Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures

HIGHLIGHTS
We developed a database, ProkAtlas, denoting the habitat preferences of prokaryotes

ProkAtlas represents a prokaryotic community using habitat preferences of its members

The powerfulness of ProkAtlas is showcased by six datasets from various environments

We provide web interface of ProkAtlas at https://msk33.github.io/prokatlas.html
Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures

Kazumori Mise and Wataru Iwasaki

SUMMARY
The recent prevalence of high-throughput sequencing has been producing numerous prokaryotic community structure datasets. Although the trait-based approach is useful to interpret those datasets from ecological perspectives, available trait information is biased toward culturable prokaryotes, especially those of clinical and public health relevance, and thus may not represent the breadth of microbiota found across many of Earth’s environments. To facilitate habitat-based analysis free of such bias, here we report a ready-to-use prokaryotic habitat database, ProkAtlas. ProkAtlas comprehensively links 16S rRNA gene sequences to prokaryotic habitats, using public shotgun metagenome datasets. We also developed a computational pipeline for habitat-based analysis of given prokaryotic community structures. After confirmation of the method effectiveness using 16S rRNA gene sequence datasets from individual genomes and the Earth Microbiome Project, we showed its validness and effectiveness in drawing ecological insights by applying it to six empirical prokaryotic community datasets from soil, aquatic, and human gut samples.

INTRODUCTION
In the era of high-throughput sequencing, huge numbers of prokaryotic community structure datasets are being produced by 16S rRNA gene amplicon and shotgun metagenomic sequencing methods (Ramirez et al., 2018; Thompson et al., 2017). These abundant datasets from diverse environments have contributed to unveiling the diversity and distributions of environmental microbial communities on Earth (Delgado-Baquerizo et al., 2018; Gilbert et al., 2018; Sunagawa et al., 2015); however, such datasets still severely hamper microbial ecologists from intuitive interpretation. Each community structure database usually contains hundreds or thousands of operational taxonomic units, species, or genera (Lynch and Neufeld, 2015; Nemer-gut et al., 2013; Sogin et al., 2006), and it is also common to be presented as compositions of high-rank taxonomies (i.e., phyla or classes). Whether such dimensionality reduction is conducted or not, it is difficult to directly obtain interpretable ecological insights from community structure databases, because ecological and physiological knowledge is not necessarily available for each of numerous members and such traits are often unconserved within high-rank clades (Martiny et al., 2015).

As a promising solution to this problem, the trait-based approach aims to enable the ecological interpretation of community structure datasets (Kearns and Shade, 2017; Krause et al., 2014). 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. To date, genome size (Barberán et al., 2014), rRNA gene copy number (Nemergut et al., 2016), growth rate, stress tolerance, capability to acquire carbon sources (Malik et al., 2020), metabolic potential (Louca et al., 2016), and pigmentation (Choudoir et al., 2018) data have been adopted for trait-based analyses of prokaryotic community structure datasets. Although these approaches performed well, their trait data were limited and biased to cultured prokaryotes with available genomic and physiological data. Prokaryotic communities, however, contain approximately 80% to more than 90% of uncultured members (Schloss et al., 2016; Steen et al., 2019).

As a promising trait-based approach that averts this issue, some studies exploited prokaryotic habitat information (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Species that inhabit seawater are more likely to have the trait of adaptability to saline environments than species that inhabit soil and freshwater (Krause et al., 2014; Luef et al., 2018; Weir et al., 2018; Zhai et al., 2018). This suggests that prokaryotic habitats are key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Prokaryotic habitats are defined as physical environments in which prokaryotes are unable to grow or remain long-term in the absence of human cultures (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Prokaryotic habitats are classified on the basis of spatial distribution and biotic conditions, and each habitat is defined as an environmental condition in which prokaryotes cannot grow (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This classification is based on the premise that prokaryotic habitats are the key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). As a promising trait-based approach that averts this issue, some studies exploited prokaryotic habitat information (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Species that inhabit seawater are more likely to have the trait of adaptability to saline environments than species that inhabit soil and freshwater (Krause et al., 2014; Luef et al., 2018; Weir et al., 2018; Zhai et al., 2018). This suggests that prokaryotic habitats are key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Prokaryotic habitats are classified on the basis of spatial distribution and biotic conditions, and each habitat is defined as an environmental condition in which prokaryotes cannot grow (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This classification is based on the premise that prokaryotic habitats are the key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This suggests that prokaryotic habitats are key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Prokaryotic habitats are classified on the basis of spatial distribution and biotic conditions, and each habitat is defined as an environmental condition in which prokaryotes cannot grow (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This classification is based on the premise that prokaryotic habitats are the key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This suggests that prokaryotic habitats are key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). Prokaryotic habitats are classified on the basis of spatial distribution and biotic conditions, and each habitat is defined as an environmental condition in which prokaryotes cannot grow (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This classification is based on the premise that prokaryotic habitats are the key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017). This suggests that prokaryotic habitats are key trait information for interpreting prokaryotic community structure datasets (Morton et al., 2017; Thompson et al., 2017; Washburne et al., 2017).
freshwater only. Likewise, species that inhabit animal gut are more likely to have the trait of adaptability to copiotrophic (eutrophic) environments than species that inhabit soil only (here we use the term trait because of the affinity to the concept of the trait-based approach, although habitat preference itself may be defined as a characteristic rather than a functional or physiological trait). Importantly, habitat preferences of each prokaryote can be obtained by comparing metagenomic datasets from different environmental samples without the reliance on cultivation and isolation experiments.

