Doctoral Dissertation
博士論文

Reconsidering representation of complex microbial community structures: Habitat-based analysis based on comparative metagenomics

（比較メタゲノミクスを用いた微生物の生息環境解析
– 微生物群集データの表現方法を再考する）

A Dissertation Submitted for the Degree of Doctor of Philosophy
December 2019
令和元年 12 月 博士（理学）申請

Department of Biological Sciences, Graduate School of Science,
The University of Tokyo
東京大学大学院理学系研究科 生物科学専攻

Kazumori Mise
美 世 一 守

Abstract

The recent prevalence of amplicon and shotgun-metagenome sequencing has been producing numerous prokaryotic community structure datasets. However, methods for interpreting those structures from ecological perspectives are still insufficient. High-rank taxonomies such as phyla and classes are often used to interpret community structures, but these prokaryotic taxonomies are often decoupled from their ecological and physiological traits. On the other hand, available prokaryotic trait databases are heavily biased to cultivated and well-characterized species, which are rare in nature. Here, I propose *habitat-based analysis* as a powerful and intuitive method for interpreting prokaryotic community structure datasets. First, I systematically processed public shotgun metagenome datasets and constructed a cultivation- and PCR-bias free database, ProkAtlas, that comprehensively links 16S rRNA gene sequences to prokaryotic habitats. Then, I developed a computational pipeline for habitat-based analysis of given prokaryotic community structure datasets. After confirmation of the method effectiveness using 16S rRNA gene sequence datasets from pure-cultured isolates, single-cell amplified genomes, and the Earth Microbiome Project, I applied my method to three datasets from environments of particular interest: coastal soil samples with salinity gradients, lake water samples with different salinity concentrations, human infant gut microbiome samples, developing soil samples along retreating glacier, and potentially polluted river-water samples. As a result, the habitat-based analysis was proved to give clear ecological interpretation of chemical characteristics, community assembly processes, and pollution sources. The ProkAtlas database and pipeline are available at https://msk33.github.io/prokatlas.html.

Table of contents

Chapter 1 | General Introduction

1-1 Exploration of microbial ecology — a brief history

Since the invention of the microscope, the astounding abundance and diversity of microbes have been reported in various environments (Barberán et al., 2017; Thompson et al., 2017). Today, microbes are recognized as the largest reservoir of genetic diversity on Earth, as well as indispensable players in maintaining ecosystems on any scale. Microbial ecology in both open-system and symbiotic environments has therefore been extensively studied in the context of fundamental and practical sciences (Antwis et al., 2017; Thompson et al., 2017).

Historically, the ecology of environmental microbes has been studied through cultivation and isolation methods. This approach is part of a long tradition, dating back to the mid-19th century (Hitchens and Leikind, 1939). Microbes are isolated from environmental samples using a solid or liquid medium, and their traits (including physiology and metabolic function) are analyzed. Cultivation-dependent studies have increased our knowledge of individual microbial community members (Barberán et al., 2017; T. Tanaka et al., 2014); however, these studies are necessarily limited to culturable species, which are in fact rather rare in the environment (Steen et al., 2019). Because of this limitation, overall microbial community structures in the environment for a long time remained unknown, and the environmental microbial community was once regarded as a "black box" (Fierer et al., 2009). Although there was no doubt about the functionality of microbial communities, their detailed compositions were a mystery; and this difficulty persisted until the development of new technology in the 1990s.

From the 1990s, the development of molecular biological technology enabled researchers gradually to gain insight into environmental microbial community structures (Liu et al., 1997; Moyer et al., 1994). The basic procedure involves chemically extracting DNA molecules from environmental samples and analyzing the extracted DNA (or its amplicons, obtained by PCR), with the assumption that the extracted DNA is representative of the microbial community in the source sample (Handelsman et al., 1998; Venter et al., 2004). This is therefore a culture-independent approach, which enabled microbial ecologists to quantitatively evaluate the overall composition of microbial community, including yet-to-be cultured microbes.

While this culture-independent approach has undoubtedly expanded our view of microbial diversity, it has also posed a new problem regarding the handling of data.

Microbial community structure data, generated by culture-independent studies, are characterized by their high dimensionality (as a consequence of extreme diversity of microbes in the environment) (Bálint et al., 2016; Lynch and Neufeld, 2015). Because of this high dimensionality, microbial ecologists have struggled to effectively summarize, use, and interpret the community structure data. In the following two sections, I review the history of culture-independent approaches and struggles for effective use of the large data.

1-2 Advances in culture-independent studies

While the basic procedure for culture-independent microbial community analysis has not been updated for decades (Fierer, 2017; Hiraoka et al., 2016; Moyer et al., 1994; Venter et al., 2004), the methodology of DNA analysis has advanced remarkably. In the course of such methodological advance, two different types of study emerged in microbial ecology.

Biodiversity-ecosystem function studies. Until around 2010, molecular methods for detecting single nucleotide polymorphisms, such as restriction fragment length polymorphism (RFLP) and denaturing gradient gel electrophoresis (DGGE), had been commonly used (Moyer et al., 1994; Muyzer et al., 1993; Muyzer and Smalla, 1998). Unlike high-throughput sequencing, these techniques do not provide detailed nucleotide sequences of microbial community members; however, they do provide quantitative information on microbial community diversity. Band patterns obtained by RFLP and DGGE clearly indicate the alpha-diversity of each microbial community, and beta-diversity between pairs of communities (Figure 1-1) (Girvan et al., 2005; Wertz et al., 2006).

Although taxonomic information was unavailable (without conducting labor-intensive clone library analysis), these simple diversity evaluations provided mechanistic insights into the interplay between microbial communities and their environment (Girvan et al., 2005; Wertz et al., 2006), which is referred to as the biodiversity-ecosystem function (BEF) relationship. Biogeochemical studies, for example, regard soil microbial community structures as potential explanatory variables of biogeochemical processes (e.g. nitrification rate, carbon flux from the ground, etc.); in fact, many of them conclude that soil microbial community structures significantly

improve the predictability of biogeochemical processes, and therefore they should be incorporated into models of global biogeochemical cycling (Fierer et al., 2010; Graham et al., 2014; Isobe et al., 2018). Human gut microbiomes are also regarded as an explanatory variable of certain aspects of host phenotype, such as obesity and immune response (Li et al., 2008).

Descriptive studies of microbial community structures. Subsequently from 2008, high-throughput sequencers, such as the Roche 454 and the Illumina MiSeq/HiSeq, gradually reigned microbial community ecology (Schuster, 2008). As the cost of high-throughput sequencing has gradually decreased, shotgun metagenomic sequencing and amplicon sequencing have seen more common use among microbial ecologists (Bálint et al., 2016). Such techniques can reveal the microbial community structures in fine-scale taxonomic resolutions (Figure 1-2), and microbial diversity in a number of environments has been investigated and described (Gilbert et al., 2018; Thompson et al., 2017). For example, in the field of soil microbial ecology, biogeographical studies have successfully documented the global and regional distribution patterns of soil microbes (Caporaso et al., 2011; Lauber et al., 2009). Similarly, three distinct patterns of global variation in human gut microbiomes, known as enterotypes, have been identified (Arumugam et al., 2011).

1-3 What microbial ecology has missed

As reviewed above, microbial ecology has been driven by technological advances in molecular biology during the past two decades. Microbial ecologists have elucidated the contribution of microbial community to ecosystem functions, as well as the patterns of microbial community structures in various environments. Despite this, they have not necessarily considered the biological and ecological traits of microbial community members (Guittar et al., 2019; Martiny et al., 2015). A greater focus on these traits should certainly give more insight into the function and dynamics of microbial ecosystems, considering that such trait-centric view has yielded insights into community dynamics of plant communities (Corlett and Westcott, 2013; Nemergut et al., 2016).

In BEF studies, microbial community structure datasets are commonly summarized into simple diversity indices (Graham et al., 2016; Wertz et al., 2006). This is reasonable for RFLP and DGGE data, which are not accompanied by detailed

information on individual community members; however, recent high-throughput sequencing data are likewise treated without discussion on the community members (presumably carrying on the practice of RFLP/DGGE era). Importantly, the taxonomic descriptions of microbial community structure suffer from a similar problem. Microbial community structures are often compressed into high-rank taxonomic classifications (i.e., phyla or classes) (Thompson et al., 2017), and these high-rank classifications are largely decoupled from their ecological and physiological traits and can hardly provide meaningful ecological insights (Martiny et al., 2015).

In conclusion, conventional microbial ecology studies only marginally consider detailed information on community members, which should be the great merit of introducing high-throughput sequencing. In other words, these studies do not take full advantage of the high-resolution data that next-generation sequencing technologies have brought about — in a sense, they may be still in the era of "first-generation sequencing" (Figure 1-2).

1-4 Trait-based approach in microbial ecology

Use of community members' trait information is not a new idea in itself; in fact, plant ecologists have developed a framework of trait-based approaches over the past four decades (Grime, 1974; Violle et al., 2007), and has been a powerful tool to solve questions in community ecology, such as predicting the transition of community in response to environmental perturbations and formulizing community assembly rules in plant community (Garnier and Navas, 2012). The basic procedure of a trait-based approach is to first classify community members according to their ecological or physiological traits and then project that trait information to community structure datasets (Garnier and Navas, 2012). The trait information can be either quantitative (e.g. leaf size, root depth) or qualitative (e.g. annual/biennial/perennial, root architecture, resistance to grazing); community structures are represented by statistical distribution of quantitative traits or compositions of qualitative traits (Garnier and Navas, 2012; Violle et al., 2007).

The prevalent use of high-throughput sequencing discussed above has paved the way for microbial ecology to follow the track of plant ecology (Krause et al., 2014), where community members can be identified without state-of-the-art technologies. Accordingly, some recent studies have adopted a trait-based approach to microbial ecology (Figure 1-3); however, the microbial traits used in this context are mostly limited

to genomic features of prokaryotes, namely rRNA gene copy number (Kearns and Shade, 2017; Nemergut et al., 2016) and genome size (Barberán et al., 2014) (note that, due to their small size, many prokaryotic whole genomes have been sequenced and documented, and are now publicly available). Prokaryotes with a high gene copy number generally present high growth rates (i.e. take *r*-strategy) (Roller et al., 2016), and they proliferate after external perturbation or nutrient-rich environment (Kearns and Shade, 2017; Mise et al., 2020; Nemergut et al., 2016). Prokaryotes with large genomes tend to be ubiquitous across different soil types (Barberán et al., 2014), presumably due to their genomes allowing adaptation to various physicochemical conditions (Guieysse and Wuertz, 2012).

Importantly, these trait-based approach works well to address traditional questions in BEF studies. By using information of rRNA copy number, for example, the predictable relationship between biodiversity (dominance of *r*-strategists) and ecosystem functioning (recovery from perturbation; degradation of nutrients) was elucidated. Therefore, trait-based approach would contribute to explain and predict the dynamics and functioning of microbial communities (Oliverio et al., 2017), which is a part of the ultimate goal of microbial ecology (Antwis et al., 2017; Griffiths and Philippot, 2013) (Figure 1-3).

Nevertheless, traits other than genomic features, especially microbial phenotypes and functions, have not been adopted for trait-based approaches, with some rare exceptions explained below. Considering the versatility of trait-based approaches in plant community ecology, expanding trait-based approaches to microbial phenotypes and functionalities should certainly contribute to microbial ecology. Until recently, however, this has been severely hampered because of the poor searchability of microbial trait information.

The cultivated microbial phenotypic/functional information is documented in the form of natural linguistic text (either published as an article in a journal, typically the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM), or included in *Bergey's Manual of Systematic Bacteriology* (Bergey's Manual)). There is currently no readily-accessible and unified platform, similar to the National Center for Biotechnology Information (NCBI), for example, for this information. Only recently, several databases containing microbial trait information have been developed (Barberán et al., 2017; S. Louca et al., 2016; Reimer et al., 2019). Thanks to this advance, microbial ecologist have

just started using a number of phenotypic/functional traits including cell shapes, pigmentations, oxygen tolerances, and metabolic potentials (Choudoir et al., 2018; Guittar et al., 2019; Stilianos Louca et al., 2016).

While the usefulness of this approach is noteworthy, its limitation could be severe. As mentioned above, a high proportion of environmental microbes are yet to be cultivated, and therefore their biological traits have not been studied. For example, phyla *Acidobacteria* and *Verrucomicrobia*, which occupy typically 5–25% of soil bacteria (Delgado-Baquerizo et al., 2018), currently harbor limited number of isolated strains. In addition, phenotypes of microbes under laboratory conditions and in the natural environment may greatly differ. In fact, a large proportion of environmental microbes are viable but non-culturable (VBNC) (Nosho et al., 2018), and microbes in VBNC states have been shown to present different gene expression profiles from culturable ones (Giagnoni et al., 2018). This indicates that currently available trait information is commonly biased towards an "active" status.

Furthermore, regarding the culturable microbes, the availability of microbial phenotypic information is strongly dependent on the microbial taxonomic system. Basic phenotypes, such as cell shapes, oxygen tolerance, and optimum growth temperature have been investigated and recorded for almost every microbial taxon. On the other hand, other useful phenotypic information, such as metabolic potentials, are not recorded unless that phenotypic information helps distinguish between the species (or subspecies, strains, etc.) in question.[1]

1-5 Novel trait-based approach: habitat-based analysis

Yet another fundamental microbial trait that has not yet been focused in the context of trait-based approach is habitat information (Thompson et al., 2017). Species that inhabit seawater are more likely to have the trait of adaptability to saline environments than species that inhabit freshwater only. Likewise, species that inhabit animal gut are more

[1] Publication policy of *International Journal of Systematic and Evolutionary Microbiology* states: "For a description of a new taxon, the following must be included with the submitted article [...]: ..... 3. A list of characteristics considered essential for membership in the taxon. 4. A list of characteristics which qualify the taxon for membership in the next higher taxon. 5. A list of diagnostic characteristics, i.e. characters which distinguish the taxon from closely related taxa. ....." (https://www.microbiologyresearch.org/journal/ijsem/scope; viewed December 12, 2019)

likely to have the trait of adaptability to copiotrophic (eutrophic) environments than species that inhabit soil only. Thus, habitat information is expected to provide insights into ecological and physiological characteristics of microbial communities.

More importantly, habitat information is free of the practical problems mentioned above. Habitat information can be obtained without the reliance on cultivation and isolation experiments because accumulating shotgun metagenomic datasets, accompanied by environmental source information, themselves provide rich information about the environmental distribution of both cultured and non-cultured species. Environmental source information is recorded as the metadata in International Nucleotide Sequence Database Collaboration (INSDC; NCBI SRA/EMBL-EBI ERA/DDBJ DRA) (Karsch-Mizrachi et al., 2018) in accordance with the NCBI ontology systems. Additionally, shotgun metagenomic datasets can be systematically obtained from the INSDC without cumbersome curation.

1-6 Overview of this dissertation

In this dissertation, I aimed to propose habitat-based analysis of microbial communities and prove its usefulness for explaining and/or predicting microbial community dynamics and/or functioning (Figure 1-3). In Chapter 2, I describe the construction of a prokaryotic habitat database, ProkAtlas, which is the pre-requisite for performing habitat-based analysis of prokaryotic community structures. I also discuss possible uses of ProkAtlas (or the concept of habitat-based analysis itself), beyond its application to prokaryotic community data. In Chapter 3, I present empirical examples of habitat-based analysis using ProkAtlas, covering different environments and various aspects of ecology. In Chapter 4, I highlight the possible implications for future microbial ecology, as well as for biology in general.

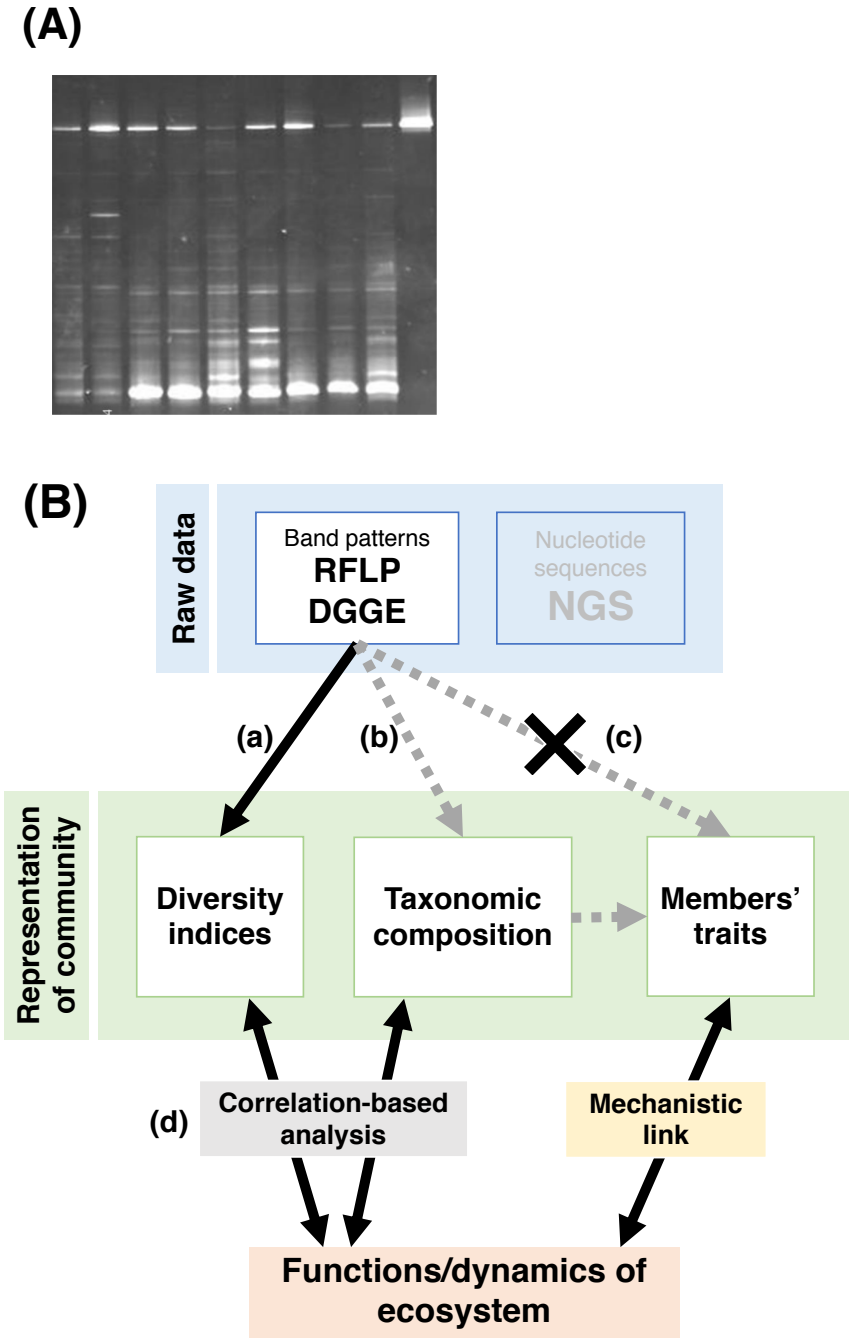


Figure 1-1 (A) An example of band pattern dataset obtained by denaturing gradient gel electrophoresis (DGGE) (Mise et al., unpublished). One lane represents one microbial community. Alpha-diversity of each community can be calculated by the number of bands observed, whereas beta-diversity between communities can be estimated by the difference in band patterns between a pair of lanes. **(B)** Schematic illustration of microbial community ecology before the prevalent use of next-generation sequencers

(NGS). Molecular methods for detecting single nucleotide polymorphisms, such as restriction fragment length polymorphism (RFLP) and DGGE, only provide diversity indices such as Shannon indices and Bray-Curtis dissimilarities (a). When combined with clone library analysis, coarse taxonomic compositions could be estimated; however, clone library analysis is costly and laborious, and therefore practically not applicable in most cases (b). Traits of community members cannot be inferred from the band patterns of RFLP or DGGE (c). The correlations between the diversity indices and functions/dynamics of ecosystems can be statistically tested (d).

