

## ORIGINAL ARTICLE

# Phylogenetic conservatism of functional traits in microorganisms

Adam C Martiny<sup>1,2</sup>, Kathleen Treseder<sup>2</sup> and Gordon Pusch<sup>3</sup>

<sup>1</sup>Department of Earth System Science, University of California, Irvine, CA, USA; <sup>2</sup>Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA and <sup>3</sup>Argonne National Laboratory, Argonne, IL, USA

**A central question in biology is how biodiversity influences ecosystem functioning. Underlying this is the relationship between organismal phylogeny and the presence of specific functional traits. The relationship is complicated by gene loss and convergent evolution, resulting in the polyphyletic distribution of many traits. In microorganisms, lateral gene transfer can further distort the linkage between phylogeny and the presence of specific functional traits. To identify the phylogenetic conservation of specific traits in microorganisms, we developed a new phylogenetic metric—consenTRAIT—to estimate the clade depth where organisms share a trait. We then analyzed the distribution of 89 functional traits across a broad range of Bacteria and Archaea using genotypic and phenotypic data. A total of 93% of the traits were significantly non-randomly distributed, which suggested that vertical inheritance was generally important for the phylogenetic dispersion of functional traits in microorganisms. Further, traits in microbes were associated with a continuum of trait depths ( $\tau_D$ ), ranging from a few deep to many shallow clades (average  $\tau_D$ : 0.101–0.0011 rRNA sequence dissimilarity). Next, we demonstrated that the dispersion and the depth of clades that contain a trait is correlated with the trait's complexity. Specifically, complex traits encoded by many genes like photosynthesis and methanogenesis were found in a few deep clusters, whereas the ability to use simple carbon substrates was highly phylogenetically dispersed. On the basis of these results, we propose a framework for predicting the phylogenetic conservatism of functional traits depending on the complexity of the trait. This framework enables predicting how variation in microbial composition may affect microbially-mediated ecosystem processes as well as linking phylogenetic and trait-based patterns of biogeography.**

*The ISME Journal* advance online publication, 13 December 2012; doi:10.1038/ismej.2012.160

**Subject Category:** Integrated genomics and post-genomics approaches in microbial ecology

**Keywords:** Phylogenomics; SEED; traits; lateral gene transfer; Biolog; consenTRAIT

## Introduction

Microorganisms control many biogeochemical processes. Therefore, the pattern of microbial biodiversity and the response to environmental changes are of great interest. Indeed, a large body of literature demonstrates that microbial community composition varies over time and space and responds to environmental changes (Horner-Devine *et al.*, 2004; Fuhrman *et al.*, 2006; Allison and Martiny, 2008). However, the overwhelming amount of microbial diversity and its possible functional redundancy suggests that the knowledge of microbial composition offers little predictive power for microbial process rates. Thus, it remains unclear how microbial community composition is related to ecosystem functioning.

The key to addressing this issue is the relationship between phylogeny and functional traits (Cadotte *et al.*, 2008; Gravel *et al.*, 2011). Predicting changes in community composition and ecosystem processes based on functional traits has received considerable attention in plant ecology. Recently, studies have found that the phylogenetic diversity of a plant community is an important predictor of plant productivity (Cadotte *et al.*, 2008; Cavender-Bares *et al.*, 2009). These results suggest that plant functional traits—especially those related to productivity—are more commonly shared among closely related species. Supporting this idea, a microcosm study showed that more closely related protists share more traits and that relatedness affected the competition among coexisting organisms (Violle *et al.*, 2011). Thus, species richness or diversity alone may be insufficient to understand how community composition and assembly affect the ecosystem function. Instead, we also need information about the phylogenetic relatedness of an organisms (Webb *et al.*, 2002). But at least some

Correspondence: AC Martiny, Department of Earth System Science, University of California, 3208 Croul Hall, Irvine, CA 92697, USA.

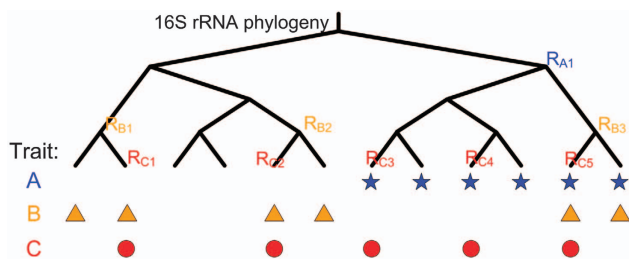
E-mail: amartiny@uci.edu

Received 1 June 2012; revised 24 October 2012; accepted 2 November 2012

studies have failed to measure a link between phylogenetic relatedness and functional traits in the plant communities (for example, Swenson *et al.*, 2012) and this can possibly be due to high variation in the functional potential among closely related lineages. Thus, the phylogenetic distribution and clustering of functional traits may have effects on both community assembly and functioning.

In a similar manner, the phylogenetic conservatism of microbial functional traits can also be considered (Philippot *et al.*, 2010). Continuous traits in microorganisms include salt tolerance or metabolic activity at different temperatures, whereas discrete traits include the metabolic ability to fix nitrogen or CO<sub>2</sub>. If the traits are highly phylogenetically conserved, then they will be shared among members of deep clades of microorganisms (Figure 1—trait A). Alternatively, traits that are less conserved will be found in small phylogenetic clades (Figure 1—trait B). Finally, gene loss, convergent evolution and lateral gene transfer can result in the distribution of many traits across multiple phylogenetic groups (Doolittle, 1999; Snel *et al.*, 2002), leading to random associations between phylogenetic and functional relatedness (Boucher *et al.*, 2003) (Figure 1—trait C).