In this study, we propose an intuitive and user-friendly method for habitat-based analysis and show its effectiveness in interpreting and investigating prokaryotic community structure datasets. We developed a database named ProkAtlas that links 16S rRNA gene sequences mined from metagenomes to prokaryotic habitats by substantially extending MetaMetaDB, which was previously developed by our group (Haider et al., 2018; Yang and Iwasaki, 2014). ProkAtlas now contains Illumina-sequencer datasets, considers differences in the sizes of research projects, and uses only shotgun metagenomic datasets to avoid biases due to the use of universal and/or clade-specific primers in amplicon sequencing (Klindworth et al., 2012) (note that ProkAtlas can be adopted for analysis of amplicon-sequencing datasets). We also developed the ProkAtlas pipeline for the habitat-based analysis of community structure datasets. As proofs of concept, we further drew a bird’s-eye network of prokaryote co-occurrences among diverse environments and analyzed datasets from the Earth Microbiome Project (EMP), soil and lake-water samples with salinity gradients, human infant gut and glacial chronosequence soil samples undergoing primary successions, and river-water samples potentially polluted with pathogenic bacteria.

RESULTS AND DISCUSSION

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 1). Unlike EMP, ProkAtlas is free of PCR bias and applicable to any (variable) region of partial/full-length 16S rRNA gene sequences (note that biases from other experimental factors such as DNA extraction methods and library preparation processes would remain). 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, we tried to balance the database comprehensiveness and size. As we show later, increasing the size of ProkAtlas marginally affects or improves the performance of habitat preference prediction, whereas 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 (Glockner et al., 2017; McDonald et al., 2012). Each sequence in ProkAtlas is labeled with one of the environmental categories listed in Table 1 for prokaryotic habitat estimation with 16S rRNA gene sequences (see Table S1 for the list of corresponding NCBI taxon IDs). Although NCBI taxon IDs contain environmental categories of different granularity, they are accompanied by all 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 rhizosphere, comprise 55.3% of all sequences, and the top 26 categories comprise more than 90% of all sequences.

We further developed a pipeline to estimate the habitats of prokaryotes based on 16S rRNA gene sequences using ProkAtlas. Basically, this pipeline uses a sequence similarity search to query every 16S
| Environmental Category       | Number of Sequences | Number of Projects |
|-----------------------------|---------------------|--------------------|
| Activated_carbon            | 44                  | 1                  |
| Activated_sludge            | 7,383               | 78                 |
| Air                         | 300                 | 3                  |
| Algae                       | 308                 | 4                  |
| Anaerobic_digest            | 500                 | 5                  |
| Annelid                     | 1,038               | 13                 |
| Ant                         | 300                 | 3                  |
| Aquatic                     | 2,667               | 29                 |
| Aquifer                     | 1,900               | 19                 |
| Bat                         | 143                 | 2                  |
| Beach_sand                  | 100                 | 1                  |
| Biofilm                     | 300                 | 3                  |
| Biofilter                   | 100                 | 1                  |
| Biogas_fermenter            | 600                 | 6                  |
| Bioreactor                  | 2,923               | 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                       | 1,100               | 11                 |
| Crab                        | 100                 | 1                  |
| Crustacean                  | 100                 | 1                  |
| Endophyte                   | 30                  | 1                  |
| Epibiont                    | 100                 | 1                  |
| Estuary                     | 200                 | 2                  |
| Feces                       | 1,306               | 14                 |
| Fermentation                | 300                 | 3                  |
| Fish_gut                    | 25                  | 1                  |
| Food                        | 604                 | 7                  |
| Food_fermentation           | 300                 | 3                  |

Table 1. Environmental Categories, Numbers of 16S rRNA Gene Sequences Labeled by These Categories, and Numbers of Research Projects of These Categories Contributing to ProkAtlas

(Continued on next page)
| Environmental Category         | Number of Sequences | Number of Projects |
|--------------------------------|---------------------|--------------------|
| Food_production                | 100                 | 1                  |
| Fossil                         | 200                 | 2                  |
| Freshwater                     | 43,216              | 437                |
| freshwater_sediment            | 13,744              | 239                |
| Fungus                         | 3,737               | 38                 |
| Glacier                        | 500                 | 5                  |
| Groundwater                    | 9,540               | 97                 |
| Gut                            | 3,167               | 36                 |
| Halite                         | 14                  | 1                  |
| Hot_springs                    | 1,056               | 12                 |
| Human_bile                     | 111                 | 2                  |
| Human_blood                    | 1                   | 1                  |
| Human_eye                      | 131                 | 2                  |
| Human_gut                      | 1,815               | 25                 |
| Human_lung                     | 7,502               | 80                 |
| Human_oral                     | 291                 | 3                  |
| Human_reproductive_system      | 500                 | 5                  |
| Human_skin                     | 400                 | 4                  |
| Hydrocarbon                    | 33                  | 1                  |
| Hydrothermal_vent              | 4,016               | 44                 |
| Hypersaline_lake               | 900                 | 9                  |
| Hypolithon                     | 113                 | 2                  |
| Indoor                         | 100                 | 1                  |
| Insect                         | 241                 | 4                  |
| Insect_gut                     | 200                 | 2                  |
| Invertebrate                   | 300                 | 3                  |
| Lake_water                     | 5,700               | 57                 |
| Landfill                       | 400                 | 4                  |
| Leaf                           | 500                 | 5                  |
| Lichen                         | 500                 | 5                  |
| Marine                         | 45,298              | 461                |
| Marine_sediment                | 3,281               | 35                 |
| Microbial_fuel_cell            | 100                 | 1                  |
| Microbial_mat                  | 1,561               | 17                 |
| Mine_drainage                  | 200                 | 2                  |

