nature ecology & evolution

ARTICLES

PUBLISHED: 5 DECEMBER 2016 | VOLUME: 1 | ARTICLE NUMBER: 0015

High taxonomic variability despite stable functional structure across microbial communities

Stilianos Louca[1,2*], Saulo M. S. Jacques[3,4], Aliny P. F. Pires[3], Juliana S. Leal[3,5], Diane S. Srivastava[1,6], Laura Wegener Parfrey[1,6,7], Vinicius F. Farjalla[3] and Michael Doebeli[1,6,8]

Understanding the processes that are driving variation of natural microbial communities across space or time is a major challenge for ecologists. Environmental conditions strongly shape the metabolic function of microbial communities; however, other processes such as biotic interactions, random demographic drift or dispersal limitation may also influence community dynamics. The relative importance of these processes and their effects on community function remain largely unknown. To address this uncertainty, here we examined bacterial and archaeal communities in replicate 'miniature' aquatic ecosystems contained within the foliage of wild bromeliads. We used marker gene sequencing to infer the taxonomic composition within nine metabolic functional groups, and shotgun environmental DNA sequencing to estimate the relative abundances of these groups. We found that all of the bromeliads exhibited remarkably similar functional community structures, but that the taxonomic composition within individual functional groups was highly variable. Furthermore, using statistical analyses, we found that non-neutral processes, including environmental filtering and potentially biotic interactions, at least partly shaped the composition within functional groups and were more important than spatial dispersal limitation and demographic drift. Hence both the functional structure and taxonomic composition within functional groups of natural microbial communities may be shaped by non-neutral and roughly separate processes.

Microbial metabolism drives the bulk of biogeochemical fluxes in virtually every natural ecosystem[1]. Microbial communities can display complex variation in composition across space or time, such as down the ocean water column[2] or across seasons[3], and this variation can have profound effects on ecosystem functions[3]. The mechanisms driving this variation remain poorly understood, because the entanglement of multiple mechanisms severely complicates the identification of causal relationships. Potential mechanisms of microbial community assembly include adaptation to local environmental conditions ('environmental filtering')[4], biotic interactions such as predation[5,6], random population drift[7], random colonization order[8] and spatially limited random dispersal[9]. Recent work suggests that the metabolic functional potential of microbial communities in the global ocean or in soil is closely related to environmental conditions, while the taxonomic variation within individual functional groups is only poorly explained by environmental conditions[10–12]. This points towards an elegant paradigm for microbial ecology, in which community metabolic function is strongly shaped by energetic and stoichiometric constraints such as the availability of electron acceptors for respiration[10], while the composition within functional groups is modulated by additional mechanisms. According to this paradigm, similar environments should promote similar microbial community function, while allowing for taxonomic variation within individual functional groups.

Trait convergence concurrent with species divergence has been reported previously for plant communities[13,14], however, it is unclear whether (and how) conclusions from plant biogeography extrapolate to microbial biogeography[9]. Microbial population sizes are typically much higher than those of plant communities (for example, ~10^9 bacterial cells per litre in lakes[15]). Consequently, stochastic demographic drift (that is, fluctuations in population sizes purely due to random birth–death events) may be less important in microbial communities compared with other mechanisms, such as priority effects or competitive exclusion[16]. Moreover, it is difficult to compare microbial metabolic traits (such as the use of various electron acceptors for respiration) to trait palettes conventionally considered in plant biogeography (for example, canopy height, dispersule shape or leaf phenology[13,14]); hence, trait convergence in plants need not imply functional convergence in microorganisms.

High microbial taxonomic variability despite functional stability has been previously observed in bioreactors[7,17,18]. In natural systems, analogous observations emerge from comparisons of β diversities of community species and gene content. Notably, human gut microbiota have been found to exhibit a core set of genes despite strong taxonomic turnover between individuals[19], and bacterial community composition on the macroalgae *Ulva australis* was best explained in terms of gene content rather than species content[8]. These studies do not, however, explicitly consider taxonomic composition within individual functional groups, perhaps because assigning shotgun environmental gene sequences to specific taxa remains a notoriously hard problem[20]. Here we circumvent this problem and compare the functional and taxonomic variability of prokaryote (that is, bacterial and archaeal) communities across 22 replicate aquatic environments, harboured within the foliage ('tanks') of bromeliads in the Jurubatiba National Park, Brazil (Fig. 1). Bromeliad tanks accumulate rain water and organic detritus from their surrounding

[1]Biodiversity Research Centre, University of British Columbia, Vancouver, V6T 1Z4, Canada. [2]Institute of Applied Mathematics, University of British Columbia, Vancouver, V6T 1Z2, Canada. [3]Department of Ecology, Biology Institute, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ 21941-590, Brazil. [4]Programa de Pós-Graduação em Ecologia e Evolugão, Universidade Estadual do Rio de Janeiro, Rio de Janeiro, 20550-013, Brazil. [5]Programa de Pós-Graduação em Ecologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941-971, Brazil. [6]Department of Zoology, University of British Columbia, Vancouver, V6T 1Z4, Canada. [7]Department of Botany, University of British Columbia, Vancouver, V6T 1Z4, Canada. [8]Department of Mathematics, University of British Columbia, Vancouver, V6T 1Z2, Canada. *e-mail: louca@zoology.ubc.ca

environment, and intense decomposition of this detritus sustains a high richness of microorganisms and macroinvertebrates[21,22]. Apart from constituting regional biodiversity hotspots, bromeliads are often used as miniature model systems for ecology[22,23]. Microbial communities in bromeliads tend to be highly distinct from the surrounding environments (for example, soil), exhibiting a strong shift towards fermenting and methanogenic organisms[21,24,25]. To ensure a high similarity between systems, we only surveyed mature plants of a single bromeliad species (*Aechmea nudicaulis*) from the same region[26].

We used amplicon DNA sequencing of the 16S ribosomal gene, a standard marker gene in microbial ecology[27], to estimate the prokaryotic taxonomic community composition in each bromeliad. Whenever possible, we assigned detected organisms to one or more metabolic functional groups of potential ecological importance, such as fermentation, dissimilatory reduction of nitrogen compounds (nitrogen respiration) or methanogenesis[21,25] (nine groups in total). This enabled us to estimate taxonomic richness and variability within each functional group. Detected taxa were assigned to these functional groups on the basis of published evidence on cultured representatives. For example, if an uncultured organism was identified within a known bacterial genus whose cultured member species have all been identified as fermenters, we considered that organism to also be a fermenter. We used environmental shotgun DNA sequencing (metagenomics) to estimate the overall stability of functional community structure, based on the relative abundances of proxy genes corresponding to the considered functional groups. For example, we used gene sequences for methyl-coenzyme M reductase (*mcr*) and heterodisulfide reductase (*hdr*) as proxies for methanogens[28]. We found that all communities exhibited a remarkably similar functional structure, which contrasts with a highly variable taxonomic composition within individual functional groups. Furthermore, we examined phylogenetic community structure and species distribution patterns, and compared taxonomic composition to abiotic environmental conditions and geographical location, to elucidate potential mechanisms driving variation within functional groups.

Results and discussion

Functional stability contrasts with taxonomic variability. We found that the metabolic functional structure of tank prokaryotic communities, in terms of the relative abundances of proxy genes, was similar between all bromeliads (Fig. 2a,b). This functional similarity is presumably promoted by strong stoichiometric balancing between coupled metabolic pathways, the majority of which serve to break down large organic compounds to simpler organic molecules and gradually move electrons from reduced organic carbon to terminal electron acceptors such as protons (H^+), carbon dioxide (CO_2), sulfate (SO_4^{2-}), nitrate (NO_3^-) and oxygen (O_2)[29]. These metabolic pathways are distributed across multiple organisms, and link the breakdown of dead organic matter captured in the bromeliads to the eventual release of carbon dioxide[30] (CO_2), methane[24] (CH_4) and presumably molecular nitrogen (N_2). Each step along these pathways thus appears to sustain highly constrained microbial productivities, resulting in specific proportions of functional groups that are conserved across bromeliads. We note that while similar functional community structure between bromeliads is highly indicative of similar productivities (and hence activities) of the functional groups, *in situ* process rate measurements are required to verify this conclusion.

On the other hand, we found that the taxonomic composition within individual functional groups was highly variable across bromeliads, in terms of the occurrence of operational taxonomic units (OTUs; at 99% 16S rDNA similarity) as well as the proportions of OTUs within each functional group (Fig. 2c–k). Within any given functional group, OTUs detected in all of the samples (core microbiome) only made up ~0–1% of total OTUs across all samples

Figure 1 | Bromeliad species used in this study. The main image is *Aechmea nudicaulis*, the bromeliad species considered in this study. The foliage forms a deep central cavity (tank, inset) that accumulates rainwater and dead organic material, such as leaves from nearby trees. The decomposition of this material sustains a highly productive and diverse food web inside the tank. Photographs by S. Louca.

(regional pool) and any two bromeliads shared only ~20–60% of their OTUs (Supplementary Table 1). This overlap between communities is significantly lower than would be expected solely due to limited sequencing depth ($P<0.001$ using a null model of random sequencing of the regional pool) or due to random independent colonization by OTUs ($P<0.05$ using a null model of OTU assignment depending on OTU abundances in the regional pool; Supplementary Table 1). Furthermore, coefficients of variation for OTU proportions within functional groups were typically much greater than coefficients of variation of relative gene abundances (~2–3 versus ~0.2–0.6, respectively; Supplementary Table 2). This taxonomic variability within functional groups persists to a considerable extent even when OTUs are combined at higher taxonomic levels (for example, genus, family, order or class level; Supplementary Figs 1–4), and is in contrast to the more constant relative gene abundances. In each bromeliad the same metabolic niches appear to be occupied by very different organisms, even if the occupancy of each niche—in terms of its relative abundance—remains almost unchanged. The variability seen within functional groups is also reflected at the community level. When we considered relative OTU abundances in the entire community we again found a significantly low overlap between samples ($P<0.001$) as well as a high coefficient of variation of relative OTU abundances (2.9 on average). Our results explain the previously observed strong variation in microbial community composition across bromeliads[22], and demonstrate that taxonomic variability between replicate ecosystems need not imply differences in community function.