The underlying mechanisms responsible for variation in the phylogenetic dispersion of functional traits are presently not clear. However, it is known that central gene systems consisting of many interacting proteins evolve more slowly and are less likely to be transferred between organisms (Jain *et al.*, 1999; Pal *et al.*, 2005; Shi *et al.*, 2005; Wellner *et al.*, 2007). In contrast, genes associated with the peripheral systems are more likely to be gained and lost. Thus, we hypothesize that functional traits based on complex genetic systems integral to the organism evolve slowly and will be more phylogenetically conserved (Figure 1—trait A). This would result in traits shared in phylogenetically deep clusters. In contrast, simple traits consisting of few interacting proteins will be more dispersed or completely randomly distributed (Figures 1—trait



**Figure 1** Hypothetical phylogenetic distribution of functional traits present in  $i$  clades including a trait A with a high  $\tau_D$ , three clades with trait B with a low  $\tau_D$ , and five randomly distributed lineages with trait C.  $R_i$  denotes the root node for each clade  $i$  with a given trait and the  $\tau_D$  is the average 16S rRNA distance between the root node and the strains in each clade sharing a trait.

B and C). Here, traits would be shared in phylogenetically shallow clusters.

Different approaches have been proposed for estimating evolutionary rate of change and phylogenetic dispersion of traits. This includes estimates of character evolution using either maximum likelihood or Bayesian approaches. Alternatively, other approaches compare the phylogenetic distribution of a trait to random or Brownian Motion models of evolution, such as Pagel's lambda for continuous traits and Purvis and Fritz's D for discrete traits (Pagel, 1999; Fritz and Purvis, 2010). Although these tests quantify the rate of evolution or evaluate the significance of phylogenetic clustering, they do not estimate the phylogenetic relatedness of the clusters of organisms sharing a trait—for example, in terms of 16S rRNA sequence similarity. For instance, if a given functional trait is associated with fine-scale clusters with a low trait depth ( $\tau_D$ ), the relative abundance of alpha- vs gamma-Proteobacteria may not accurately predict this function in a community. Thus, knowing the actual  $\tau_D$  is important for linking specific changes in microbial community composition with changes in the functional potential of a community.

The aim of this study was twofold. First, we developed a new phylogenetic metric—consenTRAIT—that directly estimates the sequence similarity of clusters of organisms sharing discrete traits. Second, we used this test to examine the distribution and degree of phylogenetic clustering of a variety of microbial traits. To address the latter, we analyzed two data sets. The first data set includes all completely sequenced prokaryotic genomes. For each genome, we identified the presence of genomic subsystems underlying many ecologically relevant functions like photosynthesis, CO<sub>2</sub> fixation, N-cycling, or growth on various organic carbon sources. The advantage of this data set is that it provides a broad overview of the metabolic potential of many strains across many phylogenetic groups. At the same time, many traits are associated with genetic changes beyond the presence or absence of a pathway, such as differences in regulation or mutations in associated genes. Therefore, we also analyzed a second data set of organic carbon usage by hundreds of bacterial strains. This data set provides phenotypic information on specific traits for many closely related bacteria, but does not cover as broad a phylogenetic range as the genomic data set.

Using our new metric and these two extensive data sets, we specifically asked three questions: Are most traits phylogenetically randomly distributed? If not, are different traits associated with different levels of phylogenetic clustering? Finally, is the degree of phylogenetic clustering of a trait correlated with its molecular complexity? If so, this result provides a biochemical basis for predicting the phylogenetic conservatism and distribution of traits in microorganisms.

## Materials and methods

### Data sets

To identify the phylogenetic distribution of traits in microorganisms, we analyzed two data sets. The first data set consisted of 2229 prokaryotic genomes (full Patric SEED database of 15 October 2010 (Snyder *et al.*, 2007)). This covered 26 phyla (Supplementary Figure S1). We then identified the distribution of 19 genomic subsystems underlying ecologically relevant functions. We used the SEED annotation database (Supplementary Table S1), which provided a uniform annotation for all genomes (Overbeek *et al.*, 1999, 2005). We used the SEED API for all analyses (Disz *et al.*, 2010). Whenever possible, we manually checked and edited the presence of each subsystem by searching literature for assigned physiologies. We estimated the number of genes underlying a trait as the genes in a given SEED subsystem. For the second data set, we retrieved data describing utilization patterns of 70 organic carbon substrates of 738 strains from Biolog Inc. (Hayword, CA, USA) (Supplementary Table S2). This data set covered five phyla (Supplementary Figure S1). These carbon sources ranged from  $C_1$  to polymeric compounds. The molecular complexity as determined by the Bertz Complexity Index (Bertz, 1981) of each carbon substrate was recorded from <http://pubchem.ncbi.nlm.nih.gov/>.