*Table 1. Continued* (Continued on next page)
| Environmental Category           | Number of Sequences | Number of Projects |
|----------------------------------|---------------------|--------------------|
| Mine_tailing                     | 200                 | 2                  |
| Mixed_culture                    | 100                 | 1                  |
| Money                            | 200                 | 2                  |
| Mosquito                         | 180                 | 2                  |
| Moss                             | 600                 | 6                  |
| Mouse_gut                        | 1,194               | 18                 |
| Oil_field                        | 3                   | 1                  |
| Oral                             | 66                  | 1                  |
| Oyster                           | 100                 | 1                  |
| Paper_pulp                       | 100                 | 1                  |
| Parasite                         | 26                  | 1                  |
| Peat                             | 7,683               | 77                 |
| Permafrost                       | 1,449               | 16                 |
| Phyllosphere                     | 12,100              | 121                |
| Pig_gut                          | 576                 | 6                  |
| Plant                            | 4,579               | 53                 |
| Plastic                          | 6                   | 3                  |
| Pollen                           | 200                 | 2                  |
| Rat_gut                          | 100                 | 1                  |
| Rhizosphere                      | 21,152              | 213                |
| Rice_paddy                       | 3,100               | 31                 |
| Rock                             | 148                 | 2                  |
| Rock_porewater                   | 100                 | 1                  |
| Root_associated_fungus           | 100                 | 4                  |
| Root                             | 400                 | 1                  |
| Salt_lake                        | 1,900               | 19                 |
| Salt_marsh                       | 5,400               | 54                 |
| Sea_squirt                       | 400                 | 4                  |
| Seawater                         | 2,977               | 30                 |
| Sediment                         | 6,431               | 66                 |
| Skin                             | 100                 | 1                  |
| Sludge                           | 100                 | 1                  |
| Soil                             | 90,158              | 984                |
| Sponge                           | 300                 | 3                  |
| Stromatolite                     | 100                 | 1                  |
| Subsurface                       | 3,020               | 32                 |

*Table 1. Continued*
rRNA gene sequence of given data against the ProkAtlas database and obtains a list of environmental categories that are labeled to the hit sequences. Compositions of the retrieved environmental categories are represented by habitat preference scores. Note that (nearly) identical sequences are included within the 361,474 sequences in ProkAtlas and one query sequence can be mapped to two or more environmental categories. Because of this one-to-many relationship, the habitat preference scores are presented as compositions of multiple environmental categories. More details on the pipeline are available in Transparent Methods and Figures S1 and S2.

Bird’s-Eye Visualization of Prokaryote Co-occurrence Network among Diverse Environments

All ProkAtlas sequences were mapped to 60,278 SILVA entries, and their composition in each environmental category was quantified. By enumerating 16S rRNA gene sequences that co-occur in different environments using ProkAtlas, we obtained a comprehensive view of co-occurrences of prokaryotes among diverse environments as a network (Figure 2A).

As expected, related environments such as soil and rhizosphere were strongly associated with each other. The betweenness centrality, which quantifies the propagation of prokaryotes among different environmental categories, of extreme environment (i.e., hydrothermal_vent) was low (Figure 2B). Regarding this observation, it is reasonable that prokaryote co-occurrences or migrations via extreme environments are rare because of their non-moderate conditions and geographical isolation. It was also found that betweenness centralities of host-associated environments were relatively low. This observation was rather unexpected because prokaryotic hosts, especially animals, may bring prokaryotes to different environments and promote their migration (Grossart et al., 2010). We 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.

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 single amplified genome (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. We then calculated the habitat preference scores (see Methods) and found that these thes 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 (Figure 3A, Mann-Whitney U test, p < 0.001). This trend was similarly observed for various sets of

| Environmental Category | Number of Sequences | Number of Projects |
|------------------------|---------------------|--------------------|
| Surface                | 100                 | 1                  |
| Symbiont               | 69                  | 1                  |
| Termite_gut            | 2,919               | 31                 |
| Terrestrial            | 2,134               | 22                 |
| Tick                   | 100                 | 1                  |
| Urban                  | 6                   | 1                  |
| Viral                  | 600                 | 6                  |
| Wastewater             | 3,228               | 35                 |
| Wetland                | 11,900              | 119                |

Table 1. Continued
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. See also Tables S1, S2, and S3; Figure S7.
environmental categories, including isolates and/or SAGs from seawater, plants, feces, groundwater, lake water, hydrothermal vent, and bioreactors (Figures 3B–3J). 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, although 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 S3). Because of this, we assume that 150-bp threshold (the default value in our pipeline) would be long enough to achieve overall accuracy, while retaining enough amount of significant hits.

