博士論文

Exploring microbial ecology using genomic information

(ゲノム情報を用いた微生物生態の探究)

Satoshi Hiraoka

平岡 聡史

A Ph.D. Thesis

博士論文

Submitted to

Department of Computational Biology and Medical Sciences,
Graduate School of Frontier Sciences, the University of Tokyo

2018

Table of contents

Chapter 1: General introduction

Data deluge in microbial ecology

Microbes play fundamental roles in various ecosystems, but the majority has not been well characterized. Bioinformatics, which aims to discover new biological concepts and laws based on large-scale data, is now expected to accelerate discovery in unexplored areas of the microbial universe. The data deluge has made bioinformatics indispensable in modern research; recent innovative technologies are producing huge amounts of data at an unprecedented pace. Observations are key to science; for example, optical and electron microscopy are important methods of observation combined with various staining methods. Among recent observational technologies, high-throughput DNA sequencing technologies have rapidly produced vast amounts of genetic information at low cost, making available thousands of microbial genomes. These genome sequences provide a comprehensive catalog of microbial genetic elements underlying diverse microbial physiology, and they also help to weave a massive tapestry of microbial evolutionary histories (1, 2).

In microbial ecology, research has been hindered by the fact that the majority of environmental microbes is unculturable. Many studies across diverse natural environments have found many microbial groups that have no axenic culture (3–6). To overcome this fundamental difficulty, culture-independent approaches, including DNA-hybridization (*e.g.*, microarray and fluorescent in situ hybridization), DNA-cloning, and PCR have been used to detect specific members and/or functional genes in microbial communities (7–18). Recently, high-throughput sequencing technologies have popularized shotgun metagenomic and (typically 16S ribosomal RNA (rRNA) gene) amplicon sequencing methods, which can identify members and/or functional genes at a greater scale and in more detail. Their use in diverse environments has revealed the presence of extremophiles (19–21), uncovered relationships between microbes and human diseases (22–30), and characterized the nutrition systems involved in symbiosis (31–33). Even more applications of these methods have occurred in agriculture (34), food science and pharmaceuticals (12), and forensics (35–38). Many large-scale metagenomic projects are now generating comprehensive microbial sequence collections for different environments (*e.g.*, human-associated (39, 40), soil (41, 42) , and ocean environments (43, 44)). Because microbial communities change as they interact with other organisms and as the environment

changes, time-series analyses have also become common (45–49).

To analyze microbial genomic, metagenomic, and amplicon sequence data, several bioinformatic tools have been developed and popularized. Web servers, such as RAST (50), MG-RAST (51), IMG/M (52), EBI Metagenomics (53), and SILVAngs (54), and pipelines, such as MEGAN (55), QIIME (56), and Mothur (57), now allow researchers to perform integrated genomic and metagenomic analyses and visualize results without command-line operations or strong computational knowledge (58–62). However, deeply exploring of huge datasets to understand the ecology and evolution of microbes in the environment is still challenging work (Fig. 1-1).

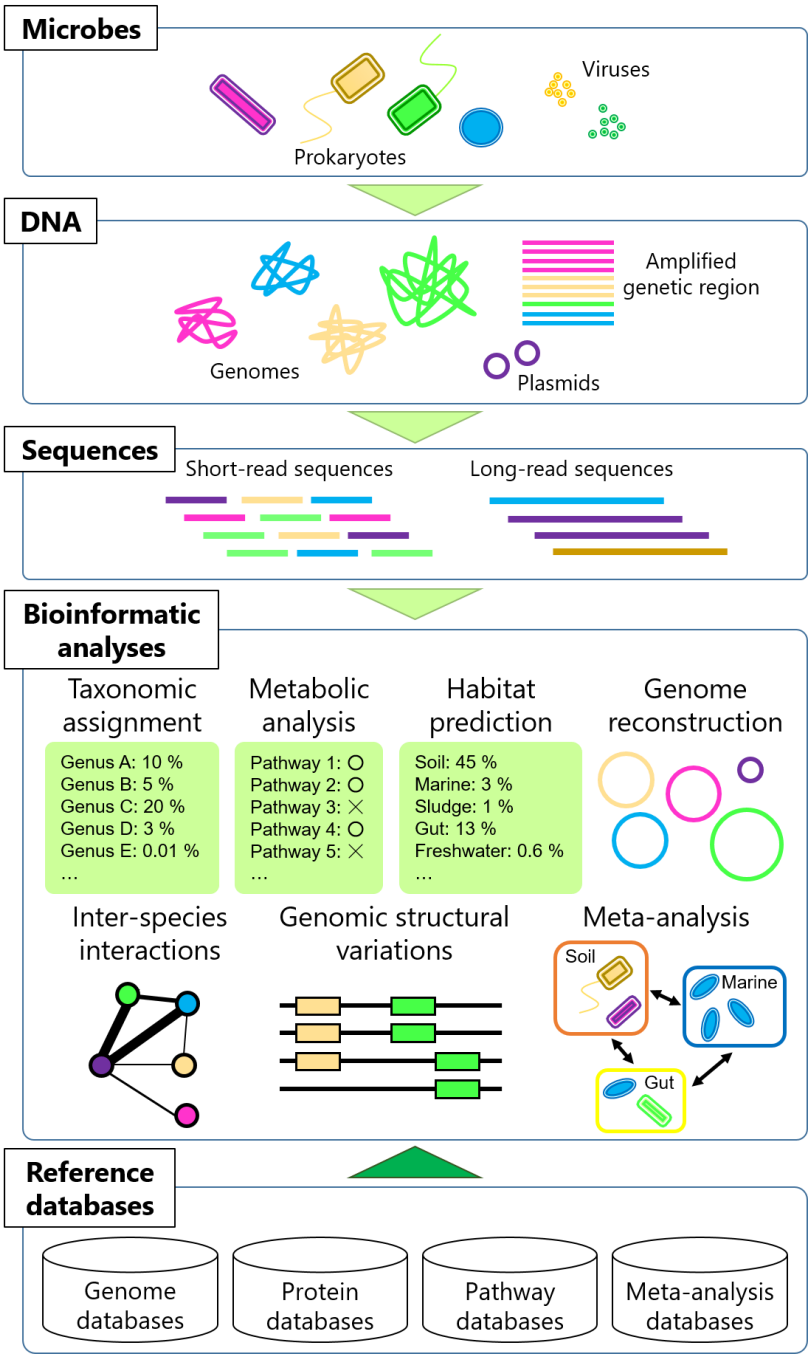

Figure 1-1 | Schematic figure of metagenomic and bioinformatic analyses in microbial ecology. The illustration covers topics that are already popular, that need further development, and that will become important in future. At the bottom of the illustration, reference databases lay foundations for various bioinformatic analyses.

Taxonomic assignment using genomic and genetic information

A fundamental step in microbial ecology is to describe the taxonomic distribution of microbial community members. Thus, precise taxonomic assignment of sequencing reads is one of the most important issues in the analysis of metagenomic and amplicon sequencing data. Reference-based methods are frequently used for this purpose, in which taxonomic assignment is based on straightforward sequence similarity searches against reference genomes (*e.g.*, RefSeq (63)) or 16S rRNA sequence databases (*e.g.*, Greengenes (64), SILVA (54), RDP (65), and Ez-Taxon (66)). These databases usually contain sequences of previously isolated and taxonomically classified strains, whereas they also contain environmental clone sequences. Many bioinformatic tools, such as TANGO (67), MetaPhlAn (68), and Kraken (69), have been developed to improve the computational efficiency, accuracy, and sensitivity of the taxonomic assignment. Recently proposed protein-sequence based taxonomic classification methods, such as Kaiju (70), are another approaches to increase sensitivity for overcoming evolutionary divergence. Although these tools perform well for many applications, discrimination between closely related species is sometimes difficult, especially in cases of highly conserved genes (*e.g.*, 16S rRNA genes). Additionally, genes that undergo horizontal gene transfer (HGT) between different taxa can cause incorrect taxonomic assignments. A more fundamental problem is taxonomic bias in the reference databases, which leads to biased taxonomic assignments. Indeed, it was shown that taxonomic assignments change greatly when different versions of reference databases are used (71). Therefore, even in this era of data deluge, further taxonomic enrichment of reference databases is a key to the improvement of reference-based methods. It may be noted that this issue would be more important in the analysis of fungal and viral sequences because less reference sequences are available and their taxonomy is under debate. To overcome this obstacle, several projects now aim to obtain a number of genomic sequences to enrich databases (72, 73). In case that amplicon sequencing data are analyzed, filtering of chimeric sequences formed during PCR is very important for accurate analysis (74). Several bioinformatic tools, such as AmpliconNoise (75), ChimeraSlayer (74), and UCHIME (76) have been proposed and commonly used to remove chimeric sequences.

As an alternative to the reference-based methods, reference-free methods can be used (*e.g.*, CD-HIT (77), UCLUST (78), and UPARSE (79)). These methods use clustering to group marker genes, such as 16S rRNA, ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), ammonia monooxygenase (*amoA*), sulfate thioesterase/thiohydrolase (*soxB*), and methyl-coenzyme M reductase genes (*mcrA*), into unique representative

sequences that can serve as operational taxonomic units (OTUs) (21). Whereas 16S rRNA genes are used to study general composition of a microbial community, RuBisCO, *amoA*, *soxB*, and *mcrA* genes are usually used to study microbes that play critical roles in carbon, nitrogen, sulfur, and methane cycles, respectively. In addition to the traditional genes, useful marker genes can be found and used for profiling metagenomic datasets (80). In the reference-free methods, OTUs often cannot be assigned to known taxa. To estimate phylogenetic information for those OTUs, PhylOTU (81), pplacer (82), and PhyloSift (83) couple the reference-free methods with phylogenetic analysis.

Cultivation-free reconstruction of genomic sequences

Currently, most metagenomic studies focus at the level of individual genes ("gene-centric" metagenomics (84)). In contrast, in some pioneering works of "genome-centric" metagenomics, microbial genomes that include those of important uncultivated taxonomic groups were successfully reconstructed by metagenomic binning and assembly from various environments including the ocean, groundwater, soil, hypersaline lake, and acid mine drainage (85–90). Although amplification bias still poses a non-negligible difficulty, single-cell genomic sequencing is expected to accelerate direct genome reconstruction from environmental samples (91–93), where combination of single cell genomic and metagenomic approaches would be a promising approach (94).

Metagenomic assembly is an important step for revealing ecology and physiology in environmental microbes, where fundamental concepts of metagenomic assembly from short-read sequences are well described already (58–62). Several tools have been developed for metagenomic assembly, and they are classified into reference-based (*e.g.*, AMOS (95)) and de novo methods (*e.g.*, MetaVelvet-SL (96), SPAdes (97), and IDBA-UD (98)). Especially in the case of de novo assembly, users need to be careful of chimeric contigs because similar genetic regions may be shared by different genomes (99–101). To improve the performance of de novo metagenomic assembly, composition-based methods use specific sequence features in a metagenomic dataset to split reads into different species. For example, CONCOCT (102), MetaBAT (103), and MaxBin (104) bin sequences based on their tetra-nucleotide frequency composition and coverages. These composition-based approaches are computationally intensive, especially in their memory usage. Thus, a fast-clustering approach using matrix decomposition with streaming singular value decomposition can be combined (105). On the other hand, sequence coverage information across different DNA extraction methods can also be used to effectively split sequences

into species, because numbers of sequence reads from the same genome should be similar regardless of the extraction method (86). A related approach bins co-abundant sequences across a series of metagenomic samples from similar environments (*e.g.*, human gut microbiome) to identify co-abundance gene groups (106).

Another information source that can improve the performance of metagenomic assembly is long-range contiguity. Recent development of methods for investigating long-range chromatin interactions (*e.g.*, Carbon-Copy Chromosome Conformation Capture (5C) (107) and Hi-C (108)) can also contribute to metagenomic assembly because these methods can ligate sequences from two different genomic regions that are in the same cell (109). The Irys system (BioNano Genomics, San Diego, USA), which also detects long-range contiguity with fluorescently labeled DNA, can be used for obtaining long contigs (110).

Estimation of community metabolism

Microbial genomes are affected by the environment during their evolution. In particular, metabolic processes encoded in the genome, from biosynthesis to biodegradation, directly link microbial communities to the environment. Because most microbes are unculturable, direct estimation of community-scale metabolic pathways is also targeted by metagenomic analysis. The most straightforward approach is to conduct sequence-similarity searches against pathway databases, such as KEGG (111), MetaCyc (112), and SEED (113), and use the results to annotate metabolic genes. Because with this naïve approach we usually detect many pathways whose component genes are only partially found in given metagenome data, MAPLE (114), MinPath (115), MetaNetSam (116), and HUMAnN (117) quantitatively or probabilistically evaluate whether those pathways likely function, enabling comparisons between samples. Again, significant biases in the databases of known pathways should be taken into consideration when interpreting the results of these methods. If shotgun metagenome data are unavailable, “virtual metagenomes” or functional gene abundances can be estimated using 16S rRNA amplicon sequencing data (118, 119). This approach takes advantage of the fact that closely related genomes tend to have similar gene content and, therefore, given a 16S rRNA sequence, the gene content of its host genome can be estimated (at least, to some extent) if a closely related genome is already sequenced. It may be noted that such estimation should become difficult when applied to microbial groups whose genomes are rarely available and that

genomic variations within closely related microbial groups cannot be precisely considered. Despite these difficulties, this approach is very cost-effective and more easily applicable to large-scale comparative analysis.

Community-level analysis of genomic structure variations and dynamics

Operon structures, which are unique to prokaryotic genomes, reflect the function of their encoded genes and should be associated with microbial ecological strategies. Thus, if we observe systematic variation in gene order (or gene cluster structures) due to gene losses, fusions, duplications, inversions, translocations, and HGTs from the analysis of metagenome data, these variations would provide important clues for linking microbial communities to the environment (Fig. 1-2). Whereas it is sometimes difficult to distinguish variations under selection pressure from those because of population changes, MaryGold (120) is a tool for visual inspection of such variations. Variations in gene order for genes of the tryptophan pathway were identified within contigs assembled from the Sargasso Sea metagenome (121). Because the availability of long sequences that encompass multiple genes greatly facilitates gene-order analysis, DNA cloning can also be used if the targeted pathways can be efficiently enriched by colony selection (122, 123). On a larger scale, gene order can be affected by genome replication mechanisms. Most prokaryotic genomes are circular with one replication origin; thus, genes close to the origin can physically exist in multiple copies, especially during an active growth phase. Thus, detection of such regions from metagenomic sequences can reveal growth dynamics of microbes in a community (124).