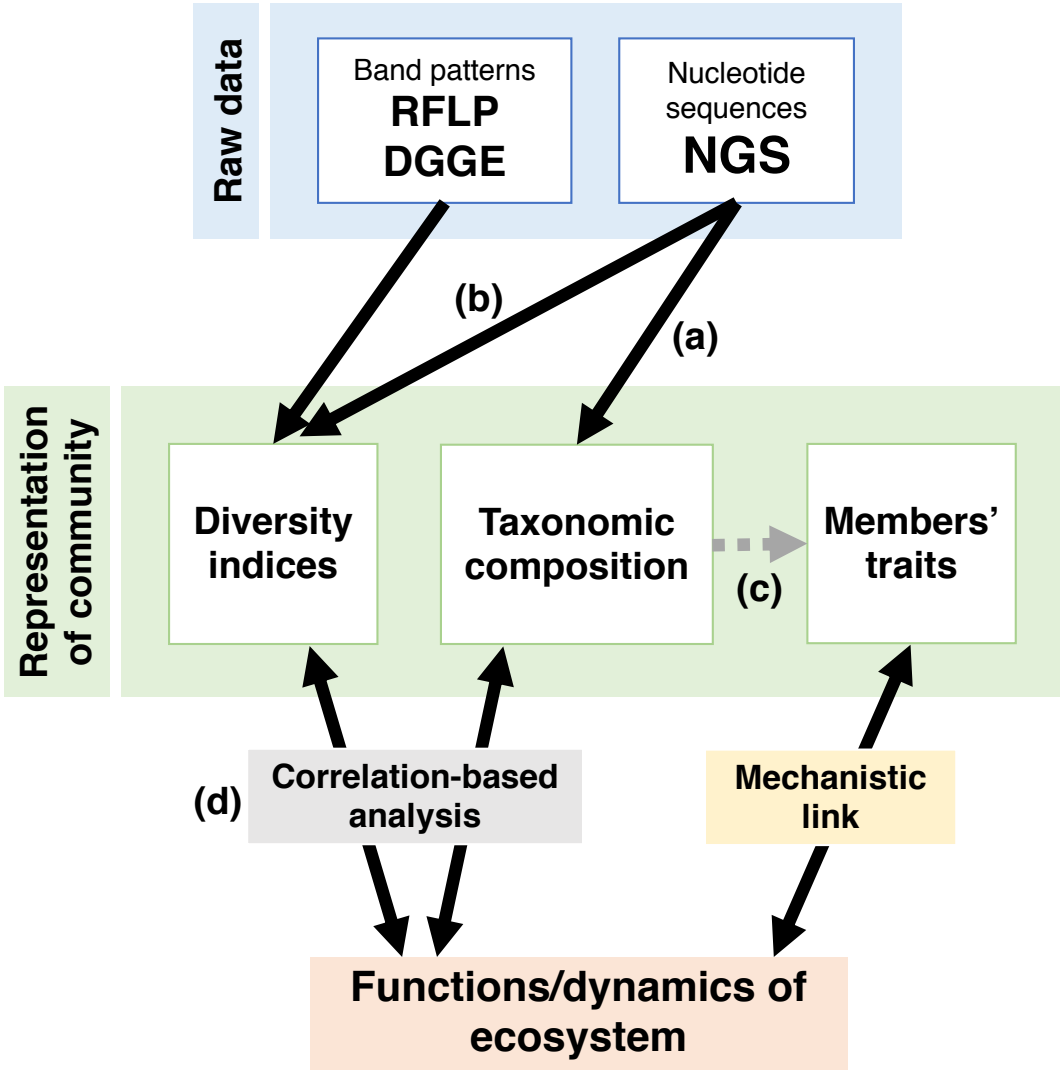


Figure 1-2 Schematic illustration of microbial community ecology in the era of next-generation sequencers (NGS). NGS provides high-resolution information on taxonomic compositions (a) and diversity patterns (b) of microbial communities. The traits of microbial community members cannot be directly inferred from the taxonomic composition in many cases, with rare exceptions of phylogenetically-conserved traits such as nitrification and methanogenesis (Isobe et al., 2019; Martiny et al., 2015) (c). Aside from such exceptions, the NGS data are ultimately subjected to correlation-based analyses, in the same way as RFLP and DGGE data (d).

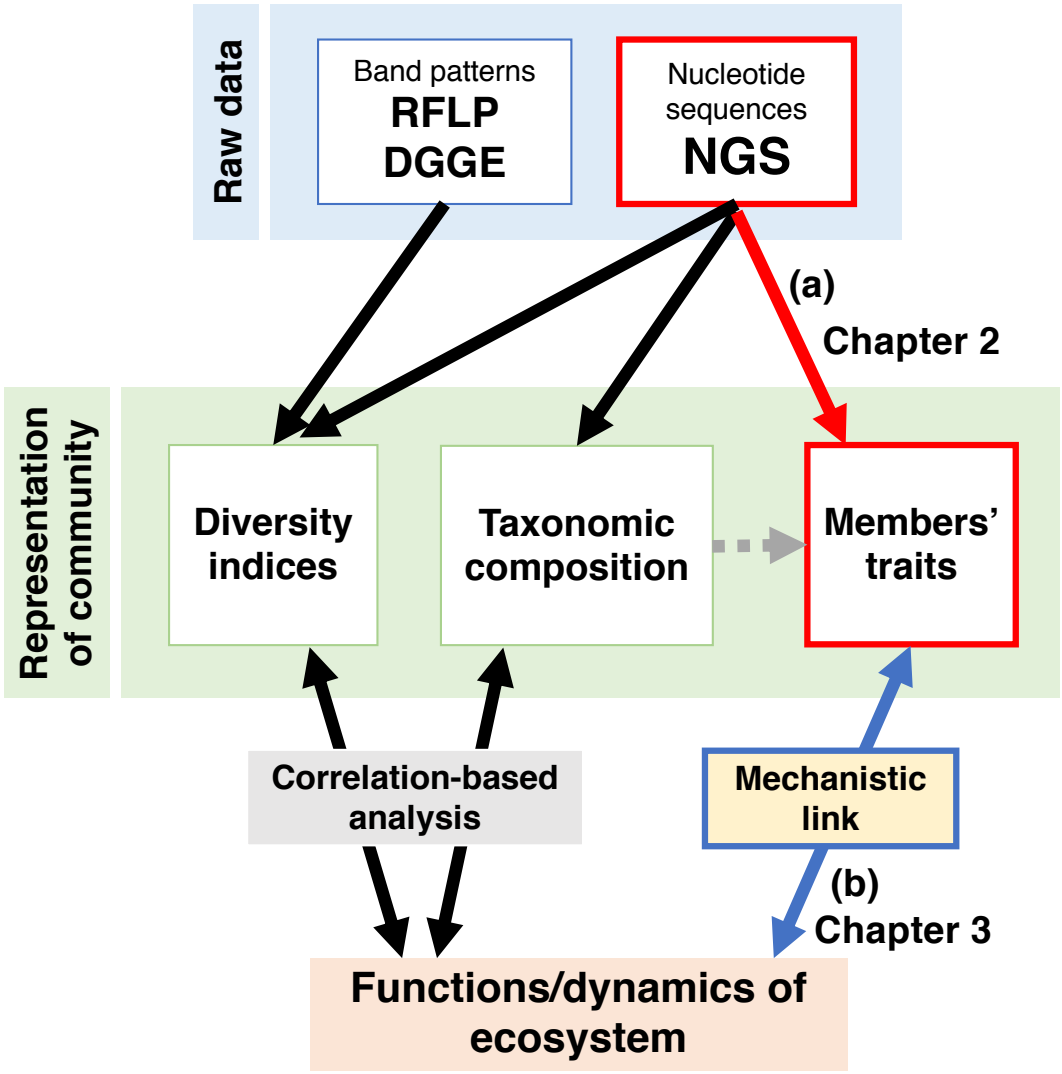


Figure 1-3 Schematic illustration of microbial community ecology and trait-based approach, presenting how this dissertation advances the academic field. High-resolution information, provided by NGS, can be used to infer the traits of microbial community members. In Chapter 2 of this dissertation, I present a novel method for this inference (a). The traits of community members can be mechanistically linked with the functions and dynamics of ecosystems, which is the strength specific to trait-based approach. In Chapter 3 of this dissertation, I provide some examples of such mechanistic links, intending to prove the significance of the method proposed in Chapter2 (b).

Chapter 2 | Construction of habitat database and its evaluation[2]

2-1 Introduction

In spite of the recent substantial increase in microbial community structure data, the methods for interpreting prokaryotic community structure datasets from ecological perspectives are still insufficient. As a promising solution to this problem, the trait-based approach can help ecological interpretation of community structure datasets (Barberán et al., 2014; Nemergut et al., 2016). The basic idea of this approach is to first classify prokaryotes according to their ecological or physiological traits and then project that trait information to community structure datasets. This means that trait-based approach is largely dependent on the quality and comprehensiveness of trait database. In fact, most of the trait-based microbial community studies focused on genomic traits such as genome size (Barberán et al., 2014) and rRNA gene copy number (Nemergut et al., 2016), which are available in authoritative public database.

The primary aim of the present study is to propose habitat-based analysis of microbial communities (see Chapter 1). For this analysis to work, ready-to-use microbial habit database is indispensable as discussed above. As an existing microbial habit database, MetaMetaDB, which links prokaryotic 16S rRNA gene sequences to environments, may be useful (Yang and Iwasaki, 2014). MetaMetaDB was constructed by curating pyrosequencing data from environmental samples, both amplicon sequencing data and shotgun metagenomic data, registered in INSDC. It contains only 16S rRNA gene sequences, and each of the entries are labeled with one environmental category representing the source of the sequenced sample. However, the current form of MetaMetaDB suffers from a number of shortcomings. First, the size of research project is not considered, meaning that large projects dealing with many samples tend to be overrepresented. Second, because amplicon sequencing data were incorporated, the database inevitably bears primer biases (Klindworth et al., 2012). Moreover, the prevalent use of non-universal primers of 16S rRNA genes targeting specific clades of bacteria (Pfeiffer et al., 2014) critically hinders the balanced representation of prokaryotic community members in the environment. Third, MetaMetaDB has not been updated since 2014, and contain no sequences generated by state-of-the-art Illumina sequencers.

---

[2] The content of this chapter has been partly published as the following paper:
Mise and Iwasaki, 2020. Environmental atlas of prokaryotes enables powerful and intuitive habitat-based analysis of community structures. *iScience* (Cell Press).

In this chapter, I report the development of extended and refined prokaryotic habitat database, named ProkAtlas that links 16S rRNA gene sequences to prokaryotic habitats, carrying on the basic structure of MetaMetaDB.

2-2 Materials and methods

2-2-1 ProkAtlas database construction: data collection

The overall procedure of constructing ProkAtlas is illustrated in Figure 2-1 (A). I fetched all metagenomic sequence entries with environmental categories (NCBI taxon IDs) under 410657 (ecological metagenomes) and 410656 (organismal metagenomes) on July 4, 2018. Those entries contained metagenomic data from diverse environments and were based on different sequencing platforms and library construction strategies. I selected entries annotated as whole-genome sequencing (WGS) (i.e., shotgun sequencing data) to avoid PCR-biased data due to amplicon sequencing (Klindworth et al., 2012). I further selected entries generated by the most popular platform, i.e., Illumina sequencers, but excluded entries generated by HiSeq 3000, 4000, or X because of their potential inaccuracies (Sinha et al., 2017). The filtered entries included 5,368 projects, which contained 1–3,693 runs each. To avoid datasets being too biased towards data from specific projects with high numbers of samples (Ramirez et al., 2018) and to keep the database size small, up to ten runs were randomly selected from each project (Figure 2-1(B)). For each single-end or paired-end sequencing run, one or two gzipped fastq file(s), respectively, were downloaded from the ftp server of the European Nucleotide Archive (ENA). In the rare cases in which a gzipped fastq file exceeded the data size of 200 MB, the first 200 MB was retrieved.

2-2-2 ProkAtlas database construction: data processing

Paired-end sequences with overlapping regions of 20 bp or longer were merged using USERACH v11.0.667 (Edgar, 2010), while single-end sequences were used as they were. Low quality regions (Q-score < 20) at the 3′-ends were pruned, and sequences with mean Q-scores of less than 30 were discarded using PRINSEQ 0.20.4 (Schmieder and Edwards, 2011). PARTIE (Torres et al., 2017) was used to remove amplicon sequence files mistakenly annotated as WGS. Sequences longer than 600 bases (the maximum read length of Illumina MiSeq and HiSeq) were removed, because they were likely artifacts. SortMeRNA 2.1 (Kopylova et al., 2012) trained with SILVA v132 Nr99 (Quast et al., 2013)

(hereafter referred to as SILVA) with the default parameter settings was used to extract 16S rRNA gene regions from query sequences. From the SILVA database used in this study, eukaryotic sequences (i.e. those annotated as "Eukaryota" at the kingdom/domain level) had been removed beforehand, retaining only prokaryotic 16S rRNA gene sequences. To filter out non-16S rRNA hits that mingled in the output sequences from SortMeRNA, they were further subjected to a BLASTn (BLAST+ 2.3.0) search against SILVA with an e-value threshold of 1E−10. For each query sequence, an alignment covering the longest part of query sequence was selected among the top 100 hits (in bitscore), and the aligned region of that query sequence was retrieved. When multiple hits tie in alignment length, one with the highest bitscore was chosen. If the longest-aligned query region was shorter than 150 bases excluding gaps, that sequence was removed. As a result, 1–43,259 rRNA gene sequences per project were obtained. Again, to prevent datasets from being biased towards data from specific big projects and to keep the database size small, I randomly sampled up to 100 sequences from each project (Figure 2-1(C)). Finally, I compiled 361,474 rRNA gene sequences, each retaining environmental category information (Table 2-1) that was accompanied by the original sequence dataset in DRA/ERA/SRA as a taxon ID. Note that each sequence in ProkAtlas is labeled by one environmental category. To test if the randomness of the sampling step affects results and if the sampling of 100 sequences from each project is enough, I prepared five additional alternative datasets: two by sampling 100 sequences (sets A and B) and three by sampling 500 sequences (sets C, D, and E; each contained 1,412,963 rRNA gene sequences).

2-2-3 ProkAtlas pipeline and its implementation

For habitat-based analysis of 16S rRNA gene sequence data, the ProkAtlas pipeline projects the associated environmental category data in ProkAtlas (originally presented as taxon IDs in DRA/ERA/SRA) to query sequences. A query can be either a single sequence from individual prokaryotic genome or a prokaryotic community dataset consisting of OTUs (or sub-OTUs, amplicon sequence variants) representative sequences and an OTU table (typically from amplicon and shotgun metagenomic sequencing). The ProkAtlas pipeline characterizes each query sequence or community with habitat preference scores, a vector denoting the composition of possible habitats inferred from compiled metagenomic sequences. A schematic illustration of the ProkAtlas pipeline is

provided as Figure 2-2. The pipeline consists of two parts, namely BLASTn search against ProkAtlas database and calculation of habitat preference scores based on the hits retrieved by the BLASTn search.

The ProkAtlas pipeline uses a BLASTn search to query each input sequence against ProkAtlas, and all hits under an e-value threshold of 1E−5 are collected (the other parameters are set to default). Partial alignments are accepted because both query sequences and ProkAtlas entries can contain partial 16S rRNA genes (Figure 2-3 (A)–(D)); however, hits harboring mismatches longer than 2 bp at either end of the alignment (Figure 2-3 (E)(F)) are ignored because they may be erroneous hits. The hits are further filtered to satisfy sequence similarity and alignment length criteria. The default value for sequence similarity and alignment length criteria are 97% and 150 bp, respectively, which is discussed later in this chapter.

The habitat preference of a prokaryote or a prokaryotic community can be represented by a composition of environmental categories within the list of significant hits (Figure 2-2); however, simply counting a number of hits that are labeled with each environmental category may incorrectly emphasize hits to environments that are frequently studied, such as human gut. Therefore, the contribution of each environmental category is weighted by the log-transformed reciprocal of the proportion of sequences in that category within ProkAtlas (Yang and Iwasaki, 2014). This diminishes and increases the habitat preference scores of overrepresented and underrepresented categories, respectively.

Mathematically, a habitat preference score of a prokaryote or a prokaryotic community for each environmental category is defined by:

$$\mathbb{E}_i = \frac{\left(n^i_{soil}, n^i_{marine}, n^i_{gut}, \ldots\right)}{\sum\left(n^i_{environment}\right)};$$

$$\mathbb{E} = \sum_i (\mathbb{E}_i \times C_i) \Big/ \sum_i C_i ;$$

$$\mathbb{W} = \left(\log\left(\frac{R_{tot}}{R_{soil}}\right), \log\left(\frac{R_{tot}}{R_{marine}}\right), \log\left(\frac{R_{tot}}{R_{gut}}\right), \ldots\right);$$

$$habitat\ preference\ score = \frac{\mathbb{E} \circ \mathbb{W}}{\sum(\mathbb{E} \circ \mathbb{W})}$$

where n^i_X is the number of significant hits to OTU i within a specific environmental category X, $\mathbb{E}_i$ is the environmental vector denoting the habitat preference of OTU i, C_i

is the number of read counts of OTU i, $\mathbb{E}$ is the average of environmental vectors weighted by the read count of each OTU (i.e. C_i), R_{tot} and R_X are the number of ProkAtlas entries in total and within the environmental category X, respectively, and $\mathbb{W}$ is the vector of weighing factors of each environmental category. The arithmetic operator $\circ$ indicates the element-wise multiplication of two vectors with the same length (Hadamard product). When applied to a single prokaryotic sequence to illustrate the habitat preference of the corresponding microbe rather than community characteristics, the query is treated as a community composed of one OTU and one read (i.e., $\mathbb{E} = \mathbb{E}_i$).

The ProkAtlas database and pipeline are available at https://msk33.github.io/prokatlas.html.

2-2-4 Bird's-eye visualization of prokaryote coappearance network

Because I constructed ProkAtlas using shotgun metagenomic sequences only, each of the sequences in ProkAtlas covers different regions of 16S rRNA genes. To compare these staggered sequences, they were mapped to SILVA using BLASTn search and subjected to closed-reference clustering. More specifically, up to 100 top hits (ranked by bitscores) were retrieved after the BLASTn search. Following the principle of parsimony, the greedy algorithm was employed to obtain the (approximately) smallest subset of SILVA entries containing at least one top hit for every query sequence (Chvatal, 1979). Then, for each environmental category, the number of sequences associated with each SILVA entry was counted. Of the 115 environmental categories, 27 categories harboring more than 2,000 sequences successfully mapped to SILVA were subjected to visualization. Bray-Curtis dissimilarities between the SILVA entry composition vectors associated with the environmental categories and betweenness centralities were calculated and their network was visualized using the *sna* package on R ver3.5.1 (R Core Team, 2017).

2-2-5 Application to 16S rRNA gene sequences of isolated and non-isolated prokaryotes

I downloaded 16S rRNA gene sequences of pure-isolated bacterial strains from manually curated IJSEM phenotypic database (https://doi.org/10.6084/m9.figshare.427239, as of October 2018) (Barberán et al., 2017). In addition, I downloaded 16S rRNA gene sequences produced from a large SAG sequencing project (Rinke et al., 2013). The

ProkAtlas pipeline with the default parameter settings was used for habitat estimation, with an exception that I used three different alignment length thresholds, namely 150 (recommended value), 200, and 250 bases, to check the robustness of the pipeline. In addition, to test whether the random sampling process in constructing ProkAtlas affects the results, I performed the same analysis using the five alternative datasets as described above.

For each set of estimated habitat compositions, consistency with the source-environment information was tested. To test if estimated habitat compositions of soil-derived isolates are actually soil-related, the scores of environmental categories related to soil (namely "soil", "rhizosphere", "rice paddy", and "wetland") were compared between soil-derived and other isolates using the Mann-Whitney U-test.

2-3 Results and discussion

2-3-1 ProkAtlas database and pipeline

ProkAtlas was developed as a comprehensive database of prokaryotic habitat traits based on a meta-analysis of metagenome shotgun sequencing datasets (Figure 2-1). It comprises 361,474 16S rRNA gene sequences from 5,368 shotgun metagenome projects registered in the INSDC SRA/ERA/DRA databases. Notably, to achieve reliable but efficient prokaryotic habitat estimation, I tried to balance the database comprehensiveness and smallness. As discussed later, increasing the size of ProkAtlas marginally affects or improves the performance of habitat preference prediction, while computational cost linearly increases. It is also notable that the number of 16S rRNA gene sequences in ProkAtlas is comparable to those in Greengenes and SILVA (Glöckner et al., 2017; McDonald et al., 2012). Each sequence in ProkAtlas is labeled with one of the environmental categories listed in Table 2-1 for prokaryotic habitat estimation with 16S rRNA gene sequences. Although NCBI taxon IDs contain environmental categories of different granularity, they are accompanied by all of the metagenomic samples in DRA/ERA/SRA and therefore suitable for constructing a database that covers a wide variety of environments. The four major environmental categories, *soil*, *marine*, *freshwater*, and *human_gut*, comprise 72.6% of all sequences, and the top 27 categories comprise 90% of all sequences.

2-3-2 Bird's-eye visualization of prokaryote coappearance network among diverse environments

By enumerating 16S rRNA gene sequences that coappear in different environments using ProkAtlas, I obtained a comprehensive view of coappearance of prokaryotes among diverse environments as a network (Figure 2-4(A)). All ProkAtlas sequences were mapped to 39,049 SILVA entries, and their composition in each environmental category was quantified. Beta-diversities between the environmental categories were then calculated using the Bray-Curtis dissimilarity metric.

As expected, related environments such as *soil* and *rhizosphere* were strongly associated with each other. The betweenness centrality, which quantify propagation of prokaryotes among different environmental categories, of extreme environment (i.e., *hydrothermal_vent*) was low (Figure 2-4(B)). Regarding this observation, it is reasonable that prokaryote coappearances or migrations via extreme environments are rare because of their non-moderate conditions and geographical isolation. It was also found that information centralities of host-associated environments were relatively low. This observation was rather unexpected because prokaryotic hosts, especially animals, are generally expected to bring prokaryotes to different environments and promote their migration (Grossart et al., 2010). I assume that strong prokaryote-host dependencies prohibit prokaryotes from settling in new environments, regardless of their hosts' movement, and that prokaryotic hosts may actually have limited roles in shaping microbial distributions across the earth.

