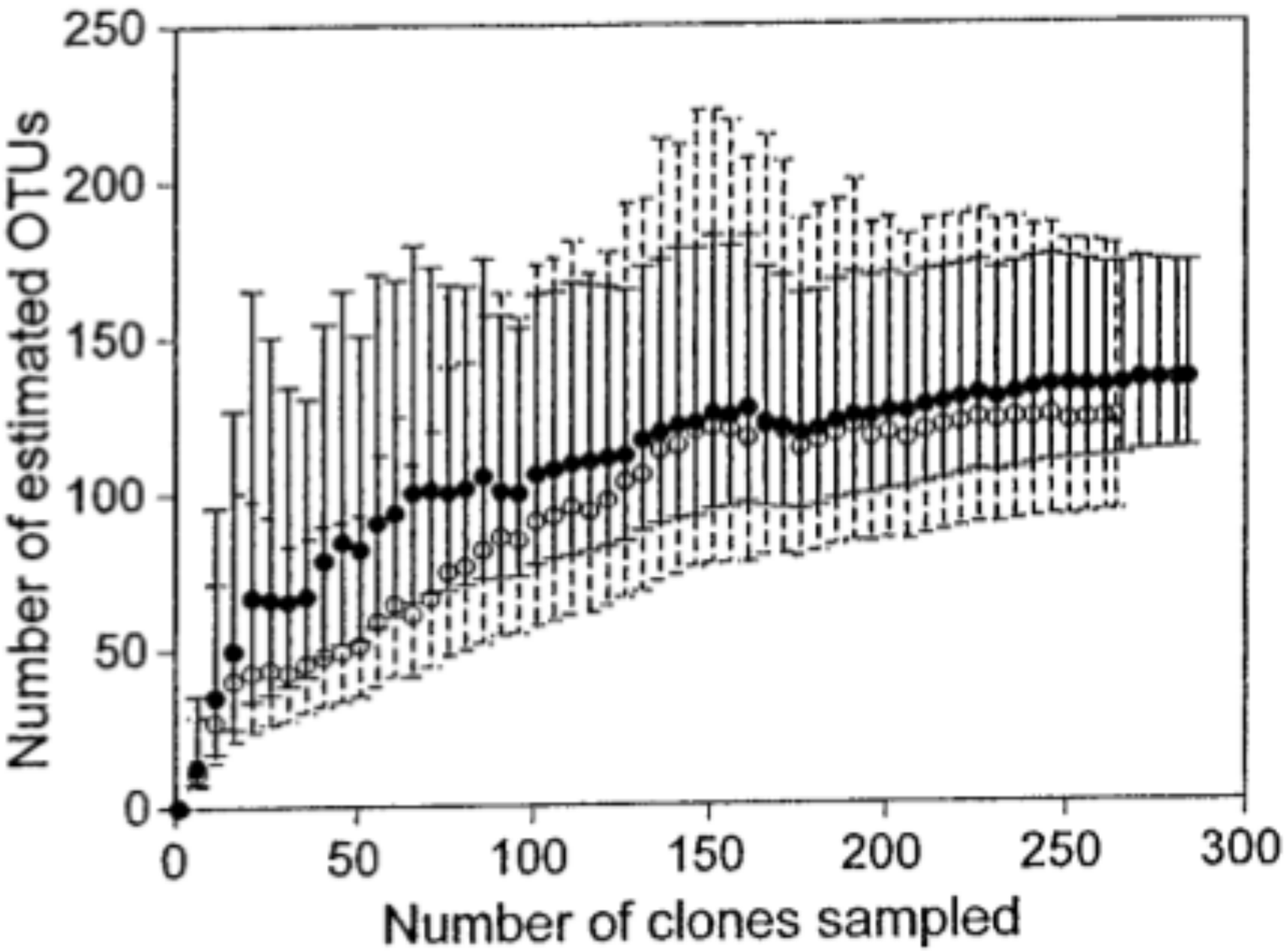


FIG. 4. Chaol estimates of human mouth (○) and gut (●) bacterial richness as a function of sample size. Error bars are 95% CIs and were calculated with the variance formula derived by Chao (8). The dashed lines are error bars for the mouth. The solid lines are error bars for the gut.

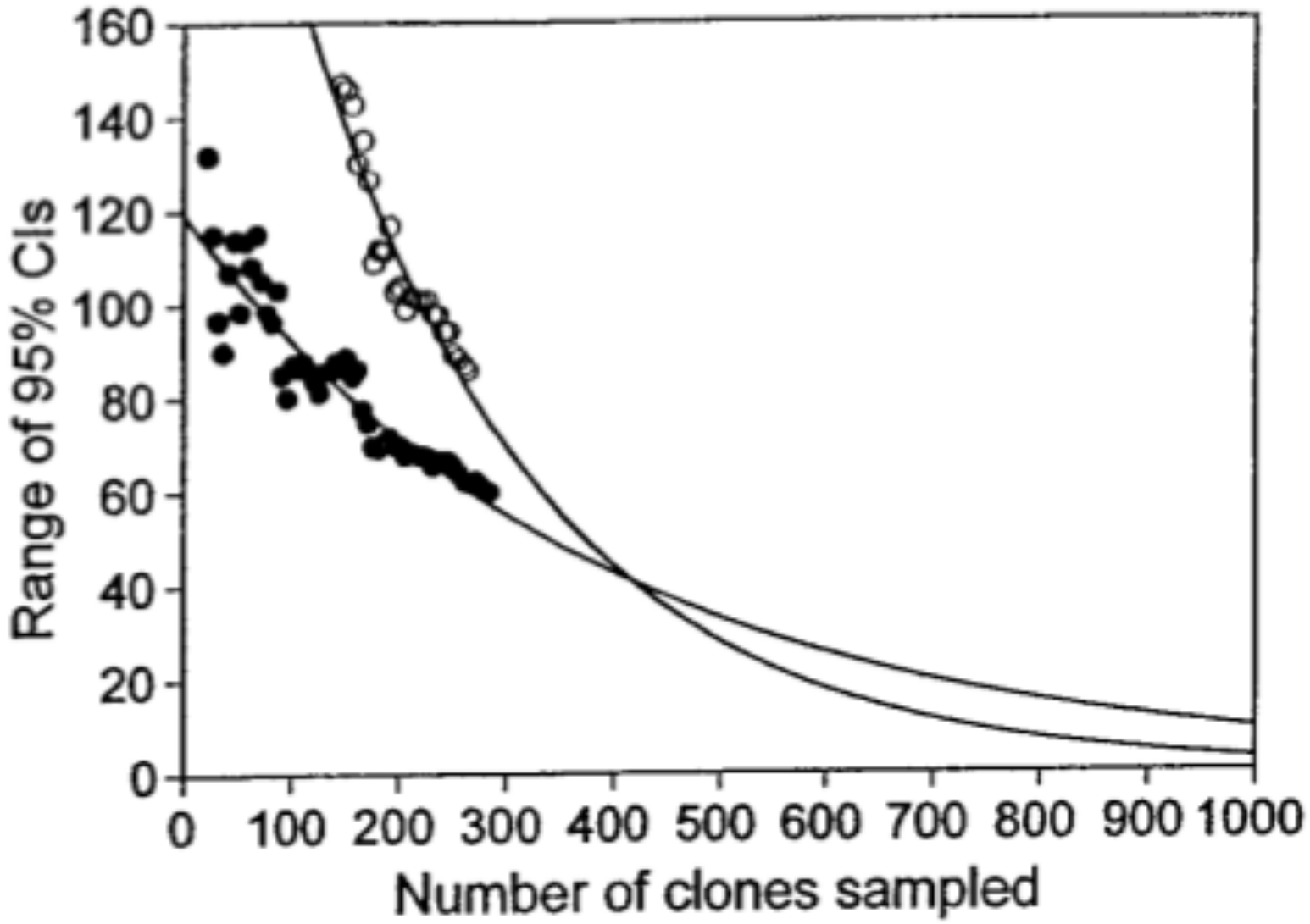


FIG. 5. Average size of the 95% CIs of Chaol estimates for bacteria in the human mouth (○) and gut (●) as sample size increases. These CIs are the same as in Fig. 4, but only the decreasing portions of the CIs are plotted. The curves are fitted negative exponential curves [mouth, $f(x) = 270e^{-0.0046x}$, $r^2 = 0.90$; gut, $f(x) = 120e^{-0.0026x}$, $r^2 = 0.87$].