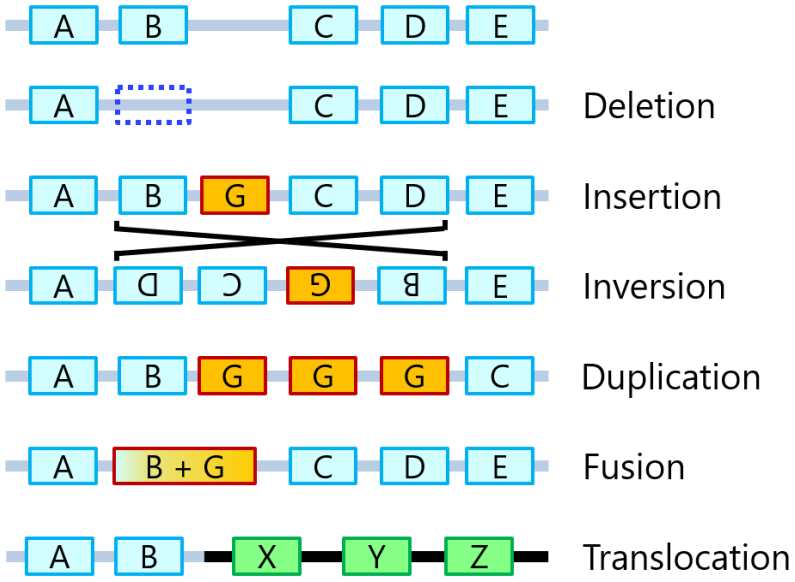

Figure 1-2 | Schematic figures of genomic structural variations in environmental microbes. Each box represents a protein-coding gene, where alphabet characters indicate homology. Boxes and thick lines in different brightness represent genes and genomic fragments, respectively, that originated from different genomic areas or genomes. Dashed lines represent lost genes or expressions.

Among various sources of genomic variations and dynamics, HGT is of particular interest because it can help microbes adapt quickly to different environments (125–127). Although many comparative genomic studies have analyzed HGT (128–130), the role of HGT in microbial communities remains less understood. A classical approach to detect HGT within communities is based on DNA cloning, which is particularly effective if colony selection can be applied to the targeted genes (16, 131). In metagenomics, assembled contigs can be used to comprehensively identify HGT events by analyzing phylogenetic incongruence and gene order differences (132–134). In addition, gene gains via plasmids are also important driving forces that accelerate microbial adaptation to their environment. In accordance with this hypothesis, plasmids are frequently found to contain genes that likely contribute to fitness, such as detoxification genes (135–137) and antibiotic resistance genes (138). Notably, plasmid-specific metagenomics, or "plasmidome" analysis, is now being conducted to directly investigate environmental plasmids without culturing or cloning (139, 140). For example, the bovine rumen plasmidome was reported to contain genes that may confer advantages to their hosts (141). Rat gut (142) and wastewater (143) plasmidomes have also been investigated.

In addition to genes, regulatory sequences in intergenic regions are related to the function of nearby genes. Thus, variations in the comprehensive set of regulatory sequences, or the regulome, for example by promoter propagation, can have important roles in microbial adaptation (144, 145). We envision direct investigations of regulomes in microbial communities, or "metaregulome" analysis, becoming an attractive research field as technical advancements occur in long-read metagenomics. By metaregulome analysis, variations and/or possible transfers of regulatory sequences, in addition to those of coding genes, can be identified from metagenomic datasets (145, 146).

Comprehensive analysis of inter-species interactions

Inter-species interactions, including mutualism and parasitism, are of general interest in microbial ecology (147). Using abundance information from large-scale metagenomic datasets, co-occurrences (or anti-occurrences) among microbes, hosts, and/or viruses have been studied and, for example, species interaction networks have been identified (148–157). Recent large-scale projects include the Tara Oceans project, which revealed interactions among all three domains and viruses (44). Because environmental samples were revealed to contain environmental DNA shed from large organisms in addition to microbial DNA (158), combinatorial analysis of microbial and environmental DNA is expected to accelerate analysis of interactions between microbes and larger organisms.

The viral metagenome is called the metavirome or simply the virome. Because viruses also play fundamental roles in ecosystems, virome analysis is becoming an important field. To date, viral communities in hypersaline (159, 160) and human gut (161) environments have been extensively studied, and antibiotic viruses have also been of interest (162). A novel bacteriophage present in the majority of published human fecal metagenomes was recently reconstructed (163), and phage-bacteria ecological networks were suggested to protect gut microflora from antibiotic stress (162). Because viruses are classified into different types of DNA and RNA viruses, different approaches must be combined for comprehensive analyses (164). Use of targeted sequence capture techniques to efficiently increase the proportion of viral reads in metagenomic samples may also be considered (165). The largest limitation of bioinformatic analyses of viromes is insufficient reference genome data. Similarity searches using viral sequences often result in no significant hits, suggesting that there are many unknown viruses. To overcome this limitation, several bioinformatic tools have been developed and used for virome studies, such

as ViromeScan (166) for taxonomy assignment and Metavir 2 (167) for viral genome reconstruction. Another difficulty lies in the fact that, in contrast to prokaryotes that have universal marker genes for phylogenetic analysis (*i.e.*, 16S rRNA), there is no such gene for viral studies. Analysis of clustered regularly interspaced short palindromic repeats (CRISPRs) is a related emerging field because these repeats represent previous exposures to (or attacks from) viruses (168–170). As CRISPRs are found in approximately 50% of bacteria and approximately 90% of archaea (171, 172), metagenomic analysis of CRISPRs will help move the field toward comprehensive analysis of viral-microbial interactions.

Meta-analysis of metagenomes

Currently, abundant metagenomic datasets containing dozens of terabytes of sequence data can be found in the Short Read Archive (SRA) database at NCBI, and the amount of the content is increasing daily (173). Whereas each metagenomic dataset provides a snapshot of the microbial community at the time of sampling, comprehensive analysis (or meta-analysis) of many datasets is expected to reveal general patterns or laws that determine how microbes interact with their environments and how their genomes have been shaped. Of course, it should be noted that different datasets have been constructed with different experimental methods and conditions.

Regarding global correlations between environments and microbial genomes, correlations involving genomic GC content (174) and genome size (175) have been reported. For the meta-analysis of different environments inhabited by a microbe and the factors that contribute to adaptation, MetaMetaDB (176) was developed. This database can be used to predict all possible habitats of microbes by searching for the presence of microbes in metagenomic and 16S rRNA amplicon sequencing datasets derived from diverse environments. Given a metagenomic or 16S rRNA amplicon sequencing dataset, researchers can find environments whose microbial community structures are similar to that dataset using MetaMetaDB (176). Meta-analysis of metagenomic datasets was also performed to examine microbial adaptation to environments in terms of metabolic flexibility (177, 178) and to examine specific functional genes that facilitate adaptation to extreme habitats, such as heavy-metal resistance genes (179, 180) and salt stress-responsive genes (181). Through meta-analysis, associations were found between membrane protein variations and oceanographic variables in the global ocean sampling expedition (182). Microbial interactions between humans and the indoor environment have also been

investigated (183).

Metagenomics and metaepigenomics with long-read sequencers

Currently, sequencers that can produce long-read data are being developed, such as the single molecule real-time (SMRT) Sequencing platforms implemented in PacBio Sequencing platforms (Pacific Biosciences, Menlo Park, USA) and nanopore-based sequencers (Oxford Nanopore Technologies, Oxford, UK). Long reads are already contributing to many types of bioinformatic analyses, including high-quality de novo assembly of bacterial and viral genomes (184, 185) and detection of genomic structural variations, such as large-scale insertions/deletions or HGTs in microbial communities (186). Long reads are expected to be helpful for reconstructing genomes from metagenome data, directly observing genomic structural variation, and analyzing metaregulomes in various microbial communities. High-density microbial habitats, such as biofilms and gut communities, would be interesting targets because their genomic structures may be changed by the frequent exchange of genetic materials.

With SMRT Sequencing, DNA chemical modifications can be captured directly and simultaneously with genome sequencing. The most major types of DNA chemical modification in prokaryotes is nucleotide base methylation including N6-methyladenine (m6A), 5-methylcytosine (m5C), and N4-methylcytosine (m4C). The DNA methylation implicated in the variety of biological processes such as defence system from phage infection and regulations of gene expression (187). Although the methylation patterns and biological significance have been well studied in culturable microbes using the SMRT sequencing technology (188, 189), little has been examined in unculturable members that overwhelmed majority in environments.

Long-read metagenomics is an emerging field, but there are still limitations to be considered. Although the PacBio system can generate reads with an average length of approximately 15 kb, less than 50,000 reads are generated per SMRT cell (*i.e.*, less than 1 Gb per SMRT cell) when using PacBio RS II system. This throughput is much lower than that of so-called massively parallel sequencers (*e.g.*, approximately 15 Gb per run of MiSeq (Illumina)) and can be insufficient for describing taxonomically diverse microbial communities. In addition, the low accuracy of PacBio reads (approximately 85%) can hinder bioinformatic analysis, unless highly redundant sequencing (*e.g.*, more than 50×

coverage) is performed to reach high accuracy in the ensemble. Along with the development of new bioinformatic methods, protocols also need to be optimized to avoid DNA fragmentation during extraction (190, 191).

Outline of this thesis

Although bioinformatics for microbial genomics and metagenomics have flourished in recent years, many regions of microbial ecology have been not understood. In this thesis, I especially grappled with the three research tasks. In Chapter 2, I investigated microbial characteristics of soil disturbed by the 2011 Great East Japan Earthquake tsunami to examine microbial changes through environmental disorder. In Chapter 3, I performed 16S rRNA amplification sequencing analysis against microbes in precipitation. In Chapter 4, I examined genomic and epigenomic characteristics of microbial community in Lake Biwa, Japan, using culture-independent metagenomic SMRT sequencing technology. In Chapter 5, conclusions of this thesis are presented with discussion and future work. A part of this thesis is based on the following publications written by the author and others (184, 192, 193).

Chapter 2: Genomic and metagenomic analysis of microbes in a soil environment affected by the 2011 Great East Japan Earthquake tsunami

Introduction

On March 11, 2011, the Great East Japan Earthquake occurred off the coast of Tohoku, Japan. The earthquake triggered large tsunami waves, which flooded broad areas of land along the Pacific coast and changed the soil environment due to seawater and sludge that originated from marine sediments (194). Previous studies showed that following the Indian Ocean tsunami of December 26, 2004, the tsunami-affected areas maintained high-salinity conditions for over eight months (195), and there were also changes in several chemical characteristics, including an increase in organic matter content (196), increase in nitrate and phosphate content (197), increase in heavy-metal ion concentrations (198–200), decrease in pH, and increase in electrical conductivity (201). Increases in salinity and organic matter were also reported at a number of places along the pacific coast following the Tohoku tsunami (194).

Such changes in the soil environment after the tsunami are also likely to have an impact on the ecosystem. There have been many studies conducted to date investigating how such changes affect plants; for example, vegetation senescence was reported after the Indian Ocean tsunami (200, 202, 203) and flora variations on sandy beaches were observed after the Tohoku tsunami (204). On the other hand, only a few studies have evaluated the effects of a tsunami on microbes. Somboonna *et al*. applied 16S ribosomal RNA (rRNA) amplicon sequencing to the soil affected by the Indian Ocean tsunami and observed changes in the microbial population structure (205). Wada *et al*. also used 16S rRNA amplicon sequencing to analyze samples of the sludge brought ashore by the Tohoku tsunami and identified several pathogenic and sulfate-reducing bacterial groups (206). However, no study has yet investigated the microbial characteristics of tsunami-affected soil at the genomic level.

In this study, I evaluated the microbial characteristics of a soil environment affected by the Tohoku tsunami, using whole-genome and shotgun metagenome sequencing approaches. Notably, whole-genome sequencing of four *Arthrobacter* strains isolated

from the tsunami-affected soil sample revealed that siderophore-synthesis genes were independently lost in each genome. Siderophores are compounds that function in iron absorption (207–209), and these gene losses were consistent with the results of soil chemical analysis and culture experiments under iron-controlled conditions. Furthermore, metagenomic analyses indicated over-representation of denitrification-related genes in the tsunami-affected soil sample, as well as the existence of pathogenic and marine-living genera and genes related to salt-tolerance.

Materials and methods

Sample collection

Soil samplings were conducted at Hiyoriyama (38°15’20”N, 141°0’42”E) and Amamiya (38°16’35”N, 140°52’16”E) in Sendai city, Miyagi, Japan, in July 2012 (Fig. 2-1). If needed, the owners of the lands gave permission to conduct the study on these sites. I and my collaborators confirm that the study did not involve endangered or protected species. The Hiyoriyama site is 0.5 km off the coastline and was affected by the tsunami, whereas the Amamiya site is 12 km off the coastline and was not affected; the two sites are 13 km apart. The surface soil was removed to a 5 cm depth before sampling. Intermingled plants were carefully removed using tweezers, and soil that passed through a 2.0-mm pore-sized sieve was collected. The collected soil samples were transported to the laboratory at 4 °C and immediately stored at -80 °C until the subsequent analysis. The sampling was conducted by Dr. Shotaro Hirase (Fisheries Lab., Graduate School of Agricultural and Life Sciences, the University of Tokyo).

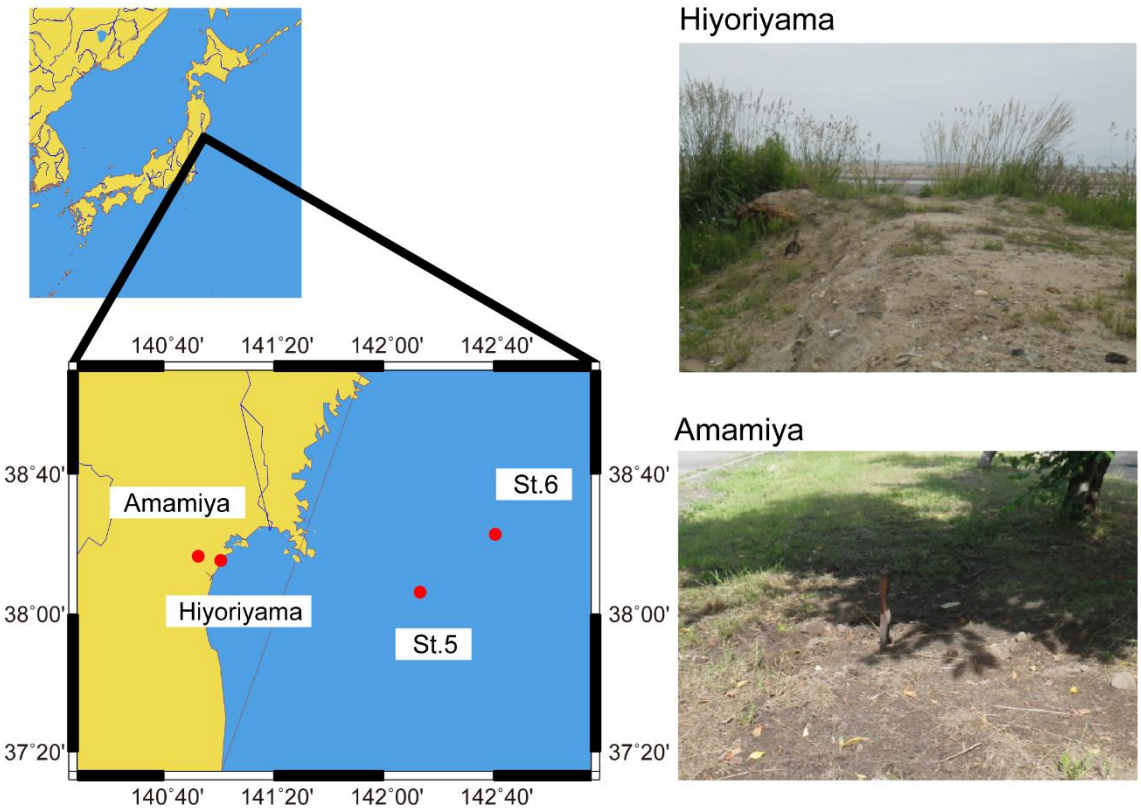

Figure 2-1 | A map and photos of the sampling sites in a coastal area of Sendai, Japan. The Tohoku tsunami reached Hiyoriyama, but not Amamiya.