The strong taxonomic variability within functional groups is presumably enabled by a high functional redundancy in the regional microbial pool (Fig. 3), allowing for potential colonization of each bromeliad by alternative, functionally similar OTUs. The precise mechanisms determining the subset of OTUs that eventually establish in each bromeliad and within each metabolic niche are, at this point, unknown. The fact that beta (β) diversity (in terms of OTU

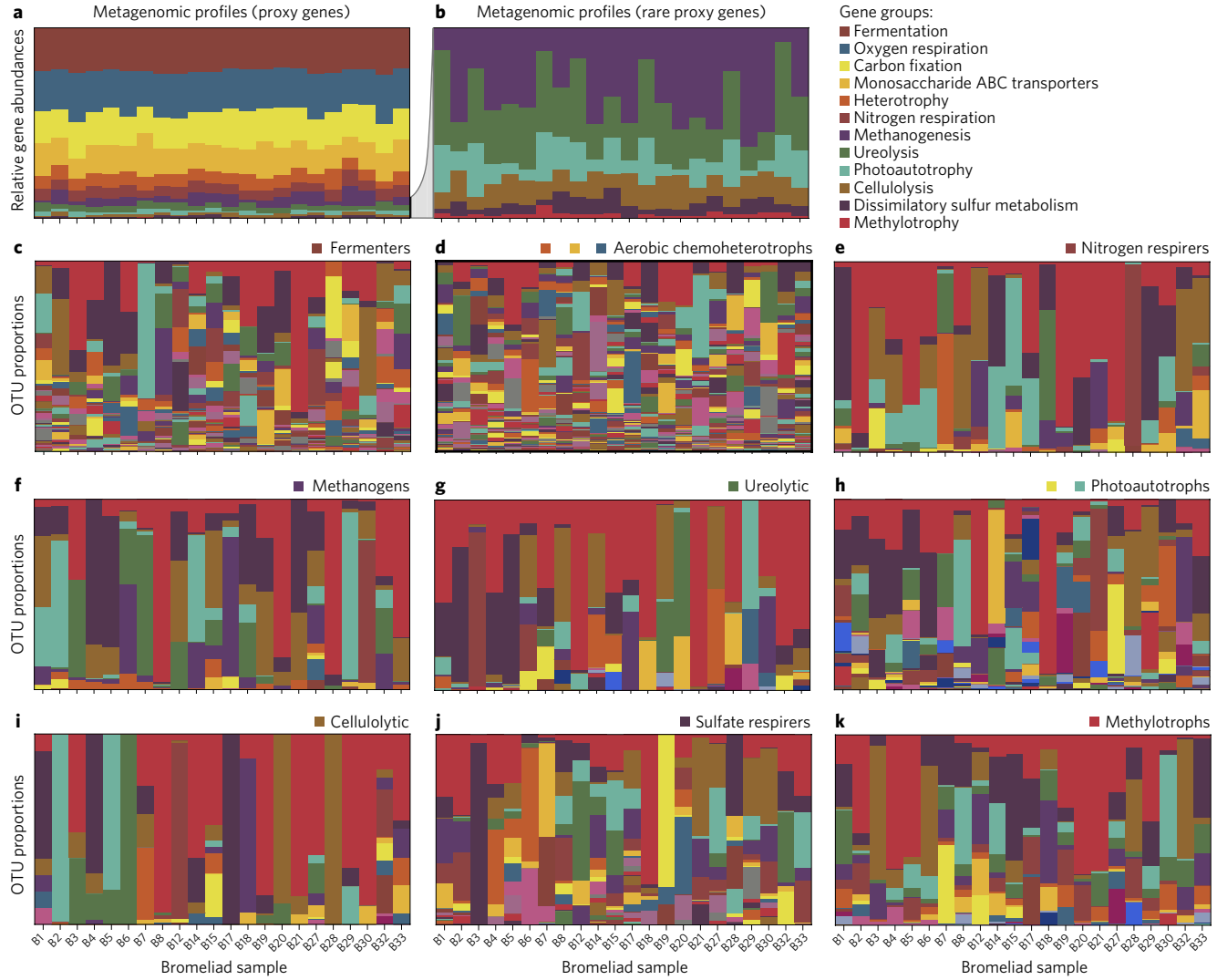

Figure 2 | Taxonomic and functional community structure. a, Relative abundances of proxy genes in prokaryotic metagenomic sequences (genes are grouped by function with one colour per gene group and one column per bromeliad). For details on associating genes with functions see Methods. **b**, Sub-plot of **a** focusing on the rarer genes for better illustration. **c**–**k**, Prokaryotic OTU proportions within individual functional groups (one colour per OTU, one column per sample, one plot per functional group), as determined from 16S rDNA sequences. Owing to ambiguities in gene function, for some functional groups (**d**,**h**) we considered multiple proxy genes. For each functional group, proxy genes are indicated via colour codes (corresponding to colours in **a** and **b**) next to the functional group's name. For more detailed metagenomic profiles see Supplementary Fig. 12. For the taxonomic composition within functional groups at higher taxonomic levels (genus, family or order) see Supplementary Figs 1–3. (Sample size: 22 bromeliads.)

overlaps) within functional groups differed significantly from the null expectation on the basis of the alpha (α) diversities within bromeliads and the regional gamma (γ) diversity, indicates that non-random (for example, niche-based or spatially structured) processes may drive OTU turnover between bromeliads[31]. To further assess the potential importance of non-random processes, we compared patterns of OTU co-occurrences, OTU detection frequencies and phylogenetic clustering to various null models of random or neutral community assembly.

Null model analysis of OTU distributions. Random colonization of bromeliads by independent OTUs would result in negligible associations between OTUs. To test this scenario for each functional group, we compared OTU co-occurrences, as defined by their *C* scores (a presence–absence-based measure for mutual OTU segregation), to a null model corresponding to random OTU sampling from the functional group's regional pool ('fixed–fixed' null model)[32]. Within six out of nine functional groups (aerobic chemoheterotrophs, cellulose degraders, fermenters, nitrogen respirers, photoautotrophs and sulfate respirers), OTUs were significantly segregated with respect to each other, that is, *C* scores were higher than expected by chance ($P < 0.05$; Supplementary Table 3). The remaining functional groups also displayed OTU segregation, although differences from the null model were not statistically significant.

Failure to detect significant co-occurrence patterns within some functional groups may reflect random assembly processes but could also be due to the coarseness of presence–absence data, and abundance-based null models may be more suitable for detecting non-random co-occurrence patterns[33]. Hence, we also considered an abundance-based metric (MA score, a measure for the congruency of OTU proportions between samples)[33], in combination with a null model that corresponds to random sampling from the regional pool and which takes into account relative OTU abundances 'IT model'[33]. We found that all functional groups exhibited an MA score that was significantly lower than expected by the null model ($P < 0.001$; Supplementary Table 4), indicating a substantial non-random segregation between samples in terms of OTU proportions. When we considered the entire community, we again found

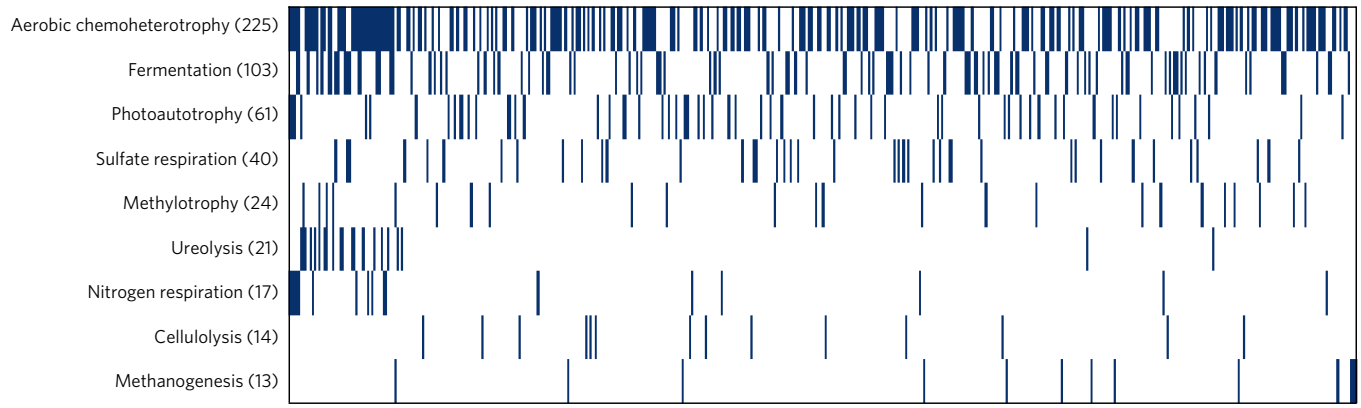

Figure 3 | Functional redundancy in the regional OTU pool. Associations of functional groups (rows) with OTUs (columns), indicated by blue cells. Functional groups are sorted according to their number of OTUs (indicated in brackets). Some OTUs were associated with more than one functional group. For analogous plots at the genus, family and class level see Supplementary Figs 13–15, respectively. (Sample size: 22 bromeliads.)

non-random segregation patterns ($P < 0.001$) both in terms of their C scores and their MA scores.

When combined with spatially limited dispersal, neutral demographic drift could in principle produce non-random (that is, spatially structured) segregation patterns[34], because bromeliads in greater proximity would tend to exhibit more similar community composition. However, as we discuss below, spatially limited dispersal is probably not important at the scales considered here. Hence, the non-random segregation patterns detected here probably reflect a mutual exclusion between OTUs that was potentially caused by environmental filtering or biotic interactions[32], rather than spatially correlated or uncorrelated neutral assembly.

To test whether community assembly was neutral with respect to phylogenetic relationships, we compared the phylogenetic distances between OTUs co-occurring in the same samples to those occurring in the regional pool. Specifically, within each functional group we assessed whether OTUs found in the same samples tend to be phylogenetically underdispersed or overdispersed in terms of their mean phylogenetic distance, when compared to the expectation based on random OTU sampling from the regional pool. Underdispersion is commonly interpreted as a sign of environmental filtering acting similarly on closely related clades[35], while overdispersion is interpreted as a sign of increased competition between close relatives, although other mechanisms may also create non-neutral patterns[36]. We found that six functional groups (aerobic chemo-heterotrophs, fermenters, nitrogen respirers, photoautotrophs, sulfate respirers and urea degraders) showed a significant tendency towards underdispersion ($P < 0.05$), while one functional group (methylotrophs) demonstrated significant overdispersion (Supplementary Table 5). The detection of a significantly non-neutral phylogenetic structure in seven out of nine functional groups is unlikely to be the result of a false positive detection rate ($P < 0.000001$). This supports the interpretation that community assembly is not neutral within these functional groups, but is subject to selection mechanisms that are sensitive to phylogenetic relationships. The absence of a statistically significant phylogenetic pattern in two functional groups could result from a weak phylogenetic signal in the processes driving OTU turnover rather than from truly neutral assembly, although on its own the test at hand cannot discriminate between the two scenarios[36].

All of the above null models aim to resemble random or neutral community assembly, but they are based on heuristic randomization algorithms that lack a clear biological mechanism[32,33]. To test whether our rejection of most of these null models is merely due to the inadequacy of the models for describing actual population dynamics, we also compared the composition within functional groups to a mechanistic dynamical model (the Sloan neutral model)[37,38]. This model was developed specifically for microbial communities and assumes that populations are solely driven by stochastic birth–death events and random immigration from the regional pool. Upon calibration, the model predicts the detection frequencies of OTUs on the basis of their mean relative abundances in the regional pool. We found that 35–60% of OTUs deviated significantly ($P < 0.05$) from the expectation, depending on the functional group (Supplementary Fig. 5 and Supplementary Table 6). These fractions are much higher than the 5% type I error rate expected under the null model. In fact, we found that the model's goodness of fit (in terms of the likelihood or the R^2) was significantly lower than if our data were generated by the model ($P < 0.001$ for all functional groups; Supplementary Table 6).

