The ISME Journal (2015) 9, 2413–2422
© 2015 International Society for Microbial Ecology All rights reserved 1751-7362/15
www.nature.com/ismej

ORIGINAL ARTICLE

Local–global overlap in diversity informs mechanisms of bacterial biogeography

Joshua A Livermore[1] and Stuart E Jones[1,2]
[1]Notre Dame Environmental Change Initiative, Notre Dame, IN, USA and [2]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

Spatial variation in environmental conditions and barriers to organism movement are thought to be important factors for generating endemic species, thus enhancing global diversity. Recent microbial ecology research suggested that the entire diversity of bacteria in the global oceans could be recovered at a single site, thus inferring a lack of bacterial endemism. We argue this is not the case in the global ocean, but might be in other bacterial ecosystems with higher dispersal rates and lower global diversity, like the human gut. We quantified the degree to which local and global bacterial diversity overlap in a diverse set of ecosystems. Upon comparison of observed local–global diversity overlap with predictions from a neutral biogeography model, human-associated microbiomes (gut, skin, mouth) behaved much closer to neutral expectations whereas soil, lake and marine communities deviated strongly from the neutral expectations. This is likely a result of differences in dispersal rate among 'patches', global diversity of these systems, and local densities of bacterial cells. It appears that overlap of local and global bacterial diversity is surprisingly large (but likely not one-hundred percent), and most importantly this overlap appears to be predictable based upon traditional biogeographic parameters like community size, global diversity, inter-patch environmental heterogeneity and patch connectivity.

The ISME Journal (2015) **9,** 2413–2422; doi:10.1038/ismej.2015.51; published online 7 April 2015

Introduction

Macroorganism species distributions are frequently constrained by geographic barriers like oceans and mountain ranges. Consequently, comparable environments on different continents can have non-overlapping biological assemblages owing to dispersal limitation (i.e., endemisim; Cifelli, 1993; Melville *et al.*, 2006). In contrast, microorganisms appear to disperse globally and rapidly in air and water currents (Caporaso *et al.*, 2012; Hanson *et al.*, 2012; Gibbons *et al.*, 2013). In the absence of geographic barriers, we expect transcontinental microbial community assemblages from similar environments to have overlapping taxa. Two pieces of evidence supporting widespread dispersal, and therefore reduced endemism, in microbial communities include the global distribution of some bacterial groups, for example, freshwater *Polynucleobacter cosmopolitanus* (Jezberová *et al.*, 2010; Jezbera *et al.*, 2011), and strong correlations between local microbial community composition and local environmental conditions, rather than geographic location (Fierer and Jackson, 2006; Newton *et al.*, 2007; Lauber *et al.*, 2009).

Despite the apparent absence of dispersal limitation, observation error and stochastic aspects of the community assembly process ensure we would not detect exactly the same set of microbial species in, even identical environments. However, we would expect environmental selection to further reduce the fraction of global diversity present in a particular location (Martiny *et al.*, 2006; Hanson *et al.*, 2012). The expectation for dispersal and environmental selection to generate emergent patterns in the fraction of global diversity found in a given local environment suggests this quantity might serve as a useful biogeographic metric. As deep DNA sequencing becomes increasingly feasible, so too does our ability to generate a complete census of local composition and diversity, as well as to quantify the fraction of global microbial diversity present at any given place and time. A recent demonstration of this approach recovered nearly 40% of the total phylogenetic richness from the global ocean in a single sample from the English Channel sampled to a depth of 10 million 16 S rRNA gene sequences (Caporaso *et al.*, 2012; Gibbons *et al.*, 2013). Albeit from a single sample, this local–global diversity comparison provides clear evidence of extensive microbial dispersal in the ocean system. In fact, the authors inferred a complete lack of dispersal

Correspondence: SE Jones, Department of Biological Sciences, University of Notre Dame, 264 Galvin Life Sciences, Notre Dame, IN 46556, USA.
E-mail: sjones20@nd.edu

Received 2 September 2014; revised 17 February 2015; accepted 1 March 2015; published online 7 April 2015

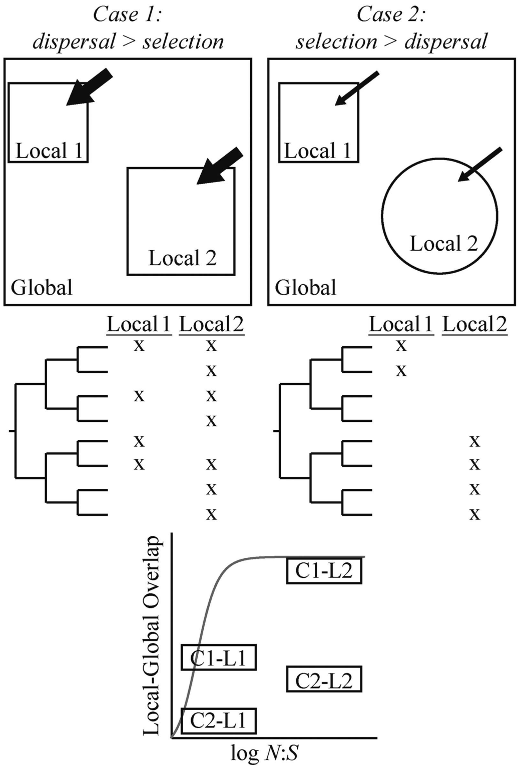

Figure 1 A conceptual figure of local–global overlap of phylogenetic diversity. Local patches receive dispersing individuals from the global species pool with richness *S*, and contain *N* individuals. In the figure, arrow size is proportional to immigration rate and patch size is proportional to N. In addition, local environmental heterogeneity (local patch shape) can impose environmental selection. The relative importance of dispersal and selection (Case 1 vs Case 2) dictate what fraction of global phylogenetic diversity is represented in a given local patch (x's indicate presence in middle panels). Our neutral model (gray line in bottom panel) generates a saturating curve of local–global overlap in phylogenetic diversity as a function of *N*:*S*. The position of observed local–global overlap (boxes labeled with case number—local patch number in lower panel) are estimated from the fraction of the global phylogeny represented in any given local patch, and may allow inference about the relative importance of dispersal vs environmental selection (Case 1 vs Case 2).

limitation, and microbial endemism, when extrapolating a log–linear rarefaction of their data to generate an estimate of the sampling depth required to achieve 100% recovery of the global ocean's bacterial richness in the English Channel sample.