2-3-3 Consistency between sources of isolated and non-isolated prokaryotes and ProkAtlas habitat estimation

ProkAtlas was applied to 1,021 (nearly) full-length 16S rRNA gene sequences of pure-isolated bacterial strains from the International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database (Barberán et al., 2017), as well as to 201 16S rRNA gene sequences retrieved from a large SAG sequencing project (Rinke et al., 2013). All sequences of pure-isolates and 183 (91.0%) sequences from SAGs had one or more significant hits in ProkAtlas. The habitat preference scores were overall consistent with their environmental sources: scores of soil-related environmental categories (namely *soil*, *rhizosphere*, *rice_paddy*, and *wetland*), for example, were significantly higher in soil-derived sequences compared with sequences of isolates from all the other environments

(Mann-Whitney U-test, $P < 0.001$). This trend was similarly observed for various sets of environmental categories, including isolates and/or SAGs from seawater, plants, feces, groundwater, lake water, hydrothermal vent, and bioreactors (Figure 2-5). On the other hand, habitats estimated by ProkAtlas were inconsistent with the actual environmental sources for a portion of individual query sequences. Such conflict may be attributed to the fact that many prokaryotic species are distributed in broad ranges of environments (Sriswasdi et al., 2017), and isolation sources of cultured strains or sampling sites of SAGs could be actually rare habitats of that prokaryotic group. That means, while estimated habitat of a specific individual prokaryote can be sometimes incorrect, habitat preference scores of a prokaryotic community consisting of multiple species can still be an informative proxy of that community. In addition, the abovementioned trends were reproduced when the alignment length thresholds were raised to 200 or 250 bases (Figure 2-6) or when one of the five alternative datasets (sets A–E) was used (Figure 2-7). Because of this, I assume that 150 bp threshold (a default value in my pipeline) and database size would be respectively long and large enough to achieve overall accuracy, while retaining enough amount of significant hits and saving computational cost.

Table 2-1 Environmental categories, numbers of 16S rRNA gene sequences labeled by these categories, and numbers of research projects of these categories contributing to ProkAtlas. The environmental categories were based on annotations in the NCBI SRA database. Note that due to data processing, several environmental categories are associated with a few sequences.

Environmental Category	Number of sequences	Number of projects
activated_carbon	44	1
activated_sludge	7383	78
air	300	3
algae	308	4
anaerobic_digester	500	5
annelid	1038	13
ant	300	3
aquatic	2667	29
aquifer	1900	19
bat	143	2
beach_sand	100	1
biofilm	300	3
biofilter	100	1
biogas_fermenter	600	6
bioreactor	2923	32
bioreactor_sludge	200	2
biosolids	200	2
bird	100	1
bovine	200	2
bovine_gut	700	7
cave	100	1
chicken_gut	731	9
compost	500	5
coral	1100	11
crab	100	1
crustacean	100	1
endophyte	30	1
epibiont	100	1
estuary	200	2
feces	1306	14
fermentation	300	3
fish_gut	25	1
food	604	7
food_fermentation	300	3
food_production	100	1
fossil	200	2
freshwater	43216	437
freshwater_sediment	13744	239
fungus	3737	38
glacier	500	5
groundwater	9540	97
gut	3167	36
halite	14	1
hot_springs	1056	12
human_bile	111	2
human_blood	1	1
human_eye	131	2
human	1815	25
human_gut	7502	80
human_lung	291	3
human_oral	500	5
human_reproductive_system	100	1
human_skin	400	4
hydrocarbon	33	1
hydrothermal_vent	4016	44
hypersaline_lake	900	9
hypolithon	113	2
indoor	100	1
insect	241	4
insect_gut	200	2
invertebrate	300	3
lake_water	5700	57
landfill	400	4
leaf	500	5
lichen	500	5
marine	45298	461
marine_sediment	3281	35
microbial_fuel_cell	100	1
microbial_mat	1561	17
mine_drainage	200	2
mine_tailings	200	2
mixed_culture	100	1
money	200	2
mosquito	180	2
moss	600	6
mouse_gut	1194	18
oil_field	3	1
oral	66	1
oyster	100	1
paper_pulp	100	1
parasite	26	1
peat	7683	77
permafrost	1449	16
phyllosphere	12100	121
pig_gut	576	6
plant	4579	53
plastic	6	3
pollen	200	2
rat_gut	100	1
rhizosphere	21152	213
rice_paddy	3100	31
rock	148	2
rock_porewater	100	1
root_associated_fungus	100	4
root	400	1
salt_lake	1900	19
salt_marsh	5400	54
sea_squirt	400	4
seawater	2977	30
sediment	6431	66
skin	100	1
sludge	100	1
soil	90158	984
sponge	300	3
stromatolite	100	1
subsurface	3020	32
surface	100	1
symbiont	69	1
termite_gut	2919	31
terrestrial	2134	22
tick	100	1
urban	6	1
viral	600	6
wastewater	3228	35
wetland	11900	119


(A)
DRA/ERA/SRA
Shotgun metagenomic data
✓ Random sampling of entries in big projects
✓ Quality filtering
✓ 16S rRNA gene selection
✓ Reduction of data from big projects
Filtered sequences
ProkAtlas
Soil
Seawater
Plant
Sediment
Human_gut
(B)
DRA/ERA/SRA
ERP002391
ERR247139 ERR247140 ERR247141 ERR247142 ERR247143 ERR247144 ERR247145 ERR247146 ERR247147
ERR247148 ERR247149 ERR247150 ERR247151 ERR247152 ERR247153 ERR247154 ERR247155 ERR247156
ERR247157 ERR247158 ERR247159 ERR247160 ERR247161 ERR247162 ERR247163 ERR247164 ERR247165
ERR247166 ERR247167 ERR247168 ERR247169 ERR247170 ERR247171 ERR247172 ERR247173 ERR247174
ERP002194
ERR248260 ERR248261 ERR248262 ERR248263
ERP009244
ERR726368 ERR726369
ERP009277
ERR732063 ERR732064
Randomly-picked entries
ERP002391
ERR247139 ERR247140 ERR247141 ERR247142 ERR247143 ERR247144 ERR247145 ERR247147
ERR247149 ERR247155 ERR247156 ERR247164 ERR247166
ERP002194
ERR248260 ERR248261 ERR248262 ERR248263
ERP009244
ERR726368 ERR726369
ERP009277
ERR732063 ERR732064
(C)
Quality-filtered metagenomic sequences
ERP002391
ERP002194
ERP009244
ERP009277
16S rRNA gene sequences
ERP002391
ERP 002194
ERP 009244
ERP 009277
Reduced dataset of 16S rRNA sequences
ERP 002391
ERP 002194
ERP 009244
ERP 009277

Figure 2-1 Schematic illustration of ProkAtlas construction. The symbol colors and sizes represent the sample sources and data sizes, respectively. Six-digit numbers starting with "ERP" and "ERR" indicate accession number of entries registered in DRA/ERA/SRA: a project (a set of data obtained under the same goal; canonically termed "BioProject") and a run (a set of nucleotide sequence file(s) that derives from one library in one experiment), respectively. **(A)** Overview of the procedure of ProkAtlas construction. To construct a high-quality, small, and comprehensive database with minimized biases, I repeatedly screened and sampled metagenomic sequences. **(B)** Random sampling of entries in big projects that harbor more than ten runs. The data of such big projects were reduced to ten runs per project by random sampling. The randomly-picked entries (lower panel) were subjected to downstream processing described in (A). **(C)** Reduction of data from big projects. 16S rRNA gene sequences were extracted from quality-filtered metagenomic sequences (upper panel). Here, sequences from all runs in one project were pooled together. The number of extracted 16S rRNA gene sequences greatly varied between projects (middle panel); therefore, to mitigate biases, the data of projects harboring more than 100 sequences at this stage were reduced by random sampling (lower panel).

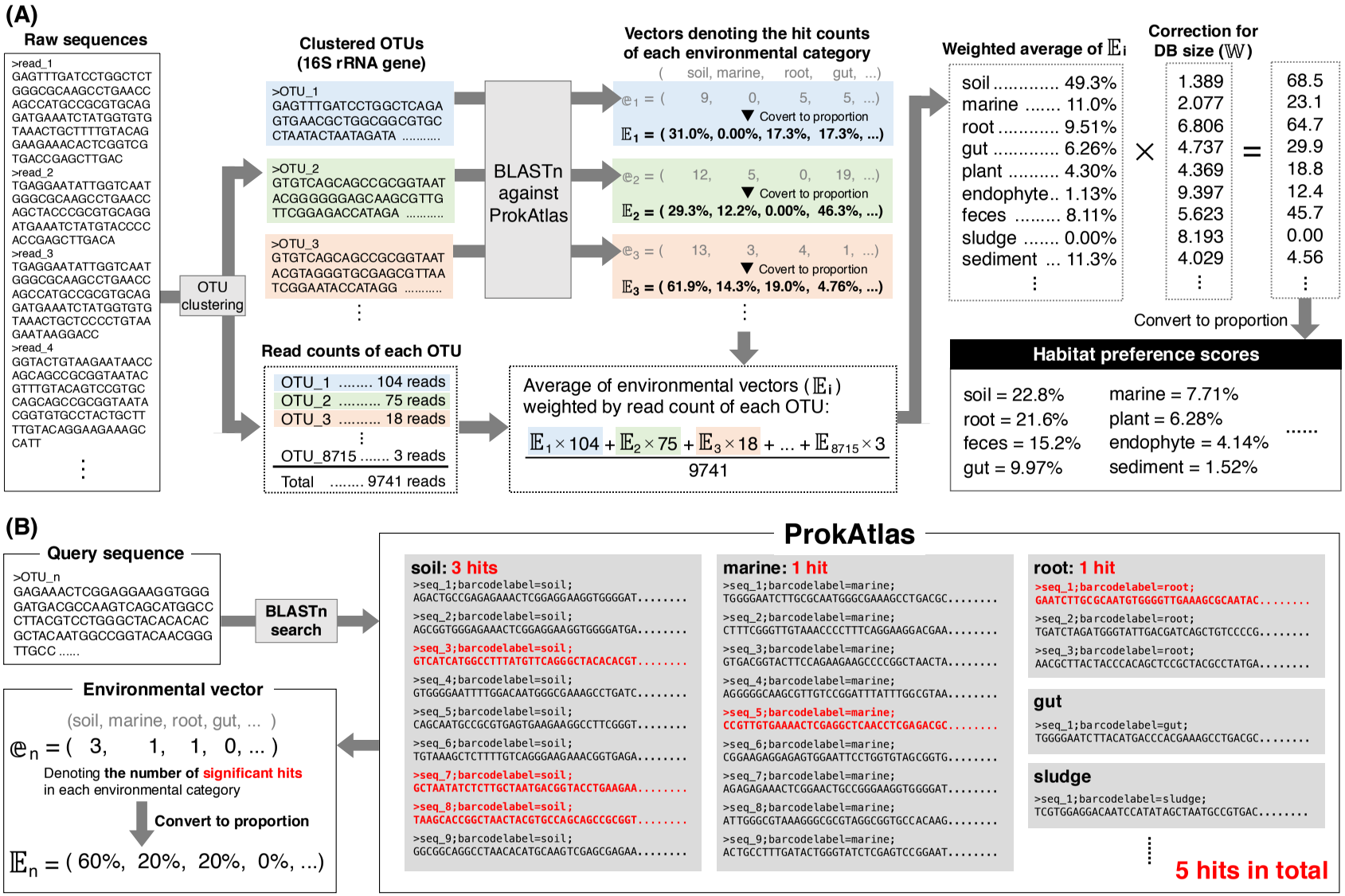

(A)
Raw sequences
>read_1
GAGTTTGATCCTGGCTCT
GGGCGCAAGCCTGAACC
AGCCATGCCGCGTGCAG
GATGAAATCTATGGTGTG
TAAACTGCTTTTGTACAG
GAAGAAACACTCGGTCG
TGACCGAGCTTGAC
>read_2
TGAGGAATATTGGTCAAT
GGGCGCAAGCCTGAACC
AGCTACCCGCGTGCAGG
ATGAAATCTATGTACCCC
ACCGAGCTTGACA
>read_3
TGAGGAATATTGGTCAAT
GGGCGCAAGCCTGAACC
AGCCATGCCGCGTGCAG
GATGAAATCTATGGTGTG
TAAACTGCTCCCCTGTAA
GAATAAGGACC
>read_4
GGTACTGTAAGAATAACC
AGCAGCCGCGGTAATAC
GTTTGTACAGTCCGTGC
CAGCAGCCGCGGTAATA
CGGTGTGCCTACTGCTT
TTGTACAGGAAGAAAGC
CATT
OTU clustering
Clustered OTUs (16S rRNA gene)
>OTU_1
GAGTTTGATCCTGGCTCAGA
GTGAACGCTGGCGGCGTGC
CTAATACTAATAGATA .....
>OTU_2
GTGTCAGCAGCCGCGGTAAT
ACGGGGGGAGCAAGCGTTG
TTCGGAGACCATAGA .....
>OTU_3
GTGTCAGCAGCCGCGGTAAT
ACGTAGGGTGCGAGCGTTAA
TCGGAATACCATAGG .....
BLASTn against ProkAtlas
Vectors denoting the hit counts of each environmental category
(soil, marine, root, gut, ...)
𝕖1 = (9, 0, 5, 5, ...)
Covert to proportion
𝔼1 = (31.0%, 0.00%, 17.3%, 17.3%, ...)
𝕖2 = (12, 5, 0, 19, ...)
Covert to proportion
𝔼2 = (29.3%, 12.2%, 0.00%, 46.3%, ...)
𝕖3 = (13, 3, 4, 1, ...)
Covert to proportion
𝔼3 = (61.9%, 14.3%, 19.0%, 4.76%, ...)
Read counts of each OTU
OTU_1 ..... 104 reads
OTU_2 ..... 75 reads
OTU_3 ..... 18 reads
OTU_8715 ..... 3 reads
Total ..... 9741 reads
Average of environmental vectors (𝔼i) weighted by read count of each OTU:
(𝔼1 × 104 + 𝔼2 × 75 + 𝔼3 × 18 + ... + 𝔼8715 × 3) / 9741
Weighted average of 𝔼i
soil 49.3%
marine 11.0%
root 9.51%
gut 6.26%
plant 4.30%
endophyte.. 1.13%
feces 8.11%
sludge 0.00%
sediment ... 11.3%
Correction for DB size (𝕎)
1.389
2.077
6.806
4.737
4.369
9.397
5.623
8.193
4.029
68.5
23.1
64.7
29.9
18.8
12.4
45.7
0.00
4.56
Convert to proportion
Habitat preference scores
soil = 22.8%
root = 21.6%
feces = 15.2%
gut = 9.97%
marine = 7.71%
plant = 6.28%
endophyte = 4.14%
sediment = 1.52%
.....
(B)
Query sequence
>OTU_n
GAGAAACTCGGAGGAAGGTGGG
GATGACGCCAAGTCAGCATGGCC
CTTACGTCCTGGGCTACACACAC
GCTACAATGGCCGGTACAACGGG
TTGCC .....
BLASTn search
ProkAtlas
soil: 3 hits
>seq_1;barcodelabel=soil;
AGACTGCCGAGAGAAACTCGGAGGAAGGTGGGGAT.....
>seq_2;barcodelabel=soil;
AGCGGTGGGAGAAACTCGGAGGAAGGTGGGGATGA.....
>seq_3;barcodelabel=soil;
GTCATCATGGCCTTTATGTTCAGGGCTACACACGT.....
>seq_4;barcodelabel=soil;
GTGGGGAATTTTGGACAATGGGCGAAAGCCTGATC.....
>seq_5;barcodelabel=soil;
CAGCAATGCCGCGTGAGTGAAGAAGGCCTTCGGGT.....
>seq_6;barcodelabel=soil;
TGTAAAGCTCTTTTGTCAGGGAAGAAACGGTGAGA.....
>seq_7;barcodelabel=soil;
GCTAATATCTCTTGCTAATGACGGTACCTGAAGAA.....
>seq_8;barcodelabel=soil;
TAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGT.....
>seq_9;barcodelabel=soil;
GGCGGCAGGCCTAACACATGCAAGTCGAGCGAGAA.....
marine: 1 hit
>seq_1;barcodelabel=marine;
TGGGGAATCTTGCGCAATGGGCGAAAGCCTGACGC.....
>seq_2;barcodelabel=marine;
CTTTCGGGTTGTAAACCCCTTTCAGGAAGGACGAA.....
>seq_3;barcodelabel=marine;
GTGACGGTACTTCCAGAAGAAGCCCCGGCTAACTA.....
>seq_4;barcodelabel=marine;
AGGGGGCAAGCGTTGTCCGGATTTATTTGGCGTAA.....
>seq_5;barcodelabel=marine;
CCGTTGTGAAAACTCGAGGCTCAACCTCGAGACGC.....
>seq_6;barcodelabel=marine;
CGGAAGAGGAGAGTGGAATTCCTGGTGTAGCGGTG.....
>seq_7;barcodelabel=marine;
AGAGAGAAACTCGGAACTGCCGGGAAGGTGGGGAT.....
>seq_8;barcodelabel=marine;
ATTGGGCGTAAAGGGCGCGTAGGCGGTGCCACAAG.....
>seq_9;barcodelabel=marine;
ACTGCCTTTGATACTGGGTATCTCGAGTCCGGAAT.....
root: 1 hit
>seq_1;barcodelabel=root;
GAATCTTGCGCAATGTGGGGTTGAAAGCGCAATAC.....
>seq_2;barcodelabel=root;
TGATCTAGATGGGTATTGACGATCAGCTGTCCCCG.....
>seq_3;barcodelabel=root;
AACGCTTACTACCCACAGCTCCGCTACGCCTATGA.....
gut
>seq_1;barcodelabel=gut;
TGGGGAATCTTACATGACCCACGAAAGCCTGACGC.....
sludge
>seq_1;barcodelabel=sludge;
TCGTGGAGGACAATCCATATAGCTAATGCCGTGAC.....
5 hits in total
Environmental vector
(soil, marine, root, gut, ...)
𝕖n = (3, 1, 1, 0, ...)
Denoting the number of significant hits in each environmental category
Convert to proportion
𝔼n = (60%, 20%, 20%, 0%, ...)

Figure 2-2 A schematic illustration of procedures to calculate habitat preference scores for prokaryotic communities. **(A)** Overview of the process. Prokaryotic community composition can be defined by the representative sequence and read count of each OTU. First, to characterize habitat preference of each OTU, the OTU representative sequences are subjected to BLASTn search against ProkAtlas database. Typically, each OTU has significant hits to sequences in multiple environmental categories, and the habitat preferences of each OTU may be represented as the environmental vector $\mathbb{E}$. Overall habitat preference of a community is denoted as the average of $\mathbb{E}$ weighted by the read count of each OTU, followed by the correction for the overrepresentation of well-studied environments like *soil* and *marine*. **(B)** Detailed schema of "one-to-many" mapping of a query sequence on ProkAtlas database. If one query sequence has multiple hits in ProkAtlas, the number of hits for each environmental category is presented as $\mathbb{e}$. Therein, $\mathbb{e}$ is converted to the environmental vector $\mathbb{e}$ denoting the habitat preference of the OTU.

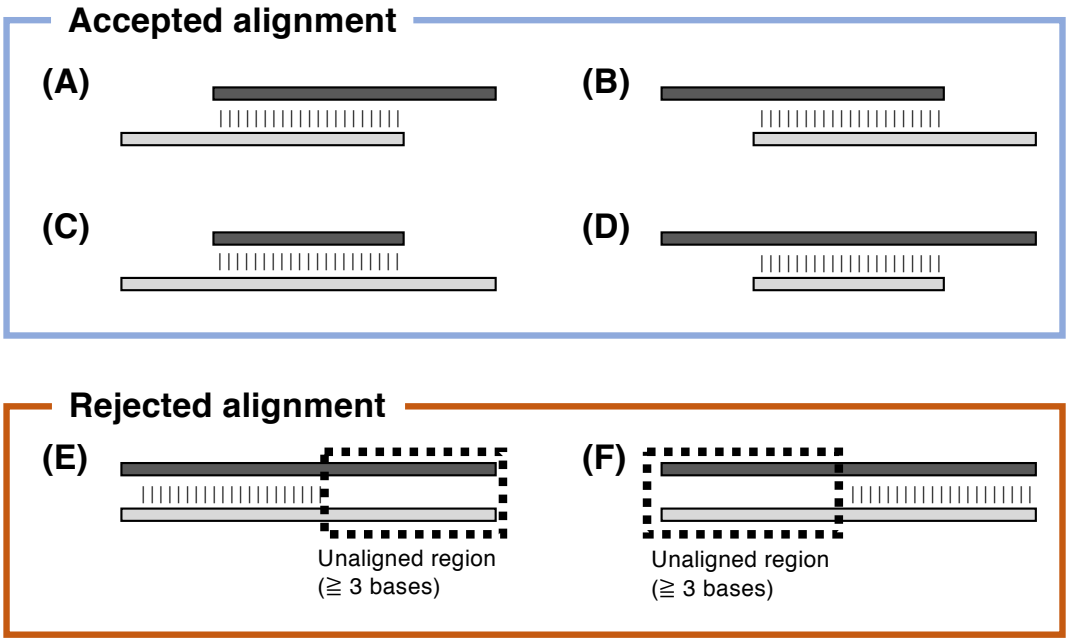


Figure 2-3 Schematic representation of alignment criteria of 16S rRNA gene sequences in the ProkAtlas pipeline. Black and gray rectangles indicate query and subject sequences, respectively. Vertical bars indicate successfully aligned regions by pairwise local alignment. While the upper four partial alignment patterns are accepted, the bottom two patterns are rejected.