Environmental filtering. Non-neutral patterns might be caused by environmental filtering, by biotic interactions, or by a combination of these, such as trade-offs between environmental stress tolerance and competition[32,36]. To determine whether environmental filtering partly drives taxonomic composition at the community level or within functional groups, we examined the predictive ability of several physicochemical variables (overview in Supplementary Table 7). We considered standard limnological variables such as pH and salinity, as well as other potentially important variables such as detrital volume and vegetative cover (shading). Using redundancy analysis[39] we found that, when considered separately, individual environmental variables only explain small fractions (~5–15%) of the overall variance in OTU proportions, both at the community level as well as within functional groups (Fig. 4a,c). When we regressed OTU proportions against multiple environmental variables using multivariate non-linear models, we found that our models had moderate predictive power, as indicated by cross-validated coefficients of determination ($R^2_{CV} \approx 0.1$–0.5; Fig. 4b). Predictive power at the community level was similar to predictive power within functional groups (Fig. 4b,d), consistent with the interpretation that the taxonomic variation at the community level mostly stems from a variation within functional groups. Furthermore, the moderate predictive power, despite the high number of available environmental variables, suggests that environmental conditions in bromeliads explain some of the variation within functional groups, but that additional factors are also important.

The potential role of dispersal limitation. To test whether microbial communities were spatially structured, we used Mantel rank correlation tests to compare geographical distances to the dissimilarities of communities in terms of OTU proportions, both at the community level as well as within functional groups[39]. Out of nine

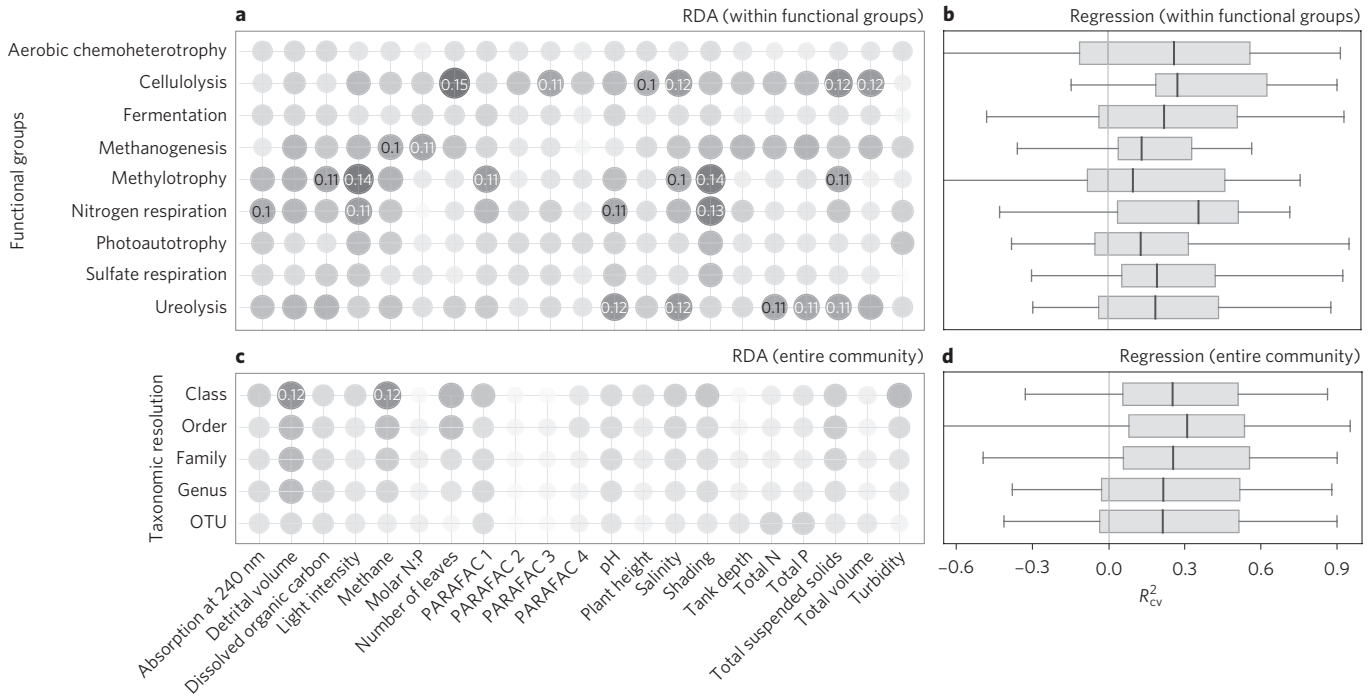

Figure 4 | Relating OTU proportions to environmental variables. a, Fractions of variance in OTU proportions explained by RDA using individual environmental variables (one column per environmental variable, one row per functional group). Larger and darker circles correspond to larger fractions of the explained variance, and indicate a stronger relation between an environmental variable and a functional group's composition. Fractions above 0.1 are written in the circles. **b**, Distribution of R^2_{CV}, a measure for a model's predictive power, for regression models of OTU proportions within each functional group using environmental variables as predictors (one box per functional group). Horizontal whiskers comprise 95% percentiles around medians. The vertical grey line at zero is shown for reference. **c**,**d**, Similar to **a** and **b**, but for relative taxon abundances at the community level, at various taxonomic resolutions. (Sample size: 22 bromeliads.)

functional groups, only aerobic chemoheterotrophs exhibited a significant positive rank correlation between geographical distance and dissimilarity ($P < 0.05$; Fig. 5b), although methylotrophs, urea degraders and the entire community also showed a weak positive correlation (Fig. 5a and Supplementary Fig. 6). To assess the extent to which these weak correlations were caused by purely spatial effects (for example, dispersal limitation) rather than spatially autocorrelated environmental filtering, we performed variation partitioning of OTU composition against spatial variables (longitude, latitude and polynomial combinations) and environmental variables[40]. In all functional groups and at the community level, the variation explained solely by spatial variables was below 5% (in terms of the R^2, adjusted for the number of variables) and much lower than the variation explained solely by environmental variables (Fig. 5d). Hence, spatial dispersal limitation probably played a negligible part in driving microbial community differences between bromeliads. These results are consistent with previous work that found negligible effects of spatial distance on bacterial communities in bromeliads at similar spatial scales[22].

Unexplained variation. Our null model analyses suggest that non-neutral selection processes at least partly shape the composition within functional groups. On the other hand, our 21 environmental variables only moderately predicted OTU proportions and pairwise dissimilarities of communities, both at the community level as well as within functional groups (Figs 4 and 5). The remaining variation that cannot be explained by our regression models could be due to unknown environmental variables further selecting for specific taxa, or due to past environmental conditions affecting current community structure. Alternatively, mechanisms other than environmental filtering, such as biotic interactions, may also have an important role in shaping microbial communities while maintaining functional similarity across bromeliads. The potential importance of biotic interactions, such as competitive exclusion or predation, in shaping microbial communities has been emphasized previously[5,6]. For example, adaptation of bacteriophages to specific hosts can influence bacterial species composition and promote spatial as well as temporal variation of microbial communities[41,42]. Consequently, unexplained taxonomic variation across locations may result from biotic interactions driving complex population dynamics. This interpretation is consistent with previous findings that the distribution of cyanobacterial taxa across coexisting bromeliads was driven by physicochemical factors as well as by protozoans and invertebrates[43]. In addition, stochastic colonization combined with biotic interactions may promote priority effects that result in multiple alternative equilibria[31]. This scenario would resemble previous findings in grassland plant communities[13], where species divergence despite trait convergence could not be attributed to dispersal limitation or neutrality, but instead appeared to be driven by priority effects.

Conclusions

We have shown that replicate natural ecosystems in close proximity can exhibit very different taxonomic composition of prokaryotic communities, despite similar metabolic functional structure. Several OTUs were not assigned to any functional group due to a lack of closely related cultured representatives, and hence the OTU richness and variability within functional groups are probably even higher in reality. Our findings point to an important difference between functional and taxonomic community structure, which arises because mechanisms leading to a convergence of metabolic function (for example, stoichiometric balancing between metabolic pathways) do not necessarily lead to a convergence of taxonomic composition. Reciprocally, strong taxonomic turnover may only weakly affect ecosystem functioning[44] (but see Strickland and