Previous work comparing local and regional richness in non-microbial systems generated much debate in the ecological literature and demonstrated the utility of comparing measures of diversity at more than one spatial scale (Cornell and Lawton, 1992; Schoolmaster, 2001; Fox and Srivastava, 2006). These local vs regional richness comparisons were accomplished by regressing local richness on regional richness with the expectation that a linear relationship indicated regional control of local species richness and a non-linear relationship suggested saturation of patch richness as a result of local ecological interactions. Although this work highlighted the potential importance of regional- to global-scale processes for local community assembly and diversity, the legitimacy of the original local vs regional regression methodology has been thoroughly discredited (Hillebrand and Blenckner, 2002; Hillebrand, 2005; Szava-Kovats *et al.*, 2012). These methodological concerns have caused many to abandon comparisons of local and regional diversity, but we argue the utility of the concept should not be conflated with methodological weaknesses. Others appear to agree with our view, given recent augmentation of the local vs regional regression approach to overcome its deficiencies (Szava-Kovats *et al.*, 2012; Gonçalves-Souza *et al.*, 2013).

A recent review of microbial biogeography called for a refocusing of microbial ecologists' efforts towards a deeper understanding of microbial biogeography process rather than continued pattern identification (Hanson *et al.*, 2012). For example Hanson *et al.* (2012), highlight the inability of distance-decay analyses to evaluate the relative importance of selection, drift, dispersal and mutation. The fraction of global diversity found in a local patch (local–global overlap) appears to be a potentially useful metric for biogeographic pattern identification (Gibbons *et al.*, 2013), but it also suffers from an inability to infer anything about process. One potential advantage of local–global overlap is that it can be quantified for a single site or sample, in contrast to the pairwise nature of distance-decay metrics, allowing relatively straightforward generation of theory-based, quantitative expectations for local–global overlap. Here, we ask whether the comparison of observed local–global overlap with the expectations from a neutral biogeography model can reveal the relative importance of dispersal and environmental selection for bacterial community assembly, and avoid the issues associated with the local vs regional regression approach (Figure 1). Although selection, drift, dispersal and mutation are all important biogeographic mechanisms, we focus on environmental selection and dispersal as these have been the primary focus of past microbial biogeography and meta-community research (Martiny *et al.*, 2006; Logue and Lindström, 2008; Jones and McMahon, 2009).

Under a neutral model, the extent of local–global overlap is dictated by dispersal, global diversity (both species richness, *S*, and species abundance distributions, SAD), and the number of individuals in the local patch (*N*), which determines how deeply the local community 'samples' global diversity. Therefore, given some knowledge of *S*, the global SAD, and *N*, a neutral expectation for local–global overlap can be generated and compared with the observations of overlap in species composition or phylogenetic diversity based upon marker gene surveys. If observations closely resemble predictions from a neutral model, dispersal is relatively strong in comparison with local selection (Figure 1—Case 1). In contrast, strong deviation of observations of local–global overlap from neutral expectations indicates local selection is more important than dispersal in driving the local community assembly (Figure 1—Case 2).

The goal of this study was to evaluate the proposed use of a neutral biogeography model and observations of local–global overlap to elucidate the relative importance of dispersal and environmental selection for microbial community assembly. To estimate local–global overlap, we combined existing deep-sequenced *local* data sets (Caporaso *et al.*, 2011) with broadly distributed *global* survey data sets from a diverse set of ecosystems, including human-associated, soil and aquatic environments. These ecosystems varied in the ratio of number of individuals in a local sample to global species richness (*N:S*), and therefore had different expected levels of local–global overlap under our neutral biogeography model. By testing whether observed levels of local–global overlap deviated strongly from neutral expectations, we evaluated our ability to infer the relative importance of dispersal and selection in these diverse ecosystem types. Although previous work has shown individual host-specific community composition in human-associated bacterial communities (Benson *et al.*, 2010; Fierer *et al.*, 2010), we hypothesized that local–global overlap for human microbiomes would more closely correspond to neutral predictions because of higher dispersal amongst patches and/or reduced inter-patch environmental heterogeneity, which could diminish the strength of environmental selection.

Materials and methods

Data sources and local–global overlap metric

To generate estimates of local–global overlap, we required both deeply sequenced samples from a single community at a single time point ('local samples') and a set of samples from spatially distributed patches from the same ecosystem ('global samples'). The local data sets we used came from a single study (Caporaso *et al.*, 2011). Six of the eight ecosystems sampled by Caporaso et al. (2011) had suitable global data sets, including human gut, tongue and skin; ocean, freshwater lake and soil. Available global data sets varied significantly in their sampling effort, and as a result we randomly downsampled the more deeply sampled data sets (Supplementary Table S1). We obtained these data sets primarily from the QiimeDB database and Earth Microbiome Project Databases (Supplementary Table S1). We selected phylogenetic overlap, the fraction of branch length in a phylogenetic tree of all global diversity recovered in a local sample, as our metric of local–global overlap because this metric was previously used in this context (Gibbons *et al.*, 2013). In addition, much recent work has highlighted the utility of including phylogenetic relationships in metrics of diversity (Helmus *et al.*, 2007; Cavender-Bares *et al.*, 2009; Kembel *et al.*, 2010). For brevity, we will refer to the fraction of global phylogenetic diversity present in a local sample as 'local–global overlap' throughout the remainder of the paper.