Seawater sampling was conducted at St.5 (38°06'00"N, 142°15'00"E) and St.6 (38°22'59"N, 142°43'01"E) off the coast of Miyagi, Japan in the Pacific Ocean, in August 2012, during the KT-12-21 cruise of R/V Tansei-Maru (JURCAOS, JAMSTEC). The St.5 and St.6 stations are located 110 km and 150 km from Sendai city, respectively. Surface seawater was collected in a prewashed bucket and immediately spread onto agar plates on a research vessel. The sampling was conducted by the member of Tansei-Maru.

Isolation and 16S rRNA sequencing

R2A medium (Wako Pure Chemical Industries) was used to cultivate microbial strains that grow under general, low-nutrient condition, and ZoBell marine medium (Becton Dickinson and Company) was used to cultivate microbial strains that have adapted to a seawater-affected condition. Soil samples were thawed at 4 °C overnight, suspended in R2A or ZoBell liquid medium, and plated to the corresponding agar medium at a density of 10^{-4} g soil per plate with five replicates. The plates were incubated at 20 °C for 7 days before colony counting and picking. To obtain strains that could grow in both media,

single colonies on the R2A agar plates were transferred to ZoBell agar plates with sterilized sticks, incubated at 20 °C for 7 days, and isolated by spread-plating on ZoBell agar at 20 °C. The seawater samples were plated to R2A and ZoBell agar at a volume of 100 μL seawater per plate with three replicates. The plates were incubated at 20 °C for 7 days before colony counting and picking.

To sequence 16S rRNA genes, seven strains that were isolated from the Hiyoriyama site and could actively grow in both R2A and ZoBell media were randomly selected. After incubation in the ZoBell liquid medium, DNA was extracted using Wizard Genomic DNA Purification Kit (Promega). The 16S rRNA genes were amplified using a standard polymerase chain reaction protocol with the primers 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492R (5'-GGCTACCTTGTTACGACTT-3') (210), and sequenced by the Sanger method. The DNA experiments were conducted by the member of the Atmosphere and Ocean Research Institute, the University of Tokyo.

Whole-genome sequencing and analysis

Four strains of the *Arthrobacter* genus that were isolated from the Hiyoriyama site and were cultivable in both R2A and ZoBell media were targeted for whole-genome sequencing. Genomic DNA was extracted by the phenol-chloroform method. Two strains (named Hiyo4 and Hiyo8) were sequenced using PacBio RS II (Pacific Biosciences) according to the manufacturer protocols. *De novo* genome assembly of the 62,608 (Hiyo4) and 65,240 (Hiyo8) raw reads obtained from the Sprai pipeline (http://zombie.cb.k.u-tokyo.ac.jp/sprai/) successfully generated one and three circular contigs, respectively, after manual curation. The other two strains (Hiyo1 and Hiyo6) were sequenced using GS FLX+ System (Roche) and Ion PGM (Thermo Fisher Scientific) according to the manufacturer protocols. *De novo* genome assembly of the 301,881 (Hiyo1) and 267,295 (Hiyo6) reads obtained from the Newbler assembler (211) generated 38 and 630 scaffolds, respectively. All GS FLX+ and Ion PGM sequencing data used in this study were obtained at the Hattory lab (Graduate School of Frontier Science, the University of Tokyo).

Coding sequences (CDSs) were predicted by applying Prodigal (212) to the contig sequences. Functional annotation was performed by blastp searches (213) against the Swiss-Prot (214) and eggNOG v4.0 (215) databases with a cut-off e-value ≤1E-5. Transfer RNA (tRNA) and rRNA sequences were predicted using tRNAscan-SE (216) and RNAmmer (217), respectively, with default settings.

For comparative genome analysis, all 21 publicly available genome sequences

(6 complete and 15 draft sequences) of the *Arthrobacter* genus were downloaded from GenBank (218) via EzGenome (http://www.ezbiocloud.net/ezgenome) in January 2015 (Table 2-1). The CDSs of the four isolated and 21 downloaded genomes were subjected to blastp searches against the eggNOG database (215) with cut-off e-value ≤1E-5 and identity ≥90%.

Table 2-1 | *Arthrobacter* dataset for comparative genome analysis.

Species Name	Strain name	NCBI taxonomy ID	Isolated environment	Status	Type strain	Total genome size (bp)	Number of scaffolds	GC content (%)	CDSs	rRNA	tRNA
Arthrobacter arilaitensis	Re117	861360	Cheese	Complete	T	3,918,192	3	59.3	3,736	18	64
Arthrobacter aurescens	TC1	290340	Soil	Complete	F	5,226,648	3	62.4	4,819	18	54
Arthrobacter castelli	DSM 16402	1121019	Mural painting	Draft	T	4,582,606	52	63.6	4,453	4	46
Arthrobacter chlorophenolicus	A6	452863	Soil	Complete	T	4,980,870	3	66.0	4,645	15	85
Arthrobacter crystallopoietes	BAB-32	1246476	Soil	Draft	F	4,348,607	347	66.7	4,387	4	51
Arthrobacter gangotriensis	Lz1y	1276920	Soil	Draft	T	4,319,900	20	63.0	4,029	4	58
Arthrobacter globiformis	NBRC 12137	1077972	Soil	Draft	T	4,954,410	125	66.2	4,544	3	50
Arthrobacter phenanthrenivorans	Sphe3	930171	Soil	Complete	T	4,535,320	3	65.4	4,246	12	50
Arthrobacter sp.	131MFCol6.1	1157944	Rhizosphere	Draft	F	4,432,383	29	67.2	3,968	5	50
Arthrobacter sp.	135MFCol5.1	1158050	Rhizosphere	Draft	F	4,453,574	37	66.1	4,083	7	50
Arthrobacter sp.	161MFSha2.1	1151118	Rhizosphere	Draft	F	4,572,124	44	63.1	4,295	6	54
Arthrobacter sp.	162MFSha1.1	1151119	Rhizosphere	Draft	F	4,399,171	55	66.1	4,107	6	51
Arthrobacter sp.	35W	1132441	Lake	Draft	F	4,652,932	6	66.8	4,103	12	54
Arthrobacter sp.	AK-YN10	1349820	Soil	Draft	F	4,839,751	107	63.3	4,614	2	53
Arthrobacter sp.	CAL618	1055770	Human	Draft	F	3,654,388	288	63.2	3,642	18	47
Arthrobacter sp.	FB4	290399	Soil	Complete	F	5,070,478	4	65.4	4,624	15	51
Arthrobacter sp.	Hiyo1	1588020	(This study)	Draft	F	5,543,883	38	63.2	5,292	2	51
Arthrobacter sp.	Hiyo4	1588021	(This study)	Complete	F	3,790,568	1	65.0	5,120	12	50
Arthrobacter sp.	Hiyo6	1588022	(This study)	Draft	F	2,594,729	630	63.3	3,767	3	33
Arthrobacter sp.	Hiyo8	1588023	(This study)	Complete	F	4,698,617	3	63.8	7,041	15	53
Arthrobacter sp.	M2012083	1197706	Soil	Draft	F	4,629,172	67	62.0	4,304	3	54
Arthrobacter sp.	Rue61a	1118963	Wastewater	Complete	F	5,081,038	3	62.2	4,723	18	53
Arthrobacter sp.	SJCon	683150	Soil	Draft	F	4,389,620	142	66.2	4,635	3	50
Arthrobacter sp.	TB 23	494419	Sponge	Draft	F	3,542,308	126	63.3	3,405	15	46
Arthrobacter sp.	TB 26	494420	Human	Draft	F	4,324,615	556	66.4	4,451	19	50

For construction of a phylogenetic tree, the 16S rRNA sequences of 56 *Arthrobacter* type strains and *Streptomyces coelicoflavus* NBRC 15399^T were additionally downloaded from the RDP webserver (65). *Streptomyces coelicoflavus* NBRC 15399^T was used as an outgroup (219). The 16S rRNA sequences of the total 82 strains were subjected to multiple alignment using MUSCLE (220) with default settings. A maximum-likelihood (ML) tree was generated by MEGA 6 (221) with the K80 substitution model with a gamma distribution and invariant sites (K2+G+I), which was the AIC-selected model, and 1000 bootstrap replicates. An ML tree of the total 17 genome-available strains was constructed on the basis of the set of 400 conserved bacterial marker genes using PhyloPhlAn (222) and MEGA 6 (221) with the WAG substitution model that incorporates gamma distribution and the amino-acid frequencies of the dataset (WAG+G+F), which

was the AIC-selected model, and 1000 bootstrap replicates.

Culture assays of iron dependency

To determine the difference in iron tolerance among strains in relation to the genetic analysis results, culture assays were conducted at different iron concentrations. In addition to the four isolated *Arthrobacter* strains, I cultivated four closely related and genome-sequenced species, *A. aurescens* Phillips 1953^T (JCM 1330^T), *A. chlorophenolicus* A6^T (JCM 12360^T), *A. globiformis* Conn 1928^T (JCM 1332^T), and *A. phenanthrenivorans* Sphe3^T (JCM 16027^T). All four species had intact siderophore-synthesis genes in their genomes. These strains were provided by the Japan Collection of Microorganisms, BioResource Center, RIKEN and National BioResource Project of Ministry of Education, Culture, Sports, Science and Technology, Japan.

Iron-controlled, modified MM9 medium was prepared as follows. A solution containing 0.3 g/L KH_2PO_4, 0.5 g/L NaCl, 1.0 g/L NH_4Cl, 6.0 g/L NaOH, and 30.24 g/L PIPES was adjusted to pH 7.0 with NaOH. After autoclaving, separately sterilized solutions of 10 mL of 20 wt% glucose, 1 mL of 1 M $MgCl_2$, and 0.1 mL of 1 M $CaCl_2$ were added to 1 L of the solution (223). Then, the iron concentration was adjusted to 0, 0.1, 1, or 10 μM with a $FeCl_3$-containing solution that was prepared in the same manner.

Each strain was precultured until its optical density at 660 nm (OD_{660}) reached 0.1 in the iron-free modified MM9 liquid medium. Then, 100 μL of the suspension was inoculated to 50-mL tubes containing 10 mL of the iron-controlled, modified MM9 medium. Among the additional four strains, only *A. phenanthrenivorans* Sphe3 showed growth in the modified MM9 medium. The tubes were incubated at 30 °C on a linear shaker at 200 rpm for 3 days, and the OD_{660} was measured periodically during the incubation period. The growth curve was fitted to the logistic model to calculate the maximum growth rate.

Soil chemical analysis

The soil samples were subjected to chemical analysis for pH, electrical conductivity, and concentrations of total organic carbon, total nitrogen, nitrate, nitrite, ammonium, effective phosphate, exchangeable ions (K^+, Ca^{2+}, Mg^{2+}, Na^+, and Mn^{2+}), available iron (Fe), chloride ion (Cl^-), sulfate ion (SO_4^{2-}), eluted heavy metals (Cd, Cr (VI), total Hg, alkyl mercury, Pb, As), eluted boron (B), contained heavy metals (Cd, Cr (VI), Hg,

Pb, As, Cu, Zn, and Ni), and contained boron (B). The analysis was conducted by Createrra Inc. (Tokyo, Japan).

Shotgun metagenome sequencing and analysis

Metagenomic DNA was extracted using PowerSoil DNA Isolation Kit (MoBio Laboratories). Shotgun metagenome sequencing was performed using the GS FLX+ System according to the supplier's protocol. Duplicated reads were removed by CD-HIT-454 (224).

Taxonomic assignment was performed using Kraken (69) against complete prokaryotic genomes from RefSeq (63). CDS prediction was performed using MetaProdigal (225). CDSs less than 30 amino acids in length were excluded from further analysis. Functional annotations were based on blastp searches against the eggNOG (215) and Swiss-Prot (214) databases with a cut-off e-value ≤1E-5.

SortMeRNA (226) was applied to the shotgun metagenome data to extract 16S rRNA sequences. For each extracted 16S rRNA sequence, a blastn search was performed against MetaMetaDB (176) and the top hit sequences with e-value ≤1E-10 and identity ≥90% were retrieved. Microbial habitability index (MHI) scores were calculated as described previously (176).

Data deposition

The whole-genome and plasmid sequence data of Hiyo1, Hiyo4, Hiyo6, and Hiyo8 were deposited in the DDBJ/ENA/GenBank database under the BioSample ID SAMD00024042, SAMD00024043, SAMD00024044, and SAMD00024045, respectively. The shotgun metagenome sequence data of Hiyoriyama and Amamiya were deposited in the DDBJ/ENA/GenBank database under BioSample ID SAMD00023516 and SAMD00023517, respectively. All data were registered under BioProject ID PRJDB3373.

Results and discussion

Isolation of microbial strains

To investigate whether the microbial community at the Hiyoriyama (tsunami-affected) site contained more microbes that are adapted to seawater-affected conditions than that at the Amamiya (tsunami-unaffected) site, we conducted culture experiments

using R2A (general low-nutrient) and ZoBell (seawater-based) media. At Hiyoriyama, the mean (± standard deviation) numbers of colony forming unit (CFU) per gram of soil were $7.0 \pm 3.9 \times 10^5$ and $3.0 \pm 2.0 \times 10^5$ on R2A and ZoBell, respectively. At Amamiya, these numbers of CFU were $21.8 \pm 4.7 \times 10^5$ and $3.6 \pm 2.3 \times 10^5$. The ZoBell/R2A CFU ratios were 0.43 and 0.17 at Hiyoriyama and Amamiya, respectively, indicating that the Hiyoriyama site would be comparatively enriched with microbes adapted to a seawater-affected condition at 10 months after the tsunami. For comparison, surface seawater samples collected at St. 5 and St. 6 in the offshore were spread onto both agar plates. The numbers of CFU per milliliter of seawater were $12.7 \pm 13.3 \times 10^1$ and $81.7 \pm 43.6 \times 10^1$ on R2A and ZoBell, respectively. As expected, the ZoBell/R2A CFU ratio (6.4) was significantly higher at the offshore sites than at Amamiya and Hiyoriyama (p-value <0.05, t-test).