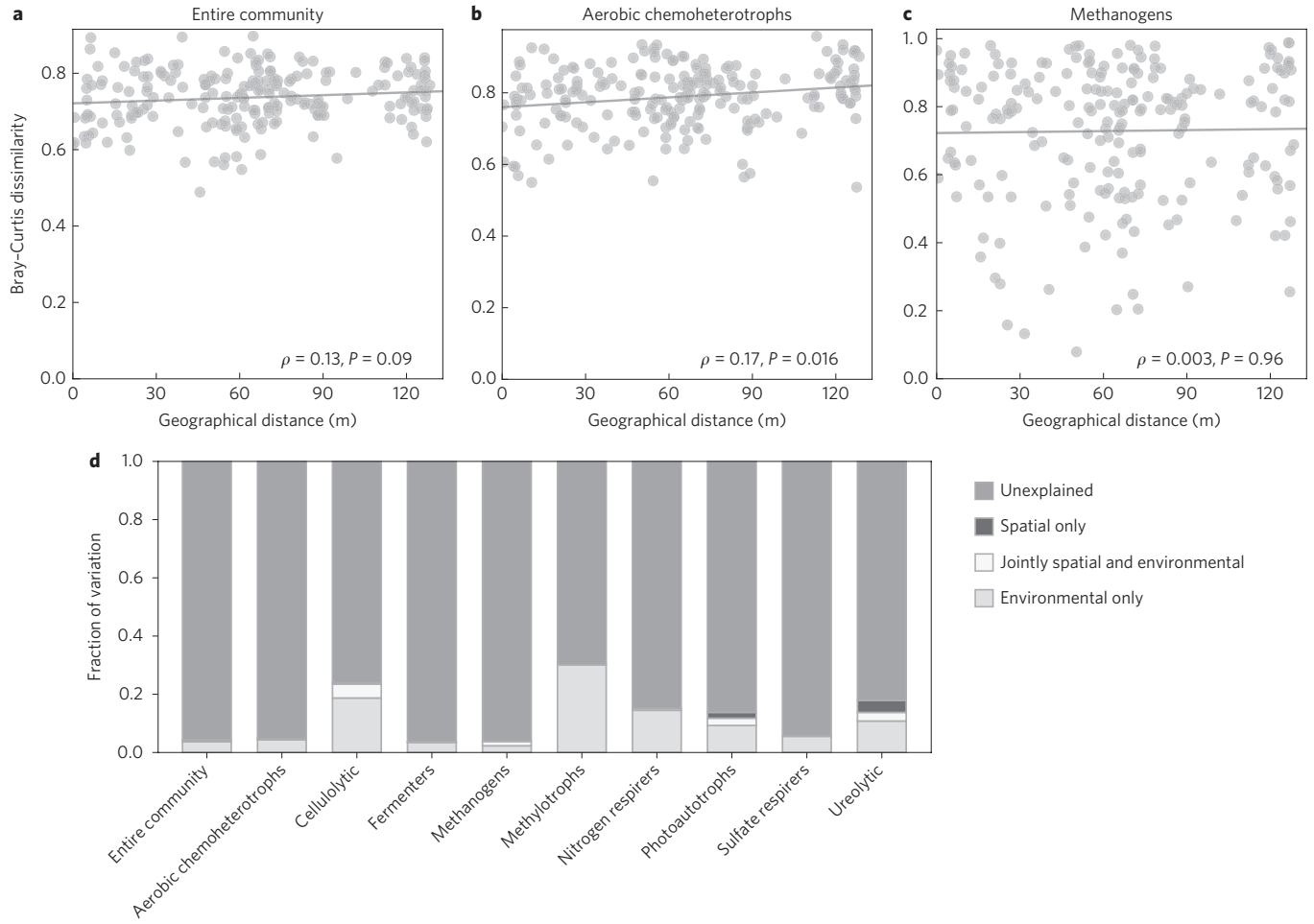

Figure 5 | Variation partitioning of OTU composition. a–c, Geographical distances between bromeliads compared to Bray–Curtis dissimilarities, in terms of OTU proportions at the community level (**a**) and within various functional groups (**b**,**c**, one point per sample pair). Least-squares regression lines are shown for reference. Rank correlations (ρ) and statistical significances (P) are written in the plots. For other functional groups see Supplementary Fig. 6. **d**, Variation partitioning of OTU proportions within individual functional groups as well as at the entire community level. Each bar segment corresponds to the fraction of variation explained in terms of the R^2 (adjusted for the number of predictors), either solely by spatial variables (that is, while controlling for environment), or solely by environmental variables (that is, while controlling for spatial structure), or jointly by spatial and environmental variables. Spatial variables included linear, quadratic and cubic combinations of latitude and longitude. (Sample size: 22 bromeliads.)

colleagues[45]). The decoupling between function and taxonomy observed here resembles previous observations in bioreactors[17] and the human gut[19]. We emphasize that, in contrast to taxonomic variation within functional groups, taxonomic variation at the community level is generally also a partial reflection of functional variation. In this study, as well as in the aforementioned studies[17,19], functional structure appeared relatively constant, presumably due to high physicochemical similarities between replicates. In more heterogeneous environments or across strong environmental gradients, the decoupling between function and taxonomy may be masked by strong metabolic niche effects. For example, functional β diversity was found to be strongly correlated with taxonomic β diversity across soil types[46]. The correlation between functional and taxonomic composition (at the community level) thus depends on the relative importance of metabolic niche effects when compared to processes causing variation within functional groups.

We suggest that functional community profiles, based on gene-centric metagenomics[10], on functional predictions for recovered genomes[47] or on a functional classification of detected taxa[11], should be the baseline of future microbial biogeography studies, particularly when the ultimate focus is on ecosystem functioning[10,48]. The residual variation within functional groups can then be analysed separately, as demonstrated here, to elucidate additional community assembly processes that act in superposition to metabolic niche effects. Our analysis suggests that in bromeliad tanks non-neutral processes, such as environmental filtering and as-yet-undetermined biotic interactions, are important drivers of the variation within individual functional groups, while spatial dispersal limitation and neutral drift appear to be less relevant. The careful separation of functional variation from the taxonomic variation within functional groups thus enables deeper insight into microbial community assembly, and will be an important step towards a truly mechanistic microbial ecology.

Methods

Biological sample collection. Detritus from the bottom of bromeliad tanks was collected and physicochemical measurements were taken from all bromeliads in the period 8–10 January 2015, within an area spanning roughly 0.2 km^2 in the Parque Nacional da Restinga de Jurubatiba, east coast Brazil. At that time, weather conditions were sunny, dry and hot, and were preceded by several weeks of extreme drought[49]. Supernatant liquid was removed from the bromeliad's central tank using a sterile serological pipette. The detritus at the bottom was then retrieved using a sterile syringe and a metal spatula, after cutting the bromeliad open for easier access. All of the retrieved detrital content was mixed before sampling. Samples were flash-frozen in liquid nitrogen within 10 min of collection and then stored in the laboratory at −80 °C until further processing. For shipment, samples were

concentrated via centrifugation (40,000*g* for 15 min, balanced using MilliQ filtered water) and removal of the supernatant fluid, and then freeze-dried for 24 h. The dried samples were shipped to the University of British Columbia, Canada, for further processing.

Chemical analysis of tank water. The water above the benthic detritus was collected using a serological pipette, stored in 25 ml centrifuge tubes on regular ice in the field and at −4 °C in the laboratory until further analysis (within 2 days). Total dissolved phosphorus concentrations were determined as the inorganic phosphorus obtained after a procedure of acid digestion and autoclaving of the water samples and the ascorbic acid–molybdate reaction[50]. Total nitrogen concentrations were determined as the concentration of nitrate obtained after an acid digestion procedure and autoclavation. Nitrate was transformed into nitrite with a cadmium column via a reduction step, and nitrite was subsequently quantified using a flow injection analysis system (FIA-Asia Ismatec)[51], yielding total nitrogen.

Water samples for CH_4 measurement were taken separately (1.5 ml per measurement) and directly from the bromeliad, fixed using formalin (4%) in 3 ml glass vials, kept on regular ice in the field and at 4 °C in the laboratory until analysis within 2 days. Air was sampled from the headspace using a syringe after shaking the vials for 1 min, and headspace CH_4 content was determined using a gas chromatograph. Methane concentrations were determined using a Shimadzu GC-2010AF chromatograph equipped with a Rt-QPLOT column (3 m × 0.32 m) and a flame ionization detector (FID-2010). Temperatures of the injection, column and detection were 120 °C, 85 °C and 220 °C, respectively. Nitrogen (N_2) was used as the carrier gas.

Conductivity, pH, temperature and total suspended solids (TSS) were measured in the field using an ExStik II EC500 (ExTech Instruments). Salinity was calculated from conductivity and temperature using the empirical formula reported previously[52]. Water turbidity was measured in the field using a Hanna Turbidimeter HI98703. Absorption spectra were measured in the laboratory using a Varian 50 Bio UV-Visible Spectrophotometer, following the manufacturer's procedures. Dissolved organic carbon (DOC) concentrations were determined using by Pt-catalysed high-temperature combustion with a Shimadzu TOC-VCPN Total Carbon Analyzer, after filtering through 0.7 µm Whatman GF/F glass fibre filters.

For one bromeliad the retrieved supernatant water was insufficient for performing all of the chemical assays in the field. That water sample was thus diluted at a ratio 1:5 using deionized water before measuring the conductivity, pH, TSS and turbidity. The resulting conductivity, salinity, TSS and turbidity were then corrected using the dilution factor. The pH was corrected using a standard curve constructed by serial dilution of water from another bromeliad. For several bromeliads the retrieved supernatant water was insufficient for measuring absorption spectra and DOC concentrations, as well as for excitation-emission spectrophotometry (EES; explained below). These water samples were thus diluted in the laboratory using deionized water as needed. All measurements were subsequently corrected for the effects of dilution.

EES of the water samples was performed using a Varian Cary Eclipse fluorescence spectrophotometer. In EES, each sample is exposed to light of several wavelengths while simultaneously measuring the resulting fluorescence spectrum[53]. The obtained excitation–emission matrices (EEMs) were analysed for organic carbon profiles using parallel factor analysis (PARAFAC) with the MATLAB package drEEM[54]. EEMs were pre-processed as follows. The EEM of pure MilliQ water was subtracted from the sample EEMs. Rayleigh (elastic) and Raman (inelastic) scatter signals were removed by replacing them with NaN. EEM entries for emission wavelengths smaller than the excitation wavelengths were set to zero. EEM entries at the excitation wavelengths 320 nm and 365 nm were ignored because of abnormal intensity troughs at all emission wavelengths, probably resulting from imperfections of the fluorometer lamp. EEMs were corrected for inner filter effects using the sample absorption spectra and the drEEM function fdomcorrect[54].

PARAFAC model fitting was attempted for various model sizes (3–9). To avoid local PARAFAC optima, fitting for each model size was repeated 50 times with random initialization using the drEEM function randinitanal. Model residuals were inspected manually to ensure that the model size was sufficient[55,56]. Split-half validation ($S_4C_4T_2$; splits: 4, combinations: 4, tests: 2)[54] failed for all considered model sizes, but was ignored because of low sample size when compared with the high richness of observed EEM profiles. Instead, to constrain the model's size and avoid overfitting, model components were inspected for physical plausibility[54] and subsequently compared to published entries in the OpenFluor fluorophore database based on Tucker's congruence coefficient[57]. We kept the model (size 4; Supplementary Fig. 7) with the highest number of plausible components represented in OpenFluor at a congruence of at least 0.98. The best matches in the OpenFluor database were 'CS-Galathea, C1' for component 1 (ref. [58]), 'Recycle_WRAMS, C5' for component 2 (ref. [59]), 'PrairieLakes, C2' for component 3 (ref. [60]) and 'FloridaKeys, C3' for component 4 (ref. [61]). The model explained 98.2% of the variance, at a core consistency of 82.9% (Supplementary Fig. 8). For each sample and for each individual PARAFAC component we determined the maximum fluorescence intensity in the component's EEM, and multiplied it by the component's score in the particular sample. This yielded four PARAFAC component intensities per sample, each in arbitrary units that are comparable across samples but not across PARAFAC components. These component intensities were subsequently used in our analysis as four additional environmental variables (PARAFAC 1–4).

Measurement of other physicochemical variables. Light intensity (the flux of photosynthetically active radiation) on bromeliads was measured using an LI-250A Lightmeter (LI-COR Biosciences), equipped with a US-SQS/L spherical micro quantum sensor (Heinz Walz GmbH). The light meter was placed on the ground next to the bromeliad at noon of a sunny day (10 January 2015), after trimming the bromeliad's foliage to avoid shading of the device by the bromeliad itself. The detrital volume was measured using the centrifuge tube scale after allowing for precipitation for 5 min, performing the read at the interface between the precipitated detritus and the supernatant transparent fluid. The total tank volume was set to the total volume of all retrieved material (detritus and water). The total tank depth was either measured using a metal wire with engraved centimetre scale, or using the serological pipette's volume scale upon calibration. Tree cover (shading) above bromeliads was measured by taking a photo from the top of a bromeliad 'face-up' on a sunny day, and processing the photo using ImageJ for contrasting objects against a blue sky background. An overview of all physicochemical environmental variables is provided in Supplementary Table 7.