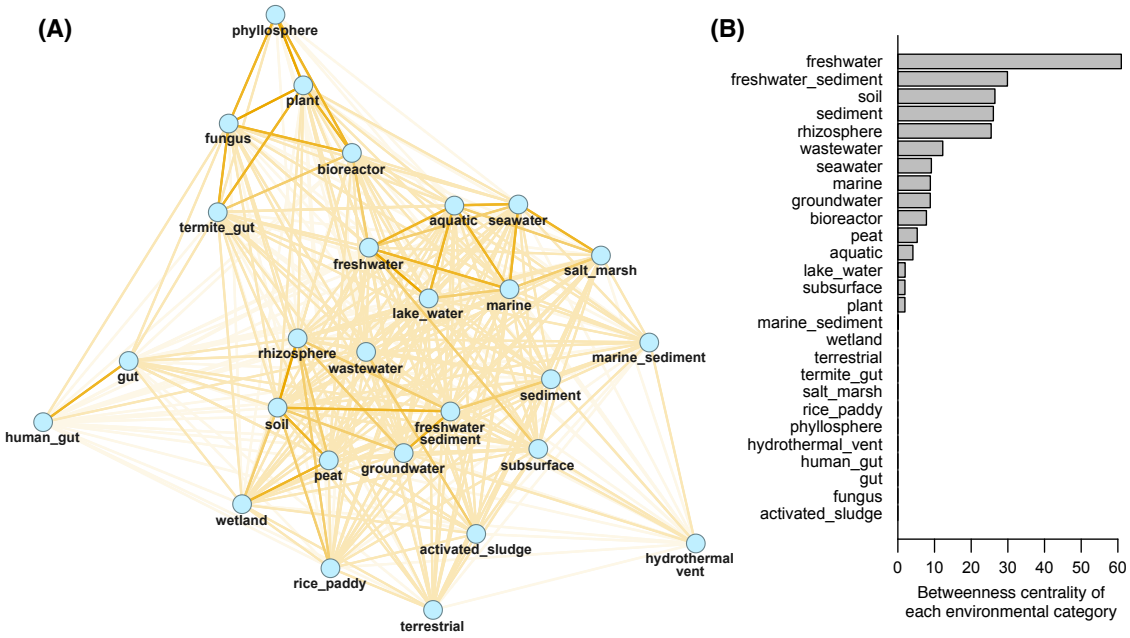


Figure 2-4 (A) A bird's-eye network visualization of the coocurrences. Nodes represent environmental categories. Edges are drawn only if Bray-Curtis dissimilarities are less than 0.9, and their color indicates the Bray-Curtis dissimilarities (smaller in dark than in bright). **(B)** A bar chart showing the betweenness centrality of each environment (i.e. the number of node pairs whose shortest paths contain the node representing that environment).

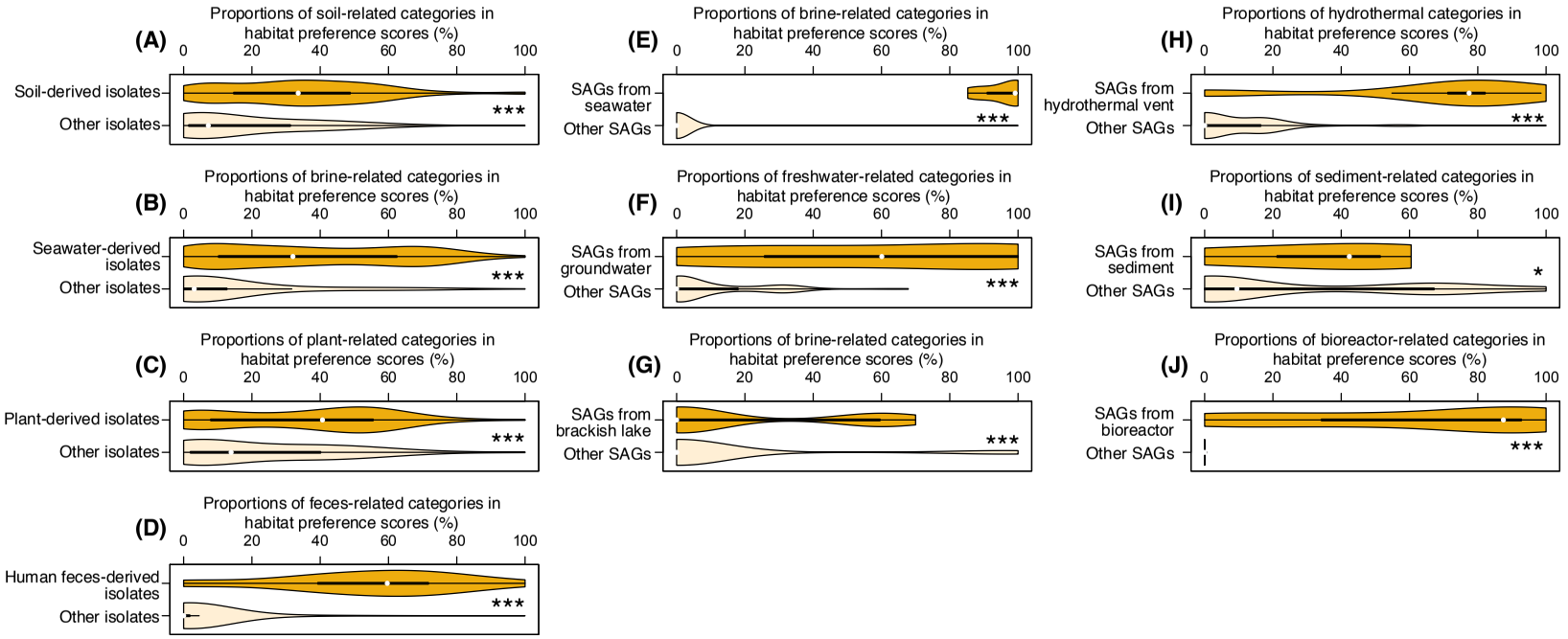

(A)
Proportions of soil-related categories in habitat preference scores (%)
0 20 40 60 80 100
Soil-derived isolates
Other isolates
***
(B)
Proportions of brine-related categories in habitat preference scores (%)
0 20 40 60 80 100
Seawater-derived isolates
Other isolates
***
(C)
Proportions of plant-related categories in habitat preference scores (%)
0 20 40 60 80 100
Plant-derived isolates
Other isolates
***
(D)
Proportions of feces-related categories in habitat preference scores (%)
0 20 40 60 80 100
Human feces-derived isolates
Other isolates
***
(E)
Proportions of brine-related categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from seawater
Other SAGs
***
(F)
Proportions of freshwater-related categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from groundwater
Other SAGs
***
(G)
Proportions of brine-related categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from brackish lake
Other SAGs
***
(H)
Proportions of hydrothermal categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from hydrothermal vent
Other SAGs
***
(I)
Proportions of sediment-related categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from sediment
Other SAGs
*
(J)
Proportions of bioreactor-related categories in habitat preference scores (%)
0 20 40 60 80 100
SAGs from bioreactor
Other SAGs
***

Figure 2-5. Habitat preference scores of isolates/SAGs derived from a specific type of environment. Each of the darker orange plot indicates the scores of isolates/SAGs from the relevant environment, while the lighter orange one indicates those of the other isolates/SAGs. **(A)** Prokaryotic isolates' habitat preference scores of soil-related environments (*soil*, *rhizosphere*, *rice_paddy*, and *wetland*). **(B)** Prokaryotic isolates' habitat preference scores of brine-related environments (*marine*, *salt_marsh*, and *seawater*). **(C)** Prokaryotic isolates' habitat preference scores of plant-associated environments (*rhizosphere*, *phyllosphere*, *root*, *soil*, and *plant*). **(D)** Prokaryotic isolates' habitat preference scores of feces-associated environments (*feces*, *gut*, and *human_gut*). **(E)** SAGs' habitat preference scores of brine-related environments (*marine*, *salt_marsh*, *salt_lake*, and *seawater*). **(F)** SAGs' habitat preference scores of freshwater-related environments (*freshwater*, *lake_water*, and *groundwater*). **(G)** SAGs' habitat preference scores of brackish lake-related environments (*marine*, *salt_marsh*, *salt_lake*, and *seawater*). **(H)** SAGs' habitat preference scores of hydrothermus-related environments (*hydrothermal_vent* and *hot_springs*). **(I)** SAGs' habitat preference scores of marine sediment-related environments (*freshwater_sediment*, *sediment*, *marine_sediment*, *marine*, *salt_marsh*, and *salt_lake*). **(J)** SAGs' habitat preference scores of bioreactor-related environments (*bioreactor* and *activated_sludge*). Asterisks denote the results of the Mann-Whitney U-tests (*P<0.05, ***P<0.001) between each pair of violin plots.

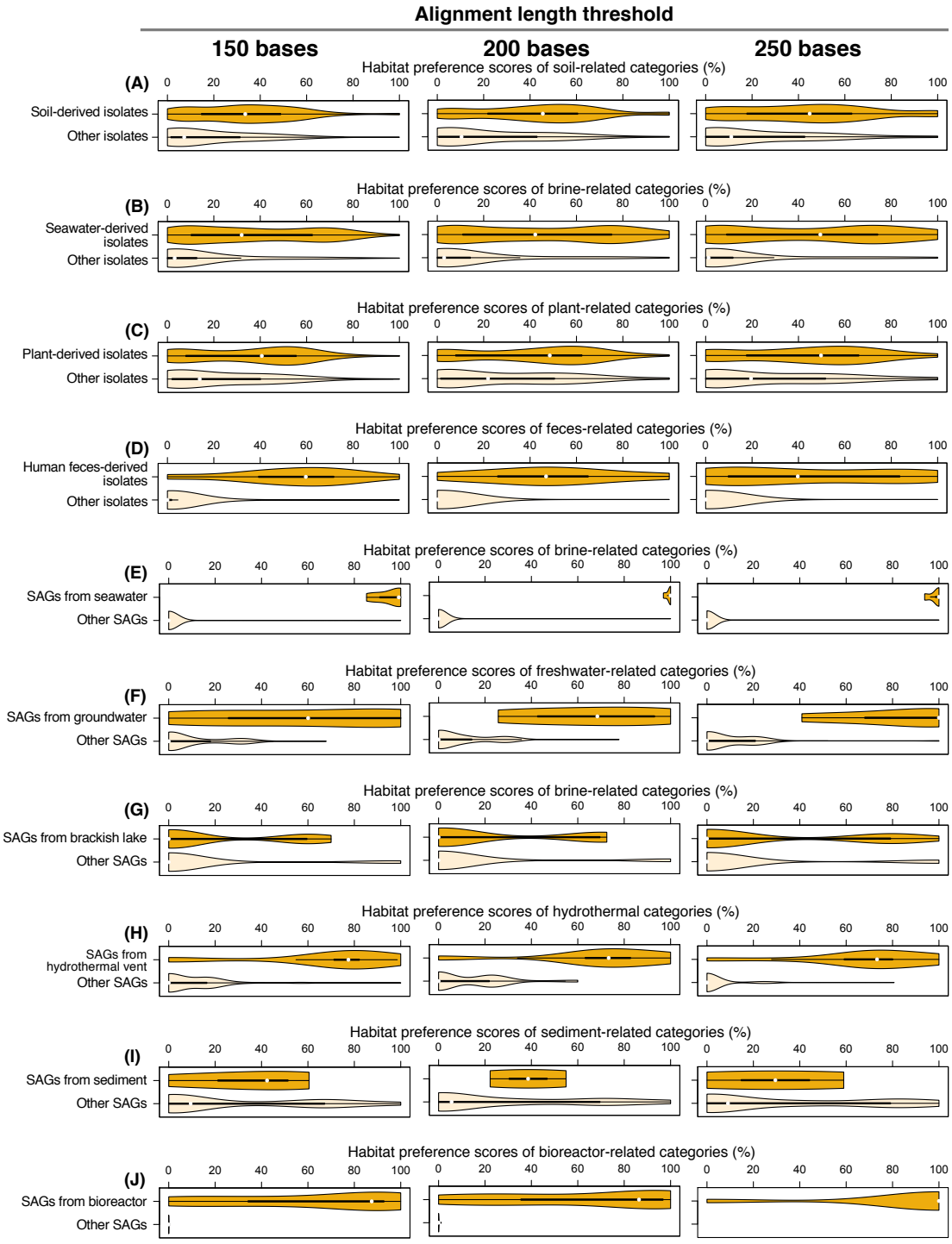
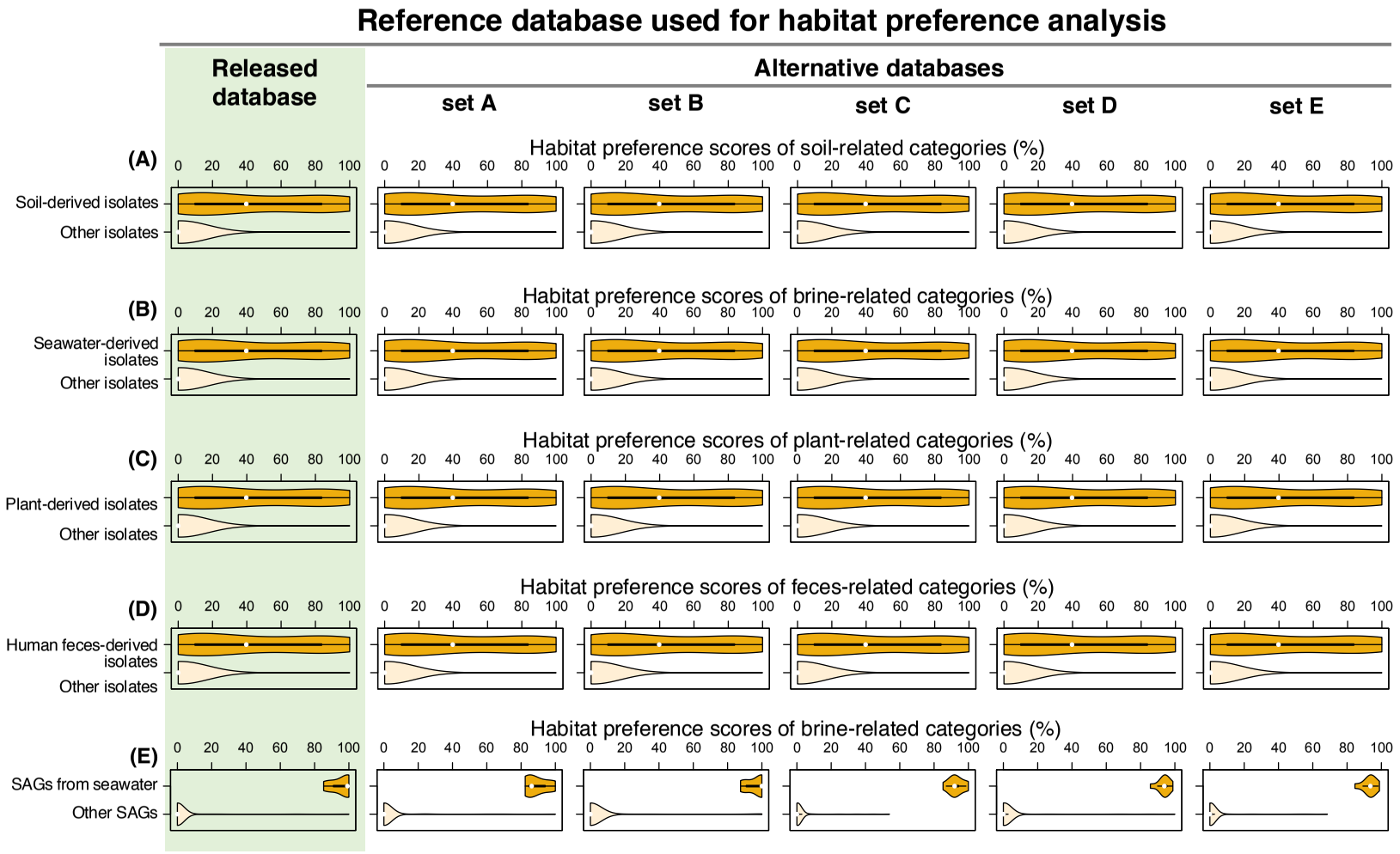


Figure 2-6. Violin plots indicating the effect of alignment length threshold on ProkAtlas-estimated prokaryotic habitat preference scores. Panels in left, middle, and right columns show the results calculated with an alignment length threshold of 150 bases, 200 bases, and 250 bases, respectively. Details on each panel are explained in Figure 2-5.


Reference database used for habitat preference analysis
Released database
Alternative databases
set A
set B
set C
set D
set E
(A)
Habitat preference scores of soil-related categories (%)
Soil-derived isolates
Other isolates
(B)
Habitat preference scores of brine-related categories (%)
Seawater-derived isolates
Other isolates
(C)
Habitat preference scores of plant-related categories (%)
Plant-derived isolates
Other isolates
(D)
Habitat preference scores of feces-related categories (%)
Human feces-derived isolates
Other isolates
(E)
Habitat preference scores of brine-related categories (%)
SAGs from seawater
Other SAGs
0 20 40 60 80 100

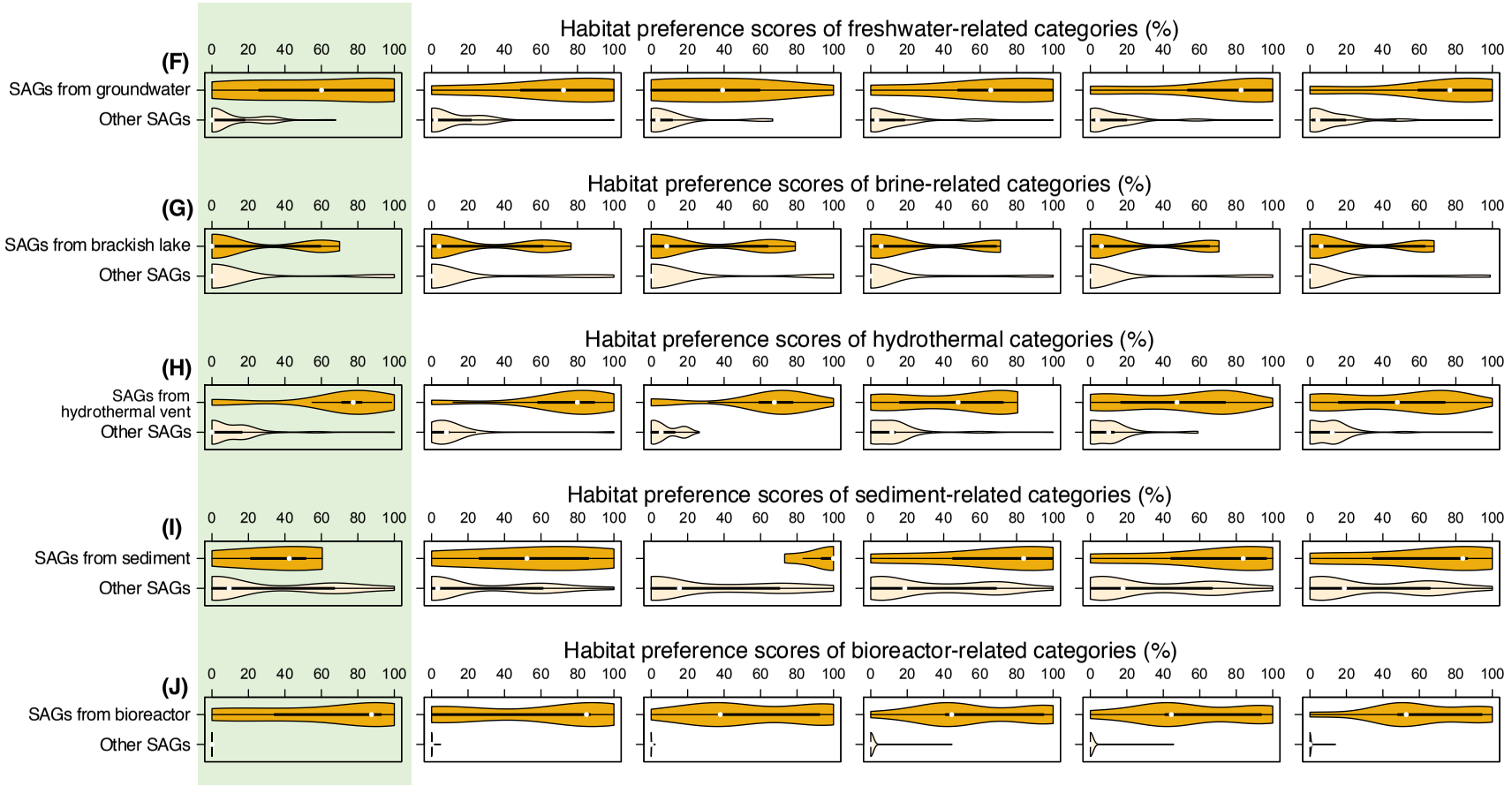

Habitat preference scores of freshwater-related categories (%)
(F)
SAGs from groundwater
Other SAGs
Habitat preference scores of brine-related categories (%)
(G)
SAGs from brackish lake
Other SAGs
Habitat preference scores of hydrothermal categories (%)
(H)
SAGs from hydrothermal vent
Other SAGs
Habitat preference scores of sediment-related categories (%)
(I)
SAGs from sediment
Other SAGs
Habitat preference scores of bioreactor-related categories (%)
(J)
SAGs from bioreactor
Other SAGs
0 20 40 60 80 100

Figure 2-7. Violin plots indicating the effect of random sampling in constructing ProkAtlas on ProkAtlas-estimated prokaryotic habitat preference scores. Panels in the leftmost column indicate the results obtained using the released version of ProkAtlas. Those in the next two columns indicate the results from alternative databases, each constructed by independent random sampling with at the same depth as ProkAtlas (up to 100 sequences per project). Panels in the last three columns indicate the results from yet other databases, each constructed by independent random sampling at deeper depth (up to 500 sequences per project). Details on each panel are explained in Figure 2-5.