Habitat-Based Analysis of the EMP Dataset

To test the versatility of habitat-based prokaryotic community analysis using ProkAtlas, we 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 4). 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

Figure 2. Network Analysis of Co-occurrences of Prokaryotes among Environments

(A) A bird’s-eye network visualization of the co-occurrences. Nodes represent environmental categories. Edges are drawn only if Bray-Curtis dissimilarities are less than 0.9 (smaller dissimilarities shown by darker colors).

(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).
SAGs’ habitat preference scores of brackish lake-related environments (Figures S3 and S4).

Prokaryotic isolates’ habitat preference scores of freshwater-related environments (Figures S3 and S4).

Prokaryotic isolates’ habitat preference scores of feces-associated environments (Figures S3 and S4).

Prokaryotic isolates’ habitat preference scores of marine sediment-related environments (Figures S3 and S4).

Prokaryotic isolates’ habitat preference scores of hydrothermal-related environments (Figures S3 and S4).

Prokaryotic isolates’ habitat preference scores of bioreactor-related environments (Figures S3 and S4).

Each of the darker orange plots indicates the scores of isolates/SAGs from the relevant environment, whereas the lighter orange one indicates those of the other isolates/SAGs. A thick black line and a white circle indicate the 25–75% range and median within each plot, respectively.

Figure 3. Habitat Preference Scores of Isolates/SAGs Derived from a Specific Type of Environment

(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 hydrothermal-related environments (hydrothermal_vent and hot_springs).

(I) SAGs’ habitat preference scores of marine sediment-related environments (freshwater_sediment, sediment, marine_sediment, marine_salt, 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.

See also Figures S3 and S4.

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 (Figures S4 and S5), 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). We thus argue that the size of ProkAtlas achieves a good balance between information content and computational usability.

Habitat-Based Analyses of Soil and Lake-Water Samples with Salinity Gradients

As a proof-of-concept of the habitat-based analysis of prokaryotic community structures, we applied ProkAtlas to six 16S rRNA gene amplicon-sequencing datasets (Table 2). 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 span 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 Gemmatimonadetes (11.7% ± 4.63%) (Figure 5A). On the other hand, when the estimated habitats were investigated, we observed a clear trend that the habitat preference compositions were affected by soil salinity concentration (Figure 5B). More specifically, saline environments such as marine, seawater, estuary, 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, $r = 0.60, p < 0.001$) as expected. Thus, ProkAtlas clearly and intuitively highlights the microbial community characteristics of high-salinity soils, which is consistent with previous knowledge on the relationship between salinity and prokaryotic community structures (Lozupone and Knight, 2007; Rath et al., 2019).

The second dataset contained saline and non-saline lake water samples (Ji et al., 2019), totaling 78 samples from 25 lakes with diverse salinity from the Tibetan Plateau. Two of the samples lacking sampling site information were excluded from analysis. The dataset contained 6,054 sOTUs, among which 5,241 (86.6%), accounting for 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 Acidobacteria, Bacteroidetes, Cyanobacteria, and/or Proteobacteria (Figure 6A). 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 6B). The proportions of saline-water-related categories were significantly and strongly correlated with salinity (Figure 6C, Spearman’s correlation test, $r = 0.88, p < 0.001$). 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).
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.

Habitat-Based Analyses of Infant Gut and Glacial Chronosequence Soil Samples

The fourth and fifth datasets contained glacial chronosequence soil samples from two different glaciers (Jiang et al., 2018; Mapelli et al., 2018), which provide transects at different developmental stages from unweathered bedrocks to matured (extensively weathered) soil (Castle et al., 2017; Delgado-Baquerizo et al., 2019). Each of these datasets contained 21 bulk soil samples collected at seven sampling sites along a retreating glacial chronosequence. One was from Midtre Lovenbreen 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 Chloroflexi in the Norwegian dataset (Figure 8A) but phyla Acidobacteria, Bacteroidetes, and Proteobacteria in the Chinese dataset (Figure 8C). 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 (Figures 8B and 8D). This suggests that prokaryotic habitat preference can be a useful trait for analyzing community successions. In addition, a notable

| 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 PRJNAS03775  | 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 Lovenbreen 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 PRJNAA354498 | Jiang et al. (2018)         |
| Water sampled along Manoa Stream (Hawaii, USA)                                     | 25                | INSD PRJNAA376213 | Kirs et al. (2017)          |

Table 2. Six 16S rRNA Gene Amplicon-Sequencing Datasets Reanalyzed Using Habitat-Based Approach
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 7B) in contrast to the soil prokaryotic communities (Figures 8B and 8D).

In summary, the three 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).