Neutral model of local–global overlap

We selected Hubbell's neutral model (Hubbell, 2001) to generate expectations for local–global overlap in microbial communities. Because we expected the ratio of local community size to global diversity to be a key regulator of local–global overlap in diversity, we systematically varied this ratio across model simulations by varying global diversity with a constant local number of individuals ($n = 1000$). Simulations were run for 1000 individual deaths after which time local communities had always reached a dynamic equilibrium of richness. Following each death of a local individual, the parameter m dictated whether local reproduction or immigration replaced the local community member. Local reproduction was proportional to local population abundances, and immigrating individuals came from a global community with richness (S) and an empirically derived, lognormal SAD. A custom function generated in the R Statistical Environment was used for these simulations (R Development Core Team, 2010). To estimate phylogenetic overlap of the local and global communities, we generated random bifurcating phylogenetic trees containing all members of the global, and therefore local, community with the rtree() function from the ape package of the R Statistical Environment (Paradis *et al.*, 2004). Local community composition (presence–absence) and the random phylogenetic tree were used as inputs to the phylosor() function in the R package Picante to estimate local–global overlap for each model simulation (Kembel *et al.*, 2010). Model results for local–global overlap based solely on OTU identity, ignoring phylogenetic relationships, were extremely similar to those using phylogenetic diversity. Ten replicate simulations were run at eight global diversities ranging from 10 to 50 000 species, which when combined with a local community size of 1000 individuals spans the range of observed *N:S* for available empirical data. All R code used to generate our local–global overlap expectations is available in Supplementary Materials.

To identify likely values for the parameters used (*N:S*, μ_{SAD} and σ_{SAD}) in our neutral biogeography model, we used our global sequence data sets (see Data Sources above) and estimates of cell abundances (n, Supplementary Table S2) in microbial biomass samples from the literature (Supplementary Table S2). We chose to use the scale of an individual sample as the spatial extent of a local patch because that is the unit of observation in DNA-sequence-based surveys. Ideally the definition of a local patch would be informed by knowledge of the spatiotemporal scales of interaction between bacterial cells and populations, but this environmental microbiology research frontier is only in its infancy (Vos *et al.*, 2013;

Lushi *et al.*, 2014; Rusconi *et al.*, 2014). Global community richness (S) was determined by maximum likelihood estimation of the asymptote for rarefaction of OTU richness from available spatial surveys (Supplementary Table S1). The mean and standard deviation of the global lognormal SAD (μ_{SAD} and σ_{SAD}) were taken as the mean of those parameters estimated by maximum likelihood for each of the deeply sequenced samples we used.

Identifying the appropriate model for asymptotic local–global overlap

We can know exactly what the local–global overlap in model simulations is, but empirical estimates are subject to issues of sampling. As such, we must use rarefaction to estimate the asymptotic local–global overlap. A previous attempt to quantify local–global overlap used a linear regression model against the misleadingly line-like plot of local–global overlap against the $\log_{10}$ of rarefaction depth (Equation 1; Gibbons *et al.*, 2013). When log transformed a number of rarefaction depths are compressed near the origin and it is difficult to observe systematic bias in the model fit. When the model predictions and residuals are viewed on an arithmetic scale, it is clear that the log–linear model generates non-random errors that make interpolation or extrapolation of the model problematic (Supplementary Figure S1). Fitting such a model without statistically evaluating its suitability yielded the conclusion that sequencing to a depth of 1.93×10^{11} would recover all phylogenetic richness in the global ocean, that is, 100% local–global overlap.

$$Percent\ overlap = m \times \log_{10}(rare\,faction\ depth) + b \tag{1}$$

To ensure we were using an appropriate model for estimation of local–global overlap, we evaluated the performance of seven models, including the log–linear model from Gibbons et al. (2013), when estimating local–global overlap. The six models in addition to the log–linear are commonly used in rarefaction procedures (Jimenez-Valverde *et al.*, 2006), and model performance was evaluated using Akaike's Information Criterion (Akaike, 1974), and visual examination of residuals (Supplementary Figure S1). The most likely model parameters were estimated in R by minimizing custom negative log-likelihood functions assuming errors conform to a beta distribution (R Development Core Team, 2010). Optimal solutions were determined using 'optim' with the Nelder-Mead algorithm repeated 100 times with varied starting values. Rarefied local–global overlap values from the English Channel were kindly provided by Sean Gibbons and Jack Gilbert.

Cross-system observations of local–global overlap

After identifying the most appropriate statistical model for estimation of asymptotic local–global overlap, we estimated this proportion for microbial communities from six diverse environments, encompassing human-associated and natural microbiomes to provide context for the English Channel result (Gibbons *et al.*, 2013) and to explore potential differences in the balance between selection and dispersal in dictating local community assembly.

OTU definition and local–global overlap. To quantify which microbial operational taxonomic units (OTUs) are shared in the local and global data sets, we have to use sequence similarity to define these groups. We used the popular Qiime software for these purposes. For each environment type, OTUs were selected using the 'pick_subsampled_reference_otus_through_otu_table.py' script in Qiime (Caporaso et al., 2010b) v1.5 (all settings default, uclust similarity cutoff = 0.97) using the Greengenes (DeSantis *et al.*, 2006) v13_5 16S rRNA gene sequence database as the reference set for the first round of OTU picking, PyNAST (Caporaso *et al.*, 2010a) alignment, and alignment masking. Full or nearly full-length Greengenes reference sequences were used preferentially as representative sequences when applicable. Otherwise, Uclust (Edgar, 2010) seed sequences were used as the representative for *de novo* OTUs.