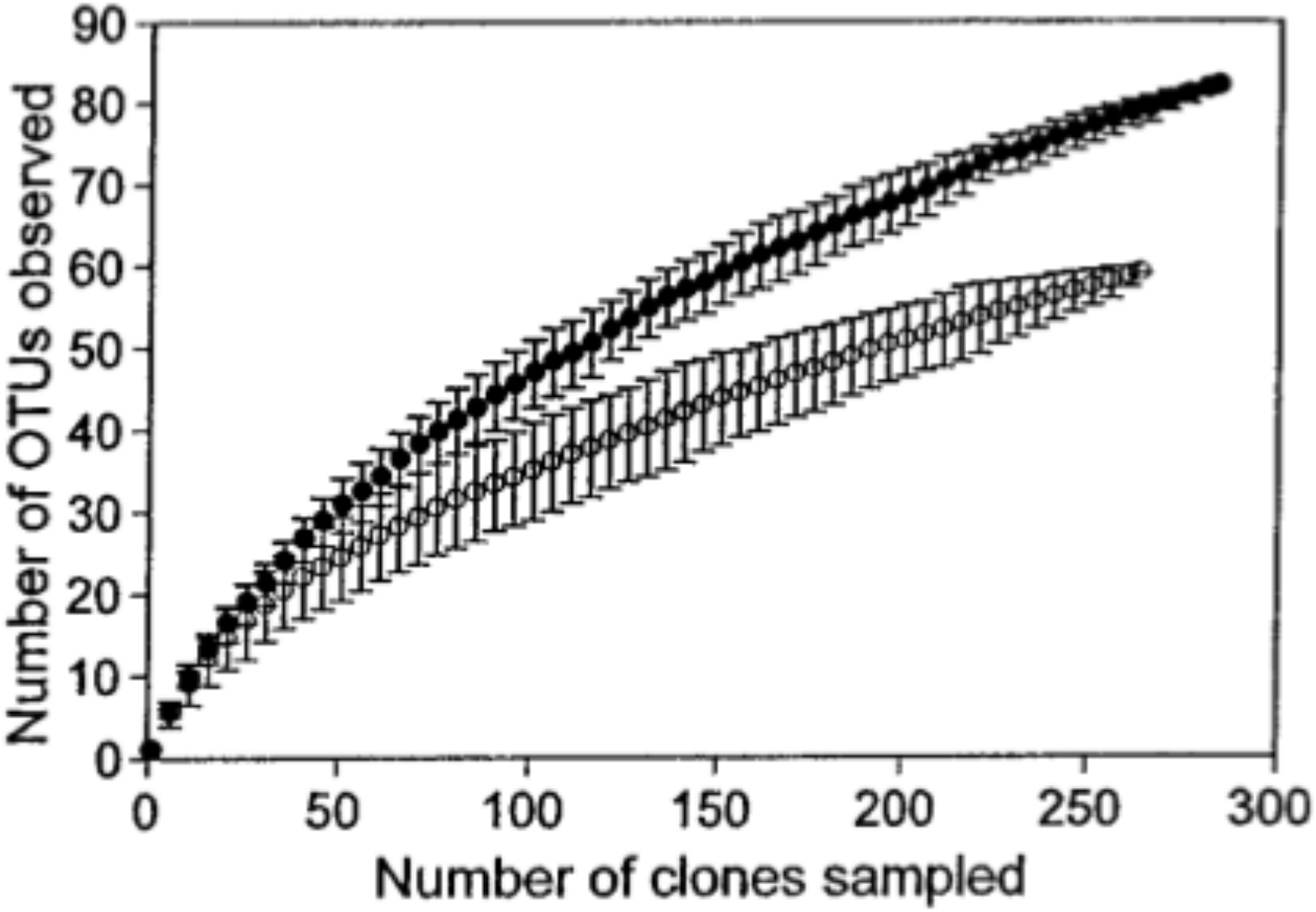


FIG. 6. Rarefaction curves of observed OTU richness in human mouth (○) and gut (●) bacterial samples. The error bars are 95% CIs and were calculated from the variance of the number of OTUs drawn in 100 randomizations at each sample size.

Rarefaction compares observed richness among sites, treatments, or habitats that have been unequally sampled. A rarefied curve results from averaging randomizations of the observed accumulation curve (25). The variance around the repeated randomizations allows one to compare the observed richness among samples, but it is distinct from a measure of confidence about the actual richness in the communities.

Aquatic Mesocosms

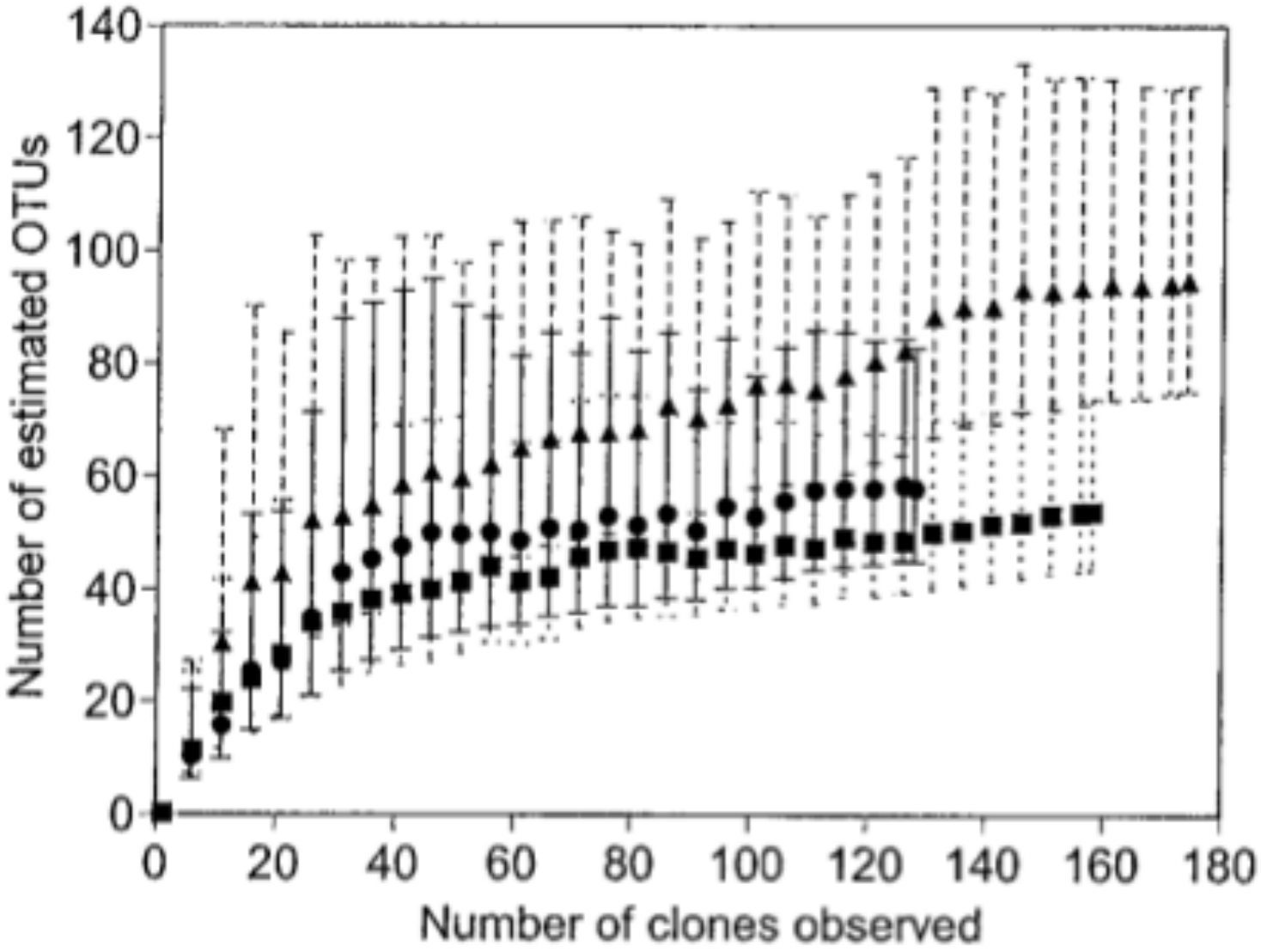


FIG. 7. Chao1 estimates of bacterial OTU richness in low- (■), intermediate- (●), and high- (▲) productivity ponds. Error bars are 90% CIs and were calculated with the variance formula derived by Chao (8). The dotted, solid, and dashed bars are error bars for the low-, intermediate-, and high-productivity mesocosms, respectively.

