Ecosystem distribution profiling of bacteria from a unique hypersaline sediment (sabkha) reveals ecological specialization among communities in the environment

Abdul-Matiin Wan [1] , **Antonio M. Martin-Platero** [2] , **Areej Alsheikh-Hussain** [3] , **Syafiq Kamarul Azman** [1] , **Lina F. Yousef** [1] , **Andreas Henschel** Corresp. [1]

[1] Masdar Institute, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates

[2] Department of Microbiology, University of Granada, Granada, Spain

[3] School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Australia

Corresponding Author: Andreas Henschel
Email address: ahenschel@masdar.ac.ae

Advances in genomic sequencing technologies resulted in massive microbial diversity data (16S ribosomal gene sequences, rDNA) being generated in every possible environment. However, the majority of microorganisms have never been cultured, and therefore, nor cataloged. This poses a problem for molecular microbial ecologists because a large portion of the marker sequences can not be taxonomically resolved past the phylum taxon level. This tells very little about who or what these microorganisms are doing in relation to their environment. Our study describes an approach to assist in drawing ecological information from a sample when the taxon resolution is poor. We generated 16S rDNA libraries from a hypersaline marine sediment (coastal Sabkha) and saline mangrove soil in Abu Dhabi and then compared the compositional features to a database of 20,470 publicly available microbial community profiles (comprising the entire Earth Microbiome Project, EMP) that were annotated with terms from the Environmental Ontology (EnvO). An accurate taxonomic classification was not possible for 80% of the Sabkha operational taxonomic units (OTUs) beyond phylum level with widely used taxonomy classification tools, but habitat profiling performed on the community revealed strong links to bacterial assemblages of soil and marine origins. To capture the notion of generalist vs. specialist formally, we developed an algorithm to derive empirical probability distributions of OTUs over ecosystems from observed occurrences in the sample database, which then give rise to OTU-specific ecosystem entropies. We observed very low average ecosystem entropy of the Sabkha in contrast to other environmental samples. Based on this concept, the Sabkha community, while of midrange alpha diversity, presented largely specialist characteristics, with most OTUs identified to be unique to the Sabkha habitat. This finding is further corroborated by the observation that the Sabkha sample is unique with respect to the EMP-derived dataset (which contains 74 hypersaline and thousands of marine samples), as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3286v1 | CC BY 4.0 Open Access | rec: 26 Sep 2017, publ: 26 Sep 2017

a comprehensive UniFrac similarity search did not yield any significant matches. Finally, we show that the ecosystem entropy formalism, which intrinsically accounts for the ability of OTUs to cross ecosystem borders according to a context database, is a novel, informative tool to describe and identify extreme environments in addition to conventional ecological diversity measures.

Ecosystem distribution profiling of bacteria from a unique hypersaline sediment (sabkha) reveals ecological specialization among communities in the environment

Abdul-Matiin Wan[1], Antonio M. Martin-Platero[2], Areej Alsheikh-Hussain[3], Syafiq Kamarul Azman[1], Lina F. Yousef[1], and Andreas Henschel[1]

[1]Khalifa University of Science and Technology, Masdar Institute, Abu Dhabi, United Arab Emirates
[2]University of Granada, Department of Microbiology, Spain
[3]University of Queensland, School of Chemistry and Molecular Biosciences, Brisbane, Australia

Corresponding author:
Andreas Henschel[1]

Email address: ahenschel@masdar.ac.ae

ABSTRACT

Advances in genomic sequencing technologies resulted in massive microbial diversity data (16S ribosomal gene sequences, rDNA) being generated for samples from wide-ranging environments. However, the majority of microorganisms have never been cultured, and therefore, are not reflected in current public databases. This poses a problem for molecular microbial ecologists because a large portion of the marker sequences can not be taxonomically resolved past the phylum taxon level. This tells very little about who or what these microorganisms are doing in relation to their environment. Our study describes an approach to assist in drawing ecological information from a sample even when the taxon resolution is poor. We generated 16S rDNA libraries from a hypersaline marine sediment (coastal sabkha) and a moderately hypersaline mangrove soil in Abu Dhabi. Intuitively, our novel algorithm identifies for each OTU in a given community, where else it occurs (i.e., in which other ecosystems). This is facilitated by a comprehensive relational database of 20,470 publicly available microbial community profiles (comprising the entire Earth Microbiome Project, EMP) with Environmental Ontology (EnvO) annotations. Analysis performed on the sabkha community revealed strong links to bacterial assemblages of soil and marine origins. Formally, the developed algorithm derives empirical probability distributions of OTUs over ecosystems from observed occurrences in the sample database, which then give rise to OTU-specific ecosystem entropies. The results are visualized in a feature rich graph. We observed very low ecosystem entropies of the sabkha constituents in contrast to other (hyper-)saline samples, indicating specialist characteristics and/or genetic isolation. This finding is further corroborated by the observation that the sabkha sample is unique with respect to the EMP-derived dataset, as a comprehensive UniFrac similarity search did not yield any significant matches. Finally, we show that the ecosystem entropy formalism, which intrinsically accounts for the ability of OTUs to cross ecosystem borders according to a context database, is a novel, informative tool to describe extreme environments complementary to conventional ecological diversity measures.

INTRODUCTION

Over the recent years, the generation of large marker gene datasets has become more common in environmental research, owing to the plummeting cost of next-generation sequencing (NGS) and the emergence of more robust bioinformatics tools (Kim et al. (2013)). Joint efforts such as the Tara Ocean Expedition, Ocean Sampling Day, Malaspina Expedition, Earth Microbiome Project, and the Human Microbiome Project (Karsenti et al. (2011); Kopf et al. (2015); Duarte (2015); Gilbert et al. (2014); Turnbaugh et al. (2007); Yutin et al. (2007)) underline the growing recognition of metagenomic and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3286v1 | CC BY 4.0 Open Access | rec: 26 Sep 2017, publ: 26 Sep 2017

marker gene datasets as a key approach in representing and cataloguing whole or near-whole microbial communities within major environmental domains. The marker gene set approach has so far proven invaluable in studying environments or systems of interest through a more realistic representation of intrinsic community composition and dynamics than was possible with classic culture-based investigations alone Su et al. (2012). Community profile reporting is now considered an important aspect of habitat characterization, especially when it comes to understanding shifts in environments or systems of interest across time and space, as seen from the growing body of marker gene datasets collected at general sequence databases (Genomes OnLine Database (GOLD), GenBank Reddy et al. (2015); Benson et al. (2013); Schloss and Handelsman (2004)) and dedicated marker gene repositories (Ribosomal Database Project-RDP, SILVA and IMNGS, Cole et al. (2009b); Pruesse et al. (2007); Lagkouvardos et al. (2016)). This revolution in microbial ecology is only expected to forge ahead with continual improvement in the processing power (number of reads, depth) of sequencing technologies and computational analysis. With thousands of community profiles contributed to public repositories through efforts targeting whole genomes or genetic markers such as the 16S rRNA gene, the current challenge hence is in distilling meaningful information from this deluge of metagenomic data (Wooley et al. (2010)) towards advancing fundamental understanding of microbial diversity, biogeography and evolution across the planet. To date, the general framework of marker gene analysis primarily utilizes existing sequence data from studied microbial taxa (identified by bioinformatics tools as operational taxonomic units or ‘OTUs’) as a means of determining the phylogenetic diversity or function of studied communities. While the wealth of marker gene sequencing data provide a robust reference for community characterization under this general framework (with a significant number of OTUs identifiable up to the genus level), our ability to derive further information on the microbial community members (e.g., ‘how unique or rare is a particular bacterium?’ and ‘what environmental niche does it occupy?’) remains limited. Existing work on the less common, low abundance OTUs (corresponding to the ’rare biosphere’) has so far shed some light on the previously overlooked populations that offer further insight on microbial communities (Huse et al. (2010)). Gleaning such insight would require integrating existing metadata accompanying each metagenomic submission (source biome, geographic location, pH, etc.) into the existing framework to support a more contextualized analysis of communities. This is especially relevant in light of previous findings of ecological importance, such as the correlation between habitat conditions and genome size (Dini-Andreote et al. (2012)), the latitudinal gradient in marine bacteria distribution (Fuhrman et al. (2008)), and the taxonomic and functional distinction of desert soil bacteria against other nondesert biomes (Fierer et al. (2012)).