Chapter 3 | Habitat-based analyses of prokaryotic communities

3-1 Introduction

As described in Chapter 2, I developed prokaryotic habitat database named ProkAtlas and showed that it provides reliable information on prokaryotic habitats. This means that ProkAtlas can work as a reference database that facilitates habitat-based prokaryotic community analysis. In this chapter, I present proofs of concept that the habitat-based analysis provides ecological insights into community assembly, enabling intuitive interpretation of community structure datasets.

To confirm the soundness of habit-based analysis at community scale, rather than for individual microbes, I first analyzed dataset from the Earth Microbiome Project (EMP) (Thompson et al., 2017). EMP is a global investigation of microbial communities in wide range of environments, providing abundant microbial community data with well-organized annotations on samples. EMP is characterized by the rigorous standardization of sampling and sequencing procedure (Gilbert et al., 2018); hence why EMP dataset is suitable for evaluating performance of ProkAtlas.

Considering the positive result of this benchmarking with EMP dataset, habit-based community analysis was regarded reliable. Thus, it was further applied to other datasets presenting specific spatial or temporal variations, namely agricultural soil samples with different salinity, lake water samples with different salinity, human infant gut microbiome samples, chronosequences of developing soils that were recently exposed after glacier retreat, and potentially polluted river-water samples, which had been obtained with different research aims. Of note, these were all 16S rRNA gene amplicon sequencing data and not included in ProkAtlas, which is composed of shotgun metagenomic sequences.

3-2 Materials and methods

EMP data were downloaded from the EMP ftp server in February 2019 (Thompson et al., 2017). The data were based on the random picking of 2,000 samples and that of 5,000 sequences per sample[3]. Sub-operational taxonomic units (sOTUs) clustered by Deblur (Amir et al., 2017) were used. Regarding the agricultural soil samples (Zhao et al., 2020),

[3] ftp://ftp.microbio.me/emp/release1/otu_tables/deblur/emp_deblur_150bp.subset_2k.rare_5000.biom

the lake water samples (Ji et al., 2019), the human infant gut microbiome (Yassour et al., 2016), chronosequences of developing soils on glacier forelands (Jiang et al., 2018; Mapelli et al., 2018), and potentially polluted river-water samples (Kirs et al., 2017), raw fastq sequence data were downloaded from public datasets (Table 3-1). All the sequences were generated by amplicon sequencing of 16S rRNA gene on Illumina MiSeq or HiSeq. Paired-end sequences with overlapping regions of 20 bp or longer were merged and quality-filtered using USEARCH v8.0.1623 (Edgar, 2010) (sequences with expected errors of 0.5 bp or less were kept), followed by removal of primer regions. sOTUs clustered by Deblur (Amir et al., 2017) with the default parameter settings were used.

The sOTUs were taxonomically annotated using RDP classifier (Wang et al., 2007) trained with SILVA with a confidence value threshold of 0.5. For one of the glacier chronosequence soil datasets (Mapelli et al., 2018), sOTUs annotated as members of phylum *Cyanobacteria* were eliminated because some samples were covered by cyanobacterial mat (Mapelli et al., 2018). Then, the sOTUs in each dataset were subjected to ProkAtlas pipeline, attributing each prokaryotic community to its estimated habitat composition. Regarding EMP dataset, which consists of short sequences (150 bases, the same as the default alignment length threshold), alignment length thresholds were set to 140 bases. In addition, habitat preference scores of EMP samples were calculated using five alternative datasets, as explained in 2-2-2.

3-3 Results and discussion

3-3-1 Habitat-based analysis of the EMP dataset

To test the versatility of habitat-based prokaryotic community analysis using ProkAtlas, I reanalyzed the EMP dataset, a large community-based project that collects and analyzes prokaryotic community samples from various natural environments (Thompson et al., 2017). Because each of the 16S rRNA gene amplicon-sequencing datasets in the EMP dataset is tagged with sampling-site metadata described by EMP Ontology, this dataset was used for assessing the validity of ProkAtlas-based analysis.
Among the 91,364 sub-operational taxonomic units (sOTUs) in the EMP dataset, 32,117 (35.2%), accounting for 65.3% of the total reads, were successfully mapped to ProkAtlas. The habitat preference scores of prokaryotic communities estimated by ProkAtlas were generally consistent with the sampling-site metadata in the EMP dataset (Figure 3-1). For example, prokaryotic communities annotated by *soil (non-saline)* showed higher

habitat preference scores of *soil* (31.7±12.9%, mean±sd) and *rhizosphere* (18.7±7.82%), compared with the other communities. Similarly, communities annotated as *water (saline)* showed higher habitat preference scores of *marine* (51.9±23.3%). This result, in line with the habitat preference scores of sequences from isolates and SAGs, highlights the reliability of the habitat preference scores. This in turn means that environmental source of a given sample may be estimated by ProkAtlas, for example to address forensic concerns (Carter et al., 2020).

When one of the five alternative datasets (sets A–E) was employed instead of ProkAtlas, the consequent habitat preference scores were only marginally affected (Figure 3-2), where larger-size databases (sets C–E) slightly improved the sequence coverages (51.5–51.7% of sOTUs, accounting for 75.4–76.8% of total reads, were mapped). This is in line with the results of habitat preference analysis of isolates and SAGs (Figure 2-7), where the alternative datasets provided habitat preference scores consistent with ones from ProkAtlas. Together, I argue that the size of ProkAtlas achieves a good balance between the information content and computational usability.

3-3-2 Habitat-based analysis of agricultural soil samples with salinity gradients

The first dataset contained 124 agricultural soil samples with different salinities sampled at 31 sites in northwest China (Zhao et al., 2020). These sampling sites spans more than 400 km in longitude, and four samples were obtained from each site. The dataset contained 12,094 sOTUs, among which 12,052 (99.6%) and 7,631 (63.1%), accounting for 99.8% and 67.9% of the total reads per sample on average, were taxonomically assigned at the phylum level and successfully mapped to ProkAtlas, respectively.

When phylum-level taxonomic structures were investigated as many amplicon-sequencing studies do, the estimated compositions were dominated by the phyla *Proteobacteria* (34.3±8.99%, mean±sd), *Bacteroidetes* (20.8±9.47%), and *Gemmantimonadetes* (11.7±4.63%) (Figure 3-3 (A)). On the other hand, when the estimated habitats were investigated, I observed a clear trend that the habitat preference compositions were affected by soil salinity concentration (Figure 3-3 (B)). More specifically, saline environments such as *marine, seawater*, *esturay*, and *salt_lake* showed substantial variation among the samples (2.44–26.4%) and a significant positive correlation with the soil salinity concentration (Spearman's correlation test, ρ=0.60, P<0.001) as expected. Thus, ProkAtlas clearly and intuitively highlights the microbial

community characteristics of high-saline soils, which is consistent with previous knowledge on the relationship between salinity and prokaryotic community structures (Lozupone and Knight, 2007; Rath et al., 2019).

3-3-3 Habitat-based analysis of non-saline, brackish, and saline lake water samples

To further test the relationship between environmental salinity and habitat-based prokaryotic community structures, a dataset comprising prokaryotic community structures in saline and non-saline lake water was re-analyzed (Ji et al., 2019). To test the versatility of habitat-based analysis, I targeted lake water microbiomes as an example of semi-closed ecosystem, which contrasts with open field soil. This dataset contains 78 prokaryotic community structure data from 25 lakes with diverse salinity. All these lakes are in Tubetan Plateau, where large number of lakes with different salinities are distributed. Two samples lacking sampling site information were excluded from the analysis. The dataset contained 6,054 sOTUs, among which 5,241 (86.6%), accounting 91.2% of the total reads per sample on average were successfully mapped to ProkAtlas. All the sOTUs were taxonomically assigned at the phylum level. All samples were dominated by phyla *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, and/or *Proteobacteria* (Figure 3-4 (A)). Phylum-level taxonomic compositions diverged highly even between samples with similar salinity concentrations and gave few ecological insights. On the other hand, the habitat-based analysis was able to differentiate the prokaryotic communities between the saline and non-saline lakes (Figure 3-4 (B)). The proportions of saline-water-related categories were significantly and strongly correlated with salinity (Spearman's correlation test, ρ=0.88, P<0.001). Here, ProkAtlas reproduces the effect of salinity on prokaryotic community structures, which was previously elucidated at global and local scales (Lozupone and Knight, 2007; Rath et al., 2019; Thompson et al., 2017), highlighting the robustness and validity of the database and pipeline. In addition, ProkAtlas relabels prokaryotic community members as salinity-tolerant/sensitive and thereby facilitates intuitive interpretation of prokaryotic community structures without cumbersome calibration or modeling.

In summary, the habitat-based analysis here gave more direct and clearer interpretations of the prokaryotic community structure datasets from an ecological perspective than typical high-rank taxonomic analyses, in answering questions such as "Do environmental factors substantially and directionally affect prokaryotic community

structures?"

3-3-4 Habitat-based analysis of human infant gut microbiome samples

The third dataset consisted of human infant gut microbiome samples (Yassour et al., 2016). In this study, 654 time-series infant feces samples were collected from Finnish infants aged 2 to 36 months. The dataset contained 2,113 sOTUs, among which 2,113 (100%) and 1,374 (65.0%), accounting for 100% and 96.8% of total reads per sample on average, were taxonomically assigned at the phylum level and successfully mapped to ProkAtlas, respectively.

When phylum-level taxonomic compositions were investigated, the compositions were highly diverse until approximately 400 days after birth, after which the compositions stabilized and were dominated by *Firmicutes* and *Bacteroidetes* (Figure 3-5 (A)). While this process was already well known (Bäckhed et al., 2015), the habitat-based analysis gave another view on the process as the convergence to *human gut-related* environmental categories (Figure 3-5 (B)). This example shows that ProkAtlas can be used to evaluate the "maturity" of prokaryotic ecosystems undergoing temporal successions toward a stable state. It may be notable that recent trait-based studies in microbial ecology have suggested that, while taxonomic compositions during primary and secondary successions are often stochastic and uninterpretable (Ferrenberg et al., 2013), traits of prokaryotic communities follow a path that is more predictable and interpretable (Kearns and Shade, 2017; Nemergut et al., 2016).

3-3-5 Habitat-based analysis of developing soils

Here I show another example of primary succession of developing microbial ecosystems: chronosequences of developing soil (Jiang et al., 2018; Mapelli et al., 2018). Retreating glacial chronosequence provides a transect of soil samples at different developmental stages, from unweathered bedrocks to matured (i.e. extensively weathered) soils (Castle et al., 2017; Delgado-Baquerizo et al., 2019). It has been acknowledged that weathering of soils affects soil microbial community structures, presumably mediated by soil chemical conditions (Castle et al., 2017; Delgado-Baquerizo et al., 2017); on the other hand, such transect is still valuable as a showcase of primary succession of an ecosystem, where temporal filtering such as priority effects or dispersal limitations strongly affects (Ferrenberg et al., 2013; Freedman and Zak, 2015).

Each of these datasets contained 21 bulk soil samples collected at seven sampling sites along a retreating glacial chronosequence. One was from Midtre Lovénbreen glacier moraine (Norway) (Mapelli et al., 2018), and the other was from Hailuogou Glacier Chronosequence (China) (Jiang et al., 2018). After the glacial retreat, those sites have been exposed to soil weathering for different time lengths. In both datasets, the phylum-level taxonomic compositions along the chronosequences presented clear gradients; however, the taxonomic clades constituting the gradients were quite different between the two — they were phyla *Bacteroidetes* and *Chroloflexi* in the Norwegian dataset (Figure 3-6 (A)) but phyla *Acidobacteria*, *Bacteroidetes*, and *Proteobacteria* in the Chinese dataset (Figure 3-6 (C)). Both of the gradients could be the outcomes of chemical condition changes (e.g., phosphorus depletion mitigation as a result of weathering) (Castle et al., 2017; Delgado-Baquerizo et al., 2017); however, their apparently different patterns hamper unified understanding of the prokaryotic community successions. On the other hand, the habitat-based analysis of the prokaryotic communities clearly illustrated similar convergence to soil-related environments during the courses of pedogenesis in both sites (Figure 3-6 (B)(D)). This suggests that prokaryotic habitat preference can be a useful trait for analyzing community successions. In addition, a notable difference was seen between the results of the bulk soil and infant gut datasets. Many of the infant gut prokaryotic communities were "matured" from the beginning possibly due to the priority effect (Figure 3-5 (B)) in contrast to the soil prokaryotic communities.

The factor driving the primary succession of microbial community is enigmatic, and habitat-based analysis may give a hint to this question. One long-discussed mechanism of primary succession is priority effects (Werner and Kiers, 2015). If this effect is prominent, young soils are be colonized by any microbe that arrived the soil earlier than other microbes. Soils under primary succession may be stochastically dominated by microbes that are not adaptive to soil environment, which are later competitively replaced with more adaptive microbes (Evans et al., 2017; Ferrenberg et al., 2013). Thus, under this effect, the overall adaptiveness of microbial community should be site-specific: some sites are dominated by adaptive microbes, while others are dominated by less adaptive ones. The results of habitat-based analysis do not support this effect in this specific case, where the proportion of soil-related (i.e. adaptive) members linearly increased along the transect.

A more plausible explanation may be that prokaryotic community structure had changed in accordance with transition in physicochemical factors. For example, soil weathering induces phosphorus depletion in the microbial community while mitigating nitrogen depletion (Bergkemper et al., 2016; Castle et al., 2017), affecting microbial diversity and functionality (Delgado-Baquerizo et al., 2017; Yao et al., 2018).

In summary, the infant gut and developing soil datasets indicate that ProkAtlas can be used to evaluate the maturity of prokaryotic ecosystems undergoing temporal successions, without prior investigation on what the "matured" state is like. Notably, such effectiveness of habitat-based analysis can be placed into the context of recent discussions on trait-based microbial community ecology: although the primary or secondary successions of microbial communities are often stochastic and unpredictable (Ferrenberg et al., 2013), trait-based patterns tend to be more conserved, predictable, and easier to interpret (Kearns and Shade, 2017; Nemergut et al., 2016).

3-3-6 Habitat-based analysis of potentially polluted river-water samples

Habit-based analysis was further applied to potentially polluted river-water samples (Kirs et al., 2017). In this study, 25 water samples were collected at nine sampling sites in the Manoa stream, which flows through urbanized areas on Oahu Island, Hawaii, USA. High levels of fecal indicator bacteria (FIB) were reported in the estuary of Manoa stream neighboring popular bathing beaches (Goto and Yan, 2011), and sources of FIB were of interest in the contexts of both environmental and health sciences. The dataset contained 4,061 sOTUs, among which 4,000 (98.5%) and 2,389 (58.8%), accounting for 99.7% and 75.1% of total reads, were taxonomically assigned at the phylum and proteobacterial class levels and successfully mapped to ProkAtlas, respectively.

The taxonomic structures showed a clear gradient from upstream (MS1–5) to downstream samples (MS7–9) (Figure 3-7 (A)). On the other hand, the investigation of the estimated habitats visualized two important ecological features. First, the transition from soil- and freshwater-related environments to seawater-related environmental categories was clearly observed from the upstream (MS1–5) to midstream (MS6) and downstream sites (MS7–9) (Figure 3-7 (B)). MS7–9 are located in a canal that is connected to the sea and directly influenced by tides, while MS6 is located approximately 500m upstream to the confluence with the canal (Kirs et al., 2017). Second, environmental categories related to anthropogenic water contamination (e.g., *human_gut* and

wastewater) showed a decrease from the upstream to midstream sites. This result was rather unexpected because potential pollution was expected to be introduced in urbanized areas (MS2–5) and increase their compositions along the river flow. Instead, the habitat-based analysis suggests that river water in the upstream, conserved forest area (MS1) already contains FIB. Notably, in line with this interpretation, some studies have claimed that riverine FIB largely come from soil instead of human pollution (Goto and Yan, 2011; Kirs et al., 2017).

The present example showcases that microbial community data may be useful for tracing pollutant source in the environment. This concept is similar to that of microbial source tracking technology for identification of pollution sources or for forensic purposes (Knights et al., 2011; Unno et al., 2018). The basic idea of microbial source tracking technology is to investigate microbial community structures in polluted sites and potential pollution sites and compare these community structures. In this context, ProkAtlas would serve as a ready-made reference database, which does not require *ad hoc* sampling and metagenomic sequencing of potential sources.

In summary, the habitat-based analysis here answered questions like "Where are prokaryotic communities from distinct environments mixed?" and "How do mixed communities develop in different environments?" without cumbersome preparation of reference datasets.

Table 3-1 Six 16S rRNA gene amplicon-sequencing datasets that underwent habitat-based analysis.

Sample description	Number of samples	Data availability	Reference
Saline agricultural soils sampled at 31 points scattered over 400 km (north-west China)	124	INSD SRP136143	Zhao et al., 2020
Saline and non-saline water samples sampled at 25 lakes (Tibet Plateau, China)	78	INSD PRJNA503775	Ji et al., 2019
Stool of newborn Finnish infants (0–36 months old)	776	DIABIMMUNE project website	Yassour et al., 2016
Bulk soil samples at different developmental stages, obtained along retreating glacier (Midtre Lovénbreen glacier, Norway)	21	INSD PRJEB12640	Mapelli et al., 2018
Bulk soil samples at different developmental stages, obtained along retreating glacier (Hailugou Glacier Chronosequences, China)	21	INSD PRJNA354498	Jiang et al., 2018
Water sampled along Manoa Stream (Hawaii, USA)	25	INSD PRJNA376213	Kirs et al., 2017

(A) Animal distal gut
(B) Plant corpus
(C) Sediment (saline)
(D) Plant rhizosphere
(E) Soil (non-saline)
(F) Water (non-saline)
(G) Water (saline)
Proportions of environmental categories in habitat preference scores (%)
human_gut
gut
human
feces
biogas_fermenter
bovine_gut
human_reproductive_system
rhizosphere
soil
food
phyllosphere
moss
rhizosphere
soil
plant
leaf
human_gut
fungus
bioreactor
groundwater
marine
freshwater_sediment
marine_sediment
sediment
seawater
soil
phyllosphere
microbial_mat
salt_marsh
hydrothermal_vent
rhizosphere
soil
freshwater_sediment
phyllosphere
groundwater
freshwater
rice_paddy
plant
wetland
peat
soil
rhizosphere
freshwater_sediment
groundwater
fungus
peat
freshwater
wetland
rice_paddy
hot_springs
freshwater
marine
lake_water
soil
rhizosphere
groundwater
freshwater_sediment
aquatic
phyllosphere
subsurface
marine
salt_marsh
freshwater
seawater
lake_water
aquatic
freshwater_sediment
subsurface
wastewater
soil

Figure 3-1 Habitat preference scores of EMP prokaryotic communities. Habitat preference scores in each sampling site represented by EMP Ontology level 3 terms are shown. **(A)** Animal distal gut, **(B)** plant corpus, **(C)** sediment (saline), **(D)** plant rhizosphere, **(E)** soil (non-saline), **(F)** water (non-saline), and **(G)** water (saline). The y-axes show the proportions of environmental categories within individual estimated habitat compositions. Means and standard deviations (by error bars) among EMP samples are shown.