To isolate microbial strains that are potentially adapted to both types of environments from Hiyoriyama, the microbial colonies grown on R2A to ZoBell agar plates were aseptically transferred. Seven isolated colonies were randomly picked up and their 16S rRNA genes were sequenced. Unexpectedly, all of the strains were found to belong to a single genus, *Arthrobacter*. The genus *Arthrobacter* is an aerobic, gram-positive member of the family Micrococcaceae, Actinobacteria (227, 228). This genus is broadly found in soils, as well as in extreme environments, including the deep subsurface (229), arctic ice (230), radioactive sites (231), and heavy metal-contaminated sites (232). Some *Arthrobacter* species were reported to tolerate drastic environmental stresses, *e.g.*, desiccation (233), starvation (234), heavy metals (235, 236), and radioactivity (237). Furthermore, at the time of analysis, 6 complete and 15 draft genome sequences were available for comparative genome analysis. Because of these characteristics, the isolated *Arthrobacter* strains were targeted as a possible platform for exploring genomic features that may be related to microbial adaptation to drastically changed environments.

Whole-genome sequencing of the isolated *Arthrobacter* strains

The whole-genome sequences of four *Arthrobacter* sp. strains were determined (Table 2-2). Assembly using the reads from PacBio RS II showed the complete genomes of two strains: Hiyo4 with one circular chromosome (3.8 Mbp), and Hiyo8 with one circular chromosome (4.7 Mbp) and two plasmids (0.3 Mbp and 15 kbp) (Fig. 2-2). Assembly using reads from GS FLX+ and Ion PGM System produced 38 and 630 scaffolds for two strains, Hiyo1 and Hiyo6, respectively.

Table 2-2 | Whole-genome sequencing of the isolated *Arthrobacter* sp. strains

	Hiyo1	Hiyo6	Hiyo4	Hiyo8
Sequencing platform	GS FLX+ & Ion PGM	GS FLX+ & Ion PGM	PacBio RS Ⅱ	PacBio RS Ⅱ
Scaffolds	38	630	1	3
Contigs	1,685	2,450	1	3
Total genome size (bp)	5,543,883[a]	2,594,729[a]	3,779,248	4,698,617[a]
N50	4,950	2,656	-	-
Coverage	20×	24×	79×	42×
GC content (%)	63.2	63.3	65.0	63.8
CDSs	5,292	3,767	5,129	7,041
rRNAs	2	3	12	15
tRNAs	51	33	50	53

[a] Plasmid sequences were not excluded

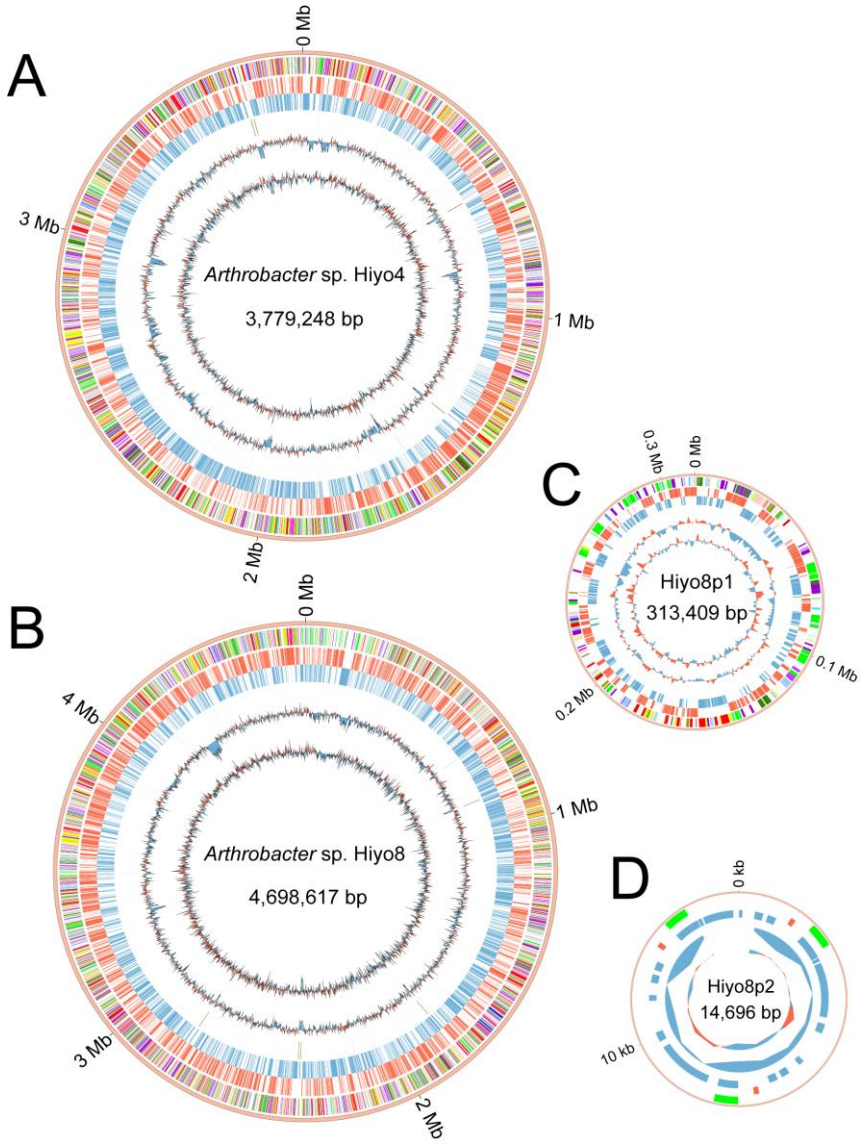

Figure 2-2 | Circular diagrams of the chromosomes and plasmids of *Arthrobacter* sp. Hiyo4 and Hiyo8. Each concentric circle represents genomic data of *Arthrobacter* sp. Hiyo4 (A) and Hiyo8 (B) chromosomes and the Hiyo8 p1 (C) and p2 (D) plasmids. The outermost circle is the contig, the 2nd circle are the coding genes colored according to the functional categories of the eggNOG database (see Fig. 2-3 for color coding), the 3rd and 4th circles are the coding genes on the leading (red) and lagging (blue) strands, respectively, the 5th circle are the rRNA (brown) and tRNA (green) genes, the 6th circle is the GC content (1-kb sliding window), and the innermost circle represents the GC skew (1-kb sliding window).

The total genome sizes of the four strains ranged from 2.6 to 5.5 Mbp. Although the genome sizes of Hiyo4 and Hiyo8 were within the range of the previously reported genomes, their CDS numbers were exceptionally large (Table 2-1), possibly because of additional genes that facilitate adaptation to different environmental conditions and/or overestimation due to the sequencing error. The GC content was 63–65%, which is similar to that of the previously reported genomes (59–67%). The functional categories of egg-NOG were assigned to 61–66% of the CDSs, and the distributions of the isolated strains were similar to those of the other close strains (Fig. 2-3).

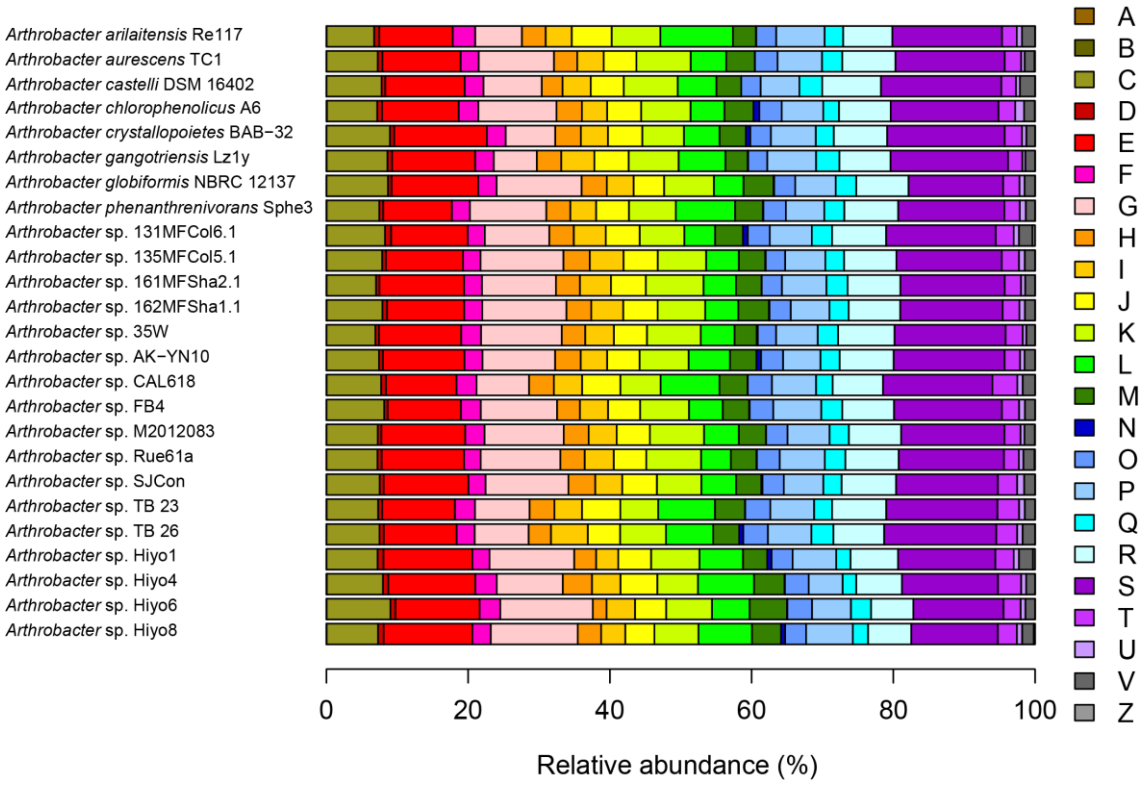

Figure 2-3 | Relative abundance of functional gene categories in the *Arthrobacter* genomes. The relative abundance of CDSs assigned to each eggNOG functional category is plotted for each *Arthrobacter* genome. The eggNOG functional categories are as follows: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation; K, transcription; L, replication; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking and secretion; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; and Z, cytoskeleton.

Phylogenetic analysis and comparative genomics

To reveal the phylogenetic relationships among the four strains Hiyo1, Hiyo4, Hiyo6, and Hiyo8, I constructed a maximum-likelihood phylogenetic tree of the *Arthrobacter* genus based on 16S rRNA gene sequences (Fig. 2-4). The tree reliably placed the four isolated strains within this genus. There was only one nucleotide base gap between the 16S rRNA sequences of Hiyo1 and Hiyo8, suggesting their close relationship. Hiyo4 and Hiyo6 were classified into different clades in the tree.

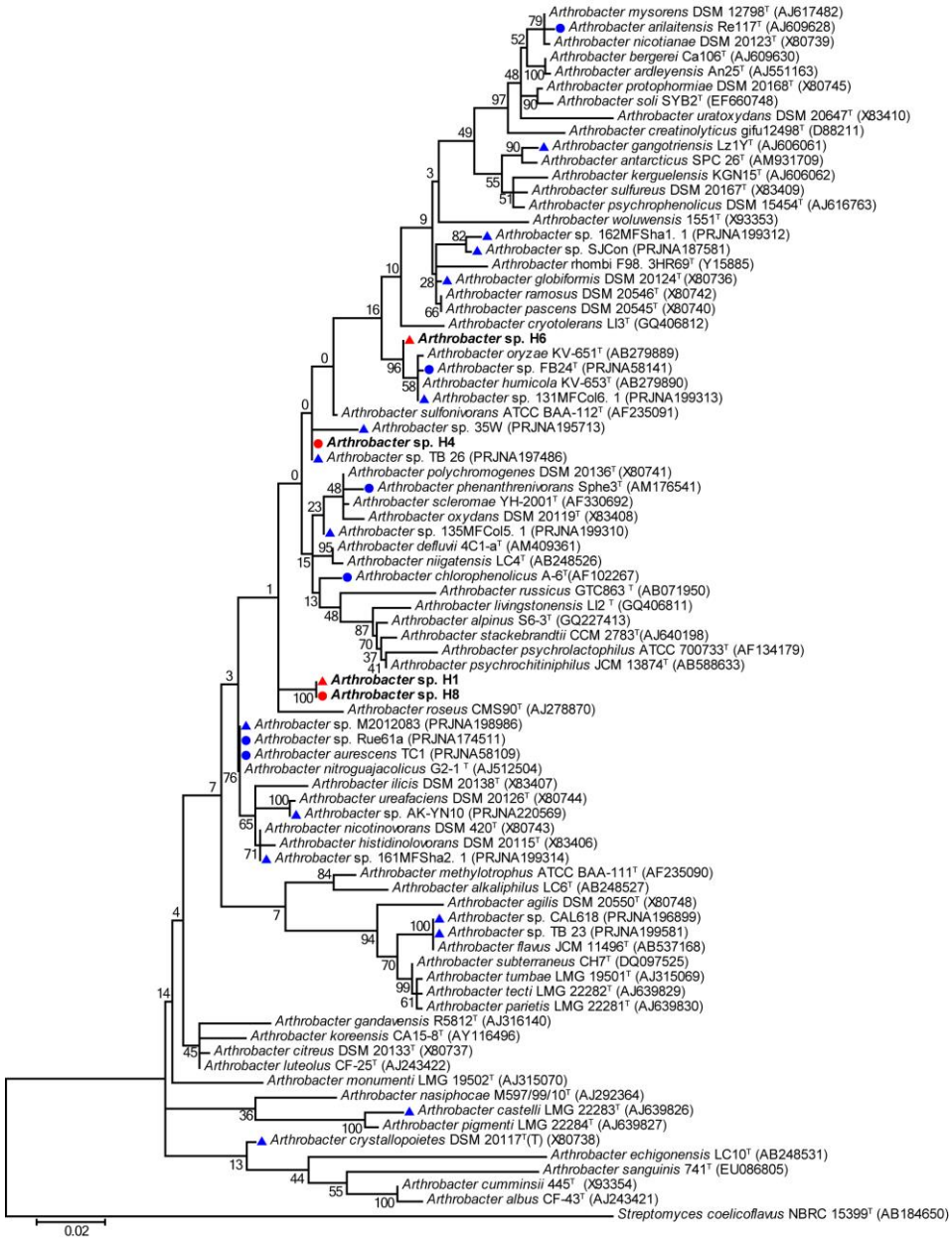

Figure 2-4 | Phylogenetic tree of *Arthrobacter* genus. The phylogenetic tree was reconstructed using the maximum-likelihood method based on 16S rRNA sequences with *Streptomyces coelicoflavus* NBRC 15399^{T} as an outgroup. Numbers adjacent to branch points are bootstrap percentages (1,000 replicates). Symbols represent the available circular genomes (circle), available draft genomes (triangle), those isolated in this study (red), and those isolated in previous studies (blue).