To address rarity and environmental niche occupation of microbial community members holistically, we considered an alternative strategy that taps into marker gene data and metadata to support the interpretation of global patterns in bacterial taxa distribution across different biomes. This can then be used to distinguish between different environmental samples based on their matching biome annotations. This approach seeks to address the aspect of biogeography in microbial ecology, which aims to reveal *where organisms live*, *at what abundance*, and *why* (Martiny et al. (2006)). Our interest is in achieving a higher-level analysis of microbial communities, moving beyond typical characterization (*who is there?*) towards understanding how a community’s ecology is related to the environmental distribution of its members. This study was designed to test the hypothesis that extreme environments select for unique microbes with a narrow range of environmental distribution (‘specialists’, here used in a more general sense wrt. observed ecosystem specificity), whereas more moderate environments would host microbes with a wider distribution (‘generalists’). Previous work investigating co-occurence patterns in soil microbes and the mechanism of environmental filtering across the terrestrial-freshwater gradient point to the potential of exploring associations between disparate communities (Barberán et al. (2012); Monard et al. (2016)), which we aim to enable at a greater scale. IMNGS is comparable to our work in that it is also capable to extract distribution patterns of community members, but requires computationally expensive sequence similarity searches and is conducted only at an individual level, without visualization of the ecosystem distribution. Our investigation involves the characterization of a microbial community from a vegetation-free, hypersaline tidal salt flat (‘sabkha’), and a grey mangrove (*Avicennia marina*) forest bed, followed by a comparison of the 16S rRNA gene libraries of these two distinctly different environments in Abu Dhabi, United Arab Emirates (UAE) against global saline samples. While true specialization in terms of genomic content can not be gleaned from 16S rRNA alone, the large number of available 16S rRNA libraries carry valuable information about the whereabouts of OTUs, which can serve as approximation for OTU specialization.

MATERIALS AND METHODS

Sequence data processing. We characterized the bacterial community of an intertidal sabkha site (N 24.146556; E 54.103194) that had previously been geochemically characterized by Bontognali et al. (2010). The site was uniformly covered with a halite layer, had no vegetation cover, and not flooded at the time of sampling. The top 10 cm layer was systematically sampled from 15 points across a 135-m^2 area, yielding a composite sample for DNA extraction, 16S Ribosomal DNA library preparation, and pair-end sequencing of 250 bases on the MiSEQ platform (Illumina; CA, USA) at the BioMicroCenter (MIT, Cambridge, MA), which produced 23,606 sequences. The mangrove forest bed sample was taken from N 24.450530; E 54.445002. Sample preparation and DNA sequencing was performed as above and yielded 46,875 amplicons. We perform 16S rRNA copy number correction as suggested by PICRUSt. We adhered to the 16S rRNA amplicon protocol recommended by the Earth Microbiome Project (Caporaso et al. (2012)), amplifying hypervariable region V4, using standard primers 515F - 806R. The 5' end fragments were analyzed using Quantitative Insights Into Microbial Ecology (QIIME 1.9) (Caporaso et al. (2010)), and closed reference OTU calling was completed using GreenGenes (DeSantis et al. (2006)) with 97% reference OTU collection (May 2013). We determined taxonomic ranks for OTU representatives using the Ribosomal Database Project (RDP version 2.2) classifier (Cole et al. (2009a)). Alpha diversity/phylogenetic distance (PD whole tree) with respect to the phylogeny that is provided by GreenGenes for its reference OTUs (clustered at 97% sequence identity) was calculated using the Qiime script `alpha_diversity.py`. All samples were rarefied to 18,000 sequences through 10-fold multiple rarefaction using QIIME's `multiple_rarefactions.py -n 10` (see Figure S10 for rarefaction curves).

Ecosystem distribution of OTUs For each OTU we identified the environments in which it occurs. To this end, we built a database of 20,472 16S rRNA profiles from 2,461 independent studies. The database contains a number of tables for sample information (meta data, sample size) as well as a table that relates sample event IDs to OTU IDs which has more than 13.5 million entries. We have indexed OTU IDs for fast retrieval of individual IDs. This table facilitates efficient OTU centric queries. The sources for the collection of profiles are a previous collection published in Chaffron et al. (2010) (henceforth referred to as Chaffron dataset), Qiime-DB/Qiita (which comprises the Earth Microbiome Project, though only as marker gene profiles, i.e. OTU abundances but no sequences) and the Sequence Read Archive (SRA). The details of the database content and construction are provided in Henschel et al. (2015). We would like to stress the suitability of this database to investigate saline/hypersaline samples such as marine sediments, as this context is represented by samples from various independent studies: 35 samples (containing at least 50 sequences) from 11 independent studies have been identified as hypersaline. Moreover, marine sediments feature prominently in our database. In total, the database contains 202 samples assigned to marine sediments from 12 independent studies. For details, please refer to Tables 1 and 2, respectively. The entire coverage of ecosystems is shown in Table For each profile, closed reference OTU calling was performed consistenly against the same reference as for the sabkha sample, GreenGenes 13.5 in consistency with the pre-picked marker gene profiles we acquired from Qiime-DB. Moreover, for all samples, we identified the ecosystem using the Environmental Ontology (EnvO) (Buttigieg et al. (2013, 2016)): EnvO (version 20-04-2012, `http://purl.bioontology.org/ontology/ENVO`) annotation was performed semi-automatically for SRA data and Chaffron's data set, whereas Qiime-DB provides EnvO annotations in mapping files accompaniying the recorded studies, according to MIMARKS guidelines (Yilmaz et al. (2011)). For a more detailed description the reader is referred to Henschel et al. (2015), Section Methods, subsection "EnvO annotation and method validation". Finally we define high-level ecosystem by grouping subtrees of EnvO classes: Biofilm, Plant, Soil, Animal/Human, Hypersaline, Geothermal, Freshwater, Marine and Anthropogenic. E.g. the ecosystem "Plant-related" is composed of EnvO-terms "plantation", "plant-associated habitat", and "plant food product" and their respective subsumed EnvO terms. As EnvO is a Directed Acyclic Graph with multiple inheritence and samples occasionally receive multiple EnvO annotations, it is possible that samples are assigned to several ecosystems simultaneously. We account for this by defining additional composite ecosystems such as Geothermal/Marine for marine hydrothermal vents. For each OTU we counted the occurrences in the above mentioned ecosystems (incl. composite ecosystems), yielding an occurrence vector of length 37. After normalization to a sum of one, the vector can be interpreted as (empirical) probability distribution for an OTU over ecosystems. We visualized all probability distributions with a stacked bar diagram, where the width of a bar corresponds to relative OTU abundance. This way, the proportion of generalists and specialists contained in a sample are immediately recognizable. As OTU bars are ordered by phylogenetic lineage, conventional taxonomic distribution is shown along the x-axis in addition to ecosystem distributions.