Tree of Life

Scottish Soil

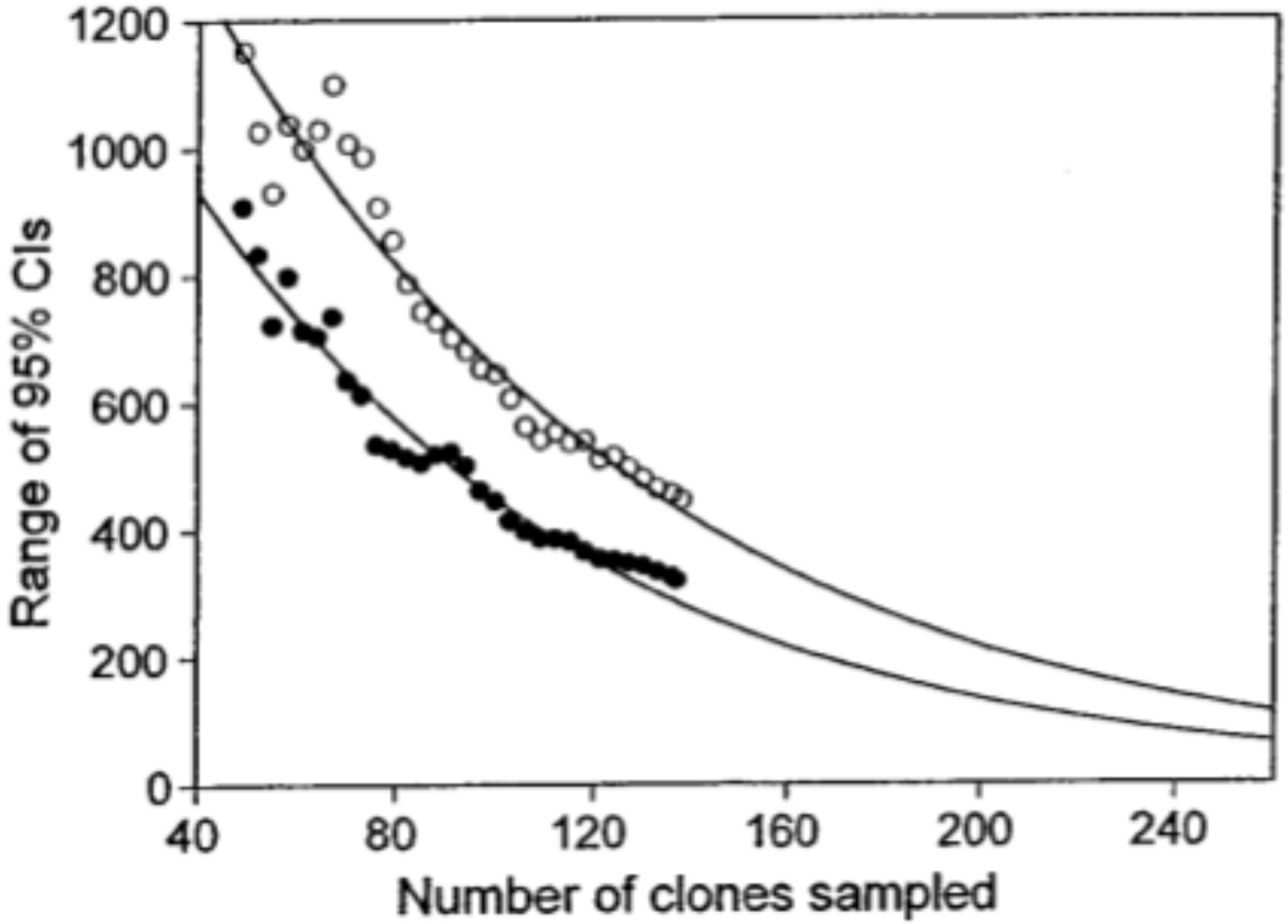


FIG. 9. Average size of the 95% CIs of Chaol estimates for the improved (○) and unimproved (●) soil as the number of clones sampled increases. These CIs are the same as in Fig. 8, but only the decreasing portions of the CIs are plotted. The curves are fitted negative exponential curves [improved, $f(x) = 1{,}500e^{-0.012x}$, $r^2 = 0.96$; unimproved, $f(x) = 2{,}000e^{-0.011x}$, $R^2 = 0.94$].

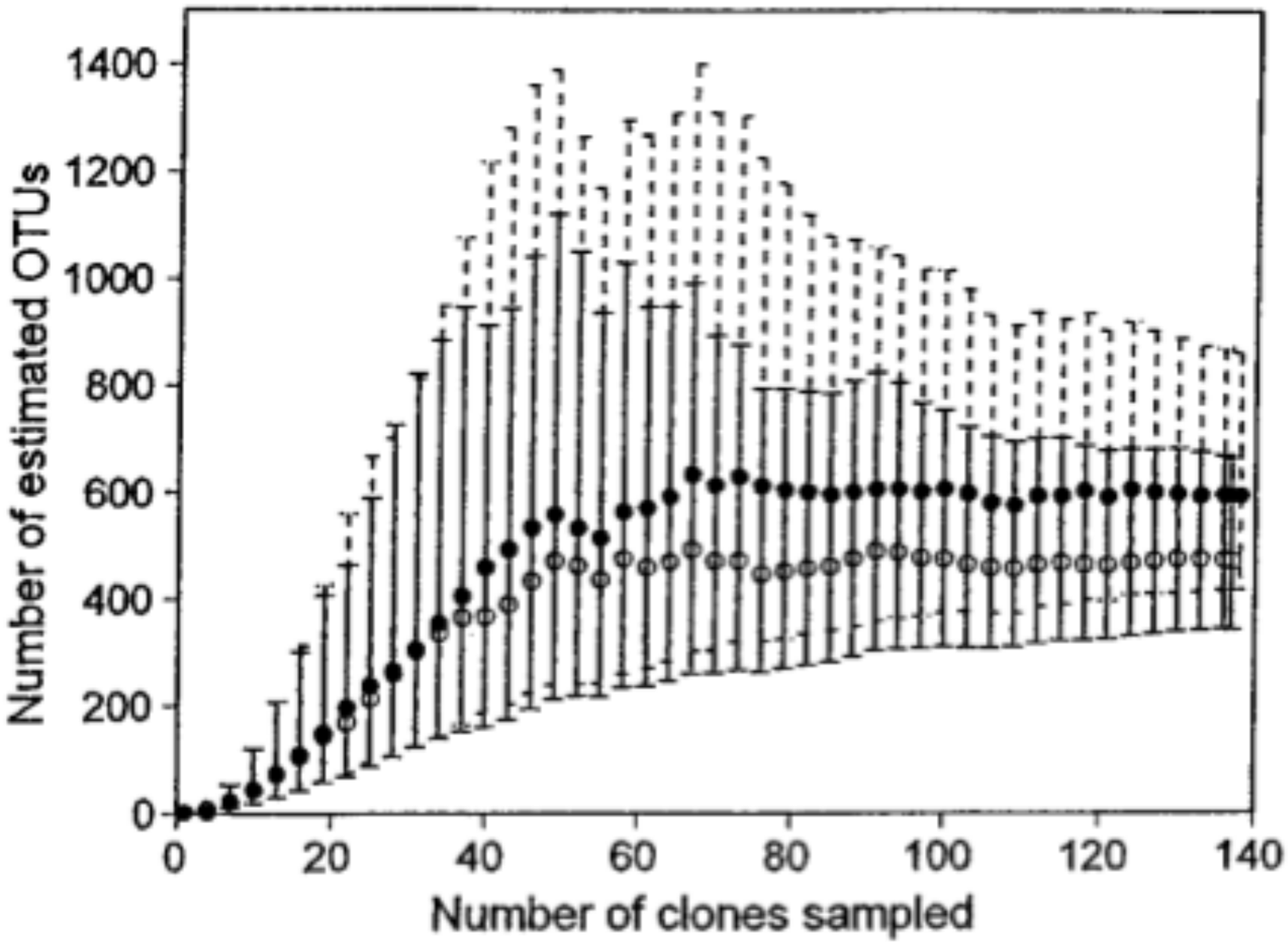
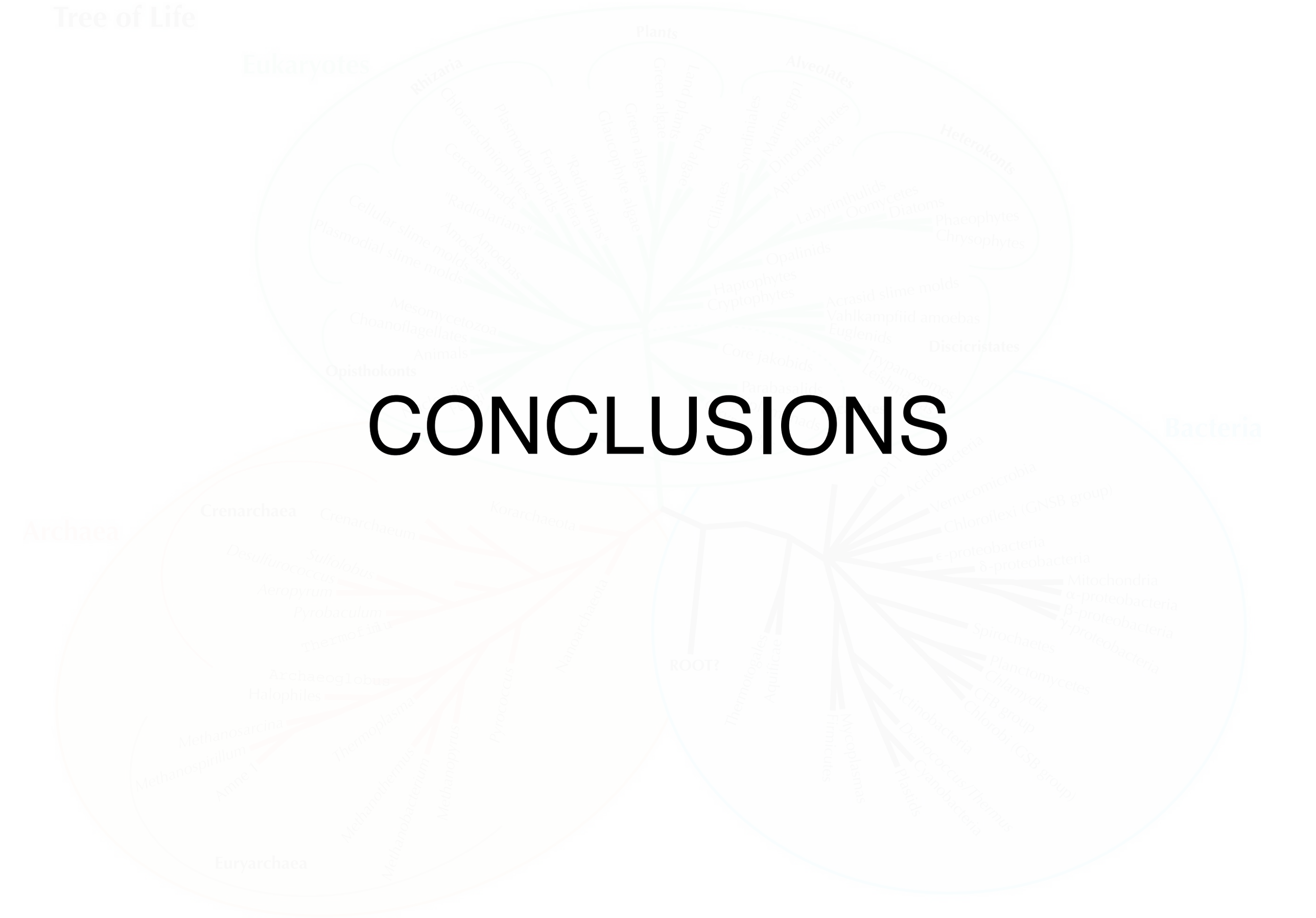


FIG. 8. Chaol estimates of bacterial OTU richness in improved (○) and unimproved (●) soil as a function of sample size. Error bars are 95% CIs and were calculated with the variance formula derived by Chao (8). The solid lines are error bars for the improved sample. The dashed lines are error bars for the unimproved sample.

CONCLUSIONS



In conclusion, while microbiologists should be cautious about sampling biases and use clear OTU definitions, our results suggest that comparisons among estimates of microbial diversity are possible. Nonparametric estimators show particular promise for microbial data and in some habitats may require sample sizes of only 200 to 1,000 clones to detect richness differences of only tens of species. While daunting less than a decade ago, sequencing this number of clones is reasonable with the development of high-throughput sequencing technology. Augmenting this new technology with statistical approaches borrowed from “macrobial” biologists offers a powerful means to study the ecology and evolution of microbial diversity in natural environments.