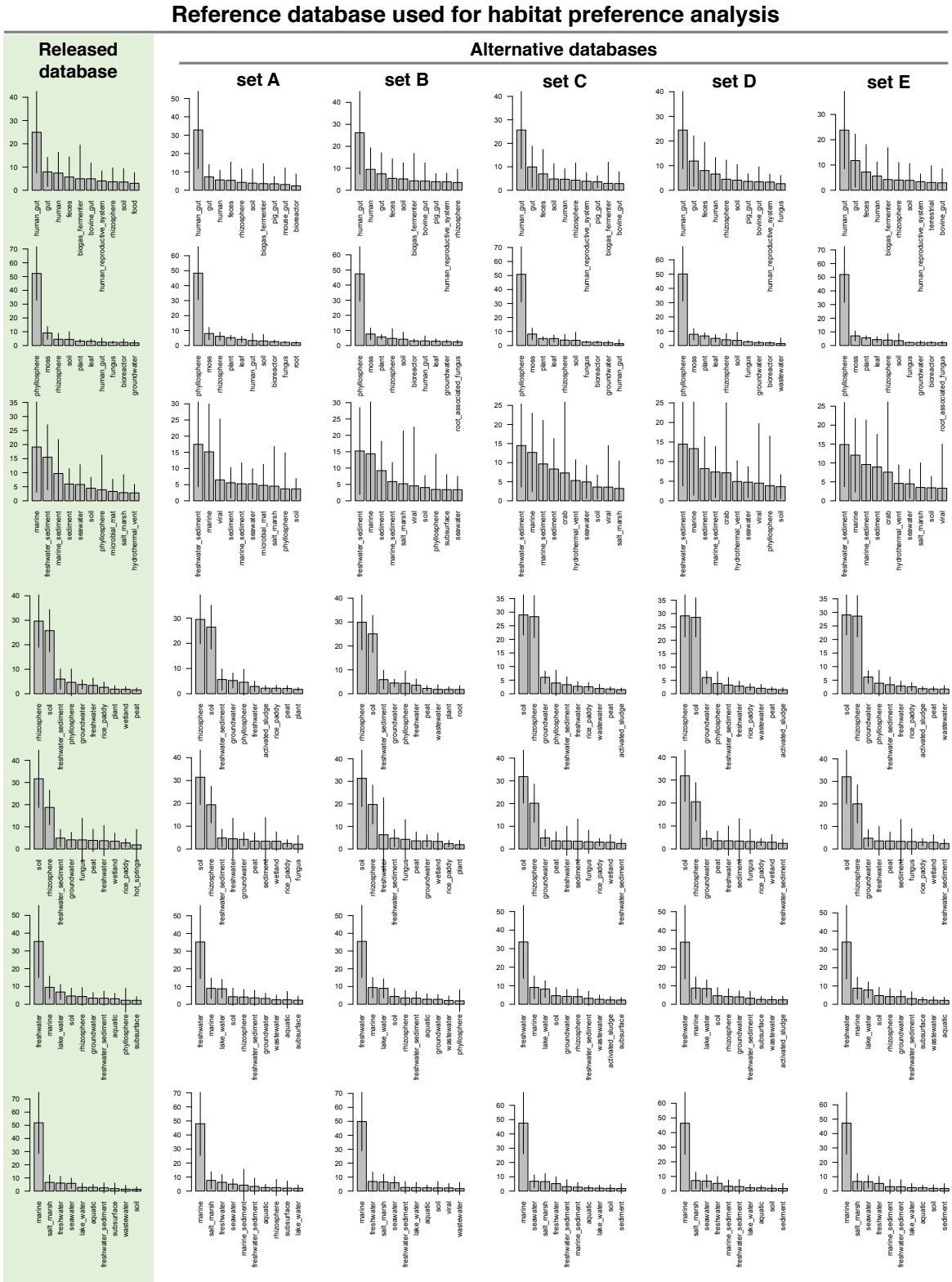

Reference database used for habitat preference analysis
Released database
Alternative databases
set A
set B
set C
set D
set E
Saline water communities
Non-saline water communities
Non-saline soil communities
Plant rhizosphere communities
Saline sediment communities
Plant corpus communities
Animal distal gut communities
Habitat preference scores of each environmental category (%)

Figure 3-2 Habitat preference scores of EMP prokaryotic communities in each sampling site represented by EMP Ontology level 3 terms, using the released version of ProkAtlas and five alternative databases obtained by repeating the random sampling of sequences. Details on each column and each panel are as explained in Figures 2-7 and 3-1, respectively.

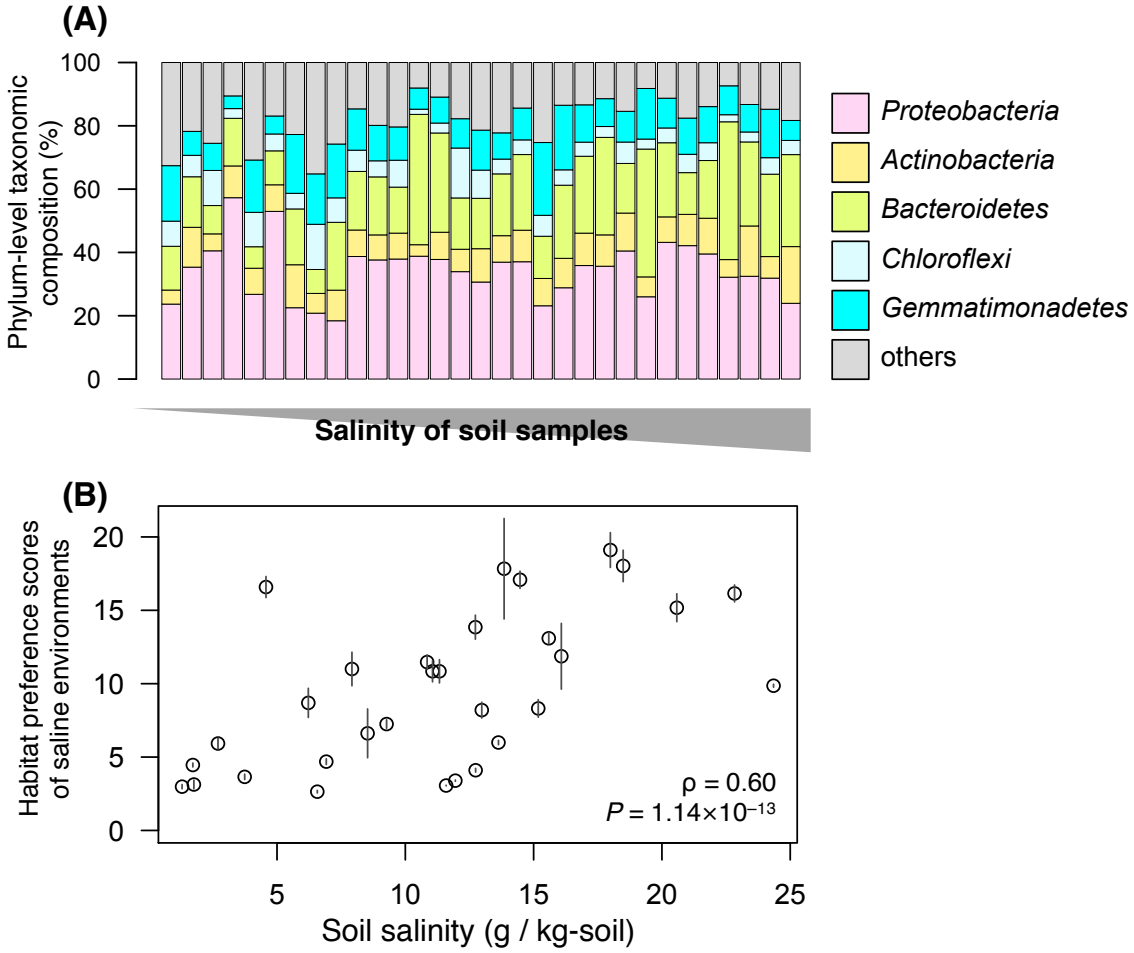


Figure 3-3 Habitat-based analysis of soil samples with salinity gradients. **(A)** Phylum-level taxonomic structures of the 31 plots ordered by soil salinity concentration (higher on the right than on the left). Each bar denotes the average of four replicates within one plot. Means among four replicates for each plot are indicated. **(B)** A scattergram of soil salinity concentrations and sum of habitat preference scores of brine-related environmental categories (*estuary*, *hypersaline_lake*, *marine*, *salt_lake*, *salt_marsh*, *seawater*, and *marine_sediment*). Means and standard deviations (by error bars) among four replicates for each plot are indicated. Result of the Spearman's correlation test is also shown.

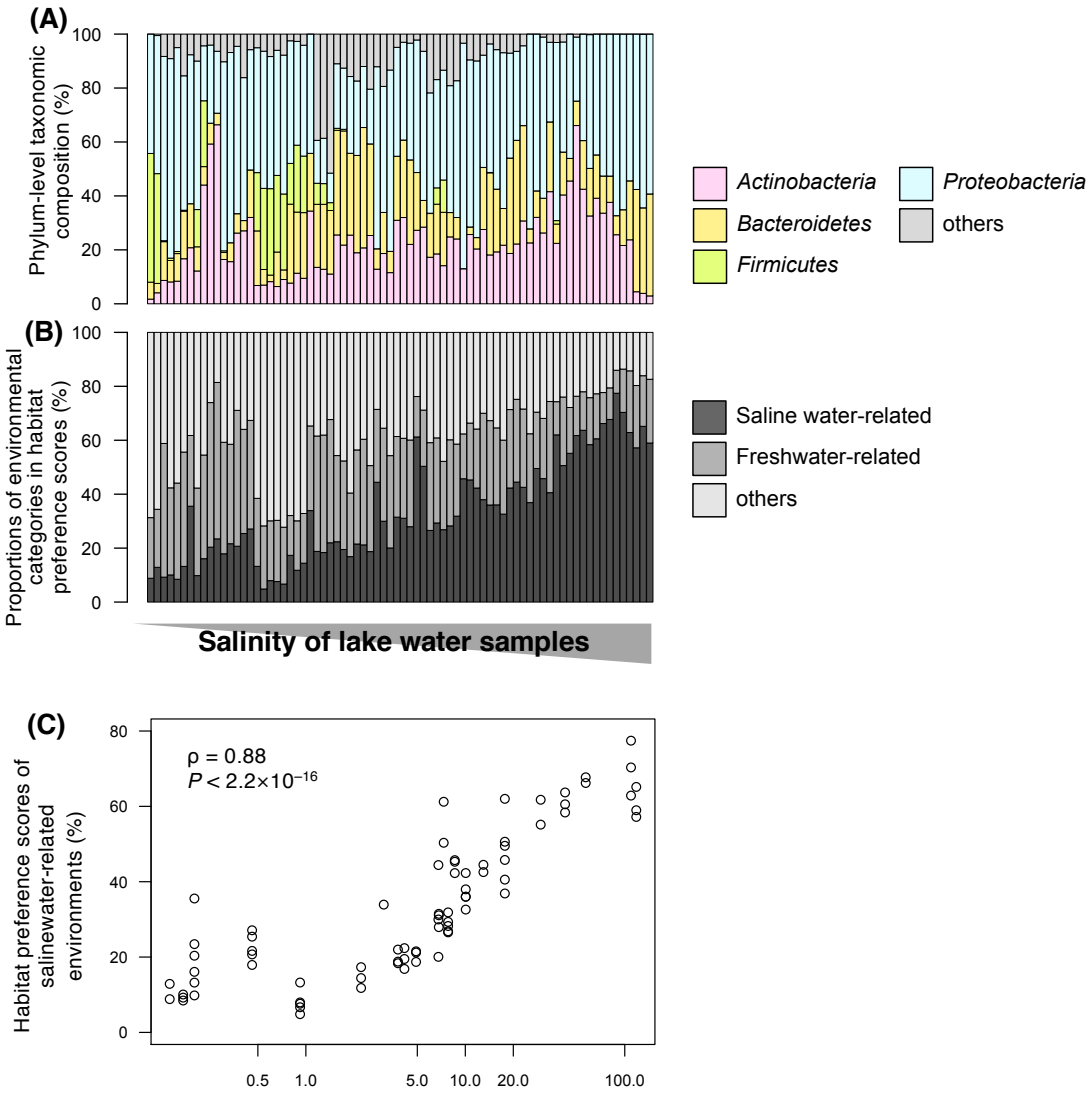


Figure 3-4 Habitat-based analysis of saline and non-saline water samples from 25 lakes. **(A)** Phylum-level taxonomic structures of the 76 samples ordered by salinity (higher on the right than on the left). **(B)** Habitat preference scores. Saline water-related: *hypersaline_lake*, *marine*, *marine_sediment*, *salt_lake*, *salt_marsh*, and *seawater*. Freshwater-related: *aquifer*, *freshwater*, *freshwater_sediment*, *groundwater*, and *lake_water*. **(C)** A scattergram of water salinity concentrations and sum of habitat preference scores of saline water-related environmental categories in the estimated habitat compositions (*estuary*, *hypersaline_lake*, *marine*, *marine_sediment*, *salt_lake*, *salt_marsh*, and *seawater*). Result of the Spearman's correlation test is shown.

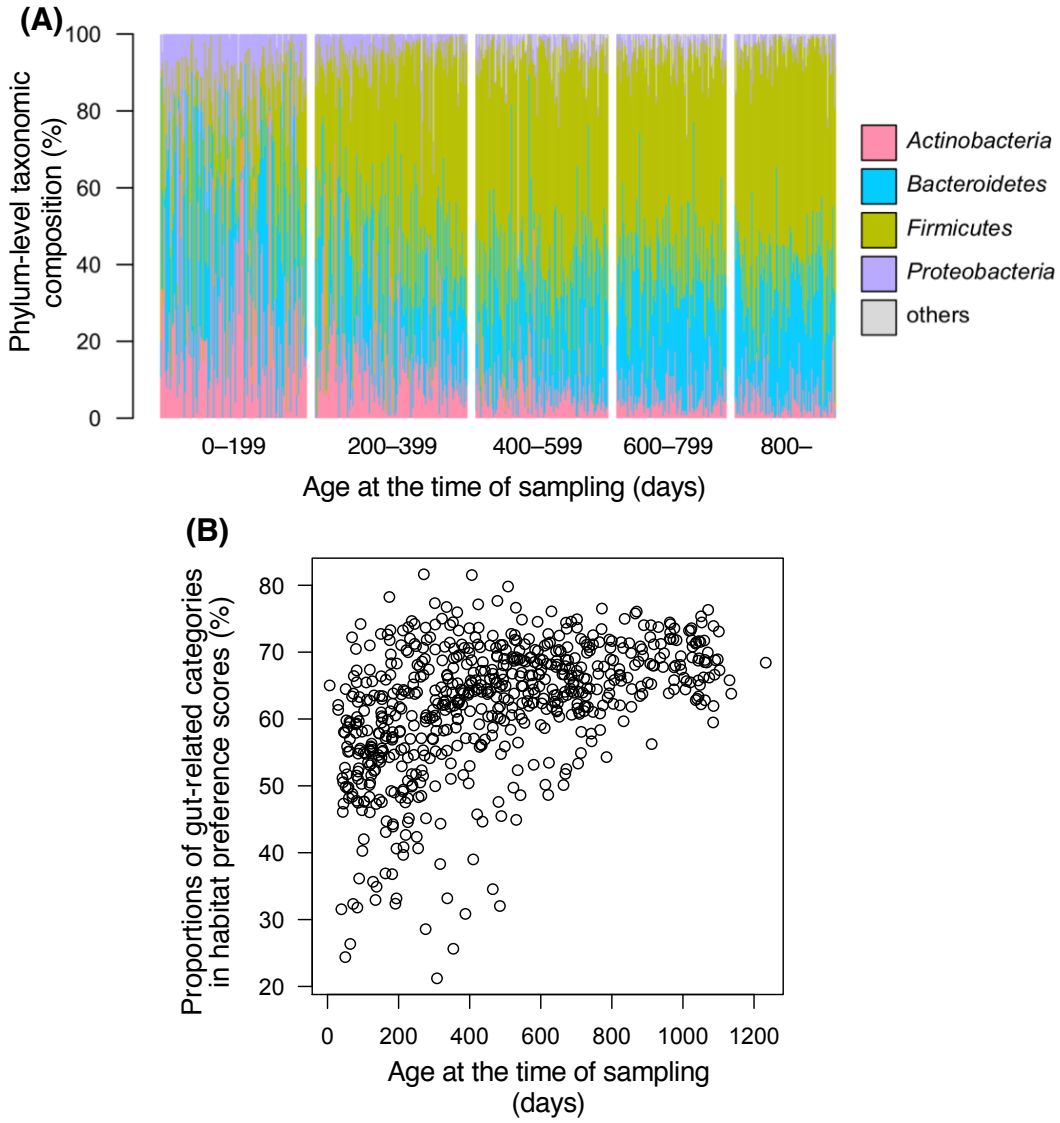


Figure 3-5 Habitat-based analysis of human infant gut microbiome samples. **(A)** Phylum-level taxonomic structures of the 654 samples ordered by sampling ages (older on the right). **(B)** A scattergram of infant ages and sum of habitabt preference scores of human gut-related environmental categories (*gut* and *human_gut*).

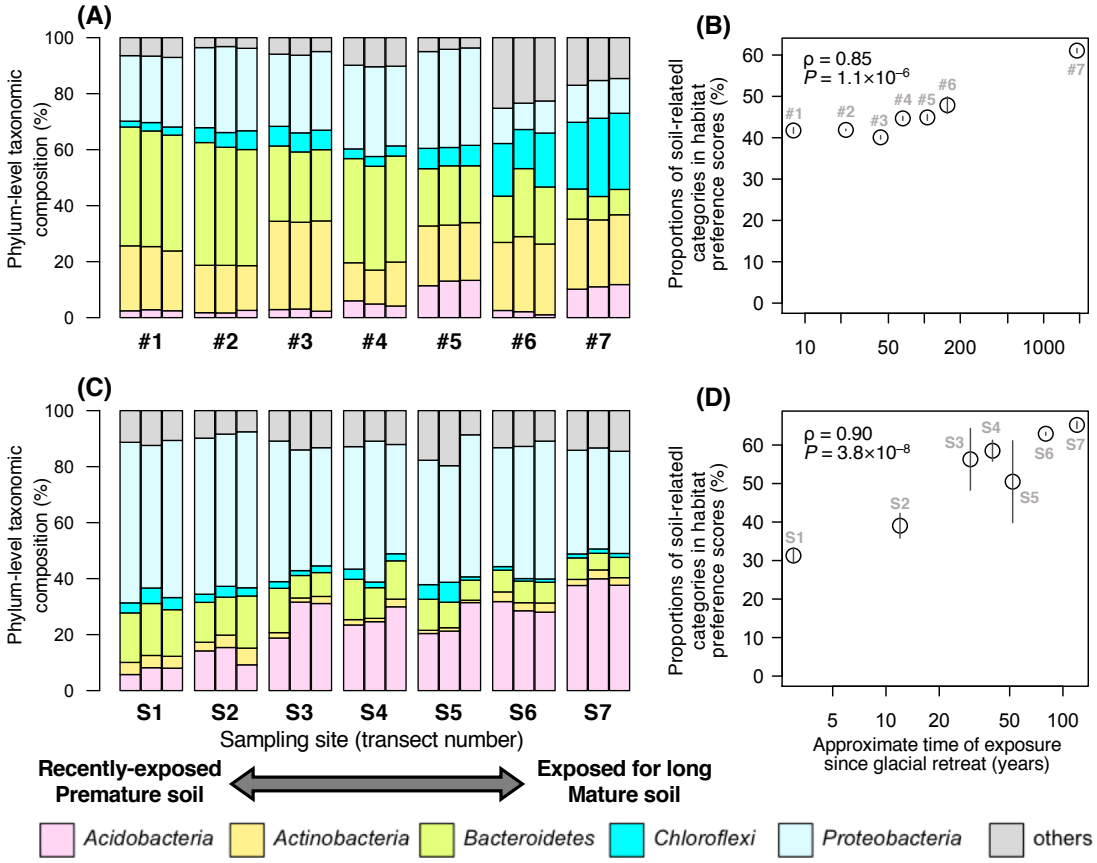


Figure 3-6 Habitat-based analysis of glacial chronosequence soil samples. **(A)** Phylum-level taxonomic structures and **(B)** sum of habitat preference scores of estimated soil-related environmental categories (*soil*, *rhizosphere*, *rice_paddy*, and *wetland*) in soil samples taken from Midtre Lovénbreen glacier moraine (Norway) (Mapelli et al., 2018). **(C)** Phylum-level taxonomic structures and **(D)** sum of habitat preference scores of estimated soil-related environmental categories (*soil*, *rhizosphere*, *rice_paddy*, and *wetland*) in soil samples taken from Hailuogou Glacier Chronosequence (China) (Jiang et al., 2018). The samples are ordered by the length of weathering time. In (B) and (D), means and standard deviations (by error bars) among three replicates for each plot are shown, along with results of the Spearman's correlation tests.