Study	Title	Isolation source	Nr
QDB_1200	Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat	microbial mat	18
QDB_1580	Saline environments that may harbor novel lignocellulolytic activities tolerant of ionic liquids	hypersaline lake	8
CHA_0507	Community composition of a hypersaline endoevaporitic microbial mat	hypersaline endoevaporitic microbial mat	1
CHA_0419	Characterization and spatial distribution of methanogens and methanogenic biosignatures in hypersaline microbial mats of Baja California	hypersaline microbial mat collected from concentrating area 4 located in Exportadora De Sal, S.A. (ESSA)	1
CHA_0742	Diversity and stratification of Archaea in a hypersaline microbial mat	hypersaline microbial mat: Guerrero Negro pond 4 near 5	1
CHA_0112	An Anaerobic Methane Oxidizing Community of ANME-1b Archaea in Hypersaline Gulf of Mexico Sediments	Gulf of Mexico sediments	1
CHA_1017	Haloarchaea and halophilic bacteria in two hypersaline soils of Jiangsu Province, China	saltern soil	1
CHA_2264	Unexpected diversity and complexity of the guerrero negro hypersaline microbial mat	hypersaline microbial mat: Guerrero Negro	1
CHA_1552	Miniprimer PCR, a new lens for viewing the microbial world	hypersaline microbial mat	1
CHA_0551	Comparison of deep-sea microbial communities in the eastern Mediterranean	sediment collected from a mound near Urania brine lake, Eastern Mediterranean, 3342m water depth: isolated from sediment layer 10-20 cm	1
CHA_0759	Diversity of Bacillus-like organisms isolated from deep-sea hypersaline anoxic sediments	Brine Lake Sediment	1
CHA_1788	Phylogenetic analysis of cultured bacteria in the deep see sediment of the east Pacific	deep sea sediment	1
CHA_0563	Comparison of the extremophiles of deep-sea and Antarctic	deep sea sediment	1
CHA_0086	Abundance and diversity of microbial life in ocean crust	deep seawater from the East Pacific Rise	1

Table 1. Hypersaline and deep sea samples in Database $EMP+$. The collection of microbial samples that Ecosystem distribution entropy (H^{EMP+}) is based on contains 35 samples from 11 independent studies. The last five samples are from independent deep sea studies. Study identifiers with QDB are taken from Qiime DB, those with CHA are from the Chaffron dataset.

The actual algorithm for ecosystem distribution is presented below: The algorithm was implemented in Python (using numerous modules such as matplotlib and numpy) in combination with SQL. The source code is available at `https://doi.org/10.5281/zenodo.847719`. The underlying database and its description including ecosystem assignment is available at `http://dx.doi.org/10.1371/journal.pcbi.1004468`.

Quantifiying ecosystem specificity Based on probability distributions over ecosystems, we calculate the Shannon entropy for each OTU:

$$H^{EMP+}(OTU) = -\Sigma_{i\in E}\ p_i \log p_i \qquad (1)$$

where E is the set of all 37 (pure and composite) ecosystems, p_i denotes the probability of an OTU to belong to an ecosystem i and $EMP+$ refers to the underlying database (as it constitutes a superset of the Earth Microbiome Project (EMP), one of the largest environmental 16S rRNA sample collections). Through Equation 1 we strive to capture the notion of OTU specialization: a specialist occurring only in one environment receives a minimal entropy of 0, whereas a generalist equally present in all environments is characterized by a high entropy value. Note that this calculation is always dependent on the suitability and completeness of the underlying database, and should therefore be regarded as an approximation. We however argue, that our database is—albeit not complete but—sufficiently comprehensive to produce valuable estimates. One desirable property of the Shannon entropy calculation is that specialists can

Study	Title	Isolation source	Nr
QDB_1046	Gulf Oil Spill Sediment	marine sediments from Gulf of Mexico	104
QDB_1198	Polluted Polar Coastal Sediments	marine sediment	57
QDB_1673	Mission Bay Sediment Viromes	marine sediment from Mission Bay	26
SRA_0011	Rich microbial communities in and around underwater springs in the Dead Sea	Dead Sea Springs Sediment (Archae)	5
QDB_1580	Saline environments that may harbor novel lignocellulolytic activities tolerant of ionic liquids	sea grass sample	3
CHA_1112	Impact of oil and higher hydrocarbons on microbial diversity, distribution and activity in Gulf of Mexico cold seep sediments	marine sediments	1
CHA_1340	Marine Derived Actinomycete Diversity	marine sediment	1
CHA_1840	Phylogenetic diversity of bacteria in marine sediments from the Arctic Ocean	marine sediments	1
CHA_1375	Microbial Communities Adherent to Sediment Particles in Heavy Metal Contaminated North Sea Surface Sediments	marine sediments	1
CHA_0096	Actinomycete and Other Gram-Positive Bacterial Diversity Cultured From Tropical Marine Sediments	marine sediment	1
CHA_2044	Seasonal variation of microbial diversity in the Yellow Sea sediment	Yellow Sea sediment	1
CHA_1560	Molecular analysis of bacterial communities in Pacific arctic surface sediment	arctic surface sediment	1

Table 2. Marine sediment samples in Database. We consider OTU presence in 202 samples annotated as marine sediment from 12 independent studies.

still get recognized as such despite occasional contaminations and artifacts, if the OTU was sampled predominantly in one environment.

We then calculated the unweighted ($\overline{H_U^{EMP+}}$) and weighted average entropy ($\overline{H_W^{EMP+}}$) for a sample S (represented as a set of OTUs with their respective relative abundances) as follows:

$$\overline{H_U^{EMP+}}(S) = \Sigma_{OTU \in S}\, H^{EMP+}(OTU)/|S| \tag{2}$$

$$\overline{H_W^{EMP+}}(S) = \Sigma_{OTU \in S}\, r_S(OTU) \times H^{EMP+}(OTU) \tag{3}$$

where r_S denotes the relative abundance of an OTU in sample S.

Percentile calculation In order to put those calculations for a particular environment into perspective, we also report the percentile of $\overline{H_W^{EMP+}}$ values. To this end, we calculated H_W^{EMP+} for all environmental samples in our database, i.e., those not related to human/animal. The reported percentiles are then the percentages of samples that achieve a lower entropy than the sample at hand.

RESULTS

Microbial community characterization of a unique salt flat and a mangrove forest bed environment

Illumina sequencing yielded 23,606 DNA sequences from the sabkha soil sample. A total of 702 closed reference OTUs (with respect to GreenGenes 13.5, 97% sequence similarity) were identified but approximately 80% of the community could not be identified beyond phylum level using the RDP classifier, confidence threshold 90%, see Figure S 1a). 36.1% of sequences subjected to OTU calling did not match any reference OTU. Identified sequences were found to be predominantly from phylum Proteobacteria (68.99%), followed by Acidobacteria (9.74%), Bacteroidetes (3.50%) and Actinobacteria (2.70%). The majority of Proteobacteria in the community were of class Gammaproteobacteria (36.05%), with up to 10.99% of these identified to belong to genus *Halomonas*. Alphaproteobacteria represent the second largest group of Proteobacteria in the sabkha community (29.86%), and up to 9.08% of these were identified to be of genus *Rhodovibrio* (Figure 1). The samples' alpha diversity (Phylogenetic Distance) is 47,916 (see Methods sections for details regarding the calculation).