As many of the *de novo* OTU representative sequences are non-overlapping (owing to sequencing of different variable regions of the 16 S rRNA gene) we generated a phylogenetic tree using FastTreeMP (Price *et al.*, 2010) with the –pseudo option which estimates distances between non-overlapping sequences based on their transitive relationships to full-length sequences (other options: -gtr, -nt, -fastest). Previous investigation has shown that nucleotide distances between different variable regions of the rRNA gene are tightly correlated (Livermore and Mattes, 2013) which indicates that this is a reasonable approximation. The use of phylogenetic diversity and previously defined reference OTUs dramatically reduces the sensitivity of our results to this potential issue. The influence of variation in OTU assignments resulting from differences in 16 S variable region is minimized because potential misclassifications are occurring at the tips of the phylogenetic tree, and therefore by definition contribute little to the total branch length of the global phylogenetic tree. In other words, small errors in OTU assignment caused by differences in sequencing approaches across studies are down-weighted by our choice to use overlap in phylogenetic diversity.

Sequences that were not alignable by PyNAST (and thus did not appear in the phylogenetic tree) were removed from the OTU tables. Single deep-sequenced samples were rarefied (three replicates at each depth) and phylogenetic overlap was computed using beta_diversity.py in Qiime with the metric 'unifrac_G' which is Faith's phylogenetic gain (Faith, 1992) or fraction of phylogenetic tree branch length

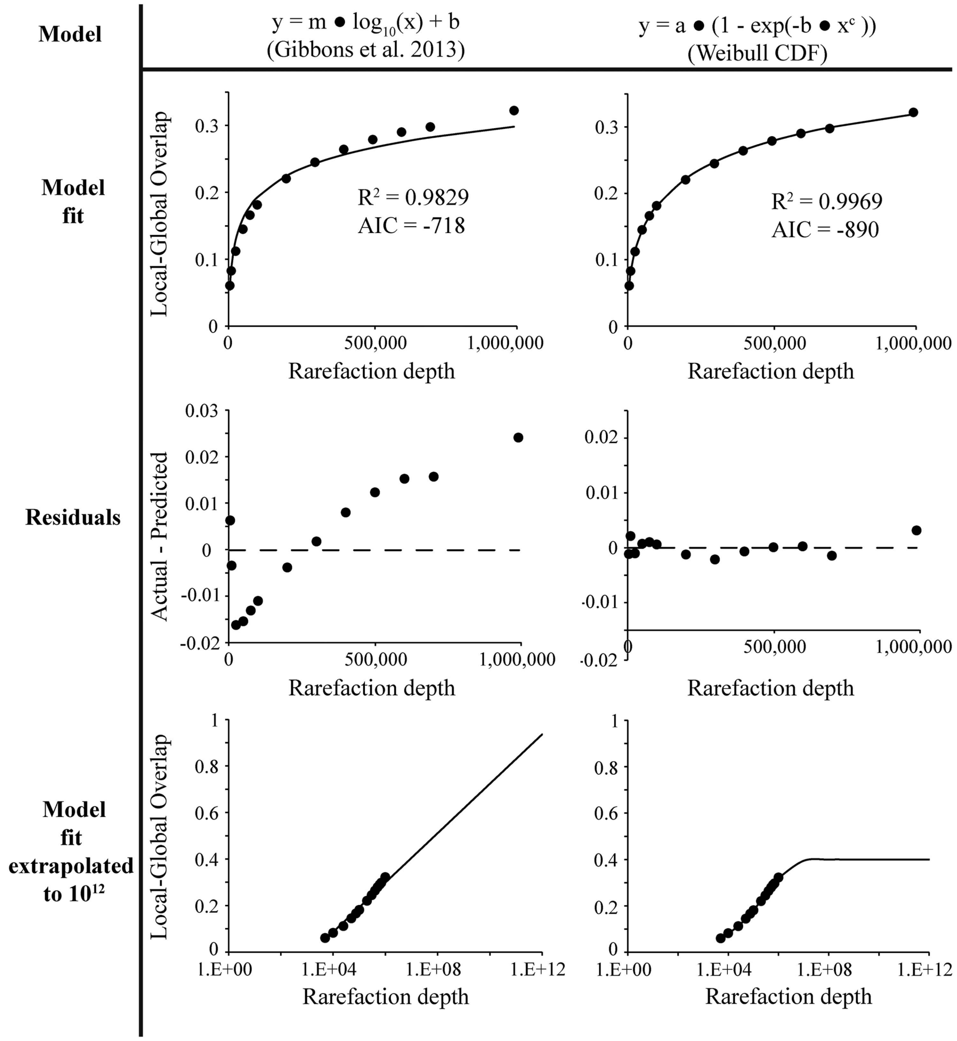

Figure 2 A log(x) linear regression was previously used to model progressively increasing local–global overlap between a single deeply sequenced sample from the English Channel and a set of many shallow samples from the global ocean as the deep sample is rarefied. A model that has previously been used to model rarefaction processes (Weibull cumulative distribution function) yields a better fit and a non-systematic error distribution compared with the log(x) regression. The central implication of this discrepancy in model selection is that the more appropriate Weibull model has an asymptote (~0.40) whereas the log(x) regression model goes to infinity. Dots on plots show data and the lines show model predictions.

unique to a sample. For each deeply sequenced sample, rarefactions were computed at 300 random depths to uniformly and non-arbitrarily sample across the entire length of the curve (Cottingham *et al.*, 2005).