**Habitat-Based Analyses of Potentially Polluted River Water**

Finally, the habitat-based analysis was applied to prokaryotic community mixture from distinct environments. The sixth dataset contained 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 the 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 9A). 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 9B). MS7–9 are located in a canal that is connected to the sea and directly influenced by tides, whereas MS6 is located approximately 500 m 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). We note that the source tracking methods (Knights et al., 2011; Unno et al., 2018) can also be adopted for a similar purpose, but ProkAtlas differs in that it is a ready-to-use and versatile database and users do not need to prepare reference datasets for themselves.

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.

Future Perspectives
In this study, we developed ProkAtlas for the habitat-based analysis of prokaryotic community structure datasets. The application of ProkAtlas to six datasets succeeded in intuitively highlighting environment-prokaryotic community structure relationships, delineating temporal succession patterns of prokaryotic communities and providing insights into environmental and ecological surveillance such as pollution sources. Notably, ProkAtlas is independent of biases toward cultured prokaryotes, which is a common problem in many tools in microbiology. In addition, as a comprehensive database, ProkAtlas is readily applicable to many prokaryotic community structure datasets without the need for additional experiments or cumbersome data processing and might also be applicable for discovering the general principles behind microbial evolution and migration. By complementing existing approaches, this habitat-based analysis will

Figure 8. Habitat-Based Analysis of Glacial Chronosequence Soil Samples
(A and B) (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 Lovenbreen Glacier moraine (Norway) (Mapelli et al., 2018).
(C and D) (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. See also Figure S6.

pollution (Goto and Yan, 2011; Kirs et al., 2017). We note that the source tracking methods (Knights et al., 2011; Unno et al., 2018) can also be adopted for a similar purpose, but ProkAtlas differs in that it is a ready-to-use and versatile database and users do not need to prepare reference datasets for themselves.

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.

Future Perspectives
In this study, we developed ProkAtlas for the habitat-based analysis of prokaryotic community structure datasets. The application of ProkAtlas to six datasets succeeded in intuitively highlighting environment-prokaryotic community structure relationships, delineating temporal succession patterns of prokaryotic communities and providing insights into environmental and ecological surveillance such as pollution sources. Notably, ProkAtlas is independent of biases toward cultured prokaryotes, which is a common problem in many tools in microbiology. In addition, as a comprehensive database, ProkAtlas is readily applicable to many prokaryotic community structure datasets without the need for additional experiments or cumbersome data processing and might also be applicable for discovering the general principles behind microbial evolution and migration. By complementing existing approaches, this habitat-based analysis will
A potential flaw of ProkAtlas is that it depends on environmental classifications defined by INSDC DRA/ERA/SRA. Although the use of this classification system enabled us to cover all the numerous metagenomic samples and enhanced the comprehensiveness of ProkAtlas, it is neither strictly defined nor hierarchically organized, and contains categories of different granularity. For example, metagenomic sequences from human gut samples may be labeled either as human_gut or human, and this choice is ultimately up to the data submitter. Similarly, sequences from rice field soils may be labeled either as soil or rice_paddy. Nevertheless, we did not adopt previously developed environmental ontologies such as MIGS/MIMS and EnvO in the present study because INSDC contains many samples lacking these systematic annotations. Here, for the sake of readers, we present an example of classification of environmental categories by referring to EMP Ontology and a supervised machine-learning of abstract texts attached to each project in DRA/ERA/SRA (Figure S7). Although such systematic annotations would restructure ProkAtlas in the future, currently it is left to ProkAtlas users to select and merge categories depending on applications and interest.

**Limitations of the Study**

A potential flaw of ProkAtlas is that it depends on environmental classifications defined by INSDC DRA/ERA/SRA. Although the use of this classification system enabled us to cover all the numerous metagenomic samples and enhanced the comprehensiveness of ProkAtlas, it is neither strictly defined nor hierarchically organized, and contains categories of different granularity. For example, metagenomic sequences from human gut samples may be labeled either as human_gut or human, and this choice is ultimately up to the data submitter. Similarly, sequences from rice field soils may be labeled either as soil or rice_paddy. Nevertheless, we did not adopt previously developed environmental ontologies such as MIGS/MIMS and EnvO in the present study because INSDC contains many samples lacking these systematic annotations. Here, for the sake of readers, we present an example of classification of environmental categories by referring to EMP Ontology and a supervised machine-learning of abstract texts attached to each project in DRA/ERA/SRA (Figure S7). Although such systematic annotations would restructure ProkAtlas in the future, currently it is left to ProkAtlas users to select and merge categories depending on applications and interest.

**Resource Availability**

**Lead Contact**

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Wataru Iwasaki (iwasaki@bs.s.u-tokyo.ac.jp).

**Materials Availability**

This study did not generate new unique reagents.
Data and Code Availability
The ProkAtlas database, pipeline, and the source code for calculating habitat preference scores are available at https://msk33.github.io/prokatlas.html.

METHODS
All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION
Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101624.