In contrast, the mangrove forest bed community comprised 2,597 OTUs, with approximately 80% of the OTUs not identified beyond the family level Figure S 1b). From the original 46,875 sequences, 25,812 (55%) could not be assigned to the GreenGenes reference OTUs. Identified sequences from the mangrove forest bed community were predominantly from phylum Proteobacteria (44.5%), followed by Bacteroidetes (8.1%), Planctomycetes (6.8%), Actinobacteria (6.71%), Chloroflexi (6.29%),

Algorithm 1: Ecosystem distribution of OTUs

```
Input  : profile 16S rRNA profile, list of tuples of OTU-ids and abundances
Output : Ecosystem distribution matrix, entropy incl. stacked barchart visualization of ecosystems
         aligned with $H(OTU)$ plot and taxonomic information

Sample preprocessing (QIIME);
Closed reference OTU picking (QIIME);
foreach OTU in profile do
    # Query Database EMP+, in which other sample OTU occurs
    otherSamples ← SELECT sample FROM otu_sample_table WHERE otuID=OTU
    # Query Database EMP+, which ecosystems other samples are assigned to
    $\mathbf{e} = [e_{Soil}, e_{Marine}, \ldots, e_{HumanAssoc}]$ ← SELECT ecosystem, COUNT (*) Frequency FROM
        ecosystem_table WHERE sampleID IN otherSamples GROUP BY ecosystem
    $\mathbf{e_N}$ =normalize ($\mathbf{e}$)
    ecoDistribution [OTU ] ← $\mathbf{e_N}$
end
Order OTUs taxonomically
foreach (OTU, abundance ) in profile do
    Calculate and plot H(OTU)
    Visualize ecoDistribution [OTU ] as stacked bar, with width proportional to
        abundance, arranged according to phylogeny
end
```

Gemmatimonadetes (5.1%) and Acidobacteria (4.9%). Proteobacteria in the community were primarily Deltaproteobacteria (14.18%), followed by Gammaproteobacteria (11.76%) and Alphaproteobacteria (6.55%)(Figure 2). The alpha diversity (PD, with respect to the GreenGenes phylogeny, see Methods) is 126,940.

The sabkha community and the mangrove forest bed community each had a distinct environmental distribution profile based on the environmental metadata analysis performed on the 16S rDNA sequences against global libraries. The total range of observed ecosystem distribution entropy is 0-2.295 (for both $\overline{H_U^{EMP+}}$ and $\overline{H_W^{EMP+}}$). The environmental distribution profile for the sabkha soil community revealed the majority of OTUs to be exclusively associated with the hypersaline marine environment, while the remainder were linked to a combination of mainly soil or marine environments, along with anthropogenic soil, geothermal and animal/human host environments (Figure 1). The exclusive occurrence of the sabkha community members in studied hypersaline marine environments indicated their narrow distribution across global environments, as represented by their low weighted mean ecosystem entropy value $\overline{H_W^{EMP+}} = 0.458$. To put this value into perspective, the quantile value is 20.36% wrt. all ecosystems and 1.06% wrt. environmental, i.e., non-human/animal associated samples (also see Figure 3). The association with hypersaline marine environments was also not restricted to any taxonomic group, but was rather widespread among the community members. The majority of OTUs identified in the community composition were found to have a limited distribution across the global libraries, as indicated by the small number of samples they were found in. It was also noted that OTUs occuring only samples tended to have ecosystem entropy values of zero or near zero in this environmental distribution profile. Potential misclassifications are further elaborated on in the Discussion section. A few exceptions were present in a relatively large number of samples ranging from 100 to 1000, with higher ecosystem entropy values ranging from 0.5 to 2.0. This may represent the minority group of generalists among the sabkha community members, indicated by their association with a more diverse range of environments compared to their specialist counterparts, and their presence in a larger number of global samples.

Conversely, the environmental distribution profile for the mangrove forest bed community revealed the majority of OTUs to be associated with a variety of environments, the most prominent being marine, followed by soil and freshwater environments (Figure 2). A few OTUs appeared to be exclusively linked to either soil or hypersaline marine environments, but their abundance was negligible relative to the entire community. Overall, the mean weighted ecosystem entropy value for the mangrove forest bed community was 0.698 (quantile wrt. environmental), considerably higher than that of the sabkha soil community.