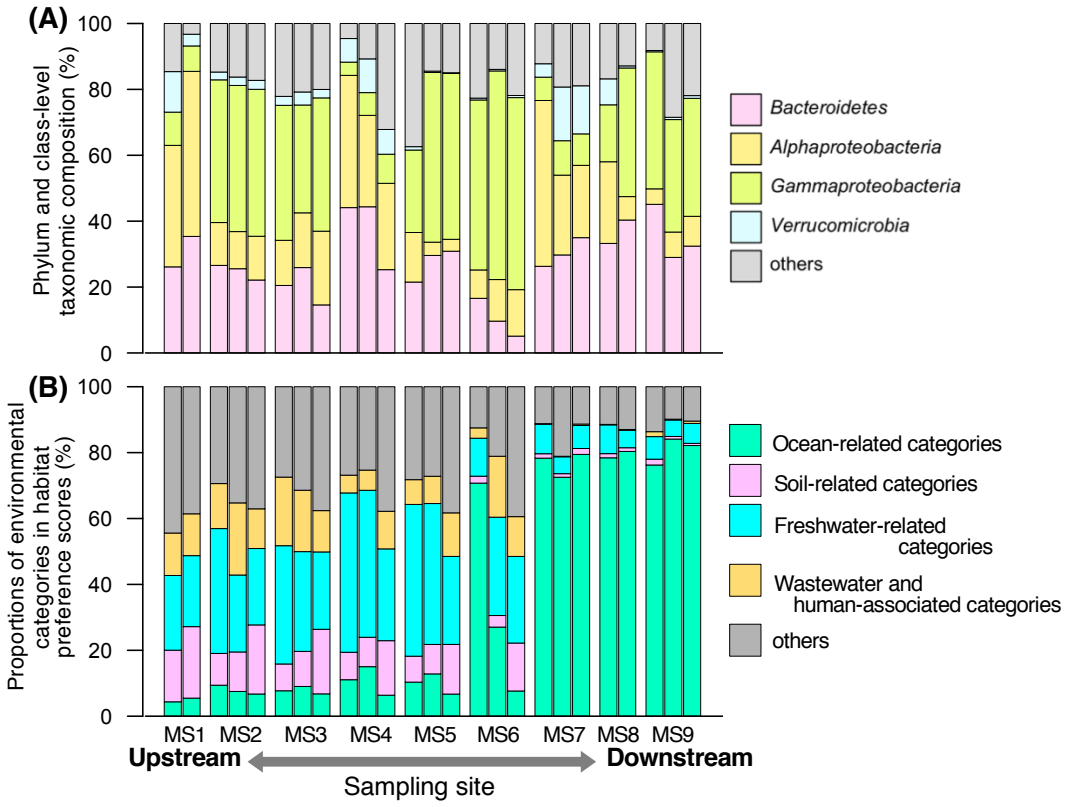


Figure 3-7 Habitat-based analysis of potentially polluted river-water samples. **(A)** Phylum- and proteobacterial-class level taxonomic structures at nine sampling points (more downstream on the right than on the left). **(B)** Habitat preference compositions. Ocean-related categories: *estuary*, *hypersaline_lake*, *marine*, *salt_lake*, *salt_marsh*, *seawater*, and *marine_sediment*. Soil-related categories: *rhizosphere* and *soil*. Freshwater-related categories: *freshwater*, *aquifer*, *groundwater*, and *lake_water*. Wastewater and human-associated categories: *gut*, *human*, *human_gut*, and *wastewater*. The community structures are ordered by the geographical locations of sampling sites.

Chapter 4 | Concluding Remarks

4-1 Overview of this dissertation

In this dissertation, I propose a habitat-based analysis of prokaryotic community structure. To facilitate this, I developed ProkAtlas, a novel database comprehensively linking 16S rRNA gene sequences to prokaryotic habitats (Chapter 2). I, then, testified the effectiveness of habitat-based analysis in explaining and predicting microbial community dynamics, using datasets of soil, lake water, human gut, and river water microbiomes (Chapter 3). Overall, habitat-based analysis clarified the mechanistic and predictable relationships between ecosystem (i.e. environment) and microbial communities. For example, the relationship between environmental salinity and microbial community has been enigmatic (Rath et al., 2019); although statistical correlations have been reported, no causal mechanisms underlying the correlations had been reported. Contrastingly, habitat-based analysis clearly indicated that high salinity favored specific clades of prokaryotes that are adapted to marine environments (Figures 3-3, 3-4). Habitat-based analysis also facilitated the evaluation of microbial community primary succession in barren soils and infant guts (Figures 3-5, 3-6). While the process of primary succession has been intensively investigated (Ortiz-Álvarez et al., 2018), the results of this study are novel in that successional processes were quantitively and intuitively evaluated. In short, habitat-based analysis well works as a novel trait-based approach to microbial community ecology, facilitating ecological interpretation of microbial community datasets.

4-2 Habitat-based analysis from the viewpoint of bioinformatics

Here, I applied a habitat-based analysis into the general context of biology, discussing the strengths and weaknesses of habitat-based analysis from a bioinformatic point of view. As discussed herein, the science of bioinformatics often provides unified solutions to problems raised in different fields of biology by taking advantage of common features of biological data. Due to the versatility of bioinformatics, I illustrated the contributions of this study to broader fields of biology beyond microbial ecology.

4-2-1 Strength of ProkAtlas as a database: sustainability

As discussed in Chapter 1, trait-based approach is fundamentally dependent on the trait database of the organisms in question. While genome-oriented traits (e.g., genome size,

rRNA gene copy number) have been available in well-maintained databases such as INSDC and JGI IMG/M (Markowitz et al., 2007), there has been no easily available database of microbial phenotypic/functional traits (Barberán et al., 2017), hindering the use of these traits in studies related to microbial community ecology. A critical problem is that these microbial phenotypic/functional traits are documented in journal articles or Bergey's Manual in the form of natural linguistic text.

In recent years, several databases of microbial phenotypic/functional traits have been developed (Barberán et al., 2017; Reimer et al., 2019). While these efforts are highly appreciated, their performances must be carefully assessed. Apart from cultivation biases (see Chapter 1), the tradeoff between the completeness and accuracy of the database is extremely severe. Automatic text parsing using natural linguistic processing technology is efficient but it is highly error-prone when applied to semantically opaque descriptions of microbial phenotypes (Barberán et al., 2017). To increase accuracy, manual compilation of data as described in journal articles or Bergey's Manual may be effective; however, this process is labor-intensive and practically incapable of covering a wide variety of microbes. Notably, the only currently available manually-curated phenotypic database, to my knowledge, seems to be largely dependent on the efforts of American undergraduate students (Barberán et al., 2017).[4] The construction and maintenance (i.e., updating) of the database is undoubtedly laborious. In comparison, ProkAtlas requires a lesser degree of human effort for its construction and updating, making it more likely to be sustainable.

Database search is an indispensable procedure in modern biology, and the quality and quantity (i.e. coverage or comprehensiveness) of databases inevitably affect every research. While popularly-used manually-curated databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and Cluster of Orthologous Groups (COG), are valuable, many researches in biology may be bound by these databases; the information not available in such mega-databases could be left unused. To foster the new budding researches, currently-unused information (for example, habitat information of microbes) would be useful. In this context, I argue that database construction without too much human effort is a valuable idea.

---

[4] The acknowledgement of Barberán et al. (2017) says: "We also thank [...] and those undergraduates at Johns Hopkins University, the University of Notre Dame, and the University of Colorado who helped compile the database."

4-2-2 Metadata in public databases

ProkAtlas consists of 16S rRNA gene sequences annotated with their source environment (i.e. habitat). The environmental information was retrieved from metadata registered in INSDC SRA/ERA/DRA: NCBI taxonomy ID (e.g., "soil_metagenome"). Thus, the core idea of habitat-based analysis is the use of metadata in INSDC as microbial traits. This suggests that other metadata may also be used for a trait-based approach: nucleotide sequences annotated with any specific metadata, instead of source environment, may constitute a microbial trait database.

In reality, however, the metadata in INSDC has a number of problems. For example, some of the soil metagenomic entries contain detailed information on sampling sites (e.g., land usage, climate, soil physicochemical properties); however, the vocabulary used for description is not standardized. Therefore, comparison between entries is difficult. More importantly, most entries lack such detailed information, which is not required to be entered by the users while registering their sequences. Although environmental ontologies such as EnvO (Thompson et al., 2017) and MIGS/MIMS (Field et al., 2008) can be registered in INSDC, few entries have one.

Apart from developing trait databases, metadata is also vital when reusing data collected by other research teams. Such reuse of data is becoming increasingly common due to recent data deluge (Hiraoka et al., 2016), and for this reason, I predict that improvement of the metadata availability will critically affect the biology in the future.

4-2-3 Short-read amplicon sequencing in the era of "fourth-generation" sequencing

ProkAtlas was designed primarily for analyzing the amplicon sequencing dataset targeting 16S rRNA genes, which is now commonly performed (Chapter 3). As pointed out by many studies, amplicon sequencing targeting 16S rRNA genes has a number of shortcomings, including PCR bias, primer bias, and low resolution of 16S rRNA gene sequences (Bálint et al., 2016; Klindworth et al., 2012). While the sequencing cost is going down and long-read sequencers such as PacBio and MinION are getting popular (Hiraoka et al., 2016), conventional amplicon sequencing is getting less attention. Nevertheless, I argue that datasets from conventional methods have unique advantages.

Because *diversity* is the nature of biology, biological hypothesis often needs to

be tested in various kinds of organisms or ecosystems. Especially in macroecology, the variance of data is inevitably higher than those obtained in controlled laboratories, and many samples are required to statistically testify the hypothesis in question. Considering this, the availability of vast amount of data should be of high priority, and therefore use of conventional approaches is justified; and furthermore, making the most of datasets obtained by conventional methods is of great importance.

4-2-4 Manipulating high dimensional data

A key concept of biology is diversity at different scales: from biomes, to species, to genes. As a natural consequence, biologists often work with data consisting of a large number of elements (i.e., high dimensional data), including microbial community structure data consisting of thousands of species (see Chapter 1), orthologous tables in genomic studies covering thousands of orthologues, and gene expression data generated by transcriptomic or microarray experiments. While high dimensionality is an essential characteristic of biological data, it often hampers the intuitive interpretation and visualization of biological data. Therefore, reducing data dimensionality has been a vital goal within the field bioinformatics motivated by the strong need to facilitate discussion on the ever-increasing high dimensional data.

In microbial ecology, multivariate statistics are popularly used. For example, microbial ecologists compress microbial community structure, consisting of thousands of species or OTUs, into two-dimensional plots using principal coordinate analysis (PCoA) or non-metric multivariate dimensional scaling. Similarly, genome-wide association studies summarize massive single nucleotide polymorphism (SNP) patterns using principal component analysis (PCA). While they serve the need for visualization, they are not useful for biological interpretation. The PCoA axes are defined *ad hoc* and do not have biological meanings; the eigenvectors of PCA are interpretable, but their high-dimensionality hampers biological interpretation of the data (of note, PCA is unsuitable for ecological data including microbial community structure, where linearity cannot be assumed). In this regard, the interpretability of trait-based summary is valuable.

4-3 Remaining problems

Here, I discuss three remaining problems that have not been addressed in the present study, mainly focusing on the limitation of habitat-based analysis and ProkAtlas.

First of all, a major limitation of this study is that it only focused on prokaryotes. Eukaryotic microorganisms are undoubtedly important players in ecosystem. Presumably, habitat-based analysis is applicable to eukaryotes, and habitat database like ProkAtlas can be constructed by identifying 18S rRNA gene sequences or ITS regions (R. Tanaka et al., 2014). A potential problem may be that metagenomic datasets contain much less eukaryotic sequences compared with prokaryotic sequences; the copy number of 18S rRNA gene is <10% of 16S rRNA gene copy number, even in fungi-rich soils (Tkacz et al., 2018). However, this could be solved by computational efforts: specifically, collecting large amount of data, part of which have been abandoned in the present study (see Section 2-2), would provide sufficient number of sequences.

Secondly, the environmental categories used in this study may look confused. Environmental categories in ProkAtlas (Table 2-1) are not mutually exclusive; for example, the three categories "*human*", "*human_gut*", and "*gut*" are obviously overlapped. In addition to aforementioned hierarchical ontology systems EnvO (Thompson et al., 2017) and MIGS/MIMS (Field et al., 2008), recently-developed Latent Environment Allocation (LEA) (Higashi et al., 2018) illustrates the relationships between environmental categories. On the other hand, the vagueness in environmental categories could merit the users of ProkAtlas. As shown in Chapter 3, several environmental categories need to be concatenated to draw interpretations from habitat compositions. This process is up to each user and the context of the research: for example, when comparing saline lakes and freshwater lakes, "*freshwater*" and "*seawater*" should be separated; on the other hand, these two may be grouped in one when comparing aquatic environments and non-aquatic environments. Supposedly, the extent of such flexibility is left to the design concept of each database or tool.

Thirdly, in the context of trait-based approach, microbial interspecies relationship should also be definitely incorporated into microbial community ecology. This is even more challenging, because the number of two-species interactions is proportional to the square of number of species. Besides in reality, microbes in complex community interact between three species or more. In fact, microbial community dynamics can be hardly predicted by simply compiling two-species interactions

(Friedman et al., 2017; although the authors' argument is opposite). Thus, the patterns of interspecies relationship to be considered are numerous, and cannot be exhaustively testified. Although a recent work preliminarily proposed methods to systematically predict interspecies interactions (DiMucci et al., 2018), their applicability to diverse microbes is still questionable. Others have struggled to model multi-species microbial ecosystems by simplifying interspecies relationships, typically using co-occurrence networks. However, the information obtained from such coarse models is currently very limited. In summary, neither of the two approaches, namely [i] compiling two-species interactions to multi-species systems, or [ii] modeling multi-species system by simplifying interspecies interaction, works well. In this dissertation, I just point out this problem for the sake of future studies.

References

(1) Amir, A., Daniel, M., Navas-Molina, J., Kopylova, E., Morton, J., Xu, Z.Z., Eric, K., Thompson, L., Hyde, E., Gonzalez, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2, e00191–16.

(2) Antwis, R.E., Griffiths, S.M., Harrison, X.A., Aranega-Bou, P., Arce, A., Bettridge, A.S., Brailsford, F.L., de Menezes, A., Devaynes, A., Forbes, K.M., Fry, E.L., Goodhead, I., Haskell, E., Heys, C., James, C., Johnston, S.R., Lewis, G.R., Lewis, Z., Macey, M.C., McCarthy, A., McDonald, J.E., Mejia-Florez, N.L., O'Brien, D., Orland, C., Pautasso, M., Reid, W.D.K., Robinson, H.A., Wilson, K., Sutherland, W.J., 2017. Fifty important research questions in microbial ecology. FEMS Microbiol. Ecol. 93, fiz044. https://doi.org/10.1093/femsec/fix044

(3) Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., Weissenbach, J., Ehrlich, S.D., Bork, P., 2011. Enterotypes of the human gut microbiome. Nature 473, 174–180. https://doi.org/10.1038/nature09944

(4) Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M.T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y.S., Kotowska, D., Colding, C., Tremaroli, V., Yin, Y., Bergman, S., Xu, X., Madsen, L., Kristiansen, K., Dahlgren, J., Jun, W., 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17, 690–703. https://doi.org/10.1016/j.chom.2015.04.004

(5) Bálint, M., Bahram, M., Eren, A.M., Faust, K., Fuhrman, J.A., Lindahl, B., O'Hara, R.B., öpik, M., Sogin, M.L., Unterseher, M., Tedersoo, L., 2016. Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genesa. FEMS Microbiol. Rev. 40, 686–700. https://doi.org/10.1093/femsre/fuw017

(6) Barberán, A., Caceres Velazquez, H., Jones, S., Fierer, N., 2017. Hiding in plain sight: Mining bacterial species records for phenotypic trait information. mSphere 2, e00237–17. https://doi.org/10.1128/mSphere.00237-17

(7) Barberán, A., Ramirez, K.S., Leff, J.W., Bradford, M.A., Wall, D.H., Fierer, N., 2014. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. Ecol. Lett. 17, 794–802. https://doi.org/10.1111/ele.12282

(8) Bergkemper, F., Schöler, A., Engel, M., Lang, F., Krüger, J., Schloter, M., Schulz, S., 2016. Phosphorus depletion in forest soils shapes bacterial communities towards phosphorus recycling systems. Environ. Microbiol. 18, 1988–2000. https://doi.org/10.1111/1462-2920.13442

(9) Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. 108, 4516–4522. https://doi.org/10.1073/pnas.1000080107

(10) Carter, K.M., Lu, M., Luo, Q., Jiang, H., An, L., 2020. Microbial community dissimilarity for source tracking with application in forensic studies. PLoS One 15, e0236082. https://doi.org/10.1371/journal.pone.0236082

(11) Castle, S.C., Sullivan, B.W., Knelman, J., Hood, E., Nemergut, D.R., Schmidt, S.K., Cleveland, C.C., 2017. Nutrient limitation of soil microbial activity during the earliest stages of ecosystem development. Oecologia 185, 513–524. https://doi.org/10.1007/s00442-017-3965-6

(12) Choudoir, M.J., Barberán, A., Menninger, H.L., Dunn, R.R., Fierer, N., 2018. Variation in range size and dispersal capabilities of microbial taxa. Ecology 99, 322–334. https://doi.org/10.1002/ecy.2094

(13) Chvatal, V., 1979. A Greedy Heuristic for the Set-Covering Problem. Math. Oper. Res. 4, 233.

(14) Corlett, R.T., Westcott, D.A., 2013. Will plant movements keep up with climate change? Trends Ecol. Evol. 28, 482–8. https://doi.org/10.1016/j.tree.2013.04.003

(15) Delgado-Baquerizo, M., Bardgett, R.D., Vitousek, P.M., Maestre, F.T., Williams, M.A., Eldridge, D.J., Lambers, H., Neuhauser, S., Gallardo, A., García-Velázquez, L., Sala, O.E., Abades, S.R., Alfaro, F.D., Berhe, A.A., Bowker, M.A., Currier, C.M., Cutler, N.A., Hart, S.C., Hayes, P.E., Hseu, Z.-Y., Kirchmair, M., Peña-Ramírez, V.M., Pérez, C.A., Reed, S.C., Santos, F., Siebe, C., Sullivan, B.W., Weber-Grullon, L., Fierer, N., 2019. Changes in belowground biodiversity during ecosystem development. Proc. Natl. Acad. Sci. 116, 6891–6896. https://doi.org/10.1073/pnas.1818400116

(16) Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-gonzález, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. Science. 325, 320–325. https://doi.org/10.1126/science.aap9516

(17) Delgado-Baquerizo, M., Reich, P.B., Khachane, A.N., Campbell, C.D., Thomas, N., Freitag, T.E., Abu Al-Soud, W., Sørensen, S., Bardgett, R.D., Singh, B.K., 2017. It is elemental: soil nutrient stoichiometry drives bacterial diversity. Environ. Microbiol. 19, 1176–1188. https://doi.org/10.1111/1462-2920.13642

(18) DiMucci, D., Kon, M., Segrè, D., 2018. Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. mSystems 3, 1–13. https://doi.org/10.1128/mSystems.00181-18

(19) Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

(20) Evans, S., Martiny, J.B.H., Allison, S.D., 2017. Effects of dispersal and selection on stochastic assembly in microbial communities. ISME J. 11, 176–185. https://doi.org/10.1038/ismej.2016.96

(21) Ferrenberg, S., O 'neill, S.P., Knelman, J.E., Todd, B., Duggan, S., Bradley, D., Robinson, T., Schmidt, S.K., Townsend, A.R., Williams, M.W., Cleveland, C.C., Melbourne, B.A., Jiang, L., Nemergut, D.R., 2013. Changes in assembly processes in soil bacterial communities following a wildfire disturbance. ISME J. 7, 1102–1111. https://doi.org/10.1038/ismej.2013.11

(22) Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M.J., Angiuoli, S. V, Ashburner, M., Axelrod, N., Baldauf, S.,

Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F.O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S.E., Li, K., Lister, A.L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., Wipat, A., 2008. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26, 541–7. https://doi.org/10.1038/nbt1360

(23) Fierer, N., 2017. Embracing the unknown: Disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579–590. https://doi.org/10.1038/nrmicro.2017.87

(24) Fierer, N., Grandy, A.S., Six, J., Paul, E.A., 2009. Searching for unifying principles in soil ecology. Soil Biol. Biochem. 41, 2249–2256. https://doi.org/10.1016/j.soilbio.2009.06.009

(25) Fierer, N., Nemergut, D., Knight, R., Craine, J.M., 2010. Changes through time: Integrating microorganisms into the study of succession. Res. Microbiol. 161, 635–642. https://doi.org/10.1016/j.resmic.2010.06.002

(26) Freedman, Z., Zak, D.R., 2015. Soil bacterial communities are shaped by temporal and environmental filtering: evidence from a long-term chronosequence. Environ. Microbiol. 17, 3208–3218. https://doi.org/10.1111/1462-2920.12762

(27) Friedman, J., Higgins, L.M., Gore, J., 2017. Community structure follows simple assembly rules in microbial microcosms. Nat. Ecol. Evol. 1, 0109. https://doi.org/10.1038/s41559-017-0109

(28) Garnier, E., Navas, M.-L., 2012. A trait-based approach to comparative functional plant ecology: concepts, methods and applications for agroecology. A review. Agron. Sustain. Dev. 32, 365–399. https://doi.org/10.1007/s13593-011-0036-y

(29) Giagnoni, L., Arenella, M., Galardi, E., Nannipieri, P., Renella, G., 2018. Bacterial culturability and the viable but non-culturable (VBNC) state studied by a proteomic approach using an artificial soil. Soil Biol. Biochem. 118, 51–58. https://doi.org/10.1016/j.soilbio.2017.12.004

(30) Gilbert, J.A., Jansson, J.K., Knight, R., 2018. Earth Microbiome Project and Global Systems Biology. mSystems 3, e00217–17. https://doi.org/10.1128/mSystems.00217-17