Subsequently, I conducted comparative genome analysis with 21 publicly available *Arthrobacter* genomes. The relative abundance of the CDSs assigned to each eggNOG functional category in each genome (Fig. 2-3) shows small difference among these *Arthrobacter* strains, *i.e.*, their genomes have similar functional composition overall. The most striking difference between the *Arthrobacter* genomes isolated from the tsunami-affected soil and those isolated from other environments was that desferrioxamine B biosynthesis genes were independently lost in each of the former genomes. Within 14 publicly available, high-quality *Arthrobacter* genomes, the desferrioxamine B biosynthesis gene cluster and surrounding synteny structures were found to be highly conserved (Fig. 2-5). On the other hand, the desferrioxamine B biosynthesis gene cluster was entirely absent in the completed Hiyo1 and Hiyo8 genomes: the desA (pyridoxal-dependent decarboxylase) and desB (L-lysine 6-monooxygenase) genes of the cluster had nonsense mutations in the Hiyo4 genome, and neither the cluster nor the surrounding synteny structure was found in the Hiyo6 genome.

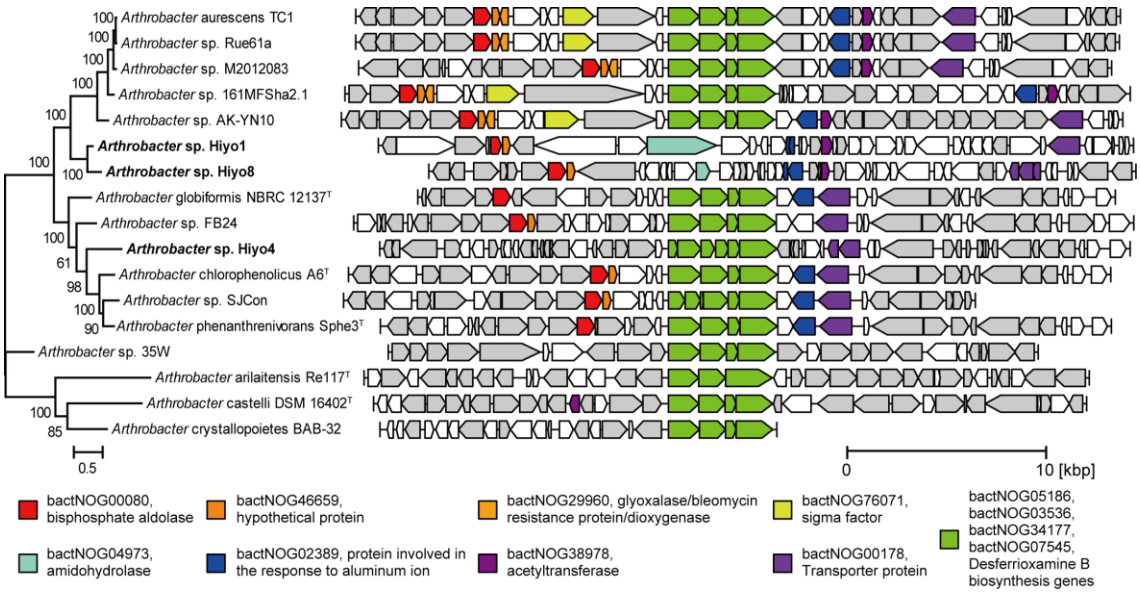

Figure 2-5 | Syntenic map around siderophore-synthesis gene clusters in *Arthrobacter* genomes. Genes are represented by arrows whose lengths are proportional to the gene lengths. Desferrioxamine B biosynthesis genes are shown in green. Other conserved genes are shown in different colors according to their annotation. Other genes annotated using the eggNOG database are shown in gray. The phylogenetic tree was reconstructed on the basis of the set of up to 400 conserved bacterial marker genes with 1,000 bootstrap replicates by the maximum-likelihood method.

Desferrioxamine B is a member of the siderophores family of molecules, which are low-molecular-weight, iron-chelating compounds secreted by many microbes and plants for the uptake of iron (207–209, 238, 239). The ability to use siderophores confers an ecological advantage when iron is limited (240). Many *Arthrobacter* strains have a desferrioxamine B biosynthesis gene cluster, which is composed of four genes, named desABCD, for siderophore production (241, 242). It should be noted that no iron-rich media were used during the isolation procedures.

The independent losses of the siderophore-synthesis genes are not likely to have occurred by chance but likely because of natural selection. Thus, these *Arthrobacter* strains were assumed to have been under weak selection pressure for iron uptake and to be at a growth disadvantage under low iron concentrations. To evaluate the growth potentials of these strains under various iron concentrations, culture experiments in iron-controlled media were conducted (Fig. 2-6). Two of the isolated strains (Hiyo1 and Hiyo8) required 10 μM Fe^{3+} iron for rapid growth, whereas a control strain (*A. phenanthrenivorans* Sphe3) that has a desferrioxamine B biosynthesis gene cluster required 1 μM Fe^{3+} iron (Fig. 2-6 A, B, E). Notably, under the 1 μM Fe^{3+} iron concentrations, the maximum growth rates of Hiyo1 and Hiyo8 ($1.08 \pm 0.14 \times 10^{-2}$ and $1.15 \pm 0.32 \times 10^{-2}$, respectively) was significantly smaller than that of Sphe3 ($2.34 \pm 0.20 \times 10^{-2}$) (p-value <0.05, t-test with Bonferroni correction). Hiyo4 and Hiyo6 showed very weak growth even with 10 μM Fe^{3+} iron, possibly because these two strains require additional nutrients for growth (Fig. 2-6 C, D).

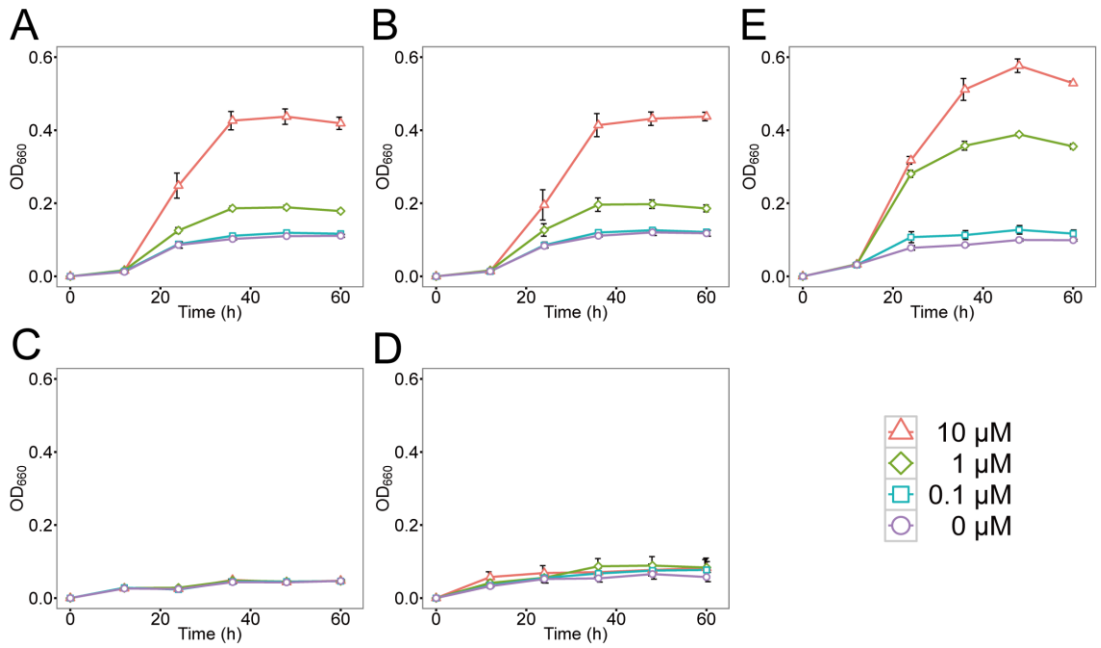

Figure 2-6 | Growth curves of the *Arthrobacter* strains at different iron concentrations. Growth was measured as optical density values at 660 nm in modified MM9 medium containing different concentrations of iron (III): 0.0, 0.1, 1, and 10 μM. Growth curves of *Arthrobacter* sp. Hiyo1 (A), Hiyo8 (B), Hiyo4 (C), Hiyo6 (D), and *A. phenanthrenivorans* Sphe3 (E) were measured.

Based on these results, I hypothesized that strains with *de novo* mutations in siderophore synthesis genes or those that originally lacked these genes would be selected under iron-enriched conditions. Notably, siderophore production by soil-living microbes has been reported to help various plants absorb iron (*e.g.*, tomato, cucumber, barley, and corn) (243–245) and has been associated with N_2 fixation (pigeon pea) (246); therefore, these observed genomic changes in the bacterial communities might also relate to plant growth.

Soil chemical analysis

To confirm that the tsunami-affected soil sample analyzed in this study did in fact have a high iron concentration and/or chemical characteristics that are similar to those reported in previous studies on tsunami-affected soils, chemical analysis of the soil samples of the Hiyoriyama and Amamiya sites was conducted (Table 2-3). As expected, Hiyoriyama contained 13 times more iron than Amamiya, which is consistent with the observed losses of the siderophore-synthesis genes. Because Hiyoriyama was also found to

be substantially rich in sulfate (*e.g.*, SO_4^{2-} levels were 169 times higher in Hiyoriyama than Amamiya), the iron was possibly provided in the form of iron-sulfur compounds (*i.e.*, FeS, FeS_2, Fe_2S_3), which are contained in seawater and sediment (247). These sulfurs can be oxidized into sulfates via biological processes in the presence of electron acceptors (248), including nitrate (NO_3^{3-}) (249–252). I propose that the substantially smaller amount of nitrate in Hiyoriyama than Amamiya (>13-fold) may reflect this process.

Table 2-3 | Chemical characteristics of the soil samples.

		Hiyoriyama	Amamiya
pH		5.9	6.1
Electrical conductivity	dS/m	0.23	0.02
Total organic carbon	g/kg	3	2.4
Total nitrogen	g/kg	0.2	0.2
Ammonium nitrogen (NH_3)	mg/kg	19.9	15.9
Nitrate nitrogen (NO_3^-)	mg/kg	6.2	84.5
Nitrite nitrogen (NO_2^-)	mg/kg	<0.05	<0.05
Effective phosphate (PO_4^{3-})	mg/kg	12	40
Exchangeable K^+	cmol(+)/kg	0.34	0.36
Exchangeable Ca^{2+}	cmol(+)/kg	6.19	4.32
Exchangeable Mg^{2+}	cmol(+)/kg	0.44	0.94
Exchangeable Na^+	cmol(+)/kg	0.42	0.1
Exchangeable Mn^{2+}	mg/kg	1.52	3.07
Available iron (Fe)	mg/kg	142	10.4
Cl^-	mg/kg	15.7	12.4
Sulfate (SO_4^{2-})	mg/kg	379	2
Cd[a]	mg/l	<0.001	<0.001
Cr(VI)[a]	mg/l	<0.005	<0.005
Total mercury (Hg)[a]	mg/l	<0.0005	<0.0005
Alkyl mercury (Hg)[a]	mg/l	<0.0005	<0.0005
Pb[a]	mg/l	<0.004	<0.004
As[a]	mg/l	<0.001	0.001
B[a]	mg/l	<0.1	<0.1
Cd[b]	mg/kg	<15	<15
Cr(VI)[b]	mg/kg	<25	<25
Hg[b]	mg/kg	<1.5	<1.5
Pb[b]	mg/kg	<15	<15
As[b]	mg/kg	<15	<15
B[b]	mg/kg	<400	<400
Cu[b]	mg/kg	<10	<10
Zn[b]	mg/kg	100	49
Ni[b]	mg/kg	<30	<30

[a] Elution amount of chemicals by water.

[b] Total amount of chemicals contained in the soil sample.

Except for these chemicals, the characteristics of the two samples were similar overall, suggesting that the two soil samples share a similar geological origin. In particular, the absence of heavy metals such as Pb, Hg, and Cu in Hiyoriyama might indicate that the soil was not completely covered or replaced with marine sediments. In addition, the similarities in electrical conductivity and Na^+ and Cl^- content between samples can be attributed to the effects of rain; in the case of the 2004 Indian Ocean tsunami, water-soluble salts derived from the tsunami were strongly reduced after a rainy season in a coastal area in Thailand (253). I note that the annual precipitations in Sendai city were 1,214 and 1,179 mm in 2011 and 2012, respectively (the Japan Meteorological Agency).

Shotgun metagenome sequencing

To investigate differences in the taxonomic compositions and protein-coding gene abundance between the two samples, shotgun metagenome sequencing was conducted (Table 2-4). After quality control, 822,865 and 961,221 reads were obtained from the Hiyoriyama and Amamiya samples, respectively.

Table 2-4 | General features of the metagenome sequences.

	Hiyoriyama	Amamiya
Raw sequence reads	1,091,366	1,177,491
After quality control	822,865 (75.40 %)	961,221 (81.60 %)
CDSs	1,170,916	1,323,575
16S rRNAs	628	633
Taxonomically classified reads	114,838 (13.96 %)	112,459 (11.70 %)
- Bacteria	113,696 (99.01 %)	111,326 (98.99 %)
- Archaea	933 (0.81 %)	707 (0.63 %)
- Viruses	209 (0.18 %)	426 (0.38 %)

Using Kraken (69), 114,838 (14.0%) and 112,459 (11.7%) shotgun reads from Hiyoriyama and Amamiya were taxonomically classified, respectively. Almost all reads were assigned to Bacteria (99.01 and 98.99%), whereas few reads were assigned to Archaea (0.81 and 0.63%) and Viruses (0.18 and 0.38%). The microbial composition of abundant genera is shown in Fig. 2-7. The most abundant genus in both samples was *Burkholderia* (4.85 and 7.20%), followed by *Bradyrhizobium* (4.66 and 6.30%), *Rhodopseudomonas* (3.27 and 3.46%), and *Pseudomonas* (2.98 and 3.13%). The similar composition of the major taxonomic groups reflects the similar overall chemical characteristics between the two soil samples. In addition, I estimated the typical habitats of the contained microbes by querying the extracted 16S rRNA genes against MetaMetaDB (176), a database that links 16S rRNA gene sequences to environments based on comprehensive analysis of published metagenomic and amplicon-sequencing datasets. The estimated habitats quantified as MHI values (176) showed that the top habitat was soil in both communities; however, the marine habitat was estimated to be modestly more abundant in Hiyoriyama, whereas the soil habitat was more abundant in Amamiya, as expected (Fig. 2-8).