16S sequencing. DNA was extracted from the rehydrated samples using the MoBio PowerSoil DNA extraction kit, by applying the manufacturer's suggested protocol. Amplification of the 16S rRNA gene was done using barcoded primers covering the V4 region (*Escherichia coli* 515F and 806R) that included Illumina adapters, and using the Earth Microbiome Project 16S amplification protocol version 4_13 (ref. 62). Amplicon DNA from all samples was pooled into a single library, at such proportions that each sample contributed a similar amount of DNA. Primer dimers and remaining PCR enzymes were removed from the amplicon library using the MoBio UltraClean PCR Clean-Up Kit. Library quantitation was performed by Genoseq Core (University of California, Los Angeles) using a high-sensitivity Agilent Bioanalyzer and Kappa Biosystems' Illumina Genome Analyze (KAPA SYBR FAST Roche LightCycler 480) kit, followed by qPCR. Sequencing was performed by Genoseq Core using an Illumina MiSeq next-generation sequencer, following the manufacturer's standard protocol.

Sequencing yielded 2,599,770 paired-end sequences (2 × 300 base pairs each). Sequence analysis was performed using the QIIME toolbox (version 1.9.1)[63]. Paired-end reads were merged after trimming forward reads at length 240 and reverse reads at length 160. Merged sequences were quality filtered using QIIME's default settings, yielding 2,393,473 sequences of median length 253. Remaining sequences were error-filtered and clustered *de novo* using cd-hit-otu[64] at a 99% 16S rDNA similarity threshold, generating 2,027 OTUs representing 1,908,183 sequences across all samples. Sample B17 yielded by far the fewest sequences (5,811 sequences corresponding to 677 OTUs). Diagnostic OTU rarefaction curves are shown in Supplementary Fig. 9.

We note that historically a lower resolution (at 97% 16S rDNA similarity) was recommended for delineating prokaryotic OTUs in biogeographical studies[65]. However, recent work shows that greater taxonomic resolution is needed to detect signals of endemism (for example, up to 99.5% for the cyanobacterium *Prochlorococcus*[66]) and signals of competitive exclusion (99–100%)[67], and that taxa defined on the basis of 97% similarity may be underspeciated[68,69].

Taxonomic assignment of representative sequences was done using uclust[70] and the SILVA reference database (release 119)[71], using the first 50 hits at a similarity threshold of at least 90% as follows: For any queried sequence, if at least one hit had a similarity $s \geq 99\%$, then all hits with similarity s were used to form a consensus taxonomy. Otherwise, if at least one hit had a similarity $s \geq 90\%$, then all hits with similarity at least $(s - 1\%)$ were used to form a consensus taxonomy. If a query did not match any reference sequence at or above 90% similarity, it was considered unassigned. A total of 1,965 OTUs (representing 1,874,361 sequences across all samples) were taxonomically annotated.

Representative sequences were aligned against the SILVA database using PyNAST[71,72], and phylogenetic relationships were calculated using the FastTree algorithm[73], at standard QIIME settings. Phylogenetic distances are in nucleotide substitutions per site. For analyses based on OTU proportions (for example, redundancy analysis and mean phylogenetic distances, described below) we normalized our OTU table by dividing each entry by the total number of sequences in a sample (this is the maximum-likelihood estimator of true OTU proportions in a sample). For analyses that depend on presence–absence data and that assume equal sampling effort, we rarefied our samples at equal sequencing depth (as described below for each case).

Functional annotation of prokaryotic taxa. To determine the taxonomic composition within each of the nine considered functional groups (aerobic

chemoheterotrophy, cellulolysis, fermentation, methanogenesis, methylotrophy, nitrogen respiration, sulfate respiration, photoautotrophy, ureolysis), we associated each taxonomically annotated OTU with one or more metabolic functions based on extensive literature search, whenever possible. Specifically, a taxon (for example, species or genus) was assigned to a function if all cultured species within the taxon are known to exhibit that function. For example, OTUs annotated at the genus (but not species) level, were only associated with functions present in all cultured species of that genus. Clades with no cultured representatives (for example, Miscellaneous Crenarcheaota Group[74]) were not functionally annotated, regardless of potential metagenomic or single-cell genomic analyses. For species with multiple known strains, we focused on type strains that are often (but not always) representative of the species in terms of metabolic function. We point out that this inference of metabolic phenotype is entirely based on cultured representatives, and as more organisms are being cultured some of the functional annotations may turn out to be false. This caveat is similar to limitations of other existing functional profiling techniques. For example, gene-centric metagenomics predicts functional genes based on their sequence similarity to genes studied in cultured organisms[20]. Furthermore, taxonomic binning of shotgun environmental sequences remains notoriously hard and coarse. In contrast, here we first identified known taxa using marker gene sequencing and then estimated their metabolic phenotype based on the experimental literature. This revealed a remarkably high taxonomic richness as well as variability within individual functional groups, although both richness and variability are probably still underestimated due to the limited coverage of cultured organisms. Our complete database for the functional annotation of prokaryotic taxa (FAPROTAX) is available online at http://www.zoology.ubc.ca/louca/FAPROTAX. A detailed evaluation of FAPROTAX, including a direct comparison with metagenomics, has been provided previously[11].

In this study, each taxonomically annotated OTU was compared against each FAPROTAX annotation rule in an automated way. In total 465 out of 1,965 OTUs (24%) were assigned to at least one functional group, yielding in total 518 functional annotations (see Supplementary Table 8 for an overview). A substantial fraction of OTUs could not be assigned to any functional group, thus OTU proportions inside a functional group only apply to the subset of functionally characterized OTUs (although this limitation does not affect the conclusions of this study). Conversely, a small number of OTUs were assigned to multiple functional groups (Supplementary Fig. 10). The complete list of functional annotations is available as Supplementary Data 1.

We note that FAPROTAX functional groups are not completely one-to-one comparable with metagenomic gene groups, due to ambiguities in the functions potentially performed by some genes[20]. To strengthen our confidence in the stability of the nine considered functional groups, we provide detailed gene-centric functional profiles for multiple related functions (Fig. 2a,b and Supplementary Fig. 12).

Metagenomic sequencing. To assess the functional stability of microbial communities across samples, we performed shotgun environmental DNA sequencing (metagenomics), which allows the detection of known genes in an environment regardless of their host organisms. Extracted DNA was sequenced in 100-bp paired-end fragments on an Illumina HiSeq 2000. Library preparation and sequencing was done by the Biodiversity Research Centre NextGen Sequencing Facility and followed standard Illumina protocols. All samples were uniquely barcoded and run together on a single lane. The resulting sequence data were processed using Illumina's CASAVA-1.8.2. Specifically, output files were converted to fastq format, and sequences were separated by barcode (allowing one mismatched base pair), using the configureBclToFastq.pl script. This yielded a total of 80,206,935 quality-filtered paired-end reads. Reads were trimmed at the beginning and end to increase average read quality, yielding an average forward and reverse read length of 97 and 98 bp, respectively. Sufficiently overlapping paired-end reads were merged using PEAR 0.9.8 with default options[75], yielding 9,757,035 merged reads. Non-merged read pairs were deduplicated using the SOFA pipeline (version 1.2)[76] and the KEGG protein reference database (release 2011.06.18)[28], in order to reduce potential double-counts during subsequent gene annotation. MetaPathways 2.5 (ref. 77) was used for open reading frame (ORF) prediction in all merged and non-merged reads (minimum peptide length 30, algorithm prodigal), yielding 111,568,314 ORFs. Predicted ORFs were taxonomically annotated in MetaPathways using LAST and the NCBI RefSeq protein database (release 2015.12.12)[78], and multiple taxonomic annotations were consolidated using a lowest common ancestor algorithm[77]. Non-prokaryotic ORFs were excluded from subsequent analysis. LAST annotation of prokaryotic ORFs against the KEGG protein reference database was performed using MetaPathways (KEGG release 18 June 2011, minimum BLAST-score ratio 0.4, maximum E-value 10^{-6}, minimum score 20, minimum peptide length 30, top hit), yielding 30,730,175 annotations. Metagenomic KEGG orthologous group (KOG) counts[28] were normalized using the total number of KEGG-annotated sequences per sample (total sum scaling). Whenever possible, multiple KOGs associated with similar metabolic functions (for example, dissimilatory nitrite reduction to ammonium, *nirBD* and *nrfAH*) were combined into a single gene group. An overview of KOGs associated with each function is provided in Supplementary Table 9. The resulting metagenomic profiles are given in Fig. 2a,b and Supplementary Fig. 12.

Comparing OTU turnover to null models. To assess the degree of OTU turnover between bromeliads, for every functional group and for any two bromeliads we measured the OTU overlap in terms of the Jaccard overlap index, defined as the number of OTUs detected in both samples, divided by the number of OTUs detected in any of the two samples[79]. Hence, a Jaccard overlap of 1 corresponds to complete overlap (regardless of OTU proportions), while a Jaccard overlap of 0 corresponds to no overlap at all. Mean Jaccard overlaps (MJO; that is, averaged over all bromeliad pairs) were within the range ~0.2–0.6 for all functional groups (Supplementary Table 1). These low MJOs indicate substantial differences in community structure across bromeliads. In principle, however, such low MJOs may also be observed between identical communities purely from stochasticity in OTU detection, that is, due to insufficient sequencing depth[80]. We thus compared each MJO with hypothetical MJOs generated under a null model of random sampling from the regional OTU pool. Specifically, for any given functional group, sequences were randomly reassigned to OTUs from a multinomial distribution corresponding to OTU proportions in the regional OTU pool, while maintaining the original total number of sequences per bromeliad and per functional group. The statistical significance (P value) of an observed MJO was defined as the probability that a random MJO would be lower than the observed MJO, and was estimated based on 1,000 iterations. All functional groups had a significantly low MJO ($P<0.001$), showing that low overlaps are not just the result of detection stochasticity. The same analysis as the above was also applied to the entire community, again yielding a significantly low MJO ($P<0.001$). We note that the Jaccard overlap of gene groups was 1 for all sample pairs, since all considered gene groups were detected in all samples.

OTUs may colonize bromeliads randomly and independently of one another, but colonization events may be so rare that population growth within bromeliads leads to amplified (that is, population-level) differences between samples. In that case, the individual-based null model described above, which re-assigns individual sequences to OTUs, may inflate the statistical significance of low overlaps between bromeliads. To test this scenario separately for each functional group, we also compared MJOs to a population-based null model in which OTUs are randomly reassigned to samples. Specifically, for each sample, OTUs were picked at probabilities corresponding to their mean relative abundances in the regional pool, while the number of OTUs assigned to each sample remained unchanged. This approach has been previously suggested[81] as a means for comparing community overlaps to the expectation based on random colonization, while accounting for α diversities in the local communities and γ diversity in the regional pool. The only difference to the model used previously[81] is that here we assign OTUs to samples based on their mean relative abundances in the regional pool, rather than their detection frequencies, to account for low community evenness and to account for the fact that rare OTUs are less likely to colonize a site than highly abundant ones. Because the null model is sensitive to variation in sampling effort[81], we rarefied all samples at the maximum possible equal sequencing depth without replacement, after omitting samples that had fewer than 100 sequences in the considered functional group. The statistical significance of an observed MJO was determined as described earlier, based on 1,000 iterations. All MJOs were again found to be significantly lower than expected by the null model ($P<0.05$), although for some functional groups significances were weaker than when using the individual-based null model (Supplementary Table 1). The corresponding mean Raup–Crick dissimilarities, which quantify the deviation of sample overlaps from the null model[81], are provided in Supplementary Table 1. The same analysis was also applied to the entire community, again revealing a significantly low MJO ($P<0.001$).