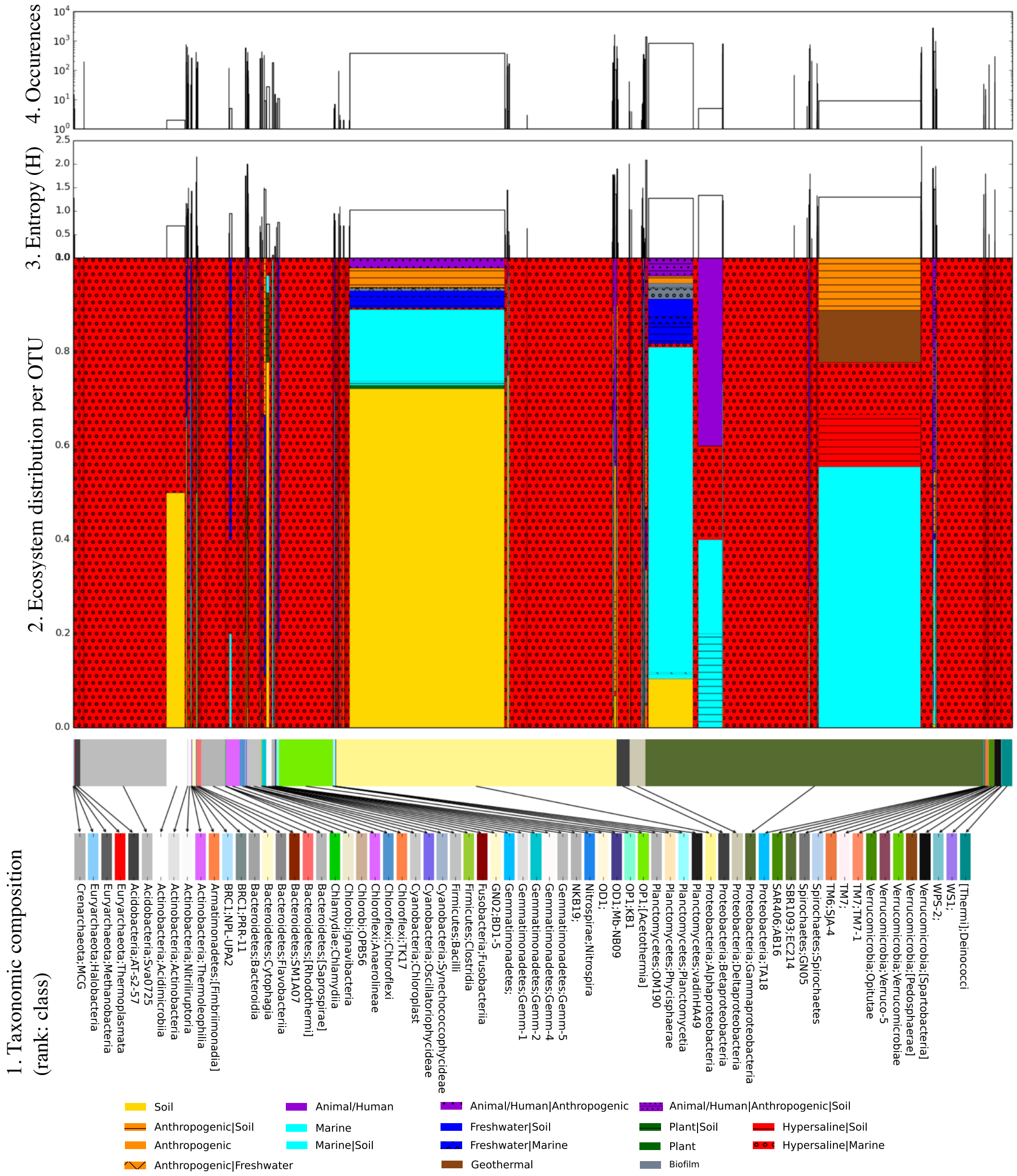

Figure 1. Ecosystem distribution profile for sabkha sample as produced by Algorithm 1. The profile contains four parts: 1. a conventional bardiagram for displaying community composition, including a legend with taxonomic categories. 2. Orthogonal to taxonomic categories, we show bar diagrams of OTUs reflecting their respective (empirical) probability distribution over ecosystems with respect to our EMP-derived database. Note that a bar for each OTU is placed above the taxonomic category it belongs to and moreover, the width of the bar corresponds to the relative abundance of the OTU in the sample. Composite ecosystems (e.g. sabkha being Marine/Soil/Hypersaline) are shown with consistent respective hatching patterns, see legend in Figure S8. 3. For each OTU, we calculate the ecosystem entropy H as described in equation 1. The entropies are horizontally aligned with the ecosystem distribution of 2. 4. Again, horizontally aligned with OTU specific information, the uppermost section displays the total occurrences of OTUs in all samples of our database.

This most likely indicates a broader distribution of the mangrove soil community members across global environments, compared to sabkha soil bacteria known so far to occur only in hypersaline marine settings.

Members of this community are also present in a greater number of global samples, averaging at 69.3 samples/OTU.

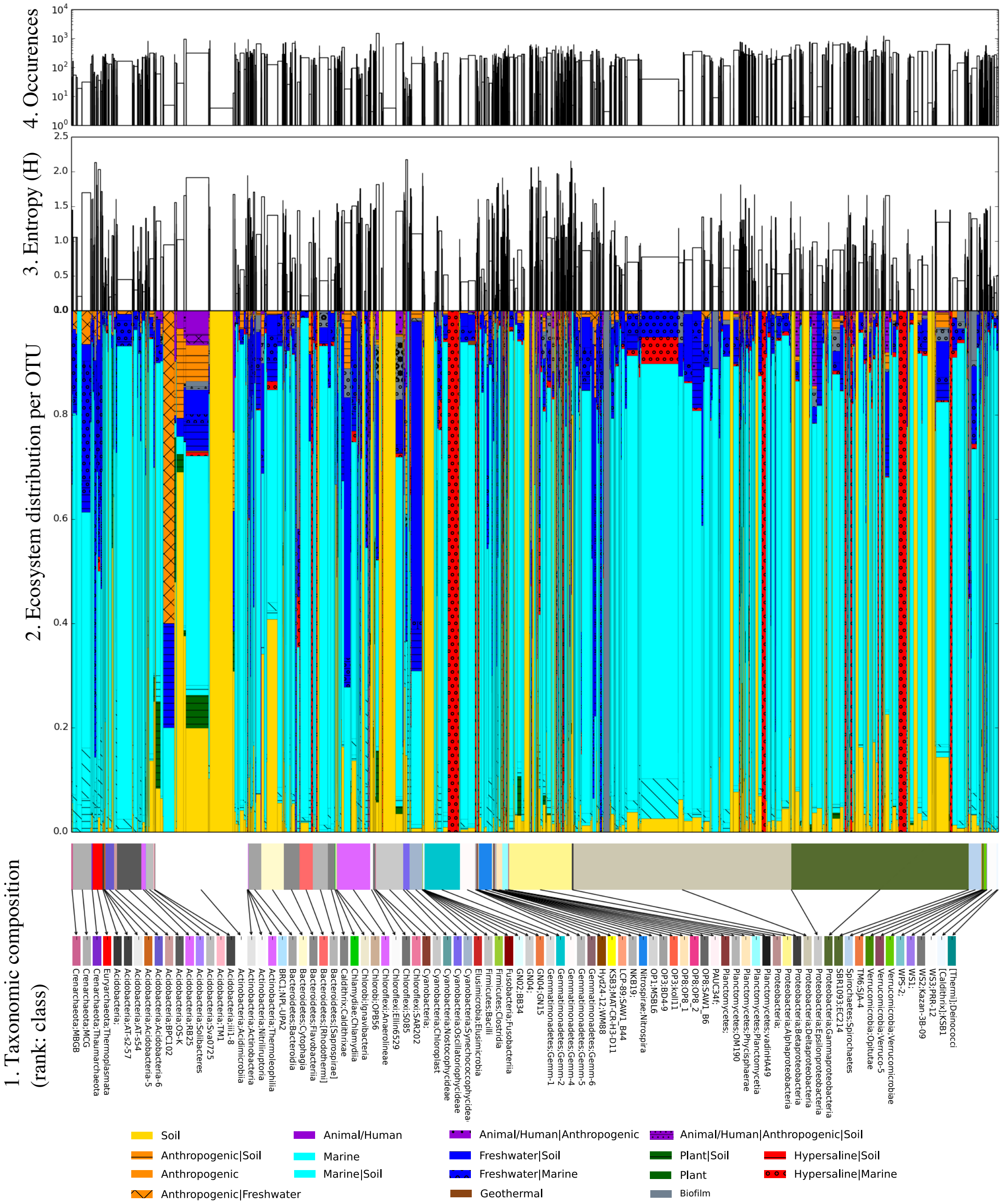

Figure 2. **Automatically generated ecosystem distribution profile for Mangrove sample.** The Mangrove sample contains OTUs predominantly found in marine environments with occasional soil and hypersaline specific specialists. The habitat-based profile gives thus not only an impression of the biogeography of its constituents, but also a sense of a more mixed background than the sabkha sample.