### Phylogenetic tree reconstruction

An aligned 16S rRNA sequence associated with each genome was retrieved from the Silva database (Pruesse *et al.*, 2007). For the carbon substrate utilization data set, we matched the strain names to entries in the Silva database in order to retrieve 16S rRNA sequences. A phylogenetic tree for each data set was estimated using PHYLIP (DNA distance with F84 correction, neighbor-joining, and 100 bootstraps) (Felsenstein, 2006) and RAxML (350 bootstraps) at the CIPRES science gateway (Stamatakis *et al.*, 2008). Unless otherwise noted, the results displayed were estimated using the DNA distance method.

### Phylogenetic mapping of traits

For each trait from the two data sets, we determined the  $\tau_D$ , phylogenetic dispersion and frequency of gene gain and loss.  $\tau_D$  was estimated using a custom designed algorithm—consenTRAIT: consensus analysis of phylogenetic trait distribution. The script was written in R using functions from the ‘ape’ and ‘adephylo’ packages (Paradis *et al.*, 2004; Jombart *et al.*, 2010). To estimate  $\tau_D$ , we first identified the root node of clades where at least 90% of the members shared the trait. We then estimated the average consensus sequence distance ( $d$ ) between the root node ( $R_j$ ) of  $n$  clades ( $j$ ) sharing a given

trait and the  $m$  members ( $i$ ) of clades (that is, the leaves,  $S_i$ ):

$$\tau_D = \frac{1}{n} \sum_j \frac{1}{m} \sum_i^m d(S_{ij})$$

This estimate was repeated for each bootstrap tree. We attributed the presence of singleton entries (that is, no neighbor sharing the trait) to undersampling. Therefore, we scored  $\tau_D$  for singleton entries using half the distance to the nearest internal node, which essentially assumes an equal likelihood of finding a neighbor organism with or without the trait. We included a non-parametric estimate for significant clustering by randomly assigning traits 1000 times (10 times to each bootstrap tree) to entries in the phylogenetic tree as a null distribution. We then compared the estimated  $\tau_D$  to this null distribution.

We used the test developed by Fritz and Purvis, (2010) to estimate the phylogenetic dispersion ( $D$ ) for discrete traits from the R package ‘caper’. We did 1000 permutation based on random or Brownian motion patterns of evolution and compared the observed phylogenetic pattern to these two distributions for each trait. Here, a  $D < 0$  suggests a highly clustered trait,  $D \sim 0$  indicates a Brownian motion mode of evolution,  $D = 1$  suggests a random mode of evolution and  $D > 1$  suggests phylogenetic overdispersion (Fritz and Purvis, 2010). A trait was assigned as non-random if the probability of matching a random distribution was less than 50%. We used the Lilliefors Test to examine for normality (Lilliefors, 1967) and Pearson and Spearman correlation analysis to test for significant correlations between variables. This was done with Matlab (Mathworks, Natick, MA, USA).

### Character evolution rate

An estimation of trait gain and loss events was performed based on a Wagner parsimony approach using Count (Csuros, 2010). We used a likelihood ratio of 1:4 for gene gain and loss events, respectively, but also tested different ratios. These gave comparable results ( $R > 0.9$ ). We also estimated the rate of trait change with a Bayesian approach using SIMMAP version 1.5 (Bollback, 2006). We estimated the rates using both the default priors as well as both overall and bias priors estimated with a MCMC analysis.

## Results

To identify the phylogenetic clustering of functional traits in microorganisms, we analyzed the phylogenetic distribution of 89 functional traits across the Bacteria and Archaea using both genotypic and phenotypic data. Both data set contained a high proportion of Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes but especially the genome data set included many phyla (Supplementary Figure S1). We first tested if any traits were randomly distributed—potentially due to gene loss,



lateral gene transfer or convergent evolution. We found that 93% (83 out of 89) of the traits were non-randomly distributed, as predicted by consenTRAIT ( $P < 0.05$ ) and the phylogenetic dispersion of discrete traits metric ( $P(D)_{\text{random}} < 0.5$ ) (Figure 2).

Given that most of the traits were non-randomly distributed, we next used consenTRAIT to identify if  $\tau_D$  varied among the traits. We found that  $\tau_D$  ranged from 0.0029–0.11 rRNA sequence distance (Figure 2, Supplementary Table S1 and S2). As expected, oxygenic photosynthesis was the trait with the highest  $\tau_D$  (median 0.11, mean 0.10) and only occurred in one clade—the phylum Cyanobacteria (Figure 3a). Methanogenesis was also found in only a few, deep clades. Sulfate reduction was present in 20 clusters with a  $\tau_D = 0.039$  16S rRNA distance.

These deeply clustered traits were the exceptions, however. The vast majority of traits examined were found in clusters of a depth less than 0.02 rRNA sequence distance. For example, nitrogen fixation was found in 54 clusters (Figure 3a) with a more intermediate level  $\tau_D$  of  $\sim 0.018$  16S rRNA distance. At the finest scale, both the analysis of genomic and phenotypic data sets revealed that the ability to grow on a particular carbon source was generally seen in shallow clusters (Figures 2 and 3). In the phenotypic data set, growth of each of the 70 organic carbon sources was associated with a  $\tau_D < 0.011$  16S rRNA distance (Figure 2b and Supplementary Table S2). The genomic data largely supported this view as we here observed the potential for growth on several carbon sources in clades with  $\tau_D < 0.017$  16S rRNA distance (Figure 2a). We also found many traits located in clusters of one, in which no neighbors shared the trait. This finding suggests that we have still significantly undersampled microbial genomics and phenotypic diversity (assuming at least some clonal replication of the genome).