Comparing functional and taxonomic variability. To compare the degree of functional variability versus taxonomic variability within functional groups, we examined the coefficients of variation (CVs; that is, the standard deviation divided by the mean) of relative gene group abundances on the one hand, and the CVs of OTU proportions within individual functional groups, on the other hand. Because each particular functional group contained multiple OTUs, we averaged the CV over all OTUs within the functional group. We note that the considered gene groups (Fig. 2a) only cover a small fraction of the total detected gene pool (~5% of annotated metagenomic sequences). Hence, to minimize the dependence of the CV of any particular gene group on the choice and coverage of other considered gene groups, we considered gene group abundances relative to the total number of annotated metagenomic sequences in each sample. An overview of CVs is provided in Supplementary Table 2. Observe that OTU CVs are generally an order of magnitude higher than gene group CVs, consistent with our conclusions based on overlap indices. We also calculated the CVs of relative OTU abundances at the community level and found a similarly high mean CV as within the functional groups (Supplementary Table 2).

Comparing OTU co-occurrences to a null model. To examine whether OTU co-occurrences across samples follow non-random patterns (for example, resulting from competitive exclusion), we considered a statistical quantity known

as the checkerboard score (*C* score) of the OTU presence-absence matrix[32]. The *C* score is defined as

$$C = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{i-1} (N_i - N_{ij})(N_j - N_{ji}) \quad (1)$$

where M is the total number of considered OTUs, N_i is the number of samples containing OTU i and N_{ij} is the number of samples containing both OTUs i and j. Hence, for fixed N_i, the C score becomes larger if species co-occur less frequently (that is, N_{ij} values are smaller). To assess whether an observed C score was likely to be due to chance (that is, if OTUs occur independently of each other), we compared it to the C score distribution of random presence–absence matrices generated under a null model, described in detail below. If random C scores generated by the null model are mostly below the observed C score, this would mean that OTUs tend to exclude each other more often than expected by chance (that is, OTUs are segregated). We calculated the C score and its deviation from the null model separately for each functional group.

We used the fixed–fixed null model to generate randomized versions of the original presence–absence matrix[32]. Specifically, the null model shuffles the cells of the matrix, while preserving the total number of samples containing each OTU as well as the number of OTUs present in each sample and in each functional group. The fixed–fixed null model was previously found particularly suitable for detecting non-random co-occurrence patterns across 'island lists', where islands can have different sizes and thus species–area relationships may lead to strong differences in the number of species detected in each island[82]. This null model is thus suitable for detecting non-random co-occurrence patterns across samples that may differ in terms of OTU richness or sequencing depth, while maintaining a low false positive error rate[82]. Randomized presence–absence matrices corresponding to the null model were generated using the 'curveball' algorithm[83]. We used 1,000 random matrices to asses the statistical significance of C scores. We also applied the same analysis to the entire community. An overview of results is given in Supplementary Table 3.

Comparing OTU co-abundances to a null model. Previous work suggests that abundance-based null models of species covariation may be more powerful than presence–absence-based null models for detecting OTU segregation or aggregation[33,84]. Hence, to test the robustness of our conclusions from the co-occurrence analysis described above, we also performed null model analysis using OTU abundances. Specifically, for each functional group we considered a statistical measure of meta-community overlap known as the generalized Morisita similarity index[85], henceforth referred to as the MA score[33]. The MA score is defined as follows:

$$\mathrm{MA} = \frac{\sum_{i=1}^{M} \left[\left(\sum_{j=1}^{S} p_{ij} \right)^2 - \sum_{j=1}^{S} p_{ij}^2 \right]}{(S-1) \sum_{i=1}^{M} \sum_{j=1}^{S} p_{ij}^2} \quad (2)$$

where S is the number of samples and p_{ij} is the proportion of OTU i within the considered functional group in sample j. Hence, lower values of the MA score indicate a lower similarity between samples in terms of OTU proportions and thus a potential segregation between OTUs. We compared the MA score to the MA score distribution of random abundance matrices generated under the so-called IT null model, as suggested previously[33]. This null model randomly assigns sequences to matrix cells proportional to the total number of sequences in each sample (within the functional group considered) and proportional to the total number of sequences assigned to each OTU across samples, until the total number of sequences per sample and per OTU is reached. The null model thus accounts for potential differences in sequencing depth between samples, and exhibits good power for detecting segregation or aggregation while maintaining a low type I error rate[33]. We used 1,000 random abundance matrices to asses the statistical significance of MA scores. When compared to the null model, all functional groups showed a significantly low MA score, suggesting segregation between OTUs ($P < 0.001$; Supplementary Table 4). When we applied the same analysis to the entire community, we again detected a highly significant segregation.

Note that the standardized effect sizes and statistical significances found here are greater than typically found for larger organisms, such as plants and animals[33], because the high number of sequences (when compared with typical plant or animal counts) leads to a smaller variance of the random MA scores under the null model. This 'law of large numbers effect' is absent in the presence–absence-based analysis described in the previous section, which may explain why patterns detected therein were less significant.

Phylogenetic dispersion. To assess whether community assembly within functional groups was neutral with respect to phylogenetic relationships, we examined the phylogenetic distances between functionally similar OTUs co-occurring in the same samples. The phylogenetic distance (PD) between any two OTUs was calculated as the sum of branch lengths needed to traverse the phylogenetic tree from one OTU to the other. For each functional group, we calculated the mean phylogenetic distance (MPD) between co-occurring OTUs as follows:

$$\mathrm{MPD} = \frac{\sum_{j=1}^{S} \sum_{i=1}^{M} \sum_{k=1}^{i-1} d_{ik} p_{ij} p_{kj}}{\sum_{j=1}^{S} \sum_{i=1}^{M} \sum_{k=1}^{i-1} p_{ij} p_{kj}}, \quad (3)$$

where d_{ik} is the phylogenetic distance between OTUs i and k. Note that this definition of MPD is almost equivalent to the 'phylogenetic diversity' of a single community introduced previously[86], with the difference that previously d_{ik} was defined as the divergence time between two OTUs (which is half of their phylogenetic distance in most cases) and here the average is taken at the metacommunity level (that is, across multiple samples). Samples with fewer than two OTUs within the considered functional group were omitted.

For each functional group, the MPD was compared with the expected MPD ($\overline{\mathrm{MPD}}$) under the null model of random phylogenetic relationships between OTUs within each sample. The distribution of MPDs under the null model was estimated by randomly and repeatedly permuting OTUs in the phylogenetic tree 1,000 times, while keeping their proportions in each sample fixed. OTUs were permuted independently for each sample, and permutations were restricted to OTUs within the same functional group. The standardized effect size (SES) of the MPD—which quantifies the deviation of the observed MPD from the expectation of the null model, was calculated as:

$$\mathrm{SES} = \frac{\mathrm{MPD} - \overline{\mathrm{MPD}}}{\sigma_{\mathrm{MPD}}} \quad (4)$$

where σ_{MPD} is the standard deviation of random MPDs generated under the null model. Hence, a strongly positive or strongly negative SES corresponds to strong phylogenetic overdispersion or underdispersion, respectively. The statistical significance of the SES was defined as the probability that the null model would yield an SES at least as large (in magnitude) as observed. The same analysis was also performed at the community level. The SESs and statistical significances are summarized in Supplementary Table 5.

Comparing OTU detection frequencies to a mechanistic neutral model. The above null models generate presence–absence matrices or abundance matrices that may be interpreted as random or neutral, however, they remain *ad hoc* heuristic randomization algorithms that lack a biological mechanism. Hence, the deviations from these null models observed here could in principle be due to their non-realistic representation of actually occurring neutral processes. A widely used mechanistic (that is, process-based) model for neutral microbial community assembly is the Sloan neutral model[37]. This model assumes that all OTUs are equivalent, and that the abundances of individual OTUs in each community are solely driven by stochastic birth–death events (leading to demographic drift) and random immigration from a regional pool (the source)[37,38]. Given the mean relative abundances of OTUs in the source and a fixed sequencing depth for each community, the Sloan model predicts (in a probabilistic sense) the frequency at which each OTU would be detected in a set of local communities. The Sloan model includes a single free immigration parameter for the entire metacommunity, which accounts for the importance of immigration compared to local demographic processes. This model is similar to the Hubbell neutral model[87], but was developed specifically for microbial communities (which typically exhibit high cell densities) and for relative abundance data from molecular techniques. Note that speciation is not included in the Sloan neutral model. Speciation probably contributes negligibly to the observed community variation (that is, at 99% OTU similarity) at the spatial (~100 m) and temporal (~1 year) scales relevant to this study[88,89].

To compare the OTU composition within functional groups to the Sloan neutral model we proceeded as follows, separately for each functional group. We first estimated the mean abundances of OTUs in the source based on their proportions in all samples, as suggested previously[37]. We then rarefied each sample at a constant depth without replacement. The rarefaction depth was chosen to be the maximum possible, after omitting samples containing fewer than 100 sequences within the functional group. The detection frequency for each OTU was set to the number of rarefied samples containing the OTU. We fitted the immigration parameter by maximizing the likelihood of the observed OTU detection frequencies (approximated as the product of likelihoods of the binomial distributions for all OTUs). This calibration method is known as maximum likelihood (ML) estimation and is widely established in statistical regression and physics[90]. A comparison of the fitted models and data is provided in Supplementary Fig. 5. The same approach as the above was also used for the entire community (Supplementary Fig. 11).

As mentioned before, the Sloan neutral model makes predictions about the probabilities at which each OTU would be detected at various frequencies in the metacommunity (this probability distribution is a binomial distribution[38]). The deviation of an observed OTU detection frequency from the corresponding expectation of the model is usually assessed in terms of the probability that such a strong deviation could occur if the fitted model was true[91,92]. Here we found that about 35–60% of OTUs deviated significantly ($P < 0.05$) from the expectation of the fitted model, depending on the functional group considered (Supplementary Table 6). At the entire community level, 56% of OTUs deviated significantly from the neutral model. These fractions are much higher than the 5% type I error rate than would be expected under the null model, indicating that the distribution of a substantial fraction of OTUs may be driven by non-neutral processes, both at the community levels as well as within functional groups. Note that no statement can be made about the mechanisms influencing the remaining OTUs, apart from the fact that their detection frequencies are within the 95% confidence interval of the fitted model.