Because of inconsistencies in how diversity is measured in individual studies, e.g., how operational taxonomic units (OTUs) are selected or which region of the rRNA gene is sequenced, it is only by integrating information from these studies into a single phylogenetic context that these important questions can be addressed

Later - add genes and genomes

- Can do alpha, beta and gamma for taxa or genes or genomes

UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

Catherine Lozupone and Rob Knight
Appl. Environ. Microbiol. **2005, 71(12):8228. DOI: 10.1128/AEM.71.12.8228-8235.2005.**

Phylogenetic distance measures can provide far more power because they exploit the degree of divergence between different sequences.

Here we introduce a new phylogenetic method, called UniFrac, that measures the distance between communities based on the lineages they contain.

Here we introduce a new phylogenetic method, called UniFrac, that measures the distance between communities based on the lineages they contain. **UniFrac can be used to compare many samples simultaneously** because it satisfies the technical requirements for a distance metric (it is always positive, is transitive, and satisfies the triangle inequality) and can thus be used with standard multivariate statistics such as unweighted-pair group method using average linkages (UPGMA) clustering (**9**) and principal coordinate analysis (**23**). Similarly, UniFrac is more powerful than nonphylogenetic distance measures because it **exploits the different degrees of similarity between sequences**. To demonstrate the utility of the UniFrac metric for comparing multiple community samples and determining the factors that explain the most variation, we compared bacterial populations in different types of geographically dispersed marine environments.

How is a UNIFRAC Distance Calculated?

UniFrac metric.

The unique fraction metric, or UniFrac, measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both (Fig. 1). This measure thus captures the total amount of evolution that is unique to each state, presumably reflecting adaptation to one environment that would be deleterious in the other. rRNA is used purely as a phylogenetic marker, indicating the relative amount of sequence evolution that has occurred in each environment.

Sum of Length of All Unique Branches

Sum of Length of All Branches

Unifrac calculation

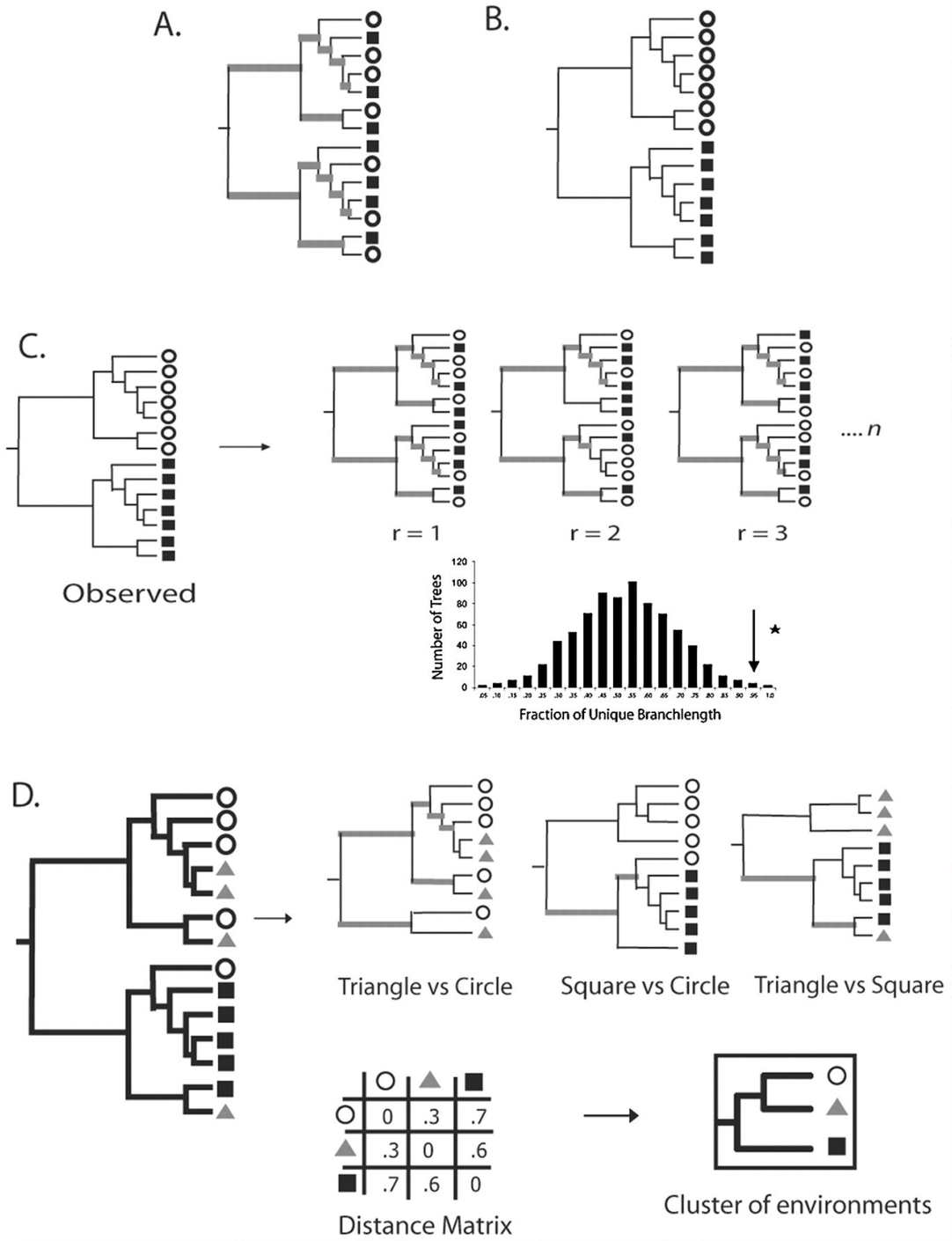
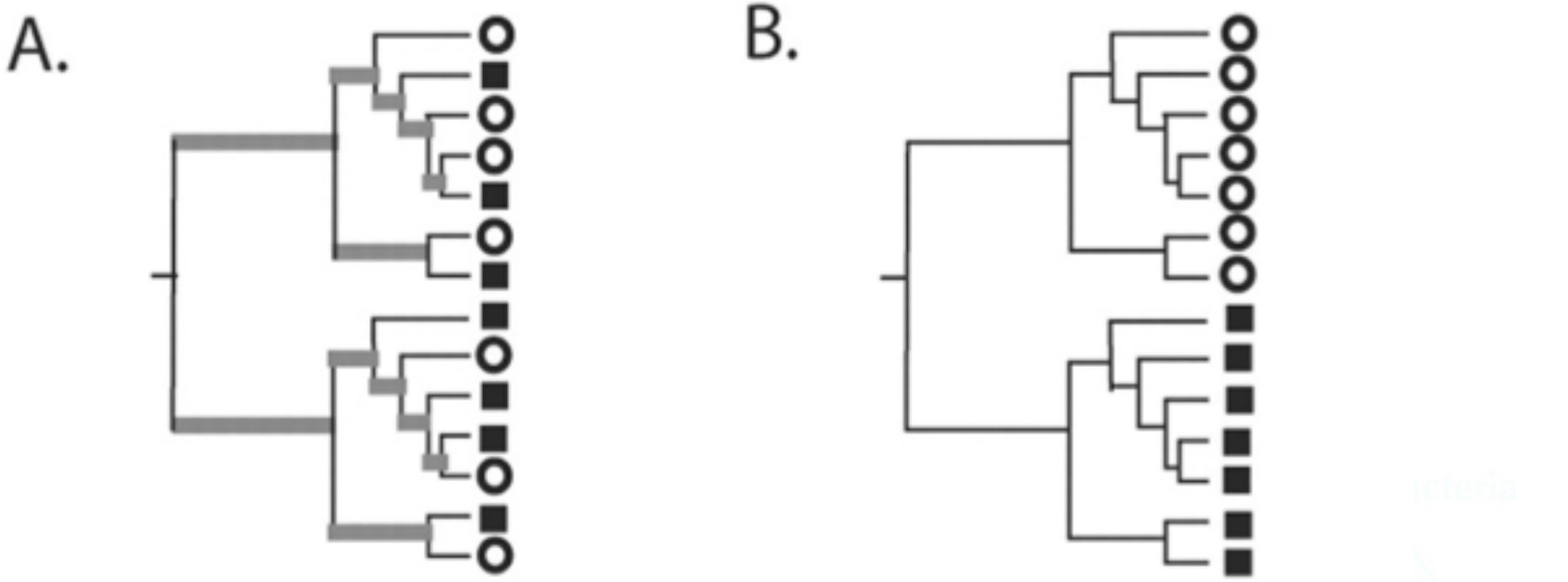
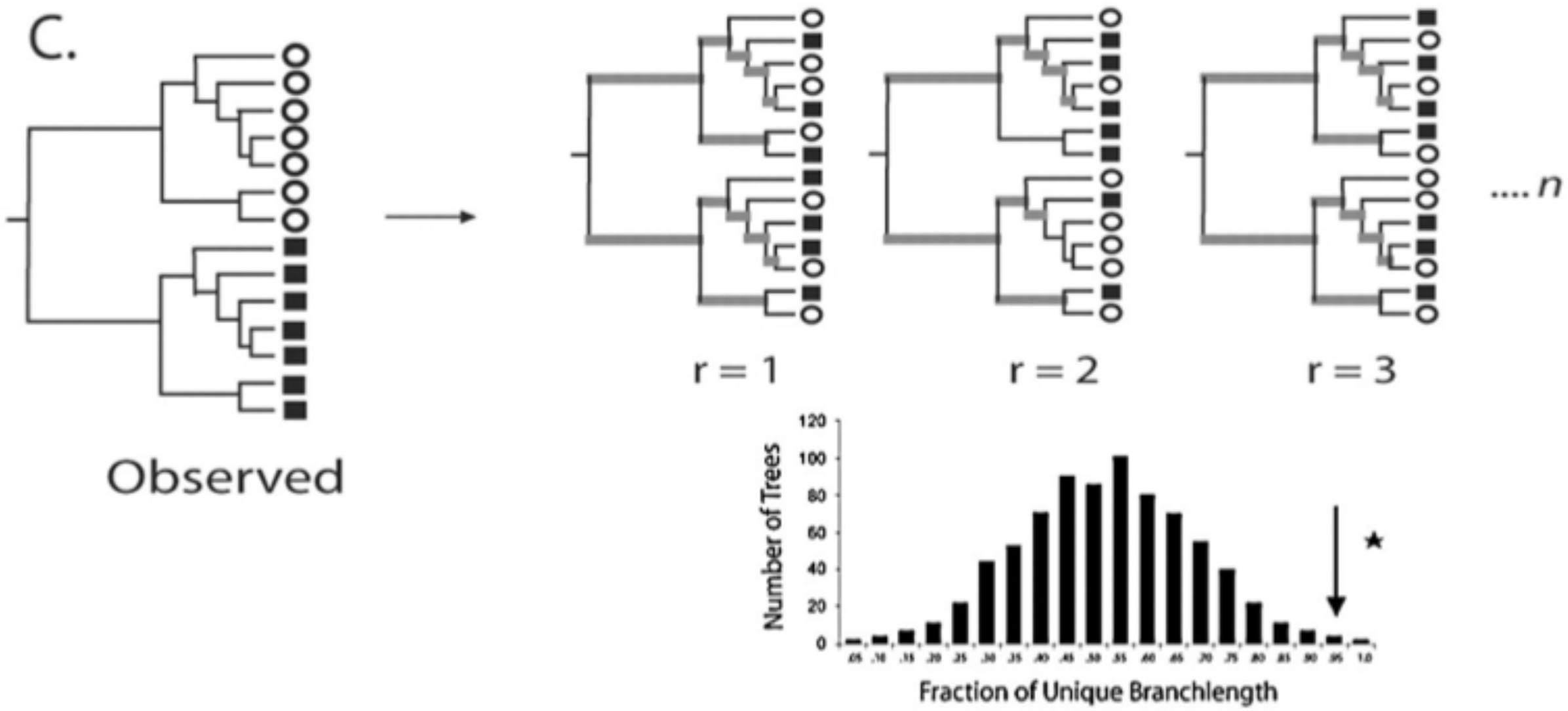