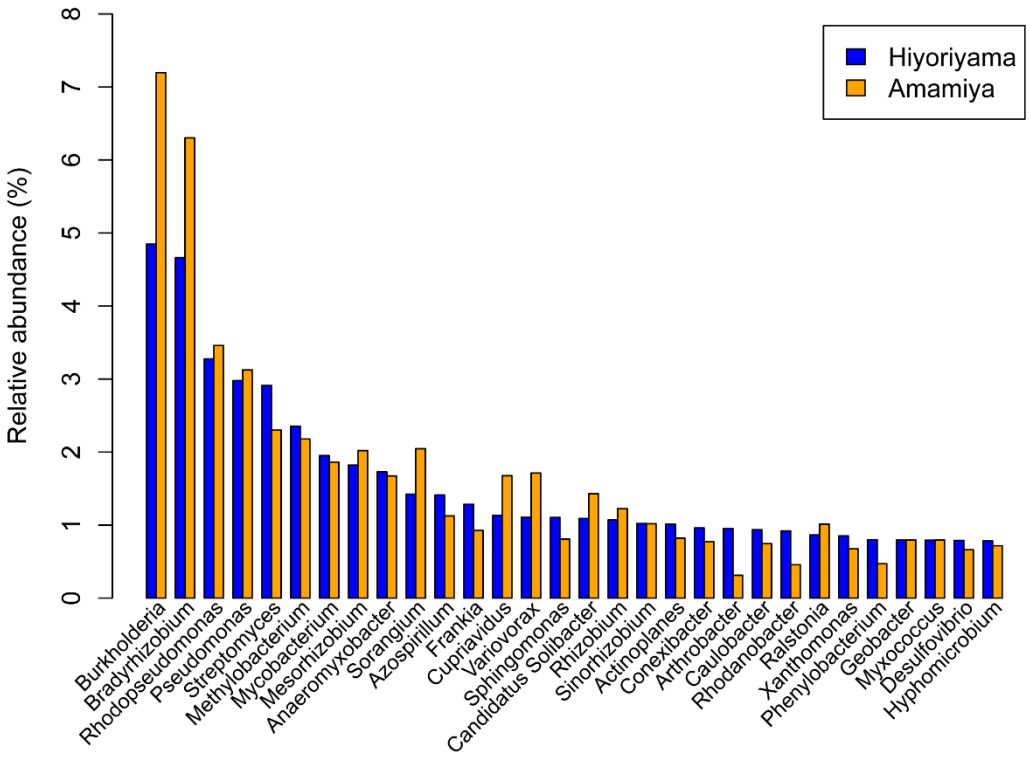

Figure 2-7 | Abundant microbial genera determined in metagenome shotgun sequencing. The 30 most abundant microbial genera at Hiyoriyama and their relative abundance at both sites are displayed. Blue and orange bars represent Hiyoriyama and Amamiya, respectively.

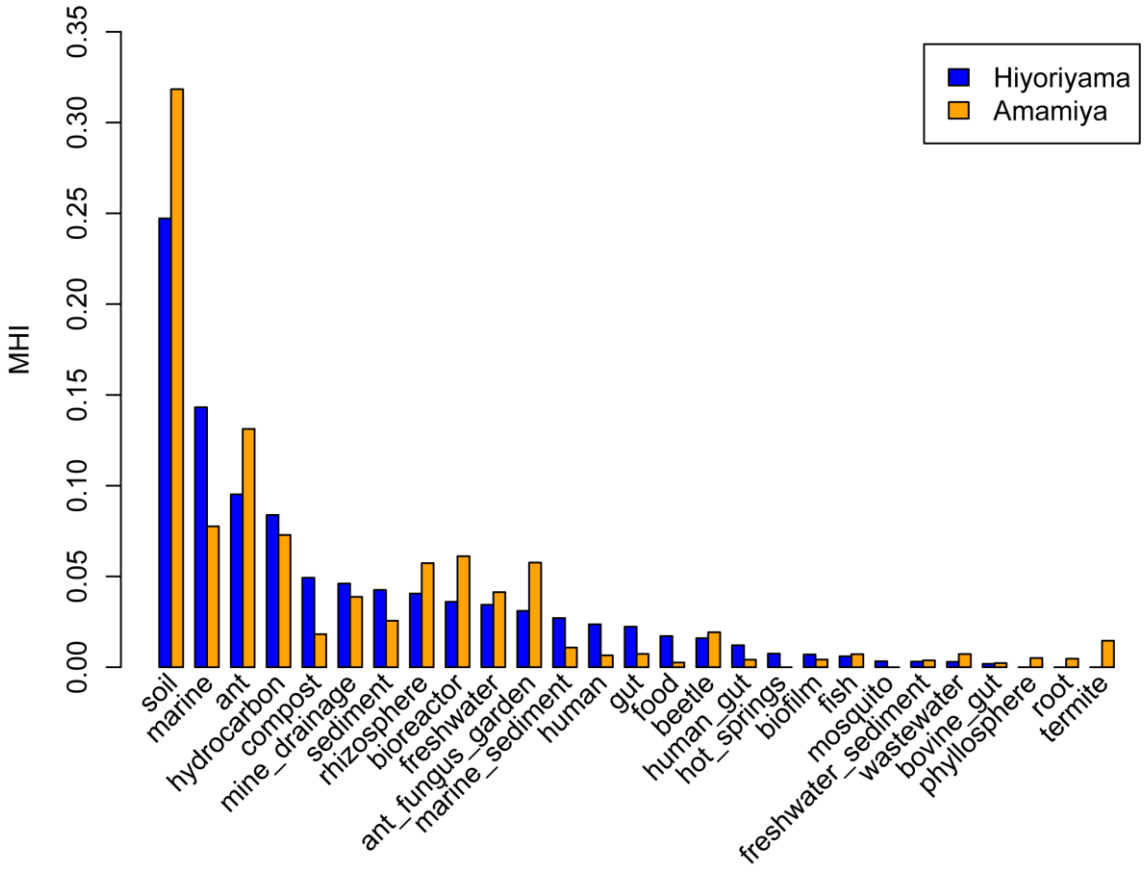

Figure 2-8 | Estimated habitats of microbes at each site. Microbial habitability index (MHI) scores were calculated using the top-hit sequences of blastn searches against MetaMetaDB, where queries were all 16S rRNA gene sequences extracted from the shotgun metagenome sequences. Blue and orange bars represent Hiyoriyama and Amamiya, respectively.

Figure 2-9 displays the genera whose relative abundance substantially differed between the two samples, including only those whose abundance in one sample was more than three times greater than that in the other. Notably, *Arthrobacter* was the only genus that was both abundant in and differed substantialy between the two samples. Considering the fact that *Arthrobacter* was the genus cultivated in both the R2A and ZoBell media, I propose that this genus likely shows a greater potential for adaptation to tsunami-affected soils. The other genera that were more abundant in Hiyoriyama included *Erysipelothrix*, where all reads were assigned to a single species, *Erysipelothrix rhusiopathiae* (254), which is known to cause erysipelas, a bacterial skin infection, in animals (255). Although previous culture-based studies reported several pathogen species (*Mycobacterium elephantis*, *Massilia timonae*, *Vibrio ichthyoenteri*, *V. natriegens*, and *V. fluvialis*) in sludge derived from tsunami-affected soil in Tohoku (206, 256), these species were not detected

in the present dataset. It may also be notable that the genera detected only in Hiyoriyama included typical marine-living groups such as *Croceibacter* (257), *Marinitoga* (258), and *Pyrococcus* (259–261), implying that the tsunami facilitated microbial immigration.

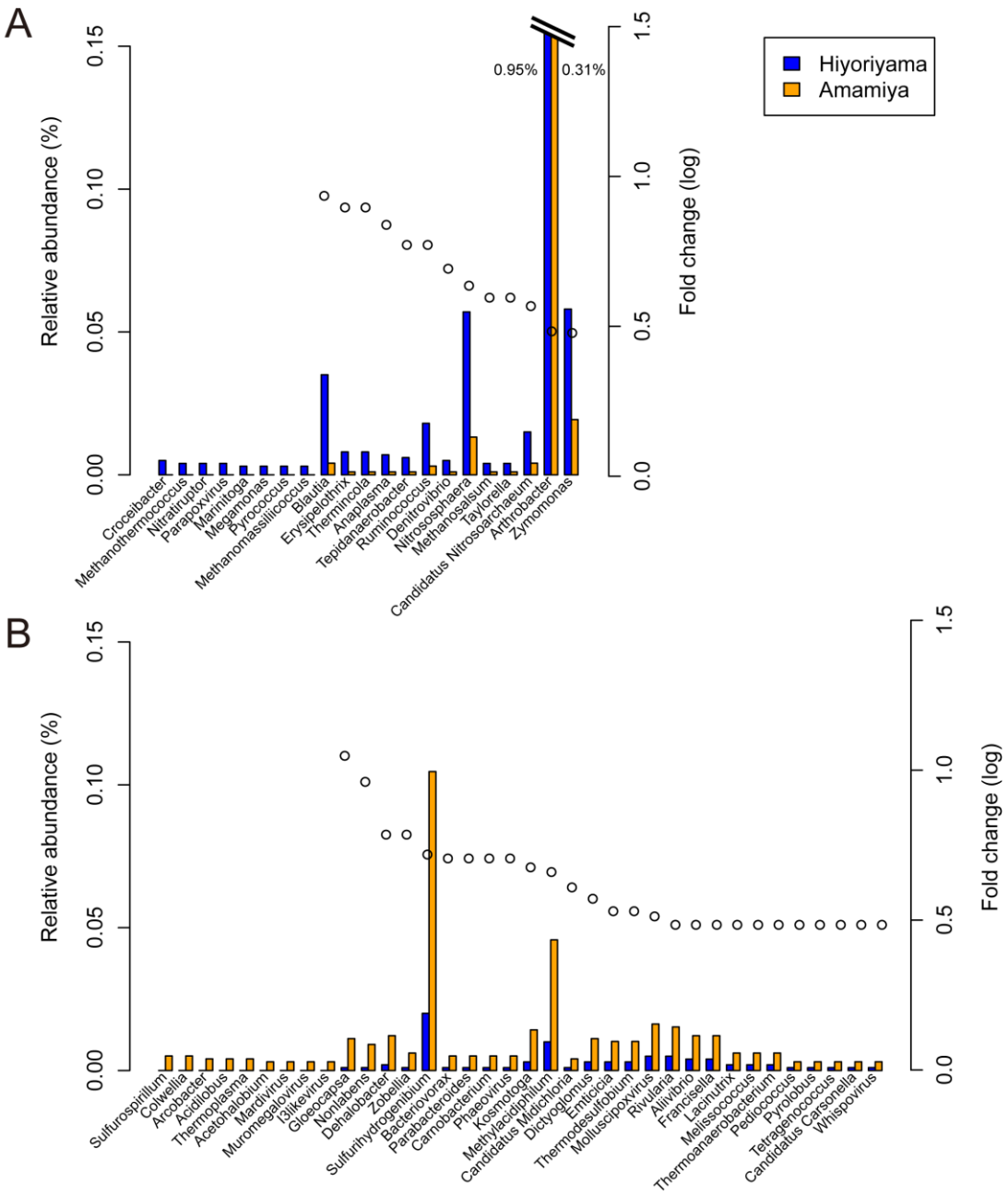

Figure 2-9 | Microbial genera with a substantial difference in abundance between the two sites. Genera whose relative abundance values differed by more than three times between the two sites and represented more than 0.003% of the total abundance in either of the two sites are displayed for genera overrepresented in (A) Hiyoriyama and (B) Amamiya. Blue and orange bars represent the relative abundance in Hiyoriyama and Amamiya, respectively. Dots represent the ratios of larger abundance values divided by smaller values, if the smaller value was not zero.

I annotated the CDSs and investigated the relative abundance of nitrogen cycle-related genes, because the taxonomic analysis identified genera known to metabolize inorganic nitrogens, such as *Bradyrhizobium, Azospirillum*, *Frankia*, *Mesorhizobium*, *Rhizobium*, and *Sinorhizobium* (Fig. 2-9), and the chemical analysis revealed differences in the amount of nitrogen compounds (Table 2-3). The abundance of functional genes showed that genes related to denitrification and nitrogen fixation were more abundant in Hiyoriyama and genes related to nitrite reduction were more abundant in Amamiya (Fig. 2-10). In addition to the oxidization of iron-sulfur compounds, this dominance of denitrification-related genes at Hiyoriyama may be another cause of the relatively small amount of nitrate observed in Hiyoriyama (Table 2-3), which might affect terrestrial vegetation indirectly.

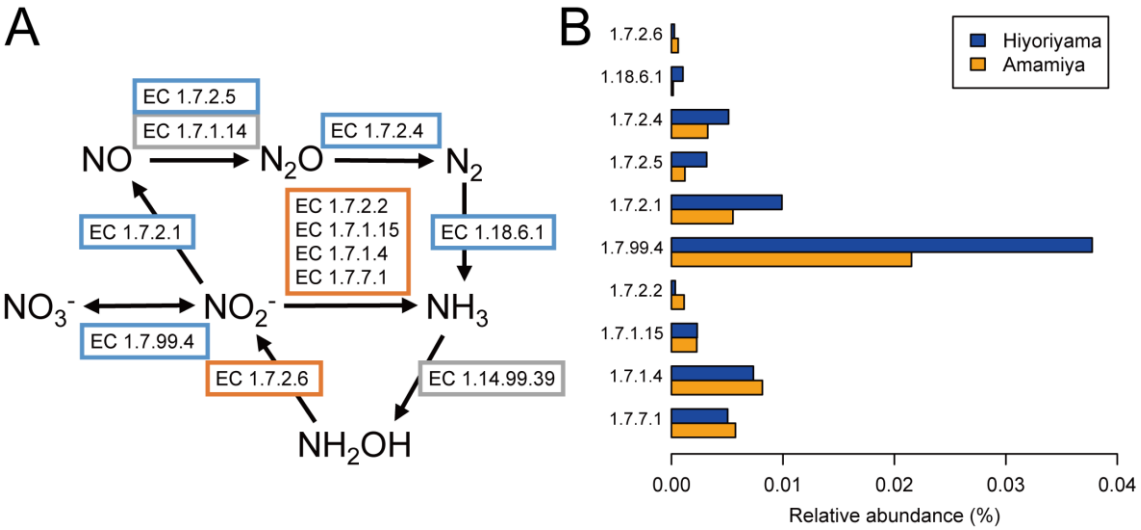

Figure 2-10 | Relative abundance of genes related to nitrogen metabolism. (A) A pathway map of nitrogen metabolism genes with Enzyme Commission numbers. Blue and orange rectangles represent genes that were found to be more abundant in Hiyoriyama and Amamiya, respectively. Gray rectangles represent genes not found in either sample. (B) A bar plot of the relative abundance for each gene represented by an Enzyme Commission number.