Next, we tested if trait complexity was correlated with the phylogenetic dispersion and depth of microbial traits. For the genomic data set, the number of genes underlying a trait was used as an indicator of trait biochemical complexity. We observed a significant relationship between trait complexity and phylogenetic dispersion (Figure 4a). Complex traits encompassing many genes like oxygenic photosynthesis and methanogenesis were extremely clumped ( $D < 0$ ) (Figure 4a and Supplementary Table S2). Other traits including nitrogen and  $\text{CO}_2$  fixation, anoxygenic photosynthesis and sulfate reduction displayed a clumped distribution consistent with a Brownian motion model of evolution ( $D \sim 0$ ). Finally, traits like carbon substrate utilization were dispersed in a mode between a Brownian motion and a random model ( $0 < D < 1$ ), suggesting that the ability to grow on different carbon compounds is quite dispersed ( $D_{\text{mean}} = 0.48$ ,  $N = 71$ , Supplementary Table S2). Further, we observed that trait complexity was significantly correlated to  $\tau_D$  for both the Pearson and Spearman correlation tests (Figure 4b). Thus,

traits encoded by many genes were shared among deeper clades than traits encoded by fewer genes. The only clear exception to this trend was aerobic methane oxidation, which was associated with a few deep clades despite requiring only three genes.

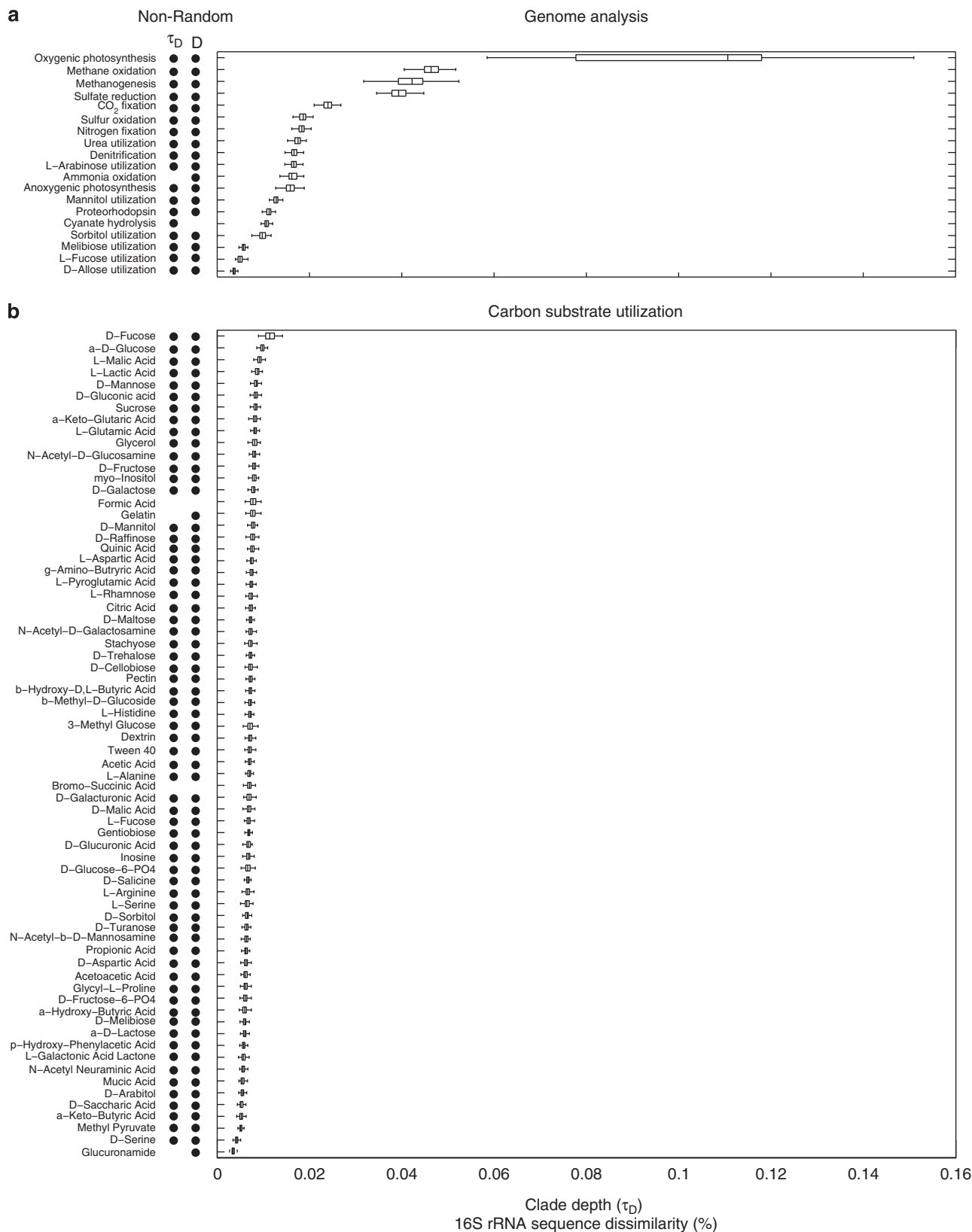
For the phenotypic carbon substrate utilization data set, we used the Bertz molecular complexity index of the carbon substrate as a measure of trait complexity (Bertz, 1981). The Bertz complexity index takes into account not only the compound size but also the bond types and overall structure. We found that molecular complexity was significantly correlated with phylogenetic dispersion ( $R^2 = 0.07$ ,  $P \sim 0.03$ )—albeit with a low degree of variance explained. In contrast, we did not find a significant relationship between  $\tau_D$  and complexity of substrate (Supplementary Table S2), likely due to low variability in  $\tau_D$  (all  $\tau_D < 0.011$  16S rRNA distance).