ACKNOWLEDGMENTS
This study was financially supported by JSPS KAKENHI (Grant Numbers 19J14142, 19H05688, 18H04136, and 16H06279). The authors thank Ching-chia Yang (National Agriculture and Food Research Organization) for supporting the study. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

AUTHOR CONTRIBUTIONS
Conceptualization, W.I.; Methodology, K.M. and W.I.; Software, K.M.; Validation, K.M.; Formal Analysis, K.M.; Investigation, K.M.; Resources, W.I.; Data Curation, K.M.; Writing – Original Draft, K.M.; Writing – Review & Editing, W.I.; Visualization, K.M.; Supervision, W.I.; Project Administration, W.I.; Funding Acquisition, K.M. and W.I.

DECLARATION OF INTERESTS
The authors declare no competing interests.

Received: June 29, 2020
Revised: August 3, 2020
Accepted: September 24, 2020
Published: October 23, 2020

REFERENCES
Backhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17, 690–703.
Barberán, A., Caceres Velazquez, H., Jones, S., and Fierer, N. (2017). Hiding in plain sight: mining bacterial species records for phenotypic trait information. mSphere 2, e00237–17.
Barberán, A., Ramirez, K.S., Leff, J.W., Bradford, M.A., Wall, D.H., and Fierer, N. (2014). Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. Ecol. Lett. 17, 794–802.
Carter, K.M., Lu, M., Luo, Q., Jiang, H., and An, L. (2000). Microbial community dissimilarity for source tracking with application in forensic studies. PLoS One 15, e0236082.
Castle, S.C., Sullivan, B.W., Knelman, J., Hood, E., Nemergut, D.R., Schmidt, S.K., and Cleveland, C.C. (2017). Nutrient limitation of soil microbial activity during the earliest stages of ecosystem development. Oecologia 185, 513–524.
Choudoir, M.J., Barberán, A., Menninger, H.L., Dunn, R.R., and Fierer, N. (2018). Variation in range size and dispersal capabilities of microbial taxa. Ecology 99, 322–334.
Delgado-Baquerizo, M., Bardgett, R.D., Vitousek, P.M., Maestre, F.T., Williams, M.A., Eldridge, D.J., Lambers, H., Neuhauer, S., Gallardo, A., Garcia-Velazquez, L., et al. (2019). Changes in belowground biodiversity during ecosystem development. Proc. Natl. Acad. Sci. 116, 6891–6896.
Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-gonzález, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., and Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. Science 352, 320–325.
Delgado-Baquerizo, M., Reich, P.B., Khachane, A.N., Campbell, C.D., Thomas, N., Freitag, T.E., Abu Al-Soud, W., Sierens, S., Bardgett, R.D., and Singh, B.K. (2017). It is elemental: soil nutrient stoichiometry drives bacterial diversity. Environ. Microbiol. 19, 1176–1188.
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., et al. (2013). Changes in assembly processes in soil bacterial communities following a wildfire disturbance. ISME J. 7, 1102–1111.
Gilbert, J.A., Jansson, J.K., and Knight, R. (2018). Earth Microbiome Project and global systems biology. mSystems 3, e00217–17.
Glöckner, F.O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Brun, G., Yarza, P., Peplies, J., Westram, R., et al. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. J. Biotechnol. 261, 169–176.
Goto, D.K., and 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.
Grossart, H.P., Ozias, C., Leuen, F., and Tang, K.W. (2010). Bacteria dispersal by hitchhiking on zooplankton. Proc. Natl. Acad. Sci. 107, 11959–11964.
Haider, M.N., Nishimura, M., Iijichi, M., Yang, C.C., Iwasaki, W., and Kogure, K. (2018). Habitability analyses of aquatic bacteria. J. Oceanogr. 74, 197–207.
Ji, M., Kang, W., Yue, L., Wang, J., Deng, Y., and Zhu, L. (2019). Salinity reduces bacterial diversity, but increases network complexity in Tibetan Plateau lakes. FEMS Microbiol. Ecol. 95, fiz190.
Jiang, Y., Lei, Y., Yang, Y., Korapelainen, H., Ninemets, U., and 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.

Keams, P.J., and 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.

Kes, M., Kisand, V., Wang, M., Caffaro-Filho, R.A., Moravcik, P., Harwood, V.J., Yoneyama, B., and Fujjoka, 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.

Klindworth, A., Peplies, J., Pruesse, E., Schweer, M., Richter, A., Zavarzin, G.A., Stenzel, U., Crameri, F., Kubal, M., Westh, H., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41, e1.

Knights, D., Kuczynski, J., Carlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., and Kelley, S.T. (2011). Bayesian community-wide culture-independent microbial source tracking. Nat. Methods 8, 761–765.

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

Louca, S., Jacques, S.M.S., Pires, A.P.F., Leal, J.S., Srivastava, D.S., Parfrey, L.W., Farjalla, V.F., and Doebeli, M. (2016). High taxonomic variability despite stable functional structure across microbial communities. Nat. Ecol. Evol. 1, 0015.

Louziane, C.A., and Knight, R. (2007). Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U. S. A 104, 11436–11440.

Lynch, M.D.J., and Neufeld, J.D. (2015). Ecology and exploration of the rare biosphere. Nat. Rev. Microbiol. 13, 217–229.

Malk, A.A., Martiny, J.B.H., Brodie, E.L., Martin, A.C., Treseder, K.K., and Allison, S.D. (2020). Defining trait-based microbial strategies with consequencess for soil carbon cycling under climate change. ISME J. 14, 1–9.