I also investigated the abundance of siderophore-synthesis genes in the shotgun metagenome data, but only three and four reads of genes that are involved in this process (bactNOG07545, bactNOG14638, and bactNOG30540) were detected in Hiyoriyama and Amamiya, respectively, which is not a sufficient sample size for statistical analysis. Differences in sulfur metabolism genes were not as large as those of nitrogen metabolism

genes. A substantial difference was observed in the numbers of cation transporter genes, where 127 and 59 monovalent cation/H^+ antiporter subunits, and 96 and 16 Na^+/Ca^{2+} antiporter family proteins (bactNOG00892) were detected in Hiyoriyama and Amamiya, respectively. These genes may have facilitated salt tolerance in the tsunami-affected soil, because cation transporters are known to function in bacterial salt tolerance (262, 263).

Conclusion

In this study, I isolated four *Arthrobacter* strains from a soil sample affected by the Tohoku tsunami and determined their whole-genome sequences. Independent losses of siderophore-synthesis genes were suggested in these genomes, which was consistent with the rich iron content detected in the tsunami-affected soil sample and the weak cultivability of the isolated strains in iron-limited media, although further experimental analysis will be needed to conclude it. The chemical and metagenomic analyses indicated that the tsunami-affected sample was largely similar to the unaffected sample, although some notable differences were observed regarding nitrogen metabolism and taxonomic composition. It should be noted that I cannot conclusively determine whether the isolated *Arthrobacter* strains were brought into the sampled area from sea and then adapted to soil, or were originally in the soil and survived under the tsunami-affected conditions. In either case, it also remains undetermined whether the siderophore-synthesis genes were mutated after the tsunami or the strains that originally lost these genes were simply favored and selected in the tsunami-affected soil.

The Pacific coast of Tohoku, Japan has been flooded by tsunamis many times in history (more than 11 tsunamis were triggered in the last 200 years, according to (264)). Because a tsunami should affect the soil and its microbial communities in diverse manners, I envision that further comprehensive analyses on microbial ecology and evolution after a tsunami will be necessary to develop a deeper understanding of the recovery processes of terrestrial microbial ecosystems.

Chapter 3: Seasonal analysis of microbial communities in precipitation in the Greater Tokyo Area, Japan

Introduction

Microbes are present and move around nearly everywhere in the Earth. Aerial microbes have received considerable attention within this context because the atmosphere not only is an unusual habitat for microbes but also likely represents a path by which microbes move exceptionally long distances (265–269). To date, several studies have investigated aerial microbial communities on airborne particles and in clouds using culture-dependent and independent techniques (270–277), and revealed that aerial microbes can originate from terrestrial habitats, including plant surfaces (271, 276, 278). The long-distance transport of aerial microbes has also been reported, for example from Chinese deserts to Japan over the east Eurasian continent and the Sea of Japan (279, 280). Pathogens in the atmosphere may be transported over long distances, as integrated simulation analyses of climate and disease propagation suggest the involvement of aerial microbes in human diseases (281, 282). Likewise, the outbreak of several plant infections due to aerial microbes transported beyond borders and seas has been hypothesized (283, 284).

Precipitation, *i.e.*, rainfall and snowfall, would bring aerial microbes in the troposphere to the ground surface. Quantitative polymerase chain reaction (PCR) has detected pathogenic bacterial sequences in precipitation samples (285), implicating that precipitation may alter microbial ecosystems on the ground (286, 287). In the reverse direction, aerial microbes impact the climate by accelerating cloud formation and precipitation, known as "bioprecipitation" (288–293). Several microbial species experimentally exhibit ice nucleation activity (INA), which is the ability to accelerate ice nucleation at relatively warm temperatures by producing so-called INA proteins (294). Such INA microbes are broadly distributed among bacteria and fungi and have been isolated from precipitation and cloud water (295, 296). In addition, microbes in clouds may affect the chemical composition of clouds via carbon (297, 298) and nitrogen metabolism (299). Thus, a basic understanding of microbial communities in precipitation provides important knowledge regarding microbial ecology, public health, and even meteorology. To date, several cloning-based (270, 287, 300, 301) and community-wide but short-term (302, 303) analyses of microbial communities in precipitation have been carried out. However, community-

wide and seasonal analyses have not been conducted.

Here, I and collaborators conducted 16S ribosomal RNA (rRNA) amplicon-sequencing analysis of 30 precipitation samples that were aseptically collected over one year in the Greater Tokyo Area, Japan. Microbial community analysis revealed seasonal variations in their composition. Notably, the estimated original habitats of precipitation microbes showed reasonable consistency with estimated air mass backward trajectories. My results support a precipitation-mediated microbial cycle model in which soil, oceanic, and animal-associated microbes are spread in the atmosphere, transported for long distances, and deposited via precipitation.

Material and methods

Precipitation sampling

Precipitation samples were collected at two sites in the Greater Tokyo Area, Japan: Kashiwa (35°54'00"N, 139°55'59"E, 50 m above sea level) and Hongo (35°42'55"N, 139°45'56"E, 30 m above sea level) (Fig. 3-1). The Kashiwa site was on the roof of a seven-story building on the Kashiwa campus, the University of Tokyo, Chiba, Japan, which is surrounded by residences, farms, and woods in a suburb of Tokyo. The Hongo site was on the roof of a five-story building on the Hongo campus, the University of Tokyo, Tokyo, Japan, which is located in downtown Tokyo. The sites are 25.5 km apart and neither geologically nor meteorologically separated in the Kanto plain. The upper areas of both sites are wide open and lack any obstructing buildings or structures that would contaminate the precipitation samples. At the Kashiwa site, precipitation was aseptically collected using a US-330 automatic precipitation sampler (Ogasawara Keiki, Tokyo, Japan) following the method of Kaushik *et al*. (285). This device consists of a sterile and disposable bottle inside a 4°C refrigerator and automatically collects precipitation by opening the lid only when a sensor detects precipitation. At the Hongo site, precipitation samples were manually collected into a sterile and disposable bottle on ice and immediately stored in a 4°C refrigerator. At both sites, every part of the collection equipment that potentially directly contacted precipitation samples (*e.g.*, disposable collection bottles and channel tubes) was sterilized by gamma rays in advance of each sample collection. The precipitation samples were pre-filtered through 5-μm membrane filters, and microbial cells were collected using 0.22-μm Sterivex filters (Millipore, USA). The Sterivex filters were promptly moved to a -20°C freezer and stored until DNA extraction. Precipitation sampling required no special permission. To prepare negative control samples, 1 L of Milli-

Q purified water was poured into the collection equipment and filtration was carried out in the same manner.

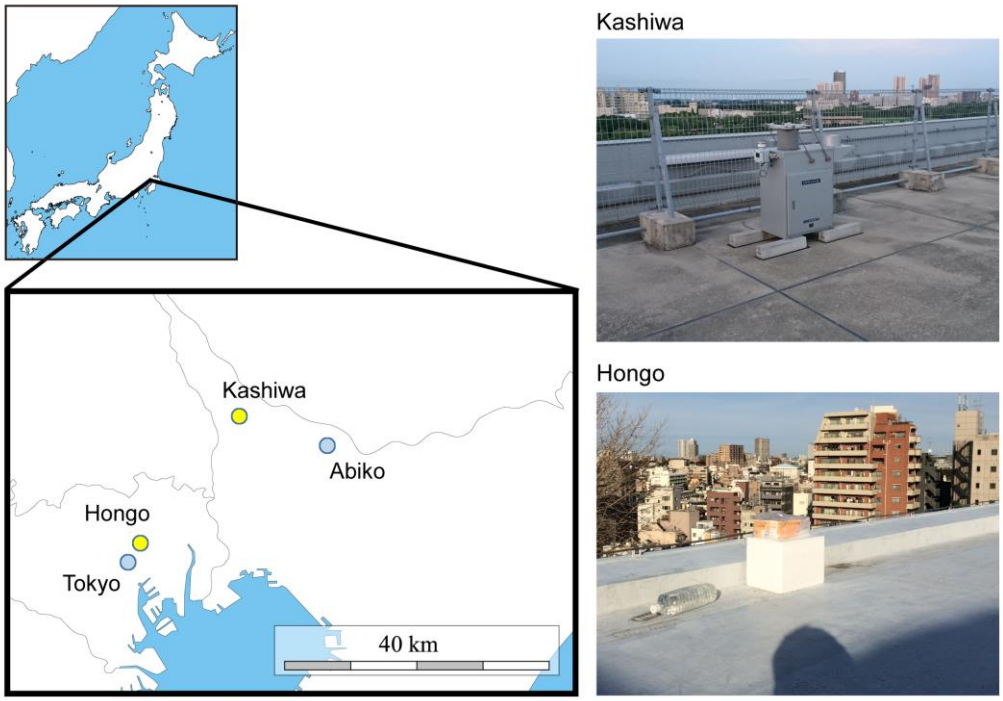

Figure 3-1 | A map of the sampling sites (Kashiwa and Hongo, yellow) and meteorological observatories (Abiko and Tokyo, blue) (left panel), with photos of the sampling sites (right panel). At the Kashiwa site, a US-330 automatic precipitation sampler (Ogasawara Keiki, Tokyo, Japan) was installed. At the Hongo site, precipitation samples were manually collected.

I and my collaborators collected 25 and 5 precipitation samples containing sufficient amounts of microbial DNA at the Kashiwa and Hongo sites, respectively. The sampling dates spanned more than one year from May 2014 to October 2015, encompassing the rainy and typhoon seasons in Japan (Table 3-1; the six digits, letter, and suffix number for each sample name represent the sampling date (YYMMDD), the site (K for Kashiwa and H for Hongo), and the volume (if multiple samples were collected during the same precipitation event). A precipitation event was defined if there was no precipi-

tation six hours before and after the event. The volumes of collected and filtered precipitation ranged from 50 to 1,000 mL. For correlation analysis with meteorological data, I excluded the data obtained from samples 140630K_50, 140630K _100, 140810K_50, and 140810K_100, which were retrieved as replicate samples with different volumes. Eight negative control samples were also collected at different dates at the Kashiwa and Hongo sites.

Table 3-1 | Sequencing statistics and meteorological characteristics of each precipitation sample.

Sample	Sampling time (YYMMDD)[a]	Note	Amount of precipitation [mm]	Temperature [°C]	Atmospheric pressure	Wind speed [m/s]	Filtration volume [mL]	Raw reads	Effective reads	OTUs	Shannon's diversity index
140521K	140521(01:00)–140521(18:00)	–	36	16.56	994.31	3.16	50	8,340	246	44	2.62
140630K_50	140628(01:00)–140630(05:00)	Rainy season	22	21.98	999.23	1.74	50	8,622	1,092	27	1.63
140630K_100	140628(01:00)–140630(05:00)	Rainy season	22	21.98	999.23	1.74	100	7,444	1,287	39	1.26
140630K_200	140628(01:00)–140630(05:00)	Rainy season	22	21.98	999.23	1.74	200	7,462	1,118	44	1.48
140810K_50	140810(00:00)–141810(23:00)	Typhoon	31.5	25.46	998.46	3.89	50	7,621	275	76	3.70
140810K_100	140810(00:00)–141810(23:00)	Typhoon	31.5	25.46	998.46	3.89	100	8,441	108	53	4.41
140810K_200	140810(00:00)–141810(23:00)	Typhoon	31.5	25.46	998.46	3.89	200	6,664	371	129	3.82
140926K	140925(02:00)–140926(04:00)	–	6.5	21.48	1,001.08	2.14	200	1,941	18	15	2.66
141014K	141013(13:00)–141014(07:00)	Typhoon	32.5	19.58	992.99	4.47	200	1,641	317	157	4.76
141023K	141021(05:00)–141023(18:00)	–	31.5	14.89	1,010.85	2.03	200	1,354	120	54	3.66
150107K	150106(16:00)–150106(18:00)	–	4	12.35	992.30	5.10	200	3,410	37	22	2.93
150116K	150115(11:00)–150116(00:00)	–	40.5	4.97	1,005.15	3.05	1,000	2,494	72	55	3.86
150202K	150130(05:00)–150130(19:00)	Snow	12.5	1.11	1,015.18	2.17	400	9,705	1,256	194	4.69
150409K	150407(03:00)–150408(17:00)	–	20.5	6.39	1,017.32	2.13	200	8,337	473	111	4.15
150412K	150410(17:00)–150411(14:00)	–	16	8.99	1,018.06	1.60	200	6,805	771	125	4.16
150414K	150413(11:00)–150414(17:00)	–	36.5	10.13	1,013.38	1.91	200	5,818	655	116	3.98
150513K	150512(21:00)–150513(01:00)	Typhoon	23	20.00	997.43	5.60	200	2,223	47	25	3.92
150604K	150603(08:00)–150603(13:00)	Rainy season	13	20.46	998.26	1.44	200	4,663	8	7	2.88
150628K	150626(19:00)–150627(12:00)	Rainy season	13.5	20.92	997.42	1.24	150	9,634	284	46	3.41
150711K	150708(15:00)–150709(20:00)	Rainy season	17.5	19.27	1,013.22	1.48	200	9,557	31	20	1.91
150718K	150716(04:00)–150717(13:00)	Typhoon	16.5	25.99	1,004.62	3.26	200	5,189	5	4	4.03
150816K	150814(05:00)–150814(22:00)	–	43	25.11	1,000.38	1.91	200	6,864	269	68	3.05
150827K	150826(00:00)–150826(17:00)	Typhoon	27	20.10	1,005.27	1.92	200	7,993	1,041	226	3.39
150926K	150924(19:00)–150926(06:00)	–	21	17.63	1,005.28	1.90	200	2,988	6	4	2.75
151014K	151011(01:00)–151011(11:00)	–	8	16.94	1,008.92	1.00	200	6,357	129	30	1.33
150414H	150413(07:00)–150414(12:00)	–	39.5	10.03	1,014.99	3.15	200	7,215	882	125	3.76
150513H	150512(20:00)–150513(06:00)	Typhoon	58.5	20.25	997.43	7.23	200	3,198	59	38	4.83
150627H	150626(15:00)–150627(10:00)	Rainy season	16	21.33	998.27	2.22	150	6,450	667	108	1.24
150710H	150708(10:00)–150710(00:00)	Rainy season	22	20.20	1,013.22	2.33	200	9,280	159	48	3.00
151014H	151011(02:00)–151011(10:00)	–	15	18.01	1,008.80	1.90	200	6,342	286	67	3.70

[a] *The six digits, letter, and suffix number in each sample name represent the sampling date (YYMMDD), the sampling site (K for Kashiwa and H for Hongo), and the filtered sample volume if prepared as a technical replicate with multiple volume sizes (50, 100, and 200 mL).*

DNA extraction and PCR amplification

Microbial DNA on the Sterivex filters was retrieved using a ChargeSwitch Forensic DNA Purification Kit (Invitrogen) according to the supplier's protocol with one exception: the filters were directly suspended in the extraction solution from the kit during the cell lysis process. The V5-V6 region of the prokaryotic 16S rRNA gene was amplified using a standard PCR protocol with TaKaRa Ex Taq (Takara) and the following high-performance liquid chromatography-purified primers: 784F (5'- RGGATTAGATACCC -3') and 1064R (5'- CGACRRCCATGCANCACCT -3') (304, 305). Amplified DNA was concatenated to multiplex identifier tags that were unique to each sample, and a mixture of ten samples on average was sequenced in one run on a 454 GS Junior System (Roche) after size selection (350 ± 50 bp). Pre-packaged sterile water for injection (in lieu of water from a laboratory water purification system) was used throughout the DNA extraction, PCR amplification, and DNA library preparation steps to avoid water-mediated contamination. The sampling and DNA sequencing were conducted by the member of the Atmosphere and Ocean Research Institute, the University of Tokyo.