Estimating global OTU richness and local density. Because the ratio of the number of individuals in the local patch to global species richness (*N:S*) structures our neutral expectations, we must estimate *S* from our global data sets. To estimate global OTU richness from each environment (*S*), we selected fully overlapping sequence sets representing diverse samples (Supplementary Table S1) and picked OTUs as described above. Sequences were rarefied, and richness estimated as asymptotic number of 97% identity OTU sequences. Typical sample sizes in terms of the number of individual bacteria for each environment (*N*) were estimated based on sampling conventions for each ecosystem (Supplementary Table S2).

Species presence–absence vs relative abundance. One simple prediction we can derive from our conceptual framework, is that compositional differences between patches in ecosystem types with high local–global overlap in diversity should be driven by differences in SAD rather than species presence–absence. To simply evaluate this prediction, we considered inter-patch, but intra-ecosystem pairwise differences in composition for all patches in the global data sets from human gut and lake ecosystems as measured with Sørensen's, presence–absence based, and Bray-Curtis, includes consideration of relative abundances, distance metrics. If the ratio of these two distance metrics (Sørensen's:Bray-Curtis) is close to one, it suggests species presence–absence drives inter-patch differences. Whereas, a ratio much below one suggests a greater importance of differences in SAD across patches. We made visual comparisons of these values by plotting frequency densities of observed ratios.

Results

Comparison of rarefaction models

Our candidate models provided widely different predictions for local–global overlap. Having said that, all candidate models with the exception of the Weibull and Chapman–Richards had extremely non-random residual distributions. We selected the Weibull over the Chapman–Richards based upon the well-accepted model selection criterion AIC (Akaike's Information Criterion). Notably, the previously used log–linear regression model exhibited non-random distribution of errors (Figure 2; Supplementary Figure S1), and resulted in much larger estimates of local–global overlap than the Weibull model. We assert that the most appropriate model for estimating asymptotic local–global overlap in diversity is the Weibull cumulative distribution function (Equation 2), and note that our findings are consistent with previous assessments of model fits for rarefaction processes (Flather, 1996; Van Rooijen, 2009).

$$Percent\ overlap = a\times(1-\exp(-b\times rare\,f\,action\ depth^{c})) \tag{2}$$

A central property of the Weibull cumulative distribution function is the presence of an upper bound (asymptote) parameterized as 'a' in the model. We interpret a as the theoretical maximum proportion of global richness of an ecosystem that can be found locally. Fitting a Weibull cumulative distribution function to the rarefied local–global overlap plot from the English Channel indicates local richness is ~40% of total ocean richness, which contradicts the assertion that with sufficient sampling 100% overlap would be observed (Gibbons *et al.*, 2013), but remains a staggering amount of local–global overlap. As a result, we recommend the use of a in the model above as an asymptotic estimate of local–global overlap.

Neutral model of local–global overlap

For the neutral simulations of local–global overlap, we estimated the parameters of a lognormal distribution describing the relative SAD of the 16 deep-sequenced samples ($\mu = -13 \pm 0.3$, $\sigma = 1.43 \pm 0.04$; mean ± s.d.). The mean and standard deviation of each SAD were strongly correlated across the 16 samples ($r = 0.82$), but the model was not sensitive to even independent changes in these parameters. The model was also relatively insensitive to the probability of an open site in the local community being replaced by an immigrant (m; Supplementary Figure S2). Simulation results were much more sensitive to the assumed form of the SAD (uniform vs lognormal), as has been shown previously (Schoolmaster, 2001). The means of these parameters were used to describe the 'global' SAD with the ratio of local density to global richness set to a range of values including empirical estimates from the six environments we consider here. Our simulations predict local–global overlap to be low at a low N:S, but rapidly rise to complete overlap as N:S increases by ~2 orders of magnitude (Figure 3b).

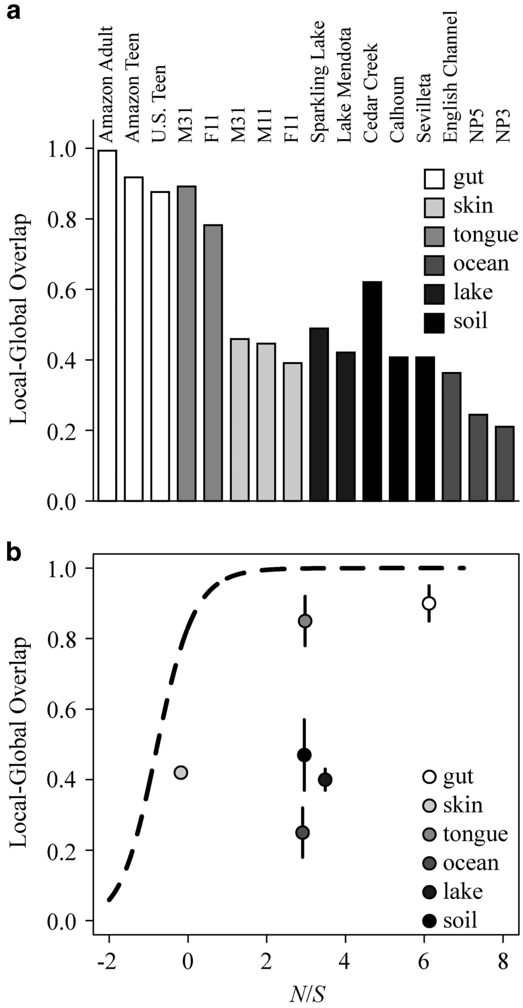

Figure 3 (**a**) Asymptotic local–global overlap (quantified as the fraction of phylogenetic diversity in many shallow sequenced samples from the same environment present in a single deeply sequenced sample) for 16 samples from 6 ecosystem types. (**b**) Observed maximum overlap is positively related to the ratio of the number of individuals sampled (N) and global environmental OTU richness (S). The dashed line is neutral theory predictions based on simulated overlap values with random sampling from a global taxa pool assuming a lognormal species abundance distribution.