As one should expect, the overall variation in dispersion and  $\tau_D$  among all traits was significantly driven by the number of trait gain and loss events as estimated using a maximum parsimony approach ( $P < 0.0001$ , Pearson correlation). In contrast, when we measured the rate of trait changes (that is, the character evolutionary rate) across the tree with a Bayesian approach, we did not observe any correlation.

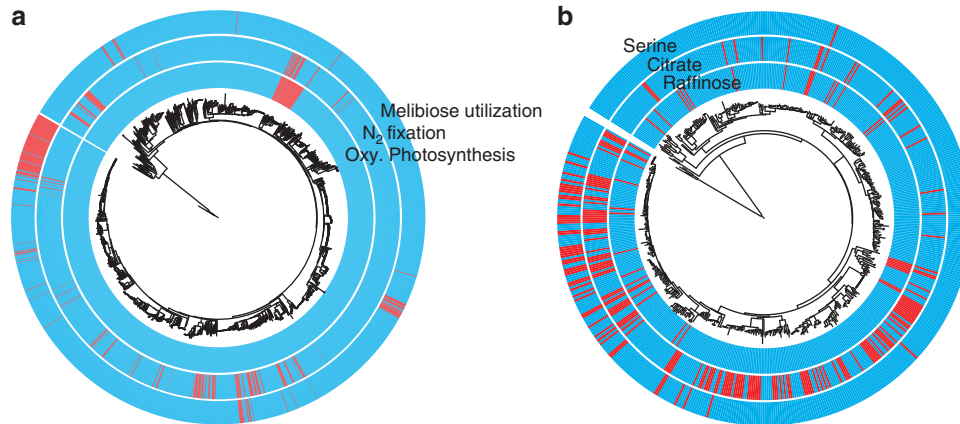
We also evaluated the influence of the phylogenetic technique on both the Purvis and Fritz's D and consenTRAIT by comparing the values based on a DNA distance and a maximum likelihood method tree. The values were highly correlated for the two phylogenetic techniques ( $R^2 > 0.9$ ,  $P < 1 \times 10^{-20}$ ) and both supported a correlation between trait complexity and either phylogenetic dispersion or  $\tau_D$ .

## Discussion

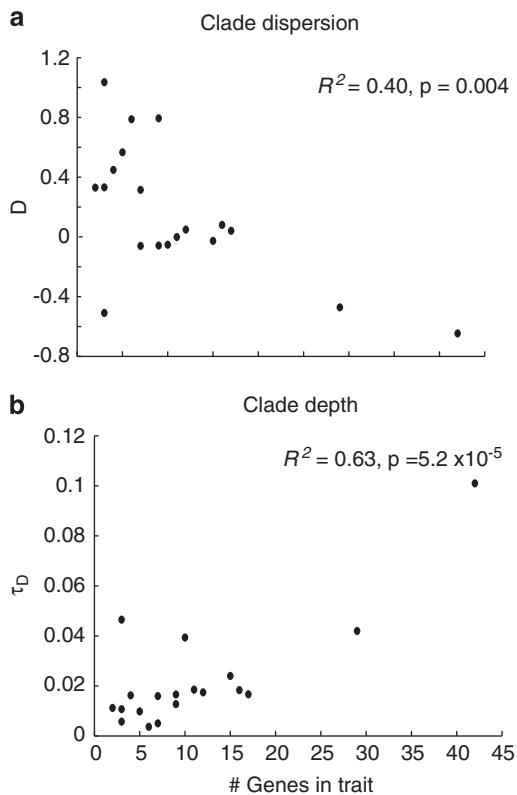
In this study, we aim to identify the distribution and degree of phylogenetic clustering of microbial traits. To address this, we developed a new algorithm—consenTRAIT. In contrast to past metrics testing for a phylogenetic conservation of traits, the aim of consenTRAIT is to estimate the phylogenetic depth of clades where members share a trait. Furthermore, it can identify the actual sequence divergence defining ecologically coherent units (at least in respect to one trait) rather than the taxonomic level associated with a trait. The latter has been done in some recent studies of plants due to the lack of a detailed phylogenetic tree (Prinzinger *et al.*, 2001; Swenson and Enquist, 2009). For microorganisms, however, there is an abundance of sequence data and organisms are mostly classified based on phylogeny and this can be used directly for the consenTRAIT analysis. This is important as it allows for a direct comparison with phylogenetic trees and can guide the threshold for defining operational taxonomic units (OTU). Instead of requiring that all



**Figure 2**  $\tau_D$  of functional traits.  $\tau_D$  is estimated using consenTRAIT as the average 16S rRNA sequence distance between members of a clade where at least 90% of the strains carry a trait and the root node of this clade. **(a)** Traits identified based on genomic subsystems. **(b)** Phenotypic traits identified based on the ability to used specific organic carbon substrates. Black dots denote non-random phylogenetic distribution based on either consenTRAIT  $\tau_D$  ( $P < 0.05$ ) or the phylogenetic dispersion test for discrete traits D,  $(P(D)_{\text{random}} < 0.5)$  (Fritz and Purvis, 2010). The box plot represents the values from 100 bootstrap trees, where the box includes values from the 25–75 percentile, the bars includes the 5–95 percentile and the line represents the median.