Mappelli, F., Marasco, R., Fusi, M., Scaglia, B., Tisani, G., Rolli, E., Fodelianakis, S., Bourgias, K., Ventura, S., Tambone, F., et al. (2018). The stage of soil development modulates rhizosphere effect along a High Arctic desert chronosequence. ISME J. 12, 1188–1198.

Martiny, J.B.H., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015). Microbiomes in light of traits: a phylogenetic perspective. Science 350, 249.

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

Morton, J.T., Sanders, J., Quinn, R.A., McDonald, D., Gonzalez, A., Vázquez-baeza, Y., Navas-molina, J.A., Song, S., Metcalf, J.L., Hyde, E.R., et al. (2017). Balance trees reveal microbial niche differentiation. mSystems 2, e00162–16.

Nemergut, D.R., Knelman, J.E., Ferrenberg, S., Bilinski, T., Melbourne, B., Jiang, L., Violle, C., Darcy, J.L., Prest, T., Schmidt, S.K., et al. (2010). Decreases in average bacterial community RNA operon copy number during succession. ISME J. 10, 1147–1156.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O’Neill, S.P., Bilinski, T.M., Stanish, L.F., Knelman, J.E., Darcy, J.L., Lynch, R.C., Wickey, P., et al. (2013). Patterns and processes of microbial community assembly. Microbiol. Mol. Biol. Rev. 77, 342–356.

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., et al. (2018). Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat. Microbiol. 3, 189–196.

Rath, K.M., Fierer, N., Murphy, D.V., and Roux, J. (2019). Linking bacterial community composition to soil salinity along environmental gradients. ISME J. 13, 836–846.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., and Gies, E.A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.

Schloss, P.D., Girard, R.A., Martin, T., Edwards, J., and Thrash, J.C. (2016). Status of the archaeal and bacterial census: an update. mBio 7, e00201–e00216.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored ‘rare biosphere’. Proc. Natl. Acad. Sci. 103, 12115–12120.

Srivasi, S., Yang, C.C., and Iwaski, W. (2017). Generalist species drive microbial dispersion and evolution. Nat. Commun. 8, 1162.

Stein, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., and Cameron Thrash, J. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. ISME J. 13, 3126–3130.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kulmina, J.R., Labadie, K., Salazar, G., Dhanshiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science 348, 1261359.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Loecey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., et al. (2017). A communal catalogue reveals Earth’s multiscale microbial diversity. Nature 551, 457–463.

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

Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., Fierer, N., and David, L.A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. Peer J. 5, e2969.

Yang, C.C., and Iwaski, W. (2014). MetaMetaDB: a database and analytic system for investigating microbial habitability. PLoS One 9, e87126.

Yassour, M., Vatanen, T., Siljander, H., Hamalainen, A.-M., Harkonen, T., Hyhani, S.J., Franzosa, E.A., Vilamakis, H., Huttenhower, C., Gevers, D., et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci. Transl. Med. 8, 343ra81.

Zhao, S., Liu, J., Banerjee, S., Zhou, N., Zhao, Z., Zhang, K., Hu, M., and Tian, C. (2020). Biogeographical distribution of bacterial communities in saline agricultural soil. Geoderma 361, 114095.
Supplemental Information

Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures

Kazumori Mise and Wataru Iwasaki
Figure S1. Schematic representation of alignment criteria of 16S rRNA gene sequences in the ProkAtlas pipeline, Related to the Transparent Methods

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 S2. A schematic illustration of procedures to calculate habitat preference scores for prokaryotic communities, Related to the Transparent Methods

(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 $E$. Overall habitat preference of a community is denoted as the average of $E$ 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 $e$. Therein, $e$ is converted to the environmental vector $e$ denoting the habitat preference of the OTU.
Figure S3. Violin plots indicating the effect of alignment length threshold on ProkAtlas-estimated prokaryotic habitat preference scores, Related to Figure 3 and the Transparent Methods.

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 3.
Reference database used for habitat preference analysis

| Released database | Alternative databases |
|-------------------|-----------------------|
| (A) | set A | set B | set C | set D | set E |
| Soil-derived isolates | Habitat preference scores of soil-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of plant-related categories (%) | Habitat preference scores of feces-related categories (%) | Habitat preference scores of brine-related categories (%) |
| Other isolates | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (B) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of plant-related categories (%) | Habitat preference scores of feces-related categories (%) | Habitat preference scores of brine-related categories (%) |
| Seawater-derived isolates | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other isolates | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (C) | [Graph] | Habitat preference scores of plant-related categories (%) | Habitat preference scores of feces-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| Plant-derived isolates | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other isolates | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (D) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| Human feces-derived isolates | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other isolates | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (E) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from seawater | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (F) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from groundwater | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (G) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from brackish lake | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (H) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from hydrothermal vent | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (I) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from sediment | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |
| (J) | [Graph] | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) | Habitat preference scores of brine-related categories (%) |
| SAGs from bioreactor | Habitat preference scores of freshwater-related categories (%) | [Graph] | [Graph] | [Graph] | [Graph] |
| Other SAGs | [Graph] | [Graph] | [Graph] | [Graph] | [Graph] |