Local–global overlap in diverse environments

In total, we estimated asymptotic, local–global overlap of 16 samples from 6 different ecosystems (lake, soil, marine, human gut, human skin and human tongue surface; Figure 3a). The largest apparent

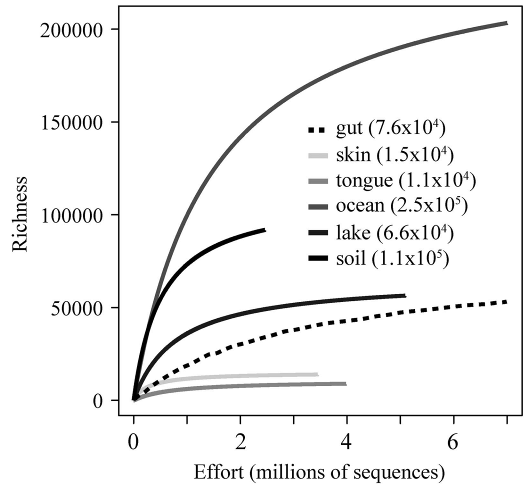

Figure 4 Rarefaction curves to estimate global 97% sequence identity OTUs richness from six environments. Estimated asymptotic richness values are in parentheses in the legend.

overlap occurred in the human gut samples (100, 91 and 87%) followed by the human tongue (90 and 78%). There was less overlap in the environmental samples with greatest observed overlap in soil samples (62, 41 and 41%), which are comparable to lake samples (49 and 42%) whereas the lowest overlap values were from the ocean samples (37, 24 and 21%). Overlap values from the human skin were more comparable with the environmental samples than other human microhabitats (46, 45 and 39%).

On the basis of our global richness estimates (Figure 4) and the number of individuals in samples (Supplementary Table S2), the *N:S* of the ecosystems we considered varied over five orders of magnitude with skin possessing low global diversity and low number of local individuals (*N:S* near 1) and the gut with tremendous local density and relatively low global diversity (*N:S* of nearly 10^6). The other four systems were fairly close with an *N:S* near 10^3. In all three cases, the human-associated systems (gut, skin and tongue) were much closer to neutral model expectations than the other three systems (marine, lake and soil).

Consistent with our predictions, gut patches, which demonstrated high local–global overlap in diversity, showed consistently greater difference based upon Bray-Curtis than Sørensen's index (a ratio much below one) suggesting an importance of differences in SAD, but not the presence or absence of species across patches (Figure 5). This difference between Sørensen's and Bray-Curtis distances was much less prominent for the global lake data set, which showed a much more modest degree of local–global overlap.

Discussion

Our results suggest deviation of local–global overlap from neutral expectations likely reflects the relative importance of dispersal and environmental selection. We base this inference on the observation that our cross-ecosystem comparison of local–global overlap supported the hypothesis that human-associated habitats, assumed to possess reduced inter-patch heterogeneity and elevated dispersal, more closely resembled expectations from our neutral model in comparison with soil, marine and freshwater samples. These findings are more-or-less consistent with our conceptual diagram (Figure 1), which depicts extremes in the relative importance of dispersal (Case 1) and local selection (Case 2). If dispersal has a dominant role in determining species composition at the local scale, we would observe a random 'sampling' of the global diversity at the local scale and the level of local–global overlap is dictated by the ratio of the number of individuals in the local patch to the global species richness (*N*:*S*). This outcome is depicted in the middle left panel of Figure 1, and yields our neutral expectation (gray line in Figure 1 bottom panel). When local selection acts as a dominant biogeographic process, only a subset of the global diversity is present in a local patch (Figure 1, middle right panel) and the observed local–global overlap is much smaller than that expected by our neutral model (Figure 1, bottom panel).

Importantly, theory provides a basis for expectations, and often is most useful when it takes the form of a null model to which more complex systems can be compared (Gotelli and McGill, 2006). Our neutral model of local–global overlap in diversity fills this role and generated results consistent with previous models of local-regional species composition (Cornell and Lawton, 1992; Schoolmaster, 2001; Fox and Srivastava, 2006). Our model generates intuitive output with increases in local density or decreases in global diversity increasing the local–global overlap and decreases in local density or increases in global diversity decreasing the local–global overlap (Figure 3). In this way, the number of individuals in a local patch dictates the depth to which global diversity is 'sampled'. This constraint indicates that for many systems a local–global overlap of 100% is impossible, and allows for ecologically relevant comparisons across diverse ecosystems. Interestingly, the region along our *N:S* axis with the most rapid change in local–global overlap of diversity is in the range of our observations, although extremely low levels of local–global overlap (<10%) were not observed (Figure 3).

When comparing local–global overlap across ecosystems, the most distinct pattern we observed was the propensity for human-associated samples to have higher local–global overlap, and to be more similar to our neutral expectations than other microbial habitats (lakes, soil and ocean; Figure 3). The human gut communities were the strongest example of this, but even the skin-associated communities were close to the neutral model simulations despite showing lower levels of absolute local–global overlap.