Comparison of ecosystem distribution entropy to other saline/hypersaline samples Environmental distribution profiles were also generated for a selection of studied communities originating from saline and hypersaline environments worldwide. These communities were selected for our analysis as we were interested in observing the biogeographic and ecological specialization patterns across communities sampled from various salt-stressed environments. Our analytical approach produced distinct environmental distribution profiles for these communities, with their respective environmental distribution patterns yielding the most visually striking indicator of their variability. For example, a sample collected from a hypersaline lake for a previous study (A23.number1.filt.D1.660399, Qiita study ID 1580, see `https://qiita.ucsd.edu/study/description/1580`, login required) was found to harbor OTUs associated with freshwater environments, along with soil, hypersaline, and plant-associated environments to a lesser extent. In contrast, another sample from the same study (WPA.filt.660391) presented a stronger association with marine environments, along with distinct links to animal/human-associated environments observed in the community's most abundant OTU (Betaproteobacteria). Two samples from another study (P.Masambaba.SA.414862 and P.Masambaba.SB.414876, Qiita study ID 1039), collected from the same depth in Lagoa Vermelha, Brazil, presented environmental distribution profiles that were largely similar in terms of community composition and their environmental distribution patterns (close links to mainly freshwater and soil environments). In Lagoa Vermelha and a number of other hypersaline sites, Gammaproteobacteria also appear to dominate, similar to the sabkha community profile we obtained. In terms of ecological specialization patterns, these samples cited from previous studies presented generalistic tendencies, indicated by their mean ecosystem entropy values ($\overline{H_W^{EMP+}}$) ranging from 0.873 (quantile: 52.40%, P.Masambaba.SB.414876) to 1.583 (quantile: 84.30%, A23.number1.filt.D1.660399), see Table 3 and Supplementary Figures S2-S7. The higher prevalence of Gammaproteobacteria in their hypersaline sites (Abu Dhabi sabkha, Lagoa Vermelha) compared to those of moderately-saline sites (mangrove forest bed) strongly hint at highly-adapted strategies for surviving salt-saturated pore waters and even entrapment in salt crystals (Ma et al. (2010)).

The OTUs constituting these samples were also quite well-represented in public databases, with an average of 398.9 libraries presenting sequences matching these studied communities. On the other hand, microbial mat communities from Yellowstone National Park presented environmental distribution profiles that contrasted remarkably against the other cited samples, in that these communities were almost exclusively associated with environmental biofilms. Based on this observation and the significant number of libraries presenting sequences matching these studied communities (680.7 samples/OTU on average), we can conclude that the Yellowstone mats hosted highly specialized bacteria with a severely limited range of habitats across the planet.

Following our targeted approach of generating environmental distribution profiles for our communities of interest, we proceeded to determine whether different habitats/environment types were characteristically generalistic or specialistic in terms of community composition. We calculated the ecosystem entropy values $\overline{H_W}$ (see section Methods, Equation 3) for the 20,472 global libraries included in our database, and generated a histogram to represent their distribution across the studied environments, see Figure 3.

Overall, there indeed appears to be correlation between a community's ecosystem entropy value and its environment type. Communities with the lowest entropy values were almost exclusively associated with the animal/human host environment. Figure 3 shows the low ecosystem entropy (in terms of the introduced formalism) of the sabkha sample in comparison to other environmental samples. Animal/Human associated samples are generally low in $\overline{H_W}$ (though with a very broad variance) and exclusively constitute the low entropy samples for $\overline{H_W} < 0.4$. On the other hand, the upper range of entropy values were represented predominantly by communities from plant-associated, soil, and anthropogenic environments. Finally, communities from marine and freshwater environments presented ecosystem entropy values that tended to be in the midrange, rather than in the lower or higher extremes.

We finally compared our local samples to the entire dataset using Visibiome (Azman et al. (2017)), a UniFrac based search engine for microbial communities. Remarkably, no matches for sabkha were found during an exhaustive search using the popular phylogeny-based distance measure, despite the database containing 35 samples from hypersaline environments, 36 of which have at least 50 OTUs (see Table 1). On the other hand, the mangrove soil sample matched against a number of samples from the Earth Microbiome Project, due to similar composition of Desulfobacteraceae, Syntrophobacteraceae, Piscirickettsiaceae and other families from the Proteobacteria phylum. In particular, the closest weighted UniFrac matches were observed for samples P.Dois.Rios.SB.414865 (Qiita 1039, UniFrac distance: 0.244), SE.20101009.GY.FF003.BC.221 (Qiita 1197, UniFrac distance 0.275) and TtA.sed.D1.660402 (Qiita 1580, UniFrac distance 0.283). These results are shown in Figure S11 and S12 and in a series of interactive visualizations (with zoom, pan and tooltip functionality) at `https://visibiome.org/`

Sample event ID	Isolation source	Title	Study	Country	Date	$\overline{H_U^{EMP+}}$	$\overline{H_W^{EMP+}}$	Percentile	Alpha-diversity
Sabkha	Sabkha soil Abu Dhabi	Sabkha soil Abu Dhabi	MI 1	UAE	10/1/2012	0.260	0.458	1.06%	47.916
Soil.Day.0	Mangroves Abu Dhabi	Microbial Diversity study in mangroves	MI 2	UAE	10/2/2012	0.699	0.698	9.51 %	126.940
P.Masambaba.SB.414876	marine sediment	Rio de Janeiro Coastline	1039	Brazil	1/24/2011	0.950	0.873	26.87%	61.926
P.Masambaba.SA.414862	marine sediment	Rio de Janeiro Coastline	1039	Brazil	1/24/2011	0.967	0.930	29.56%	47.786
WPA.filt..660391	hypersaline lake	Saline environments that may harbor novel lignocellulolytic activities tolerant of ionic liquids	1580	Puerto Rico	12/14/2011	1.222	1.467	59.69%	34.582
A23.number1.filt.D1.660399	hypersaline lake	(as above)	1580	USA	12/9/2011	1.232	1.583	67.12%	8.406
P.Dois.Rios.SB.414865	marine sediment	Rio de Janeiro Coastline	1039	Brazil	1/24/2011	1.078	0.975	31.61 %	101.46
SE.20101009.GY.FF003	marine sediment	Mexican Gulf Oil Spill Sediments	1197	USA	10/9/2010	0.873	0.788	18.89%	100.66

Table 3. Ecosystem entropy and additional information for selected saline samples. Note that for alpha diversity calculation, all samples were rarefied to 18,000 sequences.