Note that in some cases the fitted models roughly reproduce the observed OTU detection frequencies as a function of their mean relative abundances (Supplementary Figs 5 and 11). This does not, however, constitute evidence that assembly within functional groups is driven by neutral processes because a positive relationship between mean OTU abundances and OTU detection frequencies, a central prediction of the neutral model, is not unusual for meta-communities even if non-neutral mechanisms dominate community assembly[93]. Without knowledge of these specific mechanisms, we can reject the null hypothesis of neutrality if the data deviate significantly from the model predictions, but not vice versa. To assess the overall consistency of our data with the fitted model, we compared it to hypothetical data simulated according to the model itself, as follows. For each sample j and for each functional group, we chose OTU proportions randomly according to a Dirichlet distribution:

$$x_{1j}, .., x_{Mj} \sim \mathrm{Dir}(mN_T p_1, .., mN_T p_M) \quad (5)$$

where mN_T is the fitted immigration parameter, M is the total number of OTUs assigned to the functional group, p_i is the mean relative abundance of the ith OTU across samples and x_{1j} is its simulated relative abundance in the jth sample[38]. Given these simulated proportions, we then randomly assigned R sequences to any of the M OTUs according to a multinomial distribution with probabilities $x_{1j}.., x_{Mj}$, where R is the rarefaction depth described above. This yielded a random abundance matrix that is comparable in structure to the observed (rarefied) abundance matrix but whose entries are randomly distributed according to the fitted Sloan model. Using this random abundance matrix, we calculated the corresponding mean relative OTU abundances and detection frequencies, and compared these to the fitted neutral model in terms of their likelihood as well as their coefficient of determination (R^2). Note that, for a fixed number of OTUs and samples, a higher likelihood and a higher R^2 indicate a better agreement with the model. Here we found that the likelihoods and the R^2 of the simulated data were much higher than the likelihood and the R^2 of the actual data that were used to fit the model, regardless of the functional group considered ($P < 0.001$ based on 1,000 simulations; Supplementary Table 6). This further indicates that the Sloan neutral model is a poor description for the composition within functional groups. The same test was also applied at the community level and yielded similar results (Supplementary Table 6).

Comparing OTU proportions to environmental variables. To assess whether and how the taxonomic variation within functional groups can be attributed to environmental conditions, we constructed multivariate regression models for each OTU in each functional group, using 21 environmental variables as potential predictors (an overview of the environmental variables is presented in Supplementary Table 7). We used generalized linear models with a logit link function and a binomial distribution[94] to account for the fact that OTU proportions are estimated via discrete counts at finite sequencing depth and the fact that OTU proportions only take values between 0 and 1. Specifically, the number of sequences assigned to a particular OTU was modeled as a binomial distribution, whose number of trials was equal to the total number of sequences assigned to the considered functional group in each sample, and whose probability of success was the logistic function of a multivariate linear function of environmental predictors. The linear model coefficients, for any particular set of predictors, were chosen via ML estimation[90], as implemented by the Statsmodels package[95].

For each OTU we independently chose the appropriate subset of environmental variables based on the achievable cross-validated coefficient of determination (R^2_{CV}), which represents the coefficient of determination when only a random training subset (90%) of the samples are used for fitting and the remaining test subset (10%) is used to evaluate the fitted model[96]. The R^2_{CV} is typically used to assess the risk of data overfitting and inaccurate extrapolation[96], and provides a more conservative estimate of a model's predictive power than the classical coefficient of determination (R^2). For each potential subset of predictors, we estimated the R^2_{CV} using tenfold Monte Carlo cross-validation with 100 random iterations[96]. The appropriate subset of predictors was chosen in a forward stepwise manner, that is, by successively adding the predictor that maximized the R^2_{CV} in the enlarged model[97]. At each step, if the enlarged model exhibited a lower R^2_{CV} than before, the process was aborted and the previous subset of predictors was used. This process ensured that for each OTU only the most important environmental predictors were used for regression, thereby avoiding overfitting. A model's R^2_{CV} was taken as a measure for how well environmental conditions predicted the proportion of a specific OTU within a specific functional group. Consequently, the distribution of R^2_{CV} across all OTUs in a functional group provides an overview for how well environmental conditions predict the group's overall composition (Fig. 4b).

The above regression analysis enabled assessment of the overall predictive power of the environmental variables. The relative importance of specific environmental variables, however, remains unclear because each OTU was regressed on a different subset of variables that was chosen to have the best predictive power for the particular OTU. To get a general understanding of which environmental variables may be particularly relevant to the overall composition within functional groups, we used redundancy analysis (RDA)[39]. RDA was performed separately for each functional group and for each environmental variable, by regressing all OTU proportions within the functional group simultaneously against the environmental variable (all-against-one), and then calculating the fraction of variance explained by the constrained axis. This fraction provides a measure for how well a specific environmental variable explained the overall taxonomic variation within a specific functional group (Fig. 4a). RDA was performed using the scikit-bio package (v. 0.4.0). We note that in principle one could perform RDA using all environmental variables at the same time, to assess their overall explanatory power and their individual importance based on, for example, their regression coefficients. In our case, however, such an all-against-all approach would be inappropriate because the number of samples (22) would be close to the number of explanatory variables (21), and hence multiple ordinary linear regression (which underlies RDA) would grossly overfit our data. Our assessment of the overall predictability of community composition was thus performed separately using the more sophisticated regression described above.

Comparing dissimilarities to geographical and environmental distances. Pairwise dissimilarities between taxonomic community profiles reported here were calculated using the Bray–Curtis metric[39], based on OTU proportions within individual functional groups. Other dissimilarity metrics (Canberra and Hellinger) yielded similar conclusions, so they are not further discussed here. Before calculating dissimilarities, we rarefied all samples at the maximum possible equal sequencing depth within the considered functional group. Samples with fewer than 100 sequences within the functional group were omitted. The same approach was used to calculate dissimilarities at the community level.

To examine whether Bray–Curtis dissimilarities between communities were correlated to geographical distances between bromeliads, we used Mantel correlation tests to calculate correlations and their statistical significances[39]. Specifically, within each functional group, we calculated Spearman rank correlations between all pairwise dissimilarities and geographical distances. The statistical significance of correlations was estimated using 1,000 random permutations of the rows and columns in the geographical distance matrix (rows and columns permuted similarly). Out of the nine functional groups, only aerobic chemoheterotrophs displayed a significant correlation with geographical distance ($P = 0.016$; Supplementary Fig. 6). When we considered the entire community, no significant correlation was found (Fig. 5a).

Variation partitioning of OTU composition. To assess what fractions of the variation in OTU composition within functional groups are explained by purely environmental conditions or by purely spatial structure (for example, due to spatial dispersal limitation), we used variation partitioning[40,98]. Specifically, for each functional group we performed multiple linear RDA of the OTU proportions using either solely spatial variables (longitude, latitude or polynomial combinations up to third order) or solely environmental variables (Supplementary Table 7), or using both spatial and environmental variables together. In each case we quantified the fraction of explained variation based on the adjusted coefficient of determination (R^2_{adj}, adjusted to account for the number of predictor variables) according to Wherry's formula[99]. Variations explained solely by spatial or solely by environmental variables, that is, while controlling for environmental or spatial variables respectively, were calculated via basic arithmetic operations of the R^2_{adj} as described previously[40]. To reduce the risk of overfitting, only a subset of available spatial and environmental variables was used as predictors in each RDA model. Concretely, predictors were chosen separately for each functional group, using a step-wise selection algorithm that optimized the cross-validated coefficient of determination at each step, as described above for the generalized linear models. An overview of selected predictors is provided in Supplementary Table 10. An overview of explained variances is provided in Fig. 5. The same analysis was also performed at the community level.

Data availability. Molecular sequence data reported in this paper have been deposited in the NCBI Sequence Read Archive (SRX1757104 to SRX1757125

and SRX1757435 to SRX1757456), as part of BioProject PRJNA321235 (NCBI Bio-Project database http://www.ncbi.nlm.nih.gov/bioproject; SRA accession SRP074855). Environmental metadata are included with the corresponding BioSamples (SRS1433623 to SRS1433644; http://www.ncbi.nlm.nih.gov/biosample). Functional annotations of prokaryotic taxa are available as Supplementary Data.