**Figure 3** Phylogenetic distribution of functional traits (red lines) in Prokaryotes. (a) Phylogenetic distribution of the traits oxygenic photosynthesis, nitrogen fixation and utilization of melibiose based on annotated genomic subsystems in the SEED database. (b) Phylogenetic distribution of the traits raffinose, citrate, and serine utilization based on observed growth in Biolog substrate utilization plates. The phylogenetic trees are based on a 16S rRNA alignment from the Silva database and estimated in PhyIip using a distance based matrix (F84 correction), neighbor-joining, and 100 bootstraps (Felsenstein, 2006).



**Figure 4** Role of complexity on phylogenetic dispersion and  $\tau_D$ . (a) Relationship between the number of genes underlying traits and phylogenetic dispersal (Fritz and Purvis, 2010). (b) Relationship between number of genes underlying traits and  $\tau_D$ . The correlation coefficients were based on Spearman correlation, but Pearson correlations were also significant ( $P < 0.05$ ).

members in clade share a trait, we decided to use 90% membership threshold to account for unusual genome configurations for example, a cyanobacterium losing the ability to perform oxygenic photosynthesis (Zehr *et al.*, 2008). It is worth noting that

increasing the % threshold value for membership of a clade will commonly result in smaller  $\tau_D$  values, whereas a decrease will have the opposite effect.

We have deliberately not discussed the average membership number in each clade sharing a trait as we expect this number to be directly linked to sampling effort. In other words, if a group of very closely related organisms has been sequenced or typed with Biolog, then we should expect to find a high number of members in a clade sharing a trait. An example of this could be *E. coli*. On the other hand, in phyla with few analyzed strains, we will likely find clades of one with a specific trait.

We find that more than 93% of the traits are significantly non-randomly distributed (Figure 2), suggesting that even in the presence of many well-documented examples of lateral gene transfer, vertical inheritance is generally important for the distribution of functional traits in microorganisms. The traits considered are distributed in phylogenetic clusters with a continuum of depths. Thus, diversity at all phylogenetic levels is relevant for the functioning of a microbial community. Complex functions like photosynthesis, methanogenesis, and sulfate reduction are more phylogenetically conserved and found in a few deep clades. Previous studies found that habitat association, pH and soil moisture optimum are also associated with deep clades (Kowalchuk and Stephen, 2001; Fierer *et al.*, 2007; Philippot *et al.*, 2010; Lennon *et al.*, 2012; Placella *et al.*, 2012). Thus, there is a clear evidence for specific traits associated with deeper clades.

In contrast, most traits we examine arose or disappeared more recently—either by mutation or gene gain and loss—and are shared among a few closely related organisms. This includes finding many traits in individual lineages with no sister taxa sharing the trait. This suggests that we have still undersampled microbial diversity, and that many traits might be associated with even finer-scale

diversity. Thus, our estimate for  $\tau_D$  should be considered an upper bound for many functions. In particular, traits associated with the assimilation of small organic carbon sources are highly phylogenetically dispersed. This result is consistent with observations of wide variation among the strains of *E.coli*, *Acinetobacter*, *Vibrio*, etc., in their ability to use particular carbon substrates (Sarma *et al.*, 2004; Keymer *et al.*, 2007; Vieira *et al.*, 2011). Previous studies have identified other phylogenetically dispersed bacterial traits like nutrient acquisition (Martiny *et al.*, 2006) and particle colonization (Hunt *et al.*, 2008). Here, these traits are also associated clusters of a small  $\tau_D$ .

As one would expect,  $\tau_D$  of each trait is significant correlated to phylogenetic dispersion and the number of gene gains and losses. Thus, the consenTRAIT metric provides a similar ranking of phylogenetic conservatism as other approaches. However, consenTRAIT also identifies the phylogenetic level where one should expect to find ecologically coherent clusters. Thus, our approach and results have significant implications for how to analyze and interpret microbial diversity patterns. To understand, which biotic or abiotic factors control the distribution and functioning of microbial diversity, an OTU is usually defined arbitrarily. Defining an OTU can be done either explicitly by grouping lineages using a sequence similarity of 97% to cluster groups or implicitly by linking an OTU to a peak with techniques like ARISA (automated rRNA intergenic spacer analysis), T-RFLP (terminal-restriction fragment length polymorphisms), or phospholipid fatty acid profiles. Thus, most techniques rely on a fixed definition of an OTU but the advantage of consenTRAIT is that this algorithm identifies the average clade depth where organisms share a trait. Thus, it enables you to evaluate if a microbial taxon is 'defined' at a sequence similarity level that is meaningful in comparison with the phylogenetic dispersion of relevant traits. This may not always be at the phylogenetic level whereby standard techniques define taxa. It is, however, important to recognize that the consenTRAIT metric estimates the phylogenetic distance between the tips and the last common ancestor of a clade with a given trait, whereas clustering algorithms like MOTHR is based on the inter-lineage distance (Schloss *et al.*, 2009). For example, if a clade only contains two members with 3% sequence dissimilarity, the average distance between each tip and the clade root node would be 0.015. Thus, a trait with  $\tau_D = 0.015$  will be expected to show most variation among OTUs defined by 97% sequence similarity. In our study, we identify many traits associated with  $\tau_D < 0.015$  including the utilization of many small carbon molecules, whereas functions related to photosynthesis or central functions in the nitrogen or sulfur cycle are found in clusters above this threshold. Thus, OTUs will contain a mixture of lineages with and without most small carbon