Figure S4. Violin plots indicating the effect of random sampling in constructing ProkAtlas on ProkAtlas-estimated prokaryotic habitat preference scores, Related to Figure 3 and the Transparent Methods

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 3.
Reference database used for habitat preference analysis

![Graphs showing 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, Related to Figure 4 and the Transparent Methods. Details on each column and each panel are as explained in Figures S4 and 4, respectively.](image-url)

Figure S5.
Figure S6. Class-level taxonomic structures of prokaryotic communities re-analyzed in this study, Related to Figures 5–9

(A) Class-level taxonomic structures of the 31 plots ordered by soil salinity concentration (higher on the right than on the left). Related to Figure 5. (B) Class-level taxonomic structures of the 76 samples ordered by salinity (higher on the right than on the left). Related to Figure 6. (C) Class-level taxonomic structures of the 654 samples ordered by sampling ages (older on the right). Related to Figure 7. (D) Class-level taxonomic structures in soil samples taken from Midtre Lovénbreen glacier moraine. Related to Figure 8. (E) Class-level taxonomic structures in soil samples taken from Hailuogou Glacier Chronosequence. Related to Figure 8. (F) Class-level taxonomic structures at nine sampling points (more downstream on the right than on the left). Related to Figure 9.
Figure S7. An example of classification of environmental categories, Related to Table 1

(A) Manual classification of host-associated environmental categories based on the Earth Microbiome Project Ontology (EMPO). (B) Classification of host-free environmental categories using random forest modeling of word use frequency in abstract texts attached to each DRA/ERA/SRA project. The modeling was performed using randomForest package implemented in R 3.6.2. The environmental categories were classified using a random forest classifier trained with abstracts from seven representative environments: activated_sludge, air, freshwater, seawater, sediment, soil, and subsurface (default parameters were used, except for the sampling size set to a maximum of 632 for each group). When none of the seven groups were assigned to >50% of the projects within an environmental category, that environmental category was assumed to be highly heterogenous (i.e. consisting of essentially different types of samples) and therefore left unclassified. For example, abstract texts of beach_sand samples were similar to both soil and seawater ones and therefore this category was regarded unclassifiable.
Transparent Methods

**ProkAtlas database construction: Data source**

We obtained all metagenomic sequence entries in DRA/ERA/SRA with 115 environmental categories (NCBI taxon IDs) under 410657, ecological metagenomes, and 410656, organismal metagenomes, on July 4, 2018. These entries contained metagenomic data from diverse environments and were based on different sequencing platforms and library construction strategies. We 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). We 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 (listed in Table S2). To avoid datasets that were too biased towards data from specific projects with high number of samples (Ramirez et al., 2018) and to keep the database size small, up to ten runs were randomly selected from each project. 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 case in which a gzipped fastq file exceeded the data size of 200 MB, the first 200 MB was retrieved. The final dataset list is provided in Table S3.

**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 SSU 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. Because the number of rRNA gene sequences varied between projects (between 1 and 43,259 per project), we randomly sampled up to 100 sequences from each project. This prevented the datasets from being biased towards data from specific big projects, while keeping the database size small. In total, we compiled 361,474 rRNA gene sequences, each retaining environmental category information (Tables 1 and S1) 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, we 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).

ProkAtlas pipeline

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 S2. 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 S1A–D); however, hits harboring mismatches longer than 2 bp at either end of the alignment (Figure S1EF) are ignored because they may be erroneous hits. The hits are further filtered to satisfy sequence similarity and alignment length criteria (default: sequence similarity of 97% or more, alignment length of 150 bp or longer).

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 S2); 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:

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

where $n^i_k$ 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 read count of OTU $i$ (i.e. the
number of reads assigned to OTU $i$, $\mathbb{E}$ is the average of environmental vectors weighted by the read count of each OTU (i.e. $C_i$), $R_{\text{tot}}$ 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.

Bird’s-eye visualization of prokaryote cooccurrence network

Because we 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).

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

We 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, we 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 we used three different alignment length thresholds, namely 150 (default 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, we 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.

Habitat-based analyses of prokaryotic community structure datasets

ProkAtlas was applied to datasets of EMP, saline-affected agricultural soil and lake water samples, human infant gut microbiome samples, glacier chronosequence soil samples, and potentially polluted
river-water samples. Notably, these were 16S rRNA gene amplicon-sequencing data and not included in ProkAtlas.

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 (ftp://ftp.microbio.me/emp/release1/otu_tables/deblur/emp_deblur_150bp.subset_2k.rare_5000.biom). sOTUs clustered by Deblur (Amir et al., 2017) were used. Regarding the agricultural soil (Zhao et al., 2020), saline and non-saline lake water (Ji et al., 2019), infant gut microbiome (Yassour et al., 2016), glacier chronosequence soil (Jiang et al., 2018; Mapelli et al., 2018), and potentially polluted river-water (Kirs et al., 2017), raw fastq sequence data were downloaded from public databases (Table 2). 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.
Supplemental References

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

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

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217.

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

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/.

Schmieder, R. and Edwards, R.A. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864.

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., et al. (2017). Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724. https://doi.org/10.1101/125724

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

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.