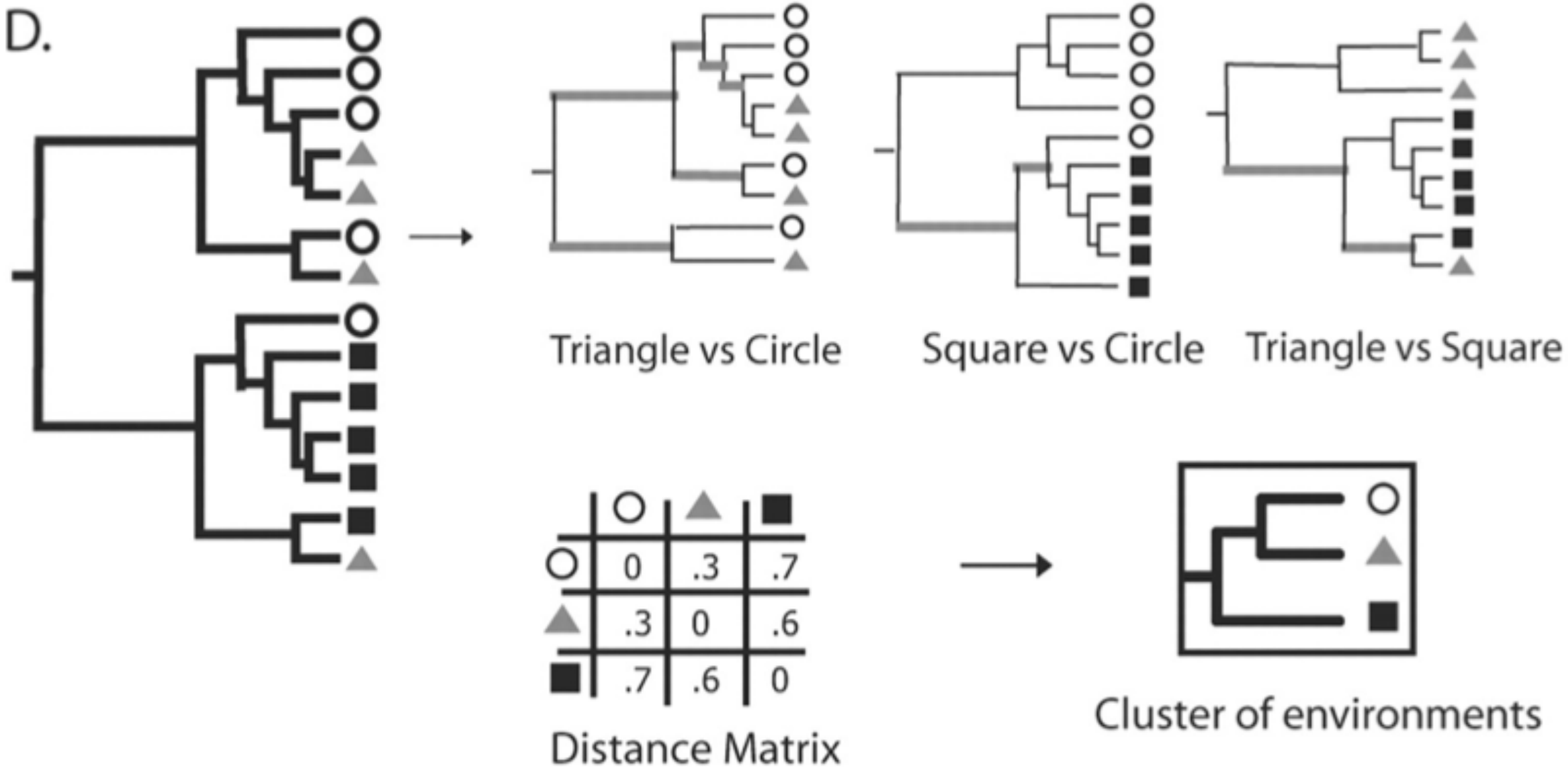
FIG. 1.
Calculation of the UniFrac distance metric. Squares, triangles, and circles denote sequences derived from different communities. Branches attached to nodes are colored black if they are unique to a particular environment and gray if they are shared. (A) Tree representing phylogenetically similar communities, where a significant fraction of the branch length in the tree is shared (gray). (B) Tree representing two communities that are maximally different so that 100% of the branch length is unique to either the circle or square environment. (C) Using the UniFrac metric to determine if the circle and square communities are significantly different. For *n* replicates (*r*), the environment assignments of the sequences were randomized, and the fraction of unique (black) branch lengths was calculated. The reported *P* value is the fraction of random trees that have at least as much unique branch length as the true tree (arrow). If this *P* value is below a defined threshold, the samples are considered to be significantly different. (D) The UniFrac metric can be calculated for all pairwise combinations of environments in a tree to make a distance matrix. This matrix can be used with standard multivariate statistical techniques such as UPGMA and principal coordinate analysis to compare the biotas in the environments.



Intuitively, if two environments are similar, few adaptations would be needed to transfer from one community to the other. Consequently, most nodes in a phylogenetic tree would have descendants from both communities, and much of the branch length in the tree would be shared (Fig. **1A**). In contrast, if two communities are so distinct that an organism adapted to one could not survive in the other, then the lineages in each community would be distinct, and most of the branch length in the tree would lead to descendants from only one of the two communities (Fig. **1B**).



Like the *P* test and the F_{ST} test, UniFrac can be used to determine whether two communities differ significantly by using Monte Carlo simulations. Two communities are considered different if the fraction of the tree unique to one environment is greater than would be expected by chance. We performed randomizations by keeping the tree constant and randomizing the environment that was assigned to each sequence in the tree (Fig. **1C**).



UniFrac can also be used to produce a distance matrix describing the pairwise phylogenetic distances between the sets of sequences collected from many different microbial communities (Fig. **1D**). We compared two samples by removing from the tree all sequences that were not in either sample and computing the UniFrac for each reduced tree. Standard multivariate statistics, such as UPGMA clustering (**9**) and principal coordinate analysis (**23**), can then be applied to the distance matrix to allow comparisons between the biotas in different environments (Fig. **1D**).