(31) Girvan, M.S., Campbell, C.D., Killham, K., Prosser, J.I., Glover, L.A., 2005. Bacterial diversity promotes community stability and functional resilience after perturbation. Environ. Microbiol. 7, 301–313. https://doi.org/10.1111/j.1462-2920.2005.00695.x

(32) Glöckner, F.O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G., Yarza, P., Peplies, J., Westram, R., Ludwig, W., 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J. Biotechnol. 261, 169–176. https://doi.org/10.1016/j.jbiotec.2017.06.1198

(33) Goto, D.K., Yan, T., 2011. Effects of land uses on fecal indicator bacteria in the water

and soil of a tropical watershed. Microbes Environ. 26, 254–260. https://doi.org/10.1264/jsme2.me11115

(34) Graham, E.B., Knelman, J.E., Schindlbacher, A., Siciliano, S., Breulmann, M., Yannarell, A., Beman, J.M., Abell, G., Philippot, L., Prosser, J., Foulquier, A., Yuste, J.C., Glanville, H.C., Jones, D.L., Angel, R., Salminen, J., Newton, R.J., Bürgmann, H., Ingram, L.J., Hamer, U., Siljanen, H.M.P., Peltoniemi, K., Potthast, K., Bañeras, L., Hartmann, M., Banerjee, S., Yu, R.Q., Nogaro, G., Richter, A., Koranda, M., Castle, S.C., Goberna, M., Song, B., Chatterjee, A., Nunes, O.C., Lopes, A.R., Cao, Y., Kaisermann, A., Hallin, S., Strickland, M.S., Garcia-Pausas, J., Barba, J., Kang, H., Isobe, K., Papaspyrou, S., Pastorelli, R., Lagomarsino, A., Lindström, E.S., Basiliko, N., Nemergut, D.R., 2016. Microbes as engines of ecosystem function: When does community structure enhance predictions of ecosystem processes? Front. Microbiol. 7, 214. https://doi.org/10.3389/fmicb.2016.00214

(35) Graham, E.B., Wieder, W.R., Leff, J.W., Weintraub, S.R., Townsend, A.R., Cleveland, C.C., Philippot, L., Nemergut, D.R., 2014. Do we need to understand microbial communities to predict ecosystem function? A comparison of statistical models of nitrogen cycling processes. Soil Biol. Biochem. 68, 279–282. https://doi.org/10.1016/j.soilbio.2013.08.023

(36) Griffiths, B.S., Philippot, L., 2013. Insights into the resistance and resilience of the soil microbial community. FEMS Microbiol. Rev. 37, 112–129. https://doi.org/10.1111/j.1574-6976.2012.00343.x

(37) Grime, J.P., 1974. Vegetation classification by reference to strategies. Nature 250, 26–31. https://doi.org/10.1038/250026a0

(38) Grossart, H.-P., Dziallas, C., Leunert, F., Tang, K.W., 2010. Bacteria dispersal by hitchhiking on zooplankton. Proc. Natl. Acad. Sci. 107, 11959–11964. https://doi.org/10.1073/pnas.1000668107

(39) Guieysse, B., Wuertz, S., 2012. Metabolically versatile large-genome prokaryotes. Curr. Opin. Biotechnol. 23, 467–473. https://doi.org/10.1016/j.copbio.2011.12.022

(40) Guittar, J., Shade, A., Litchman, E., 2019. Trait-based community assembly and succession of the infant gut microbiome. Nat. Commun. 10, 512.

(41) Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245-9. https://doi.org/10.1016/s1074-5521(98)90108-9

(42) Higashi, K., Suzuki, S., Kurosawa, S., Mori, H., Kurokawa, K., 2018. Latent environment allocation of microbial community data. PLOS Comput. Biol. 14, e1006143. https://doi.org/10.1371/journal.pcbi.1006143

(43) Hiraoka, S., Yang, C.C., Iwasaki, W., 2016. Metagenomics and Bioinformatics in Microbial Ecology: Current Status and Beyond. Microbes Environ. 31, 204–212. https://doi.org/10.1264/jsme2.ME16024

(44) Hitchens, A.P., Leikind, M.C., 1939. The introduction of agar into bacteriology. J. Bacteriol. 37, 485–93.

(45) Isobe, K., Ikutani, J., Fang, Y., Yoh, M., Mo, J., Suwa, Y., Yoshida, M., Senoo, K., Otsuka, S., Koba, K., 2018. Highly abundant acidophilic ammonia-oxidizing archaea

causes high rates of nitrification and nitrate leaching in nitrogen-saturated forest soils. Soil Biol. Biochem. in press. https://doi.org/10.1016/j.soilbio.2018.04.021

(46) Isobe, K., Ise, Y., Kato, H., Oda, T., Vincenot, C.E., Koba, K., Tateno, R., Senoo, K., Ohte, N., 2019. Consequences of microbial diversity in forest nitrogen cycling: diverse ammonifiers and specialized ammonia oxidizers. ISME J. https://doi.org/10.1038/s41396-019-0500-2

(47) Ji, M., Kong, W., Yue, L., Wang, J., Deng, Y., Zhu, L., 2019. Salinity reduces bacterial diversity, but increases network complexity in Tibetan Plateau lakes. FEMS Microbiol. Ecol. 95, fiz190. https://doi.org/10.1093/femsec/fiz190

(48) Jiang, Y., Lei, Y., Yang, Y., Korpelainen, H., Niinemets, Ü., Li, C., 2018. Divergent assemblage patterns and driving forces for bacterial and fungal communities along a glacier forefield chronosequence. Soil Biol. Biochem. 118, 207–216. https://doi.org/10.1016/j.soilbio.2017.12.019

(49) Karsch-Mizrachi, I., Takagi, T., Cochrane, G., 2018. The international nucleotide sequence database collaboration. Nucleic Acids Res. 46, D48–D51. https://doi.org/10.1093/nar/gkx1097

(50) Kearns, P.J., Shade, A., 2017. Trait-based patterns of microbial succession in dormancy potential and heterotrophic strategy: case studies of resource-based and post-press succession. ISME J. 12, 2575–2581. https://doi.org/10.1038/s41396-018-0194-x

(51) Kirs, M., Kisand, V., Wong, M., Caffaro-Filho, R.A., Moravcik, P., Harwood, V.J., Yoneyama, B., Fujioka, R.S., 2017. Multiple lines of evidence to identify sewage as the cause of water quality impairment in an urbanized tropical watershed. Water Res. 116, 23–33. https://doi.org/10.1016/j.watres.2017.03.024

(52) Klindworth, A., Peplies, J., Pruesse, E., Schweer, T., Glöckner, F.O., Quast, C., Horn, M., 2012. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41, e1. https://doi.org/10.1093/nar/gks808

(53) Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. Nat. Methods 8, 761–765. https://doi.org/10.1038/nmeth.1650

(54) Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

(55) Krause, S., Le Roux, X., Niklaus, P.A., van Bodegom, P.M., Lennon T., J.T., Bertilsson, S., Grossart, H.P., Philippot, L., Bodelier, P.L.E., 2014. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Front. Microbiol. 5, 251. https://doi.org/10.3389/fmicb.2014.00251

(56) Lauber, C.L., Hamady, M., Knight, R., Fierer, N., 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. 75, 5111–20. https://doi.org/10.1128/AEM.00335-09

(57) Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen,

J., Pang, X., Zhang, M., Wei, H., Chen, Y., Lu, H., Zuo, J., Su, M., Qiu, Y., Jia, W., Xiao, C., Smith, L.M., Yang, S., Holmes, E., Tang, H., Zhao, G., Nicholson, J.K., Li, L., Zhao, L., 2008. Symbiotic gut microbes modulate human metabolic phenotypes. Proc. Natl. Acad. Sci. 105, 2117–2122. https://doi.org/10.1073/pnas.0712038105

(58) Liu, W.T., Marsh, T.L., Cheng, H., Forney, L.J., 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl. Environ. Microbiol. 63, 4516–22.

(59) Louca, Stilianos, Jacques, S.M.S., Pires, A.P.F., Leal, J.S., Srivastava, D.S., Parfrey, L.W., Farjalla, V.F., Doebeli, M., 2016a. High taxonomic variability despite stable functional structure across microbial communities. Nat. Ecol. Evol. 1, 0015. https://doi.org/10.1038/s41559-016-0015

(60) Louca, S., Parfrey, L.W., Doebeli, M., 2016b. Decoupling function and taxonomy in the global ocean microbiome. Science. 353, 1272–1277. https://doi.org/10.1126/science.aaf4507

(61) Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U. S. A. 104, 11436–11440. https://doi.org/10.1073/pnas.0611525104

(62) Lynch, M.D.J., Neufeld, J.D., 2015. Ecology and exploration of the rare biosphere. Nat. Rev. Microbiol. 13, 217–229. https://doi.org/10.1038/nrmicro3400

(63) Mapelli, F., Marasco, R., Fusi, M., Scaglia, B., Tsiamis, G., Rolli, E., Fodelianakis, S., Bourtzis, K., Ventura, S., Tambone, F., Adani, F., Borin, S., Daffonchio, D., 2018. The stage of soil development modulates rhizosphere effect along a High Arctic desert chronosequence. ISME J. 12, 1188–1198. https://doi.org/10.1038/s41396-017-0026-4

(64) Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M.A., Grechkin, Y., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., Hugenholtz, P., Kyrpides, N.C., 2007. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 36, D534–D538. https://doi.org/10.1093/nar/gkm869

(65) Martiny, J.B.H., Jones, S.E., Lennon, J.T., Martiny, A.C., 2015. Microbiomes in light of traits: A phylogenetic perspective. Science. 350, 649. https://doi.org/10.1126/science.aac9323

(66) McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., Desantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618. https://doi.org/10.1038/ismej.2011.139

(67) Mise, K., Maruyama, R., Miyabara, Y., Kunito, T., Senoo, K., Otsuka, S., 2019. Time-series analysis of phosphorus-depleted microbial communities in carbon/nitrogen-amended soils. Appl. Soil Ecol. https://doi.org/10.1016/j.apsoil.2019.08.008

(68) Moyer, C.L., Dobbs, F.C., Karl, D.M., 1994. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl. Environ. Microbiol. 60, 871–9.

(69) Muyzer, G., de Waal, E.C., Uitterlinden, A.G., 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase

chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. 59, 695–700.

(70) Muyzer, G., Smalla, K., 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek 73, 127–41. https://doi.org/10.1023/a:1000669317571

(71) Nemergut, D.R., Knelman, J.E., Ferrenberg, S., Bilinski, T., Melbourne, B., Jiang, L., Violle, C., Darcy, J.L., Prest, T., Schmidt, S.K., Townsend, A.R., 2016. Decreases in average bacterial community rRNA operon copy number during succession. ISME J. 10, 1147–1156. https://doi.org/10.1038/ismej.2015.191

(72) Nosho, K., Fukushima, H., Asai, T., Nishio, M., Takamaru, R., Kobayashi-Kirschvink, K.J., Ogawa, T., Hidaka, M., Masaki, H., 2018. cAMP-CRP acts as a key regulator for the viable but non-culturable state in escherichia coli. Microbiology 164, 410–419. https://doi.org/10.1099/mic.0.000618

(73) Oliverio, A.M., Bradford, M.A., Fierer, N., 2017. Identifying the microbial taxa that consistently respond to soil warming across time and space. Glob. Chang. Biol. 23, 2117–2129. https://doi.org/10.1111/gcb.13557

(74) Ortiz-Álvarez, R., Fierer, N., De Los Ríos, A., Casamayor, E.O., Barberán, A., 2018. Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. ISME J. 12, 1658–1667. https://doi.org/10.1038/s41396-018-0076-2

(75) Pfeiffer, S., Pastar, M., Mitter, B., Lippert, K., Hackl, E., Lojan, P., Oswald, A., Sessitsch, A., 2014. Improved group-specific primers based on the full SILVA 16S rRNA gene reference database. Environ. Microbiol. 16, 2389–2407. https://doi.org/10.1111/1462-2920.12350

(76) Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, 590–596. https://doi.org/10.1093/nar/gks1219

(77) R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

(78) Ramirez, K.S., Knight, C.G., De Hollander, M., Brearley, F.Q., Constantinides, B., Cotton, A., Creer, S., Crowther, T.W., Davison, J., Delgado-Baquerizo, M., Dorrepaal, E., Elliott, D.R., Fox, G., Griffiths, R.I., Hale, C., Hartman, K., Houlden, A., Jones, D.L., Krab, E.J., Maestre, F.T., McGuire, K.L., Monteux, S., Orr, C.H., Van Der Putten, W.H., Roberts, I.S., Robinson, D.A., Rocca, J.D., Rowntree, J., Schlaeppi, K., Shepherd, M., Singh, B.K., Straathof, A.L., Bhatnagar, J.M., Thion, C., Van Der Heijden, M.G.A., De Vries, F.T., 2018. Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat. Microbiol. 3, 189–196. https://doi.org/10.1038/s41564-017-0062-x

(79) Rath, K.M., Fierer, N., Murphy, D. V, Rousk, J., 2019. Linking bacterial community composition to soil salinity along environmental gradients. ISME J. 13, 836–846. https://doi.org/10.1038/s41396-018-0313-8

(80) Reimer, L.C., Vetcininova, A., Carbasse, J.S., Söhngen, C., Gleim, D., Ebeling, C.,

Overmann, J., 2019. Bac Dive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. Nucleic Acids Res. 47, D631–D636. https://doi.org/10.1093/nar/gky879

(81) Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437. https://doi.org/10.1038/nature12352

(82) Roller, B.R.K., Stoddard, S.F., Schmidt, T.M., 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat. Microbiol. 1, 16160. https://doi.org/10.1038/nmicrobiol.2016.160

(83) Schmieder, R., Edwards, R.A., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. https://doi.org/10.1093/bioinformatics/btr026

(84) Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. Nat. Methods 5, 16–18. https://doi.org/10.1038/nmeth1156

(85) Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T., Morganti, R.M., Conley, S.D., Chaib, H., Red-Horse, K., Longaker, M.T., Snyder, M.P., Krasnow, M.A., Weissman, I.L., 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724. https://doi.org/10.1101/125724

(86) Sriswasdi, S., Yang, C.C., Iwasaki, W., 2017. Generalist species drive microbial dispersion and evolution. Nat. Commun. 8, 1162. https://doi.org/10.1038/s41467-017-01265-1

(87) Steen, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., Cameron Thrash, J., 2019. High proportions of bacteria and archaea across most biomes remain uncultured. ISME J. 13, 3126–3130. https://doi.org/10.1038/s41396-019-0484-y

(88) Tanaka, R., Hino, A., Tsai, I.J., Palomares-Rius, J.E., Yoshida, A., Ogura, Y., Hayashi, T., Maruyama, H., Kikuchi, T., Poulin, R., Morand, S., Sukhdeo, M., Bansemir, A., Korallo, N., Vinarski, M., Krasnov, B., Shenbrot, G., Mouillot, D., Combes, C., Waters, A., Higgins, D., McCutchan, T., Desai, N., Antonopoulos, D., Gilbert, J., Glass, E., Meyer, F., Weinstock, G., Yatsunenko, T., Rey, F., Manary, M., Trehan, I., Dominguez-Bello, M., Sogin, M., Morrison, H., Huber, J., Welch, D.M., Huse, S., Amir, A., Zeisel, A., Zuk, O., Elgart, M., Stern, S., Jumpponen, A., Jones, K., Porazinska, D., Giblin-Davis, R., Powers, T., Thomas, W., Porazinska, D., Giblin-Davis, R., Faller, L., Farmerie, W., Kanzaki, N., Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Webster, J., Macdonald, D., Holterman, M., Wurff, A. van der, Elsen, S. van den, Megen, H. van, Bongers, T., Waeschenbach, A., Webster, B., Bray, R., Littlewood, D., Katoh, K., Misawa, K., Kuma, K., Miyata, T., Castresana, J., Stamatakis, A., Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Huntley, J., Amaral-Zettler, L., McCliment, E., Ducklow, H., Huse, S., Vestheim, H., Jarman, S., Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Turnbaugh, P., Ley, R., Hamady, M.,

Fraser-Liggett, C., Knight, R., Bik, H., Fournier, D., Sung, W., Bergeron, R., Thomas, W., Hugerth, L., Muller, E., Hu, Y., Lebrun, L., Roume, H., Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E., Zhao, X., Duszynski, D., 2014. Assessment of helminth biodiversity in wild rats using 18S rDNA based metagenomics. PLoS One 9, e110769. https://doi.org/10.1371/journal.pone.0110769

(89) Tanaka, T., Kawasaki, K., Daimon, S., Kitagawa, W., Yamamoto, K., Tamaki, H., Tanaka, M., Nakatsu, C.H., Kamagata, Y., 2014. A hidden pitfall in the preparation of agar media undermines microorganism cultivability. Appl. Environ. Microbiol. 80, 7659–7666. https://doi.org/10.1128/AEM.02741-14

(90) Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. https://doi.org/10.1038/nature24621

(91) Tkacz, A., Hortala, M., Poole, P.S., 2018. Absolute quantitation of microbiota abundance in environmental samples. Microbiome 6, 110. https://doi.org/10.1186/s40168-018-0491-7

(92) Torres, P.J., Edwards, R.A., McNair, K.A., 2017. PARTIE: A partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. Bioinformatics 33, 2389–2391. https://doi.org/10.1093/bioinformatics/btx184

(93) Unno, T., Staley, C., Brown, C.M., Han, D., Sadowsky, M.J., Hur, H.G., 2018. Fecal pollution: new trends and challenges in microbial source tracking using next-generation sequencing. Environ. Microbiol. 20, 3132–3140. https://doi.org/10.1111/1462-2920.14281

(94) Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304, 66–74. https://doi.org/10.1126/science.1093857

(95) Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., Garnier, E., 2007. Let the concept of trait be functional! Oikos 116, 882–892. https://doi.org/10.1111/j.2007.0030-1299.15559.x

(96) Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267. https://doi.org/10.1128/AEM.00062-07

(97) Werner, G.D.A., Kiers, E.T., 2015. Order of arrival structures arbuscular mycorrhizal colonization of plants. New Phytol. 205, 1515–1524. https://doi.org/10.1111/nph.13092

(98) Wertz, S., Degrange, V., Prosser, J.I., Poly, F., Commeaux, C., Freitag, T.,

Guillaumaud, N., Roux, X. Le, 2006. Maintenance of soil functioning following erosion of microbial diversity. Environ. Microbiol. 8, 2162–2169. https://doi.org/10.1111/j.1462-2920.2006.01098.x

(99) Yang, C.C., Iwasaki, W., 2014. MetaMetaDB: A database and analytic system for investigating microbial habitability. PLoS One 9, e87126. https://doi.org/10.1371/journal.pone.0087126

(100) Yao, Q., Li, Z., Song, Y., Wright, S.J., Guo, X., Tringe, S.G., Tfaily, M.M., Paša-Tolić, L., Hazen, T.C., Turner, B.L., Mayes, M.A., Pan, C., 2018. Community proteogenomics reveals the systemic impact of phosphorus availability on microbial functions in tropical soil. Nat. Ecol. Evol. 2, 499–509. https://doi.org/10.1038/s41559-017-0463-5

(101) Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.-M., Härkönen, T., Ryhänen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D., Lander, E.S., Knip, M., Xavier, R.J., 2016. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci. Transl. Med. 8, 343ra81-343ra81. https://doi.org/10.1126/scitranslmed.aad0917

(102) Zhao, S., Liu, J., Banerjee, S., Zhou, N., Zhao, Z., Zhang, K., Hu, M., Tian, C., 2020. Biogeographical distribution of bacterial communities in saline agricultural soil. Geoderma 361, 114095. https://doi.org/10.1016/j.geoderma.2019.114095

Acknowledgement

First of all, I thank my supervisor, Dr. Wataru Iwasaki for his overall support for my Ph.D. work. He helped me develop my vague ideas into solid and concrete research questions, showed me how to design a research in a goal-driven manner, and made me realize why diversity is important in conducting an academic research. In short, he has managed the best environment for me to perform this study.

I also thank current and former members of Iwasaki Laboratory for helpful discussions. In particular, technical advices from Dr. Motomu Matsui, Dr. Ching-chia Yang, Ken Kuroki, and Mitsuki Usui were invaluable for handling a vast amount of metagenomic data. Dr. Masaki Hoso and Tomoyuki Mikami expertly commented on my study from the perspective of "classical" ecology, which helped me notice many gaps between classical and microbial community ecology. Regarding trait-based microbial ecology, Shun Yamanouchi provided basic and useful suggestions that helped me draw future perspectives of this field. I also thank Dr. Cosentino Salvatore for helping me improve ProkAtlas Online and Yuki Sakamoto for conducting preliminary survey on the EMP datasets and environmental ontologies. Lastly, I am grateful to Satoko Fukuda and Kaori Motoki, who enrolled in Ph.D. course at the same time with me, for making me recognize my potential strength and weakness as a researcher. I would also like to note that the fundamental motivation of this study had been fostered through discussion with Dr. Shigeto Otsuka, my supervisor in the master course.

My Ph.D. work was financially supported by JSPS KAKENHI [Grant Numbers 19J14142, 19H05688, 18H04136, and 16H06279] and JSPS Research Fellowship for Young Scientists (DC2). Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.