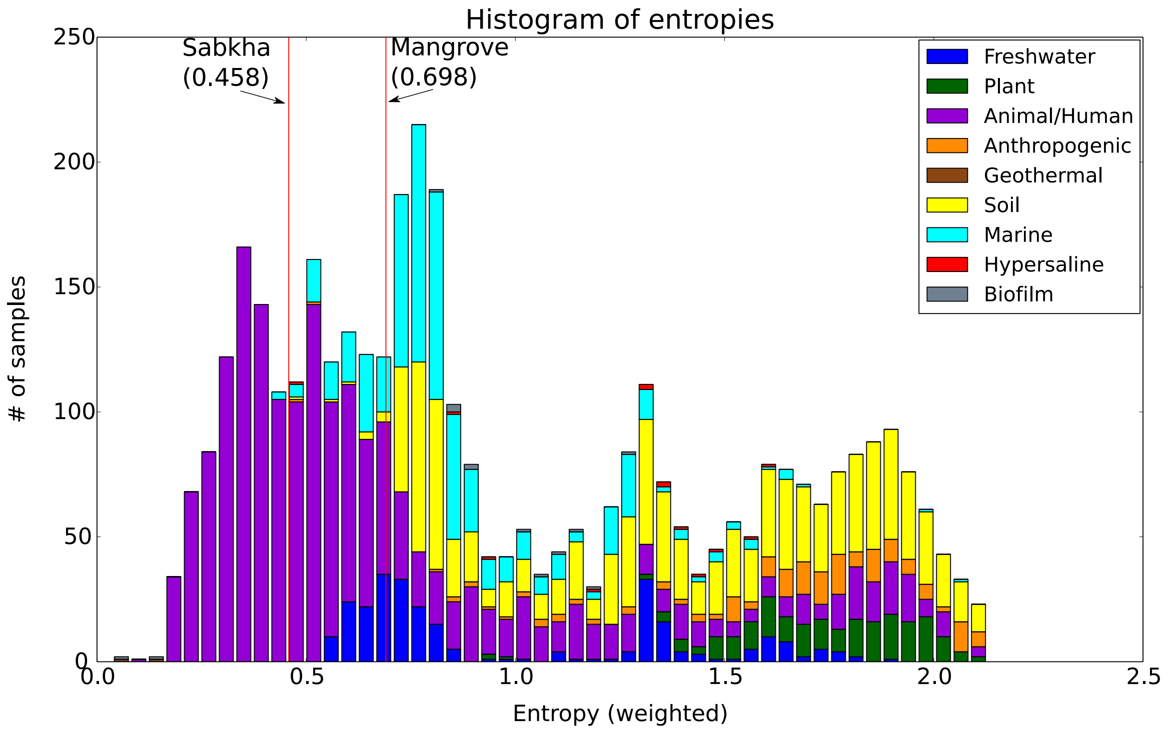

Figure 3. Histogram of weighted average entropies for all samples in our EMP-derived dataset. Remarkably the weighted average ecosystem entropy of the sabkha sample ($\overline{H_W}(sabkha)$ as defined in Equation 3) is very low, in particular wrt. marine or soil samples, owing to the high number of specialists with low ecosystem entropy.

`public/jobs/1950/ranking` in respective tabs (SabkhaAD and User_Soil.Day.0). The figures also contain the contextualization of our samples against their closest matches using Hierarchical Clustering and Principal Coordinate Analysis (PCoA).

CONCLUSIONS

In this work we designed and show cased a new type of analysis that is directed at microbial ecologists who wish to characterize samples from harsh environments and want to understand the biogeography of the constituent extremophiles. Our method visualizes microbial communities in a compact figure that captures not only the commonly provided taxonomic information, but adds an orthogonal dimension for ecosystem distribution. We demonstrate how to formally account for ecosystem specificity of OTUs in a community. To this end, we have created an algorithm and a relational database that includes most 16S rRNA profiles (composed of closed reference OTUs) from the Earth Microbiome Project. When taxonomic identification is low, i.e. it is not known who the constituents of the community exactly are, it is helpful to know at least where they occur. The visualization of OTU ecosystem distribution allows the viewer to infer the general nature of a sample and what the environmental drivers for community composition are. The rationale behind the comparison of a hypersaline and a moderately hypersaline sample was to demonstrate the differences in constituent specialization. For example, the ecosystem distribution profile of the investigated sabkha pointed to highly exclusive environmental factors that would permit only very well adapted OTUs that rarely occur elsewhere, but not facilitate circulation of animal- and plant-associated bacteria beyond short lived source-sink dynamics. It must be stated that due to low occurrence of constituent OTUs in other global samples, genetic isolation in the sabkha is an alternative explanation for low entropy. In general, however, we maintain that observed ecosystem specificity is a suitable indicator for habitat adaptation and specialization. In contrast, the mangrove soil, albeit more saline than normal seawater (45 ppt), is a more forgiving environment compared to the coastal sabkha and presents OTUs that can be found in a variety of different ecosystems, as witnessed by the majority of

OTUs exhibiting high entropy. This is most likely an indication of these OTUs' specialistic tendencies, judging from their rare occurrence across the widely-sampled global libraries and limited range of host environments.

We argue that ecosystem distribution of microbial community members are a reasonable proxy for dispersal and as such can support biogeographic studies. Likewise, we maintain that ecosystem specificity of OTUs, which is a purely descriptive measure, facilitates the identification of specialists.

Applying the ecosystem distribution formalism to the entire dataset at hand helps to put the ecosystem entropy of our local samples into perspective. Moreover, general emerging trends can be gleaned from the ecosystem distribution histogram ($\overline{H_W^{DB}}$, Figure 3): ecosystems are remarkably continuous as opposed to random, discontinuous or more pronounced multimodal distributions.

It is worth noting that alpha-diversity is not necessarily correlated with ecosystem distribution entropy for extreme environments: for the sabkha sample, we observe a community that is relatively complex given the harsh circumstances (OTUs from 29 different phyla and 67 different taxonomic classes were observed). However most OTUs seem to be adapted to a saline environment with rare occurences in other environments. Finally, we observe that the strong compositional differences in hypersaline microbial communities in our database as reflected by high beta-diversity amongst hypersaline 16S rRNA profiles indicate that many taxonomically different bacterial species evolved convergently in order to adapt to hypersaline environments.

Limitations. Certain limitations persist, e.g., non standardized protocols for 16S rRNA amplification. Closed reference OTU-calling facilitates the comparison of 16S rRNA profiles with different amplicon regions, but often fails to recognize a part of the library, which then has to be discarded. Under certain circumstances, open-reference and denovo OTU calling methods could be applied but they are not suitable for the large scale database screen we presented here. In our case, we showed that the majority of sequences can be called against the reference, and that they carry a strong signal as to where they occur and how specific they are to ecosystems. Moreover, many acquired 16S rRNA profiles were not stored as sequences but only their prepicked OTU profiles.

One possible source for misclassification is a generalist that escapes observation (e.g. due to low abundance) in the samples of the underlying database. In this light it should be noted that at this stage, some habitats such as biofilms, hypersaline environments and geothermal settings are represented by a much smaller number of samples compared to others which are more well-studied. Hence, the distribution of community entropy across these particular environment types remains unclear and may potentially be determined with greater certainty with improvements in sampling effort targeting these settings. One way of accounting for ecosystem sampling bias to normalize the probabilities p_i accordingly. While entropy can be calculated on unnormalized probabilities, the figures are impacted by sampling bias but consistently so, such that they still allow visual comparisons. Many sabkha OTUs, despite matching GreenGenes references, occur exclusively in this particular sample. As a result, their ecosystem entropy is $1 \log 1 = 0$, identifying them as extremophiles, but an alternative explanation would be underrepresentation in our database (as a consequence of being underrepresented in the Earth Microbiome Project), which still reflects their rare nature. We anticipate that the increasingly comprehensive volume of environmental available samples will mitigate this phenomenon in the future. In essence, for these cases the notion of extremophile should be relaxed to "extremophile and/or rare and/or genetically isolated".

Conversely, a specialist can be mistaken for a generalist due to contamination. Again, a growing data platform is expected to contain sufficient samples to outnumber these artefacts: the application of the Ecosystem Entropy (Equation 1) to an OTU o with some spurious ecosystem information (i.e., a entropy distribution vector of near-zero probabilities and one near-one probability) still yields near-zero values for $H(o)$.

Finally, other shortcomings such as EnvO misannotation of samples might impact the accuracy of entropy estimation negatively and efforts of improvement are underway (ten Hoopen et al. (2016)). While currently trading off ecosystem coverage and annotation quality in favor of the former, these efforts will be a suitable replacement for our dataset as soon as they reach the critical mass needed for the task at hand. Finally, we have previously shown that microbial communities cluster by ecosystem and that this way, misannotations can be removed (Henschel et al. (2015)).

REFERENCES

Azman, S. K., Anwar, M. Z., and Henschel, A. (2017). Visibiome: an efficient microbiome search engine based on a scalable, distributed architecture. *BMC Bioinformatics*, 18(1):353.

Barberán, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2):343.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). Genbank. *Nucleic Acids Research*, 41(D1):D36–D42.