Tree of Life

Methods

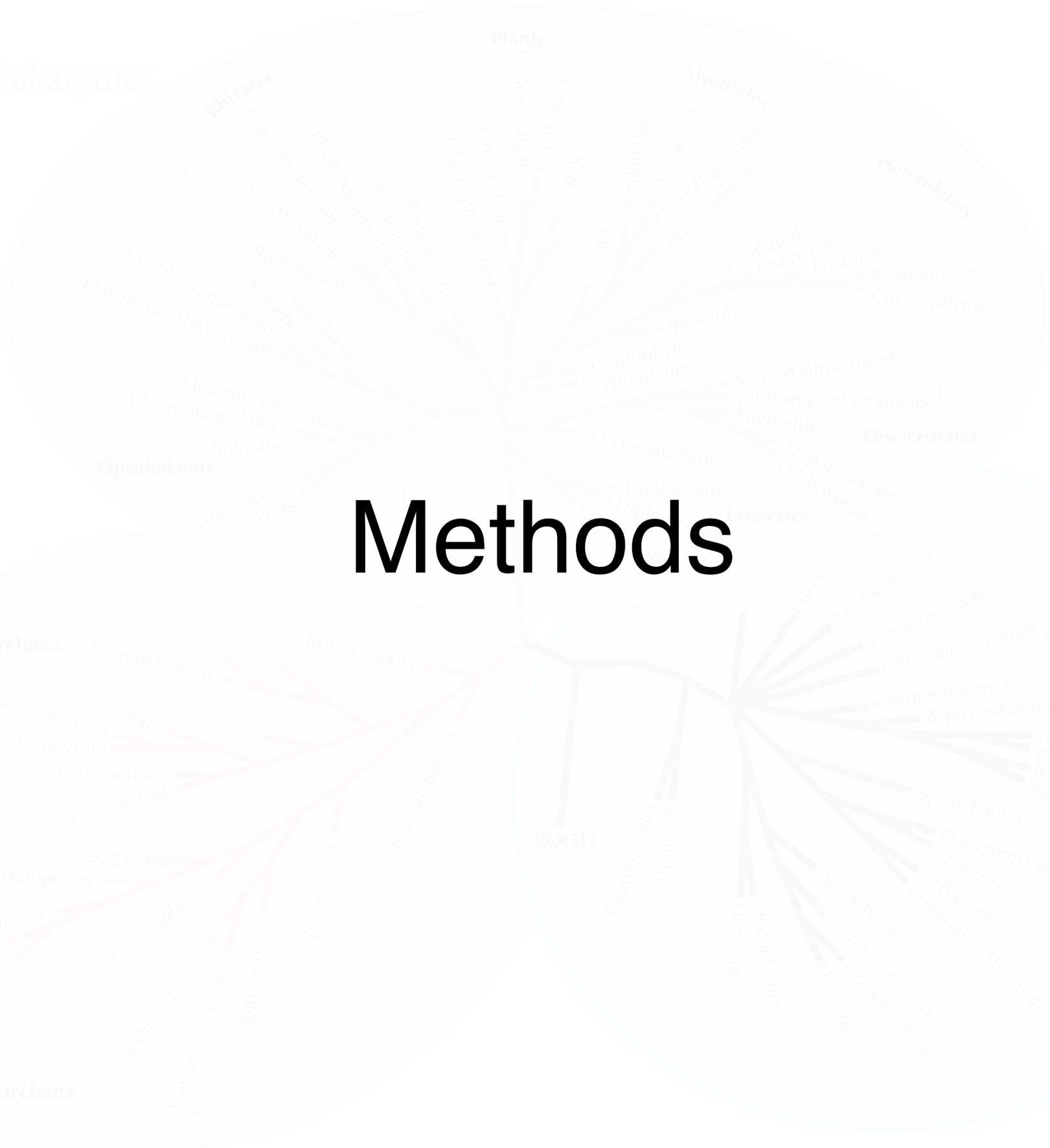


TABLE 1. Gene library information

Sample[a]	Reference	No. of sequences	Water column depth (m)	Sediment depth (cm)	Latitude, longitude	Temp (°C)
SRU1	38	79	155	0–1.1	76°58′N, 15°34′E	2.6
STU2	25	33	6,400		40°06′N, 144°11′E	
SNU3	4	36	709–940	1–2	66°S, 143°E	
SNC4	4	31	709–940		66°S, 143°E	
SNU5	5	101	761	0–0.4	66°32′S 143°38′E	
SNU6	5	146	761	1.5–2.5	66°32′S 143°38′E	
SNU7	5	231	761	20–21	66°32′S 143°38′E	
WRU8	2	87	55, 131		72–88°N, 51–356°E	
WTC9	8	36	1		54°09′N, 7°52′E	
WTU10	22	75	1–2		40°N, 73°E	23.8–29.2
WTC11	22	21	1–2		40°N, 72°E	23.8–29.2
WTU12	1	544			42°N, 71°E	16
WPU13	11	17	10		32°37′N 64°57′W	
WPU14	11	40	100, 500		31°49′N 64°57′W	
INC15	3	58			68°S, 78°E	
IRU16	6	62			80°N, 0°E	
IRC17	6	109			80°N, 0°E	
INU18	6	20			70°S, 15°E	
INC19	6	87			70°S, 15°E	
INU20	7	75			62–77°S, 74–165°E	

[a] The first character in the sample name designates the environment type (S, marine sediment; W, water; and I, ice). The second character indicates the geographic location (R, Arctic; N, Antarctic; T, temperate; and P, tropical). The third character indicates whether the sequences were derived from cultured isolates (C) or environmental clones (U).

Data analysis. We implemented UniFrac and associated analyses in Python 2.3.4 and ran all calculations on a Macintosh G4 computer running OSX 10.3.8. All code is available at **http://bayes.colorado.edu/unifrac.zip**. We implemented UPGMA clustering (**9**) and principal coordinate analysis (**23**) as described previously.

We downloaded small-subunit-rRNA sequences generated in the 12 different studies of marine environments (Table **1**) from GenBank, imported them into the Arb package (**26**), and aligned them using a combination of the Arb auto-aligner and manual curation. Because several studies used bacterium-specific primers, we excluded all nonbacterial sequences from the analysis. We added the aligned sequences to a tree representing a range of phylogenetic groups from the Ribosomal Database Project II (**29**) by Phil Hugenholtz (**15**). This sequence addition used the parsimony insertion tool and a lane mask (lanemaskPH) supplied in the same database so that only phylogenetically conserved regions were considered. We exported the tree from Arb and annotated each sequence with 1 of 20 sample designations (Table **1**). We then performed significance tests, UPGMA clustering, and principal coordinate analysis using UniFrac.

Jackknifing. We used jackknifing to determine how the number and evenness of sequences in the different environments affected the UPGMA clustering results. Specifically, we repeated the UniFrac analysis with trees that contained only a subset of the sequences and measured the number of times we recovered each node that occurred in the UPGMA tree from the full data set. In each simulation, we evaluated 100 reduced trees in which all of the environments were represented by the same specified number of sequences, using sample sizes of 17, 20, 31, 36, 40, and 58 sequences. These thresholds reflect the sample sizes from different environments in our original data set. If an environment had more than the specified number of sequences, we removed sequences at random; environments with fewer sequences were removed from the tree entirely.

Results

Results

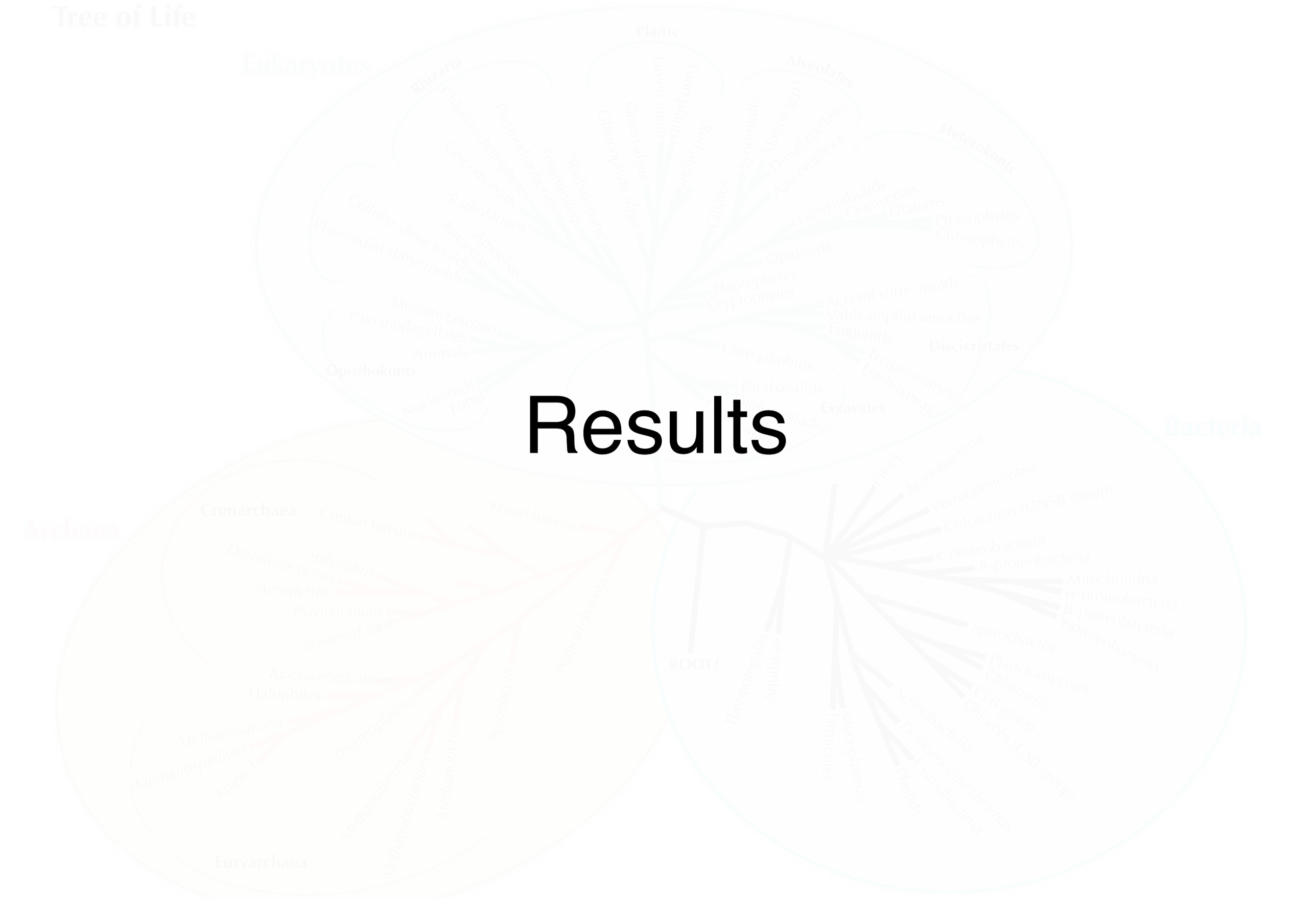


TABLE 2. UniFrac *P* values[a]