molecule usage traits if defined using a 97% 16S rRNA sequence similarity cutoff. A result of this is that studies that aim to investigate how microbial diversity varies among gradients of these kinds of carbon substrates should only expect to observe patterns using a much finer genetic resolution. Similarly, traits associated with broad clades may not come out significant if your taxa are defined at a much finer level. Our study suggests that diversity at a range of phylogenetic levels can be relevant for the functioning of the community, and different functions are related to different phylogenetic cluster sizes. Thus, broad changes in the taxonomic composition of a community might influence the functioning of traits associated with deep clusters, but unlikely to have a big effect on traits associated with shallow microdiverse clusters and vice versa. To account for this, we need a flexible classification of microbial taxa to understand the distribution and functioning of the full spectrum of phylogenetic diversity.

We hypothesized that genetic complexity of a trait could possibly influence the trait dispersion and depth. We find support for this relationship in the genomic data set examined even though many other factors beyond complexity-likely influence the phylogenetic distribution of traits (for example, selective pressures and interactions between traits). The association of specific functional traits with different phylogenetic levels has previously been described in microorganisms. For example, the marine cyanobacterium *Prochlorococcus* can broadly be divided into two groups based on their ability to grow under high and low light levels (Moore *et al.*, 1998). Light adaptation is likely a complex trait as it involves many interaction proteins. The high-light adapted clade can be further divided into a low and high iron group (Rusch *et al.*, 2010), and the high iron group can be subdivided into high and low temperature adapted cells (Johnson *et al.*, 2006). Finally, nitrate and phosphate assimilation genes are associated with shallow clades (Martiny *et al.*, 2009a,b). Nutrient uptake capabilities are possibly more simple traits as they involve clusters of a few genes. Thus for *Prochlorococcus*, specific traits are clearly dispersed differently phylogenetically, and this variation might be linked to difference in the complexity of the trait. The variation is not exclusive to microorganisms. Prinzing and co-workers find that among plants in Europe, tolerance of extremes in soil moisture was associated with higher taxonomic levels than were those for temperature or light (Prinzing *et al.*, 2001). Similarly, tree height and seed mass are linked to broader phylogenetic groups in comparison with other functional traits in trees (Swenson and Enquist, 2009).

The linkage between traits and complexity offers a framework for predicting the phylogenetic distribution of specific traits as well as the resilience and robustness of specific microbe-mediated ecosystem



processes to environmental changes. If the traits are distributed among few phylogenetic clades, shifts in community composition may strongly alter the associated ecosystem process (Schimel *et al.*, 2005). In contrast, communities may be resilient to changes if a trait is distributed among many groups. However, there are potentially millions of microbial taxa (Torsvik *et al.*, 1990) and a multitude of important traits, so it is likely impossible to map all relevant traits to all phylogenetic lineages of microorganisms. Our findings offers a way to begin to estimate how broad or narrow phylogenetically distributed a trait is based on the genetic and biochemical complexity of the function. This can help to predict how changes in microbial biodiversity may affect ecosystem functioning and how resilient a community is to a particular perturbation. Considering the important role microorganisms has in many ecosystem processes, this understanding is critical for predicting future changes in global biogeochemical cycles.

## Conflict of Interest

The authors declare no conflict of interest.

## Acknowledgements

We thank Steven Allison, Simon Levin, Jennifer Martiny, Ross Overbeek, and Martin Polz for many helpful comments on the manuscript and Anthony Amend for suggesting the consenTRAIT name. This work was supported in part by the National Science Foundation–Dimensions of Biodiversity program (ACM) and the US Department of Energy, Office of Science (BER) GtL program (ACM and KT).