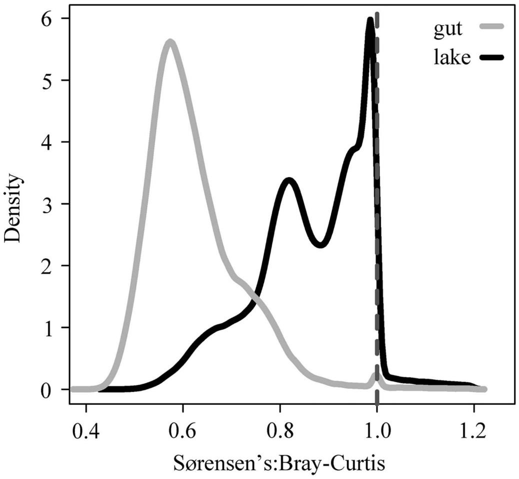

Figure 5 Frequency densities of the ratio of Sørensen's:Bray-Curtis distance metrics. Microbial community differences between lakes are primarily driven by presence–absence of OTUs as shown by the preponderance of Sørensen's:Bray-Curtis ratios near one (black dense line). In contrast, differences between gut bacterial communities are more driven by abundance differences (low Sørensen's:Bray-Curtis ratio; gray dense line). The vertical dashed line indicates a ratio of one.

In contrast, soil, lake and ocean communities consistently showed local–global overlap much lower than the neutral expectations. These results seem to indicate that human-associated microbial habitats have greater dispersal and/or less inter-patch variability in selective factors, owing to low inter-patch environmental heterogeneity, relative to the other environments.

Consider dispersal amongst human-associated microbial habitats relative to soil, lake and marine habitats. Because microorganisms disperse through air, all patches of every environment type are connected by the atmosphere establishing a baseline microbial dispersal rate common to all human and environmental ecosystems. However, in addition to air, microbes disperse directly between humans through close contact. When combined with the global human travel and high connectivity in human social networks, it is arguable that microorganisms have relatively rapid dispersal among all available patches (people). Evidence supporting this idea includes the observation that skin–skin contact causes microbial transfer (Meadow *et al.*, 2013), as well as pandemic pathogen transmissions that occur over short time scales (Mutreja *et al.*, 2011). In stark contrast, environmental ecosystems like lakes and soils are intrinsically immovable and thus globally distant ecosystem patches are hypothetically connected by atmospheric dispersal alone.

Another consideration is inter-patch environmental heterogeneity. Because humans rely on very rigid homeostatic conditions to maintain biochemical processes, each healthy human gut is likely a near identical ecosystem from a microbial perspective. In contrast, soils, lake and ocean patches are highly heterogeneous in temperature, nutrients, light availability and many other variables that affect microorganism colonization, growth and survival. A recent comparison of global soil microbial diversity to the diversity recovered in Central Park, NY, USA highlights the importance of site-to-site variation in environmental conditions. Ramirez *et al.* (2014), were able to recover comparable levels of soil microbial diversity in Central Park as that observed across broad continental gradients, and this was attributed to the tremendous heterogeneity of soil conditions observed in Central Park. We view inter-patch heterogeneity in environmental conditions, and therefore selection, as perhaps the most likely explanation for lower observed single sample overlap in aquatic and soil communities relative to the human gut and tongue. Analogous to the Central Park study (Ramirez *et al.*, 2014), intra-patch environmental heterogeneity may enhance local–global overlap, but additional modeling and empirical work would be required to fully investigate this added complexity.

It is important to note that we are not suggesting that the microbial community is completely uniform across human habitats as a great wealth of research demonstrates systematic differences between individual microbiomes (Turnbaugh *et al.*, 2008; Kuczynski *et al.*, 2010; Faust *et al.*, 2012). Rather, our analysis predicts that most bacterial OTUs can be found in a single host and differences between individuals are based on relative abundance differences as opposed to more predominant presence–absence differences, as would be expected in non-host-associated ecosystems like lakes given the levels of local–global overlap we observed, as can be observed for global human gut and lake samples from our analysis (Figure 5).

Our paired theoretical–empirical approach provided novel insight into how local microbial diversity scales to the global extent in a diverse set of ecosystems. We find an intriguing dichotomy between human-associated and non-human-associated habitats that seems to be consistent with what we know about cross-patch environmental heterogeneity and dispersal amongst patches in those types of microbial ecosystems. However, further work is required to rigorously test this hypothesis. In addition, our work relied on OTUs defined based upon 16 S ribosomal-RNA gene sequences, and other approaches are now available to resolve much finer genetic differences. Of course these more resolved approaches would likely find greater levels of endemicity, but we have no reason to expect our qualitative patterns would not be robust to exploration with more genetically resolved techniques.

Conclusions

In summary, we argue that comparison of local–global overlap of diversity to neutral model expectations can provide insight into the relative importance of biogeographic processes (e.g., dispersal vs selection). When evaluated with a conceptually and statistically robust approach, local–global overlap in diversity appears to be a property that non-randomly varies between ecosystem types. In addition, our observed cross-ecosystem patterns in local–global overlap are consistent with ecosystem properties such as dispersal rate and inter-patch environmental heterogeneity. Given the theoretical basis and initial observations we report here, this metric may prove useful in the future in addressing fundamental questions about the drivers of microbial community assembly in much the same way that distance-decay relationships have helped to identify patterns consistent with island biogeography theory.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We acknowledge Notre Dame's Environmental Change Initiative for support of JAL, members of the Jones Lab for comments on an earlier version of this manuscript, and three anonymous reviewers. SEJ was partially supported by NSF award DEB-1442230.

References

Akaike H. A new look at the statistical model identification. *Trans Automat Contr* 1974; **19**: 716–723. doi: 10.1109/TAC.1974.1100705.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J *et al.* (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci USA* **107**: 18933–18938.

Caporaso JG, Bittinger K, Bushman FD, Desantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2010a; **26**: 266–267.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010b; **7**: 335–336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108**: 4516–4522.