Sample	Compared sample(s) (*P* value)
SRU1	SNU3 (0.118), STU2 (0.111)
STU2	SNU3 (0.201), SRU1 (0.111), SNU5 (0.066), SNU6 (0.107)
SNU3	SNU6 (0.802), SNU7 (0.070), SRU1 (0.118), STU2 (0.201)
SNC4	WTC11 (0.105), SNU5 (0.053)
SNU5	SNC4 (0.053), STU2 (0.066)
SNU6	SNU7 (0.394), SNU3 (0.802), STU2 (0.107)
SNU7	SNU6 (0.394), SNU3 (0.070)
WRU8	
WTC9	WTC11 (0.639), INU20 (0.076), INC19 (0.155), INU18 (0.097)
WTU10	
WTC11	SNC4 (0.105), WTC9 (0.639), INC19 (0.055)
WTU12	
WPU13	WPU14 (0.238)
WPU14	WPU13 (0.238)
INC15	
IRU16	INU18 (0.257)
IRC17	
INU18	WTC9 (0.097), INU20 (0.055), INC19 (0.233), IRC17 (0.257)
INC19	WTC11 (0.055), WTC9 (0.155), INU18 (0.233)
INU20	WTC9 (0.076), INU18 (0.055)

[a] UniFrac *P* values were based on comparisons to 1,000 randomized trees. Results are listed only if the *P* value (listed in parentheses) is ≥0.05. All other pairwise comparisons indicated that the communities were significantly different.

We used UniFrac to determine which of the microbial communities represented by the 20 different samples were significantly different (Table 2)

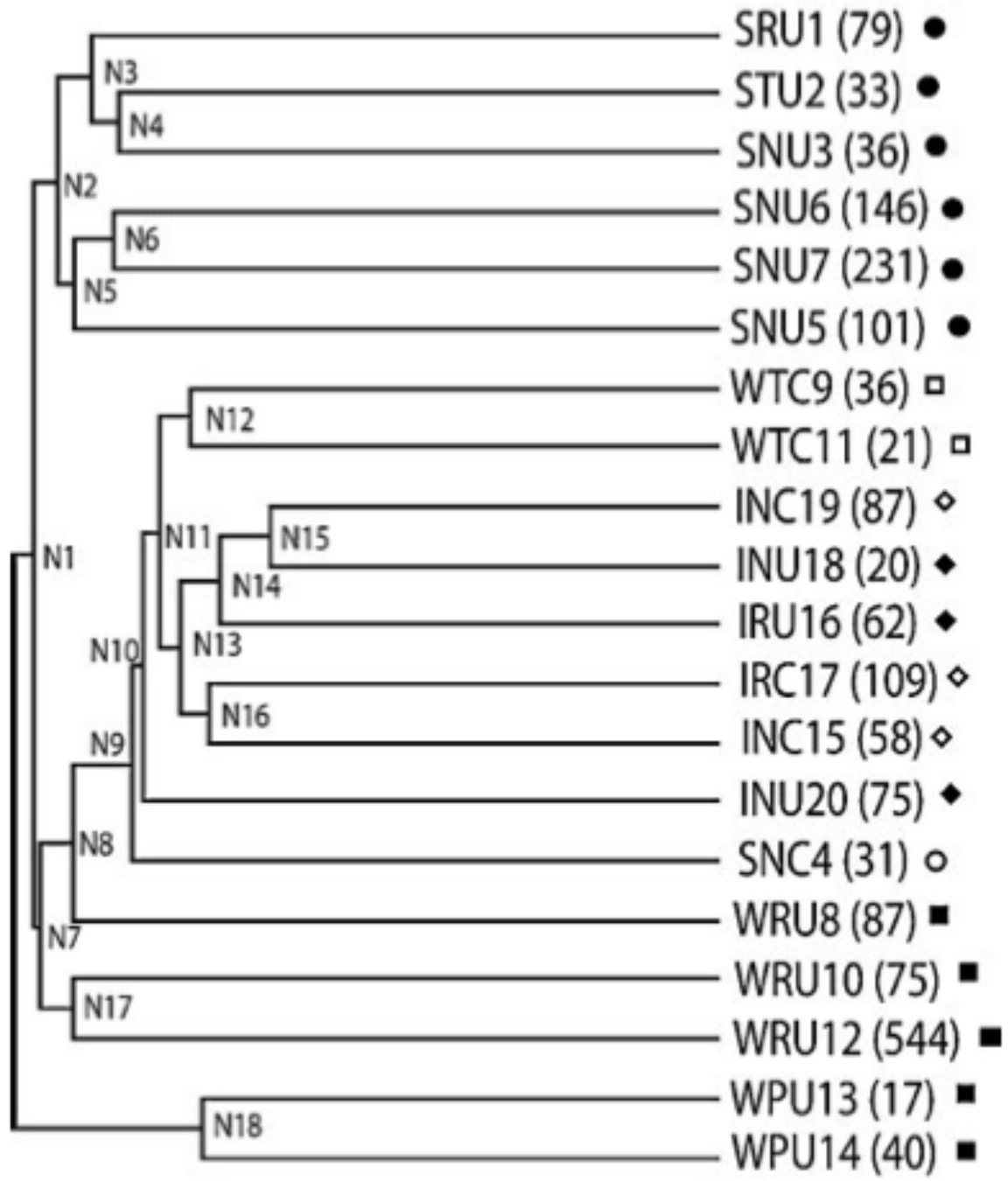


FIG. 2. UPGMA cluster of marine samples. The number of sequences that represent each environment is indicated next to the sample name, as well as the symbol with which the sample is represented in Fig. 3.

and as the basis for a distance matrix to cluster the samples using UPGMA (Fig. 2)

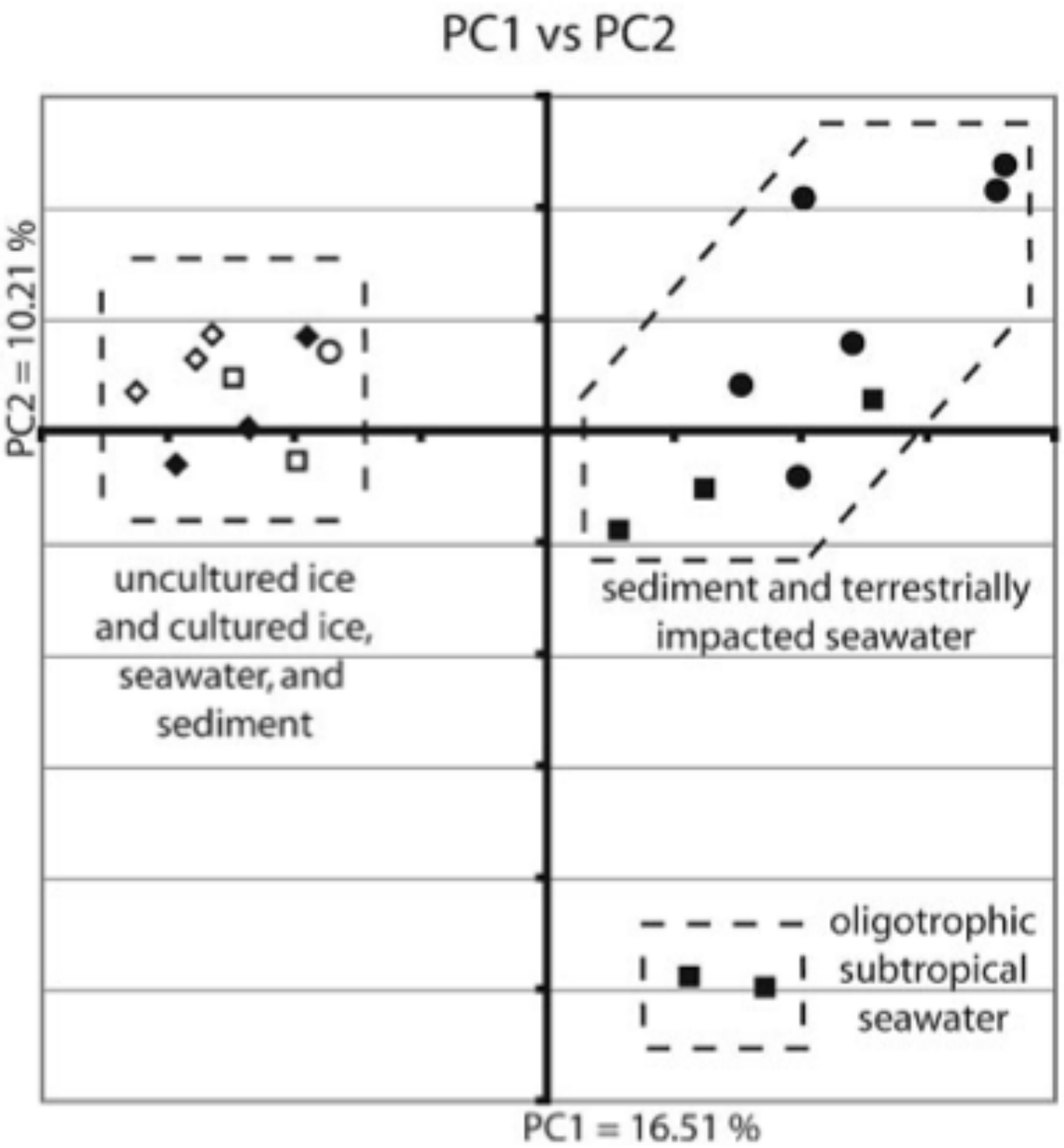


FIG. 3. First four principal coordinates from a principal coordinate analysis of marine samples. Samples from marine ice are represented by diamonds, sediment samples are represented by circles, and water samples are represented by squares. Shapes representing samples de- rived from cultured isolates are open, and those representing samples from environmental clones are filled. The percentages in the axis labels represent the percentages of variation explained by the principal co- ordinates.