## References

- Allison SD, Martiny JB. (2008). Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci USA* **105**: 11512–11519.
- Bertz SH. (1981). The first general index of molecular complexity. *J Am Chem Soc* **103**: 3599–3601.
- Bollback JP. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**: 88.
- Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbo CL *et al.* (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37**: 283–328.
- Cadotte MW, Cardinale BJ, Oakley TH. (2008). Evolutionary history and the effect of biodiversity on plant productivity. *Proc Natl Acad Sci USA* **105**: 17012–17017.
- Cavender-Bares J, Kozak KH, Fine PV, Kembel SW. (2009). The merging of community ecology and phylogenetic biology. *Ecol Lett* **12**: 693–715.
- Csuros M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**: 1910–1912.
- Disz T, Akhter S, Cuevas D, Olson R, Overbeek R, Vonstein V *et al.* (2010). Accessing the SEED genome databases via Web services API: tools for programmers. *BMC Bioinformatics* **11**: 319.
- Doolittle WF. (1999). Phylogenetic classification and the universal tree. *Science* **284**: 2124–2129.
- Felsenstein J. (2006). *PHYLIP (Phylogeny Inference Package)*, 3. 65 edn. Department of Genome Sciences, University of Washington: Seattle.
- Fierer N, Bradford MA, Jackson RB. (2007). Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354–1364.
- Fritz SA, Purvis A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv Biol* **24**: 1042–1051.
- Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* **103**: 13104–13109.
- Gravel D, Bell T, Barbera C, Bouvier T, Pommier T, Venail P *et al.* (2011). Experimental niche evolution alters the strength of the diversity-productivity relationship. *Nature* **469**: 89–U1601.
- Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ. (2004). A taxa-area relationship for bacteria. *Nature* **432**: 750–753.
- Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.
- Jain R, Rivera MC, Lake JA. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**: 3801–3806.
- Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.
- Jombart T, Balloux F, Dray S. (2010). adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* **26**: 1907–1909.
- Keymer DP, Miller MC, Schoolnik GK, Boehm AB. (2007). Genomic and phenotypic diversity of coastal *Vibrio cholerae* strains is linked to environmental factors. *Appl Environ Microbiol* **73**: 3705–3714.
- Kowalchuk GA, Stephen JR. (2001). Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annu Rev Microbiol* **55**: 485–529.
- Lennon JT, Aanderud ZT, Lehmkuhl BK, Schoolmaster DR. (2012). Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology* **93**: 1867–1879.
- Lilliefors HW. (1967). On Kolmogorov-Smirnov test for normality with mean and variance unknown. *J Am Stat Assoc* **62**: 399–402.
- Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.
- Martiny AC, Huang Y, Li WZ. (2009a). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340–1347.
- Martiny AC, Kathuria S, Berube PM. (2009b). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci USA* **106**: 10787–10792.
- Moore LR, Rocap G, Chisholm SW. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**: 464–467.
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M *et al.* (2005). The subsystems approach



- to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* **96**: 2896–2901.
- Pagel M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.
- Pal C, Papp B, Lercher MJ. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**: 1372–1375.
- Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.
- Philippot L, Andersson SG, Battin TJ, Prosser JI, Schimel JP, Whitman WB *et al*. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* **8**: 523–529.
- Placella SA, Brodie EL, Firestone MK. (2012). Rainfall-induced carbon dioxide pulses result from sequential resuscitation of phylogenetically clustered microbial groups. *Proc Natl Acad Sci USA* **109**: 10931–10936.
- Prinzing A, Durka W, Klotz S, Brandl R. (2001). The niche of higher plants: evidence for phylogenetic conservatism. *Proc R Soc Lond, Ser B: Biol Sci* **268**: 2383–2389.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al*. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.
- Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* **107**: 16184–16189.
- Sarma PM, Bhattacharya D, Krishnan S, Lal B. (2004). Assessment of intra-species diversity among strains of *Acinetobacter baumannii* isolated from sites contaminated with petroleum hydrocarbons. *Can J Microbiol* **50**: 405–414.
- Schimel JP, Bennett J, Fierer N. (2005). Microbial community composition and soil N cycling: Is there really a connection? In: Bardgett RD, Hopkins DW, Usher MB (eds). *Biological Diversity and Function in Soils*. Cambridge University Press: Cambridge, pp 171–188.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al*. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
- Shi T, Bibby TS, Jiang L, Irwin AJ, Falkowski PG. (2005). Protein interactions limit the rate of evolution of photosynthetic genes in Cyanobacteria. *Mol Biol Evol* **22**: 2179–2189.
- Snel B, Bork P, Huynen MA. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**: 17–25.
- Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M *et al*. (2007). PATRIC: the VBI pathoSystems resource integration Center. *Nucleic Acids Res* **35**: D401–D406.
- Stamatakis A, Hoover P, Rougemont J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* **57**: 758–771.
- Swenson NG, Enquist BJ. (2009). Opposing assembly mechanisms in a Neotropical dry forest: implications for phylogenetic and functional community ecology. *Ecology* **90**: 2161–2170.
- Swenson NG, Erickson DL, Mi XC, Bourg NA, Forero-Montana J, Ge XJ *et al*. (2012). Phylogenetic and functional alpha and beta diversity in temperate and tropical tree communities. *Ecology* **93**: S112–S125.
- Torsvik V, Goksoyr J, Daae FL. (1990). High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**: 782–787.
- Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fevre F, Mornico D *et al*. (2011). Core and panmetabolism in *Escherichia coli*. *J Bacteriol* **193**: 1461–1472.
- Violle C, Nemergut DR, Pu Z, Jiang L. (2011). Phylogenetic limiting similarity and competitive exclusion. *Ecol Lett* **14**: 782–787.
- Webb CO, Ackerly DD, McPeck MA, Donoghue MJ. (2002). Phylogenies and community ecology. *Annu Rev Ecol Syst* **33**: 475–505.
- Wellner A, Lurie MN, Gophna U. (2007). Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol* **8**: R156.
- Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T *et al*. (2008). Globally distributed uncultivated oceanic N<sub>2</sub>-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**: 1110–1112.

Supplementary Information accompanies the paper on The ISME Journal website (<http://www.nature.com/ismej>)