Bontognali, T. R. R., Vasconcelos, C., Warthmann, R. J., Bernasconi, S. M., Dupraz, C., Strohmenger, C. J., and McKenzie, J. A. (2010). Dolomite formation within microbial mats in the coastal sabkha of Abu Dhabi (United Arab Emirates). *Sedimentology*, 57(3):824–844.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., et al. (2013). The environment ontology: contextualising biological and biomedical entities. *J. Biomedical Semantics*, 4:43.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., and Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of biomedical semantics*, 7(1):57.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Pena, A., Goodrich, J., and Gordon, J. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., and Bauer, M. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*, 6(8):1621–1624.

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A., McGarrell, D., Marsh, T., and Garrity, G. M. (2009a). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(suppl 1):D141–D145.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje, J. M. (2009b). The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Research*, 37(1):D141–D145.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 72(7):5069–5072.

Dini-Andreote, F., Andreote, F. D., Araújo, W. L., Trevors, J. T., and van Elsas, J. D. (2012). Bacterial genomes: habitat specificity and uncharted organisms. *Microbial ecology*, 64(1):1–7.

Duarte, C. M. (2015). Seafaring in the 21st century: The malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin*, 24(1):11–14.

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., and Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, 109(52):21390–21395.

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., and Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences*, 105(22):7774–7778.

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The earth microbiome project: successes and aspirations. *BMC biology*, 12(1):69.

Henschel, A., Anwar, M. Z., and Manohar, V. (2015). Comprehensive meta-analysis of ontology annotated 16s rrna profiles identifies beta diversity clusters of environmental bacterial communities. *PLoS Comput Biol*, 11(10):e1004468.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved otu clustering. *Environmental Microbiology*, 12(7):1889–1898.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., et al. (2011). A holistic approach to marine eco-systems biology. *PLoS biology*, 9(10):e1001177.

Kim, M., Lee, K.-H., Yoon, S.-W., Kim, B.-S., Chun, J., and Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & Informatics*, 11(3):102–113.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R., Abdallah, R. Z., Sonnenschein, E. C., Cariou, T., O'Gara, F., Jackson, S., Orlic, S., Steinke, M., Busch, J., Duarte, B., Caçador, I., Canning-Clode, J., Bobrova, O., Marteinsson, V., Reynisson, E., Loureiro, C. M., Luna, G. M., Quero, G. M., Löscher, C. R., Kremp, A., DeLorenzo, M. E., Øvreås, L., Tolman, J., LaRoche, J., Penna, A., Frischer, M., Davis, T., Katherine, B., Meyer, C. P., Ramos, S., Magalhães, C., Jude-Lemeilleur, F., Aguirre-Macedo, M. L., Wang, S., Poulton, N., Jones, S., Collin, R., Fuhrman, J. A., Conan, P., Alonso, C., Stambler, N., Goodwin, K., Yakimov, M. M., Baltar, F., Bodrossy, L., Kamp, J. V. D., Frampton, D. M., Ostrowski, M., Ruth, P. V., Malthouse, P., Claus, S., Deneudt, K., Mortelmans, J., Pitois, S., Wallom, D., Salter, I., Costa, R., Schroeder, D. C., Kandil, M. M., Amaral, V., Biancalana, F., Santana, R., Pedrotti, M. L., Yoshida, T., Ogata, H., Ingleton, T., Munnik, K., Rodriguez-Ezpeleta, N., Berteaux-Lecellier, V., Wecker, P., Cancio, I., Vaulot, D., Bienhold, C., Ghazal, H., Chaouni, B., Essayeh, S., Ettamimi, S., Zaid, E. H., Boukhatem, N., Bouali, A., Chahboune, R., Barrijal, S., Timinouni, M., Otmani, F. E., Bennani, M., Mea, M., Todorova, N., Karamfilov, V., Hoopen, P. t., Cochrane, G., L'Haridon, S., Bizsel, K. C., Vezzi, A., Lauro, F. M., Martin, P., Jensen, R. M., Hinks, J., Gebbels, S., Rosselli, R., Pascale, F. D., Schiavon, R., Santos, A. d., Villar, E., Pesant, S., Cataletto, B., Malfatti, F., Edirisinghe, R., Silveira, J. A. H., Barbier, M., Turk, V., Tinta, T., Fuller, W. J., Salihoglu, I., Serakinci, N., Ergoren, M. C., Bresnan, E., Iriberri, J., Nyhus, P. A. F., Bente, E., Karlsen, H. E., Golyshin, P. N., Gasol, J. M., Moncheva, S., Dzhembekova, N., Johnson, Z., Sinigalliano, C. D., Gidley, M. L., Zingone, A., Danovaro, R., Tsiamis, G., Clark, M. S., Costa, A. C., Bour, M. E., Martins, A. M., Collins, R. E., Ducluzeau, A.-L., Martinez, J., Costello, M. J., Amaral-Zettler, L. A., Gilbert, J. A., Davies, N., Field, D., and Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience*, 4(1):1–5.

Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., and Clavel, T. (2016). Imngs: a comprehensive open resource of processed 16s rrna microbial profiles for ecology and diversity studies. *Scientific reports*, 6:33721.

Ma, Y., Galinski, E. A., Grant, W. D., Oren, A., and Ventosa, A. (2010). Halophiles 2010: life in saline environments. *Applied and environmental microbiology*, 76(21):6971–6981.

Martiny, J. B. H., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, 4(2):102–112.

Monard, C., Gantner, S., Bertilsson, S., Hallin, S., and Stenlid, J. (2016). Habitat generalists and specialists in microbial communities across a terrestrial-freshwater gradient. *Scientific reports*, 6.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21):7188–7196.

Reddy, T., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A., and Kyrpides, N. C. (2015). The genomes online database (gold) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*, 43(D1):D1099–D1106.

Schloss, P. D. and Handelsman, J. (2004). Status of the microbial census. *Microbiology and Molecular Biology Reviews*, 68(4):686–691.

Su, C., Lei, L., Duan, Y., Zhang, K.-Q., and Yang, J. (2012). Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Applied microbiology and biotechnology*, 93(3):993–1003.

ten Hoopen, P., Amid, C., Luigi Buttigieg, P., Pafilis, E., Bravakos, P., Cerdeño-Tárraga, A. M., Gibson, R., Kahlke, T., Legaki, A., Narayana Murthy, K., et al. (2016). Value, but high costs in post-deposition data curation. *Database*, 2016.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project: Exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2):e1000667.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B. W., Blaser, M. J., Bonazzi, V., Booth, T., Bork, P., Bushman, F. D., Buttigieg, P. L., Chain, P. S. G., Charlson, E., Costello, E. K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J. A., Gallery, R. E., Gevers, D., Gibbs, R. A., Gil, I. S., Gonzalez, A., Gordon, J. I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A. L., Kelley, S. T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C. L., Legg, T., Ley, R. E., Lozupone, C. A., Ludwig, W., Lyons, D., Maguire, E., Methé, B. A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K. E., Nemergut, D., Neufeld, J. D., Newbold, L. K., Oliver, A. E., Pace, N. R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D. A., Assunta-Sansone, S., Schloss, P. D., Schriml, L., Sinha, R., Smith, M. I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J. M., Ward, D. V., Weinstock, G. M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J. R., Yatsunenko, T., and Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29(5):415–420.

Yutin, N., Suzuki, M. T., Teeling, H., Weber, M., Venter, J. C., Rusch, D. B., and Béjà, O. (2007). Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ Microbiol*, 9(6):1464–1475.