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

Cavender-Bares J, Kozak KH, Fine PVA, Kembel SW. (2009). The merging of community ecology and phylogenetic biology. *Ecol Lett* **12**: 693–715.

Cifelli RL. (1993). Early Cretaceous mammal from North America and the evolution of marsupial dental characters. *Proc Natl Acad Sci USA* **90**: 9413–9416.

Cornell H V, Lawton JH. (1992). Species interactions, local and regional processes, and limits to the richness of ecological communities : a theoretical perspective. *J Anim Ecol* **61**: 1–12.

Cottingham KL, Lennon JT, Brown BL. (2005). Knowing when to draw the line: designing more informative ecological experiments. *Front Ecol Environ* **3**: 145–152.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Faith DP. (1992). Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**: 1–10.

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J *et al.* (2012). Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput Biol* **8**: e1002606.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. (2010). Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* **107**: 6477–6481.

Flather C. (1996). Fitting species–accumulation functions and assessing regional land use impacts on avian diversity. *J Biogeogr* **23**: 155–168.

Fox JW, Srivastava D. (2006). Predicting local-regional richness relationships using island biogeography models. *Oikos* **113**: 376–382.

Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci USA* **110**: 4651–4655.

Gonçalves-Souza T, Romero GQ, Cottenie K. (2013). A critical analysis of the ubiquity of linear local-regional richness relationships. *Oikos* **122**: 961–966.

Gotelli NJ, McGill BJ. (2006). Null versus neutral models: what's the difference? *Ecography* **29**: 793–800.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**: 497–506.

Helmus MR, Bland TJ, Williams CK, Ives AR. (2007). Phylogenetic measures of biodiversity. *Am Nat* **169**: E68–E83.

Hillebrand H. (2005). Regressions of local on regional diversity do not reflect the importance of local interactions or saturation of local diversity. *Oikos* **110**: 195–198.

Hillebrand H, Blenckner T. (2002). Regional and local impact on species diversity—from pattern to processes. *Oecologia* **132**: 479–491.

Hubbell SP. (2001). *The Unfied Neutral Theory of Biodiversity and Biogeography*. Princeton University Press: Princeton, New Jersey.

Jezbera J, Jezberová J, Brandt U, Hahn MW. (2011). Ubiquity of Polynucleobacter necessarius subspecies asymbioticus results from ecological diversification. *Environ Microbiol* **13**: 922–931.

Jezberová J, Jezbera J, Brandt U, Lindström ES, Langenheder S, Hahn MW. (2010). Ubiquity of Polynucleobacter necessarius ssp. asymbioticus in lentic freshwater habitats of a heterogenous 2000 km2 area. *Environ Microbiol* **12**: 658–669.

Jimenez-Valverde A, Mendoza SJ, Cano JM, Munguira ML. (2006). Comparing relative model fit of several species-accumulation functions to local Papilionoidea and Hesperioidea butterfly inventories of Mediterranean habitats. In: Hawksworth DL, Bull AT (eds). *Arthropod Diversity and Conservation*. Springer: Netherlands, pp 163–176.

Jones SE, McMahon KD. (2009). Species-sorting may explain an apparent minimal effect of immigration on freshwater bacterial community dynamics. *Environ Microbiol* **11**: 905–913.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD *et al.* (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463–1464.

Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D *et al.* (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* **11**: 210.

Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75**: 5111–5120.

Livermore JA, Mattes TE. (2013). Phylogenetic detection of novel Cryptomycota in an Iowa (United States) aquifer and from previously collected marine and freshwater targeted high-throughput sequencing sets. *Environ Microbiol* **15**: 2333–2341.

Logue JB, Lindström ES. (2008). Biogeography of bacterioplankton in inland waters. *Freshw Rev* **1**: 99–114.

Lushi E, Wioland H, Goldstein RE. (2014). Fluid flows created by swimming bacteria drive self-organization in confined suspensions. *Proc Natl Acad Sci USA* **111**: 9733–9738.

Martiny JBH, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.

Meadow JF, Bateman AC, Herkert KM, O'Connor TK, Green JL. (2013). Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ* **1**: e53.

Melville J, Harmon LJ, Losos JB. (2006). Intercontinental community convergence of ecology and morphology in desert lizards. *Proc Biol Sci* **273**: 557–563.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S *et al.* (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**: 462–465.

Newton RJ, Jones SE, Helmus MR, McMahon KD. (2007). Phylogenetic ecology of the freshwater *Actinobacteria* acI lineage. *Appl Environ Microbiol* **73**: 7169–7176.

Paradis E, Claude J, Strimmer K. (2004). APE: analysis of phylogenetics and evolution in R. *Bioinformatics* **20**: 289–290.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

R Development Core Team (2010), R: A Language and Environment for Statistical Computing. R Foundation Statistical Computing, Vienna, Austria. ISBN3-900051070.

Ramirez KS, Leff JW, Barberan A, Bates ST, Betley J, Crowther TW *et al.* (2014). Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proc R Soc B* **281**: 20141988.

Rusconi R, Guasto JS, Stocker R. (2014). Bacterial transport suppressed by fluid shear. *Nat Phys* **10**: 212–217.

Schoolmaster DR Jr. (2001). Using the dispersal assembly hypothesis to predict local species richness from the relative abundance of species in the regional species pool. *Community Ecol* **2**: 35–40.

Szava-Kovats RC, Zobel M, Pärtel M. (2012). The local-regional species richness relationship: new perspectives on the null-hypothesis. *Oikos* **121**: 321–326.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2008). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Van Rooijen J. (2009). Estimating the snake species richness of the Santubong Peninsula (Borneo) in two different ways. *Contrib to Zool* **78**: 141–147.

Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. (2013). Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiol Rev* **37**: 936–954.

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