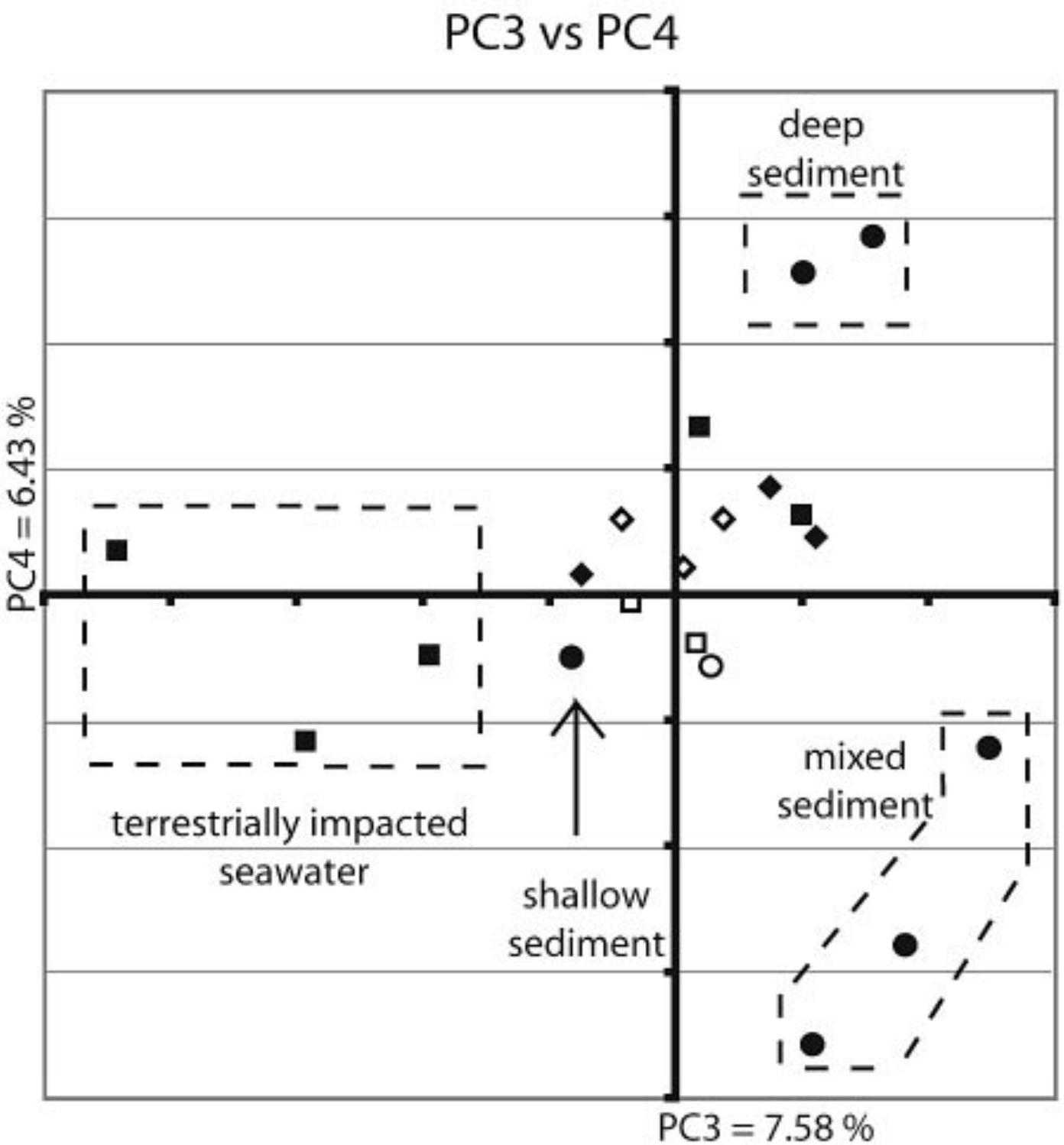


FIG. 3. First four principal coordinates from a principal coordinate analysis of marine samples. Samples from marine ice are represented by diamonds, sediment samples are represented by circles, and water samples are represented by squares. Shapes representing samples de- rived from cultured isolates are open, and those representing samples from environmental clones are filled. The percentages in the axis labels represent the percentages of variation explained by the principal co- ordinates.

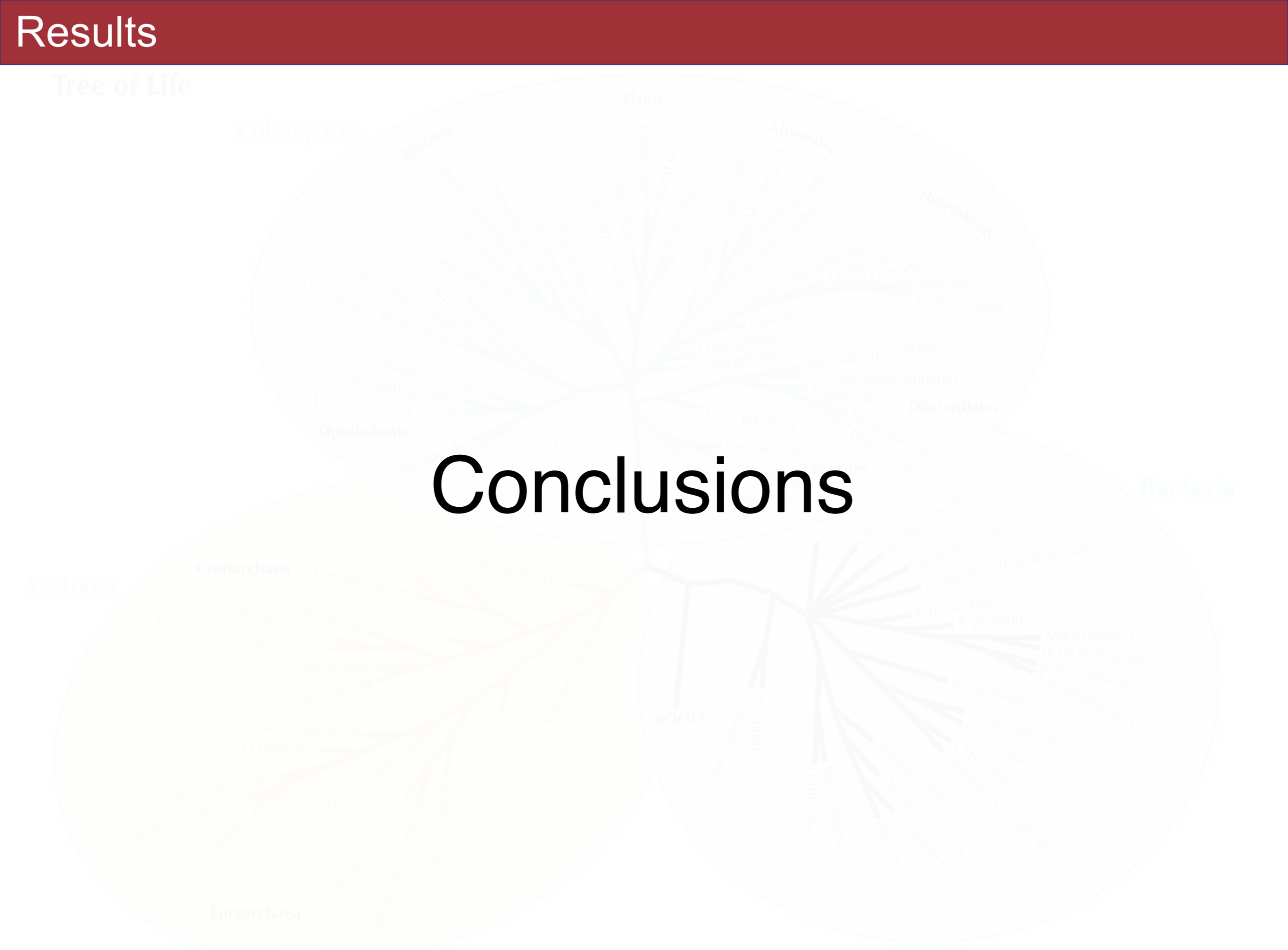
We used jackknifing to assess confidence in the nodes of the UPGMA tree (Table **3**). The results show biologically meaningful patterns that unite many individual observations in the literature and reveal several striking features of microbial communities in marine environments.

TABLE 3. UPGMA jackknifing results

Node	% of trials with node[a]					
	17	20	31	36	40	58
N1	3	14	31	27	12	NA
N2	8	1	29	33	48	63
N3	1	8	7	11	NA	NA
N4	14	16	11	NA	NA	NA
N5	1	0	0	1	27	37
N6	27	36	57	67	53	63
N7	23	23	36	44	52	66
N8	22	17	17	39	31	37
N9	52	58	64	NA	NA	NA
N10	8	16	79	96	94	100
N11	6	12	40	46	NA	NA
N12	13	31	NA	NA	NA	NA
N13	16	38	41	38	64	79
N14	34	50	29	23	12	6
N15	69	77	NA	NA	NA	NA
N16	18	40	27	28	28	21
N17	24	35	43	46	37	50
N18	97	NA	NA	NA	NA	NA

[a] For each node in the UPGMA tree (Fig. 2) (rows), the numbers show the percentages of trials ($n = 100$) that the node occurred in when each environment was represented by only 17, 20, 31, 36, 40, or 58 sequences (columns). The node names correspond to the node labels in Figure 2. NA, not available.

Results



Conclusions

Conclusion. The utility of UniFrac for making broad comparisons between the biotas of different environments based on 16S rRNA sequences has enormous potential to shed light on biological factors that structure microbial communities. The vast wealth of 16S rRNA sequences in GenBank and of environmental information about these sequences in the literature, combined with powerful phylogenetic tools, will greatly enhance our understanding of how microbial communities adapt to unique environmental challenges.