Received 3 May 2016; accepted 30 September 2016; published 5 December 2016

References

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
2. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).
3. Zaikova, E. *et al.* Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ. Microbiol.* **12,** 172–191 (2010).
4. Powell, J. R. *et al.* Deterministic processes vary during community assembly for ecologically dissimilar taxa. *Nat. Commun.* **6,** 8444 (2015).
5. Strom, S. L. Microbial ecology of ocean biogeochemistry: a community perspective. *Science* **320,** 1043–1045 (2008).
6. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348,** 1262073 (2015).
7. Ofijeru, I. D. *et al.* Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl Acad. Sci. USA* **107,** 15345–15350 (2010).
8. Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. Bacterial community assembly based on functional genes rather than species. *Proc. Natl Acad. Sci. USA* **108,** 14288–14293 (2011).
9. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4,** 102–112 (2006).
10. Raes, J., Letunic, I., Yamada, T., Jensen, L. J. & Bork, P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7,** 473 (2011).
11. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353,** 1272–1277 (2016).
12. Nelson, M. B., Martiny, A. C. & Martiny, J. B. H. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc. Natl Acad. Sci. USA* **113,** 8033–8040 (2016).
13. Fukami, T., Martijn Bezemer, T., Mortimer, S. R. & Putten, W. H. Species divergence and trait convergence in experimental plant community assembly. *Ecol. Lett.* **8,** 1283–1290 (2005).
14. Helsen, K., Hermy, M. & Honnay, O. Trait but not species convergence during plant community assembly in restored semi-natural grasslands. *Oikos* **121,** 2121–2130 (2012).
15. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95,** 6578–6583 (1998).
16. Lande, R., Engen, S. & Saether, B. *Stochastic Population Dynamics in Ecology and Conservation* (Oxford Univ. Press, 2003).
17. Fernandez, A. *et al.* How stable is stable? Function versus community composition. *Appl. Environ. Microbiol.* **65,** 3697–3704 (1999).
18. Vanwonterghem, I. *et al.* Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME J.* **8,** 2015–2028 (2014).
19. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
20. Prosser, J. I. Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbialecology. *Nat. Rev. Microbiol.* **13,** 439–446 (2015).
21. Goffredi, S. K., Kantor, A. H. & Woodside, W. T. Aquatic microbial habitats within a neotropical rainforest: bromeliads and pH-associated trends in bacterial diversity and composition. *Microb. Ecol.* **61,** 529–542 (2011).
22. Farjalla, V. F. *et al.* Ecological determinism increases with organism size. *Ecology* **93,** 1752–1759 (2012).
23. Srivastava, D. S. *et al.* Are natural microcosms useful model systems for ecology? *Trends Ecol. Evol.* **19,** 379–384 (2004).
24. Martinson, G. O. *et al.* Methane emissions from tank bromeliads in neotropical forests. *Nat. Geosci.* **3,** 766–769 (2010).
25. Goffredi, S. K., Jang, G., Woodside, W. T. & Ussler, W. Bromeliad catchments as habitats for methanogenesis in tropical rainforest canopies. *Front. Microbiol.* **2,** 256 (2011).
26. Marino, N. A. C., Srivastava, D. S. & Farjalla, V. F. Aquatic macroinvertebrate community composition in tank-bromeliads is determined by bromeliad species and its constrained characteristics. *Insect Cons. Div.* **6,** 372–380 (2013).
27. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12,** 635–645 (2014).
28. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
29. Canfield, D. E. & Thamdrup, B. Towards a consistent classification scheme for geochemical environments, or, why we wish the term 'suboxic' would go away. *Geobiology* **7,** 385–392 (2009).
30. Atwood, T. B. *et al.* Predator-induced reduction of freshwater carbon dioxide emissions. *Nat. Geosci.* **6,** 191–194 (2013).
31. Chase, J. M. & Myers, J. A. Disentangling the importance of ecological niches from stochastic processes across scales. *Phil. Trans. R. Soc. Lond. B* **366,** 2351–2363 (2011).
32. Gotelli, N. J. Null model analysis of species co-occurrence patterns. *Ecology* **81,** 2606–2621 (2000).
33. Ulrich, W. & Gotelli, N. J. Null model analysis of species associations using abundance data. *Ecology* **91,** 3384–3397 (2010).
34. Ulrich, W. Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules. *Oikos* **107,** 603–609 (2004).
35. Horner-Devine, M. C. & Bohannan, B. J. M. Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* **87,** S100–S108 (2006).
36. Pausas, J. G. & Verdu, M. The jungle of methods for evaluating phenotypic and phylogenetic structure of communities. *Bioscience* **60,** 614–625 (2010).
37. Sloan, W. T. *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8,** 732–740 (2006).
38. Sloan, W. T., Woodcock, S., Lunn, M., Head, I. M. & Curtis, T. P. Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb. Ecol.* **53,** 443–455 (2007).
39. Legendre, P. & Legendre, L. *Developments in Environmental Modelling* 2nd edn (Elsevier, 1998).
40. Legendre, P. Studying beta diversity: ecological variation partitioning by multiple regression and canonicalanalysis. *J. Plant Ecol.* **1,** 3–8 (2008).
41. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
42. Shapiro, O. H. & Kushmaro, A. Bacteriophage ecology in environmental biotechnology processes. *Curr. Opin. Biotechnol.* **22,** 449–455 (2011).
43. Carrias, J.-F. *et al.* Two coexisting tank bromeliads host distinct algal communities on a tropical inselberg. *Plant Biol.* **16,** 997–1004 (2014).
44. Langenheder, S., Lindstrom, E. S. & Tranvik, L. J. Weak coupling between community composition and functioning of aquatic bacteria. *Limnol. Oceanogr.* **50,** 957–967 (2005).
45. Strickland, M. S., Lauber, C., Fierer, N. & Bradford, M. A. Testing the functional significance of microbial community composition. *Ecology* **90,** 441–451 (2009).
46. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci. USA* **109,** 21390–21395 (2012).
47. Vanwonterghem, I., Jensen, P. D., Rabaey, K. & Tyson, G. W. Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion. *Environ. Microbiol.* **18,** 3144–3158 (2016).
48. Reed, D. C. *et al.* Predicting the response of the deep-ocean microbiome to geochemical perturbations by hydrothermal vents. *ISME J.* **9,** 1857–1869 (2015).
49. Nazareno, A. G. & Laurance, W. F. Brazil's drought: beware deforestation. *Science* **347,** 1427–1427 (2015).
50. Golterman, H. Clymo, R. & Ohnstad, M. *Methods for Physical and Chemical Analysis of Freshwaters* 2nd edn, 169 (IBP Handbook No. 8, Blackwell 1978).
51. Zagatto, E., Jacintho, A., Mortatti, J. & Bergamin, F. H. An improved flow injection determination of nitrite in waters by using intermittent flows. *Anal. Chim. Acta* **120,** 399–403 (1980).
52. Fofonoff, N. P. & Millard-Junior, R. *Algorithms for Computation of Fundamental Properties of Seawater*. UNESCO technical paper in marine science 44 (UNESCO, 1983).
53. Andrade-Eiroa, A., Canle, M. & Cerda, V. Environmental applications of excitation-emission spectrofluorimetry: an in-depth review II. *Appl. Spec. Rev.* **48,** 77–141 (2013).
54. Murphy, K. R., Stedmon, C. A., Graeber, D. & Bro, R. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal. Meth.* **5,** 6557–6566 (2013).
55. Stedmon, C. A., Markager, S. & Bro, R. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Mar. Chem.* **82,** 239–254 (2003).
56. Stedmon, C. A. & Bro, R. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnol. Oceanogr. Meth.* **6,** 572–579 (2008).
57. Murphy, K. R., Stedmon, C. A., Wenig, P. & Bro, R. Openfluor—an online spectral library of auto-fluorescence by organic compounds in the environment. *Anal. Meth.* **6,** 658–661 (2014).
58. Jørgensen, L. *et al.* Global trends in the fluorescence characteristics and distribution of marine dissolved organic matter. *Mar. Chem.* **126,** 139–148 (2011).
59. Murphy, K. R. *et al.* Organic matter fluorescence in municipal water recycling schemes: Toward a unified PARAFAC model. *Environ. Sci. Technol.* **45,** 2909–2916 (2011).

60. Osburn, C. L., Wigdahl, C. R., Fritz, S. C. & Saros, J. E. Dissolved organic matter composition and photoreactivity in prairie lakes of the U.S. Great Plains. *Limnol. Oceanogr.* **56,** 2371–2390 (2011).
61. Yamashita, Y., Boyer, J. N. & Jaffe, R. Evaluating the distribution of terrestrial dissolved organic matter in a complex coastal ecosystem using fluorescence spectroscopy. *Cont. Shelf Res.* **66,** 136–144 (2013).
62. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
63. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
64. Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.* **13,** 656–668 (2012).
65. Gevers, D. *et al.* Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3,** 733–739 (2005).
66. Martiny, A. C., Tai, A. P., Veneziano, D., Primeau, F. & Chisholm, S. W. Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environ. Microbiol.* **11,** 823–832 (2009).
67. Koeppel, A. F. & Wu, M. Species matter: the role of competition in the assembly of congeneric bacteria. *ISME J.* **8,** 531–540 (2014).
68. Keswani, J. & Whitman, W. B. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. *Int. J. Syst. Evol. Microbiol.* **51,** 667–678 (2001).
69. Stackebrandt, E. & Ebers, J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today* **33,** 152 (2006).
70. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
71. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
72. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26,** 266–267 (2010).
73. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).
74. Kubo, K. *et al.* Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J.* **6,** 1949–1965 (2012).
75. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).
76. Hahn, A., Hanson, N., Kim, D., Konwar, K. & Hallam, S. Assembly independent functional annotation of short-read data using SOFA: short-ORF functional annotation. In *IEEE Conf. Comput. Intel. Bioinformatics Comput. Biol.* 1–6 (IEEE, 2015).
77. Konwar, K. M. *et al.* MetaPathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics* **31,** 3345–3347 (2015).
78. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42,** D553–D559 (2014).
79. Wolda, H. Similarity indices, sample size and diversity. *Oecologia* **50,** 296–302 (1981).
80. Hester, E. R., Barott, K. L., Nulton, J., Vermeij, M. J. & Rohwer, F. L. Stable and sporadic symbiotic communities of coral and algal holobionts. *ISME J.* **10,** 1157–1169 (2015).
81. Chase, J. M., Kraft, N. J. B., Smith, K. G., Vellend, M. & Inouye, B. D. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. *Ecosphere* **2,** 1–11 (2011).
82. Connor, E. F. & Simberloff, D. The assembly of species communities: chance or competition? *Ecology* **60,** 1132–1140 (1979).
83. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5,** 4114 (2014).
84. Hausdorf, B. & Hennig, C. Null model tests of clustering of species, negative co-occurrence patterns and nestedness in meta-communities. *Oikos* **116,** 818–828 (2007).
85. Chao, A., Jost, L., Chiang, S., Jiang, Y.-H. & Chazdon, R. L. A two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* **64,** 1178–1186 (2008).
86. Chave, J., Chust, G. & Thebaud, C. in *Scaling Biodiversity* (eds Storch, D. *et al.*) 150–167 (Cambridge Univ. Press, 2007).
87. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* Vol. 32 (Princeton Univ. Press, 2001).
88. Ochman, H. & Wilson, A. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26,** 74–86 (1987).
89. Kuo, C.-H. & Ochman, H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol. Direct* **4,** 35 (2009).
90. Eliason, S. R. *Maximum Likelihood Estimation: Logic and Practice* (SAGE, 1993).
91. Burns, A. R. *et al.* Contribution of neutral processes to the assembly of gut microbial communities in thezebrafish over host development. *ISME J.* **10,** 655–664 (2015).
92. Venkataraman, A. *et al.* Application of a neutral community model to assess structuring of the human lung microbiome. *mBio* **6,** e02284-14 (2015).
93. Holyoak, M., Leibold, M. & Holt, R. *Metacommunities: Spatial Dynamics and Ecological Communities* (Univ. Chicago Press, 2005).
94. McCullagh, P. & Nelder, J. *Generalized Linear Models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability* 2nd edn (Taylor & Francis, 1989).
95. Seabold, J. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python Sci. Conf.* (eds Jones, E. & Millman, J.) 57–61 (SciPy, 2010).
96. Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **88,** 486–494 (1993).
97. *P.S.A.S. SAS/STAT 9.1 User's Guide: The REG Procedure* (SAS Institute, 2008).
98. Borcard, D., Legendre, P. & Drapeau, P. Partialling out the spatial component of ecological variation. *Ecology* **73,** 1045–1055 (1992).
99. Field, A., Miles, J. & Field, Z. *Discovering Statistics Using R* (SAGE, 2012).

Acknowledgements

We thank M. Chen for help with the molecular work. We thank A. L. Gonzalez and A. MacDonald for discussions and comments on our paper. We thank T. Benevides for helping with the absorption measurements. S.L. acknowledges the financial support of the Department of Mathematics, University of British Columbia. S.L. and M.D. acknowledge the support of Natural Sciences and Engineering Research Council (NSERC). V.F.F. is grateful to the Brazilian Council for Research, Development and Innovation (CNPq) for research funds (Pesquisador Visitante Especial, PVE, Research Grant 400454/2014-9) and productivity grants. S.M.S.J. acknowledges the post-graduate scholarship provided by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). J.S.L. acknowledges the financial support of Coordenacao de Aperfeicoamento de Pessoal de Ensino Superior (CAPES). We thank M. P. F. Barros, A. R. Soares, J. L. Nepomuceno and their research groups of the Nucleus of Ecology and Socio-Environmental Development of Macae (NUPEM/UFRJ) for proving field and laboratory assistance during the samplings.

Author contributions

V.F.F., S.L., S.M.S.J., A.P.F.P. and J.S.L. performed the field work. V.F.F. and S.M.S.J. performed the chemical measurements in the laboratory. S.L. performed the molecular work in the laboratory, the DNA sequence analysis and the statistical analyses. S.L., M.D., V.F.F., D.S.S. and L.W.P. interpreted the statistical findings. S.L. wrote a first draft of the manuscript, and all authors contributed to the final preparation of the manuscript. M.D. and V.F.F. supervised the project.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.L.

How to cite this article: Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nat. Ecol. Evol.* **1,** 0015 (2016).

Competing interests

The authors declare no competing interests.