Bioinformatic analysis

For raw sequence data from both precipitation and negative control samples, sequence regions at both ends that contained low-quality bases (quality score <20) were trimmed using DynamicTrim (306), chimeric sequences were filtered out using UCHIME with default settings (76), and sequences whose lengths were shorter than 150 bp were discarded. All remaining high-quality sequences were clustered with a 97% identity threshold using CD-HIT (77). After discarding clusters that contained negative control sequences (303), each cluster was designated as an operational taxonomic unit (OTU). For hierarchical cluster analysis of the precipitation samples, the Ward method was used based on Bray-Curtis dissimilarities between their OTU compositions. Nonmetric multidimensional scaling (NMDS) analysis was conducted using Bray-Curtis dissimilarities. The taxonomic assignment of each OTU was performed by conducting a blastn search (213) against the SILVA database (54) and retrieving the top hit sequence that showed e-values ≤1E-15. To estimate ordinary habitats for each 16S rRNA sequence, a blastn search was performed against MetaMetaDB (176), and the top hit sequence with an e-value ≤1E-10 and an identity ≥90% was retrieved. Microbial habitability index (MHI) scores were calculated as previously described (176).

Amplicon-sequencing data of aerosol and cloud water samples were downloaded

from NCBI SRA database (the accession numbers are shown in Table 3-2). Their ordinary habitat analyses were conducted as described above after quality filtering.

Table 3-2 | Rarefaction curves for each precipitation sample.

Organism	Bioproject	Biosample	SRR	Sequencer	Reference
Clouds	PRJNA170715	SAMN01091733	SRR521983	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Aerosol	PRJNA170715	SAMN01091732	SRR521980	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Clouds	PRJNA170715	SAMN01091731	SRR521978	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Clouds	PRJNA170715	SAMN01091729	SRR521975	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Clouds	PRJNA170715	SAMN01091719	SRR521974	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Clouds	PRJNA170715	SAMN01091718	SRR521972	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Aerosol	PRJNA170715	SAMN01091716	SRR521970	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Aerosol	PRJNA170715	SAMN01090408	SRR521967	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Clouds	PRJNA170715	SAMN01091717	SRR521965	454 GS FLX Titanium	DeLeon-Rodriguez, et al. (2013)
Aerosol	PRJNA271181	SAMN03273277	SRR1735301	454 GS Junior	Xia, et al. (2015)
Aerosol	PRJNA271181	SAMN03273276	SRR1735300	454 GS Junior	Xia, et al. (2015)
Aerosol	PRJNA271181	SAMN03273275	SRR1735299	454 GS Junior	Xia, et al. (2015)
Aerosol	PRJNA271181	SAMN03273274	SRR1735298	454 GS Junior	Xia, et al. (2015)

Meteorological data analysis

The data on the amount of precipitation, temperature, wind speed, and atmospheric pressure were retrieved from the website of the Japan Meteorological Agency (http://www.jma.go.jp/jma/menu/menureport.html), Ministry of Land, Infrastructure, and Transport of Japan. Precipitation, wind speed, and temperature data from the Abiko (35°51'48"N, 140°06'36"E; 16.4 km from Kashiwa) and Tokyo (35°41'30"N, 139°45'00"E; 2.9 km from Hongo) observatories were used for the analyses of the Kashiwa and Hongo sites, respectively (Fig. 3-1). Atmospheric pressure data from the Tokyo observatory were used (this observatory is the closest to both sites that records atmospheric pressure data). The wind speed, temperature, and atmospheric pressure data were averaged over the period of each precipitation event. To analyze long-range transport paths of air masses that caused precipitation by providing water vapor, I estimated backward trajectories of an air mass at 2,000 m altitude for 240 h prior to all precipitation events for each sampling site. The trajectories were calculated based on the hybrid single-particle Lagrangian integrated trajectory (HYSPLIT) model (http://ready.arl.noaa.gov/HYSPLIT.php) provided by the Global Data Assimilation System of National Oceanic and Atmospheric Administration, USA (307). The HYSPLIT

model uses gridded meteorological data and considers advection and diffusion of air parcels in calculation of their trajectories. This model has been used in a variety of atmospheric simulations focusing on the atmospheric transport, dispersion, and deposition of pollutants and hazardous materials (307), while it has also been adopted for estimation of sources of airborne microbes (*e.g.*, (303, 308–311)).

Mock precipitation

Ten bacterial strains were selected to prepare artificial mock precipitation samples. These strains belonged to bacterial groups that were frequently detected in previous studies investigating aerosolized microbes (265, 273, 276, 309) (Table 3-3). All strains were independently pure-cultured and suspended in 50 mL of sterile water at a cell density of 8.36×10^2 cell/mL/strain (total density: 8.36×10^3 cell/mL). This density was determined by direct cell counting of a precipitation sample collected on August 6, 2013 at the Kashiwa site (after pre-filtration with a 10-μm membrane filter, bacterial cells were captured with a 0.22-μm pore-size Sterivex filter and counted by performing 4',6-diamidino-2-phenylindole (DAPI) staining).

Table 3-3 | Strains in mock precipitation sample.

	Phylum	Class	Order	Family	Genus	Spices	Strain	Type strain
1	Acidobacteria	Holophagae	Acanthopleuribacterales	Acanthopleuribacteraceae	*Acanthopleuribacter*	*pedis*	NBRC101209	T
2	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Croceibacter*	*atlanticus*	HTCC2559	F
3	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	*subtilis*	NBRC13719	T
4	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	*Staphylococcus*	*aureus*	NBRC15035	F
5	Proteobacteria	α-Proteobacteria	Sphingomonadales	Sphingomonadaceae	*Erythrobacter*	*longus*	NBRC14126	T
6	Proteobacteria	α-Proteobacteria	Rhodobacterales	Rhodobacteraceae	*Roseobacter*	*litoralis*	NBRC15278	T
7	Proteobacteria	β-Proteobacteria	Burkholderiales	Burkholderiaceae	*Burkholderia*	*plantarii*	NBRC104885	F
8	Proteobacteria	γ-Proteobacteria	Pseudomonadales	Pseudomonadaceae	*Pseudomonas*	*stutzeri*	ATCC17588	F
9	Proteobacteria	γ-Proteobacteria	Alteromonadales	Pseudoalteromonadaceae	*Pseudoalteromonus*	*espejiana*	NBCR102222	T
10	Proteobacteria	γ-Proteobacteria	Vibrionales	Vibrionaceae	*Vibrio*	*fischeri*	ATCC700601n	T

Data deposition

The amplicon sequence data were deposited in the DDBJ/ENA/GenBank database under BioSample IDs SAMD00059585-SAMD00059614 and SAMD00060461-SAMD00060471. All data were registered under BioProject ID PRJDB5087.

Results and discussion

Amplicon sequencing of precipitation samples

A total of 64,100 high-quality sequences 231 ± 45 bp in length were generated from 30 precipitation and eight negative control samples. The precipitation samples included typhoon rain, rainy season rain, and snow. After removing sequences exhibiting >97% similarity to the negative control samples, 12,089 "effective" sequences comprising 1,297 OTUs remained. To make the analyses based on reads that were not likely from contamination as much as possible, I took a conservative and strict filtering approach, whose extent of read number reduction was similar to that in a previous study (303). The number of OTUs per sample ranged from 4 to 226 (Table 3-1). Based on rarefaction curves, the obtained OTUs represented their microbial communities well for some samples, although several samples required additional sequences (Fig. 3-2).

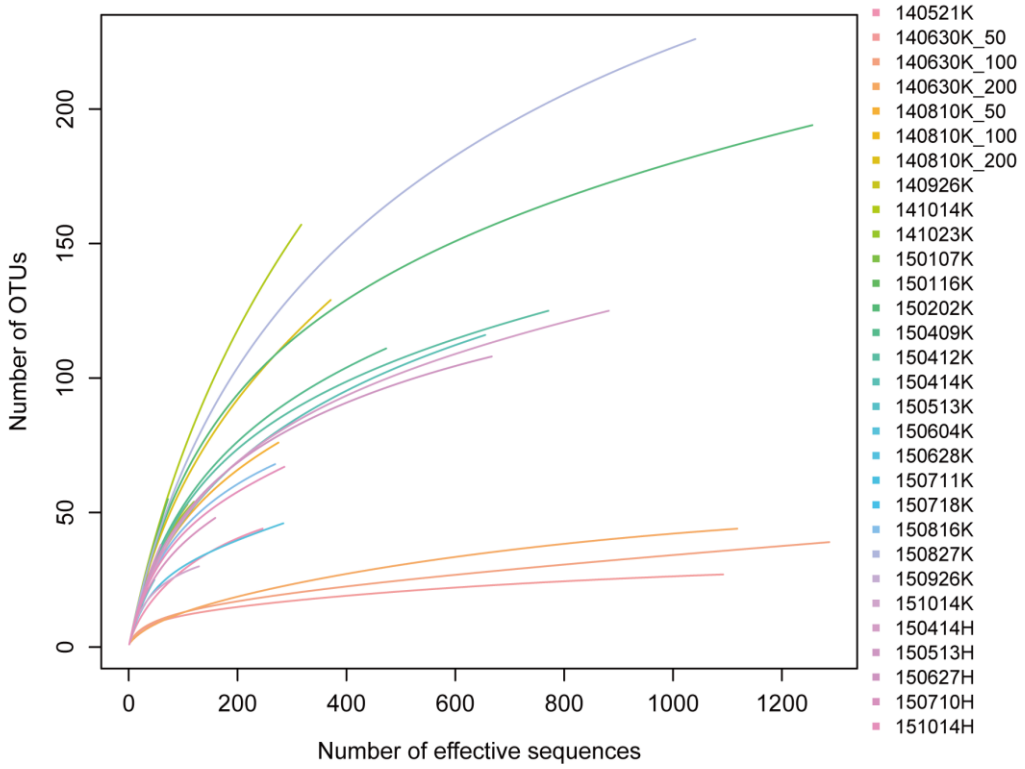

Figure 3-2 | Rarefaction curves for each precipitation sample.

Hierarchical cluster analysis of OTU composition in the precipitation samples indicated samples collected during the same precipitation event with different volumes (50, 100, and 200 mL) that were highly similar to each other (Fig. 3-3, open symbols),

suggesting that differences in volume have little effect on analysis in the 50-200 mL range. Moreover, microbial communities in samples that were collected on the same day at different sampling sites (Kashiwa and Hongo) were closely positioned in the dendrogram (Fig. 3-3, closed symbols), indicating that the observed OTU compositions reflect the microbial populations in precipitation rather than those in the atmosphere near the ground surface or equipment- or reagent-mediated contamination at each site. NMDS analysis did not show any clear trend, although samples of close dates tended to be clustered together (Fig. 3-4).

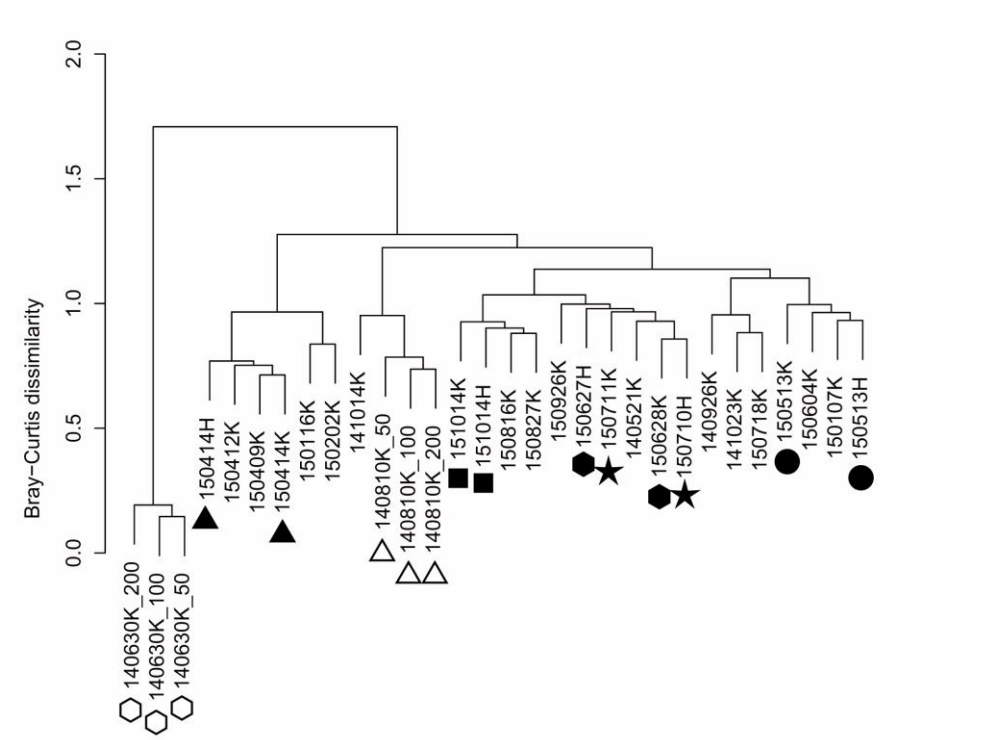

Figure 3-3 | Hierarchical clustering of precipitation samples based on OTU composition. The distance matrix was calculated based on the Bray-Curtis dissimilarity, and clusters were calculated using Ward's method. Open symbols indicate samples that were collected during the same precipitation event with different volumes. Closed symbols indicate samples that were collected on the same day at different sites (Kashiwa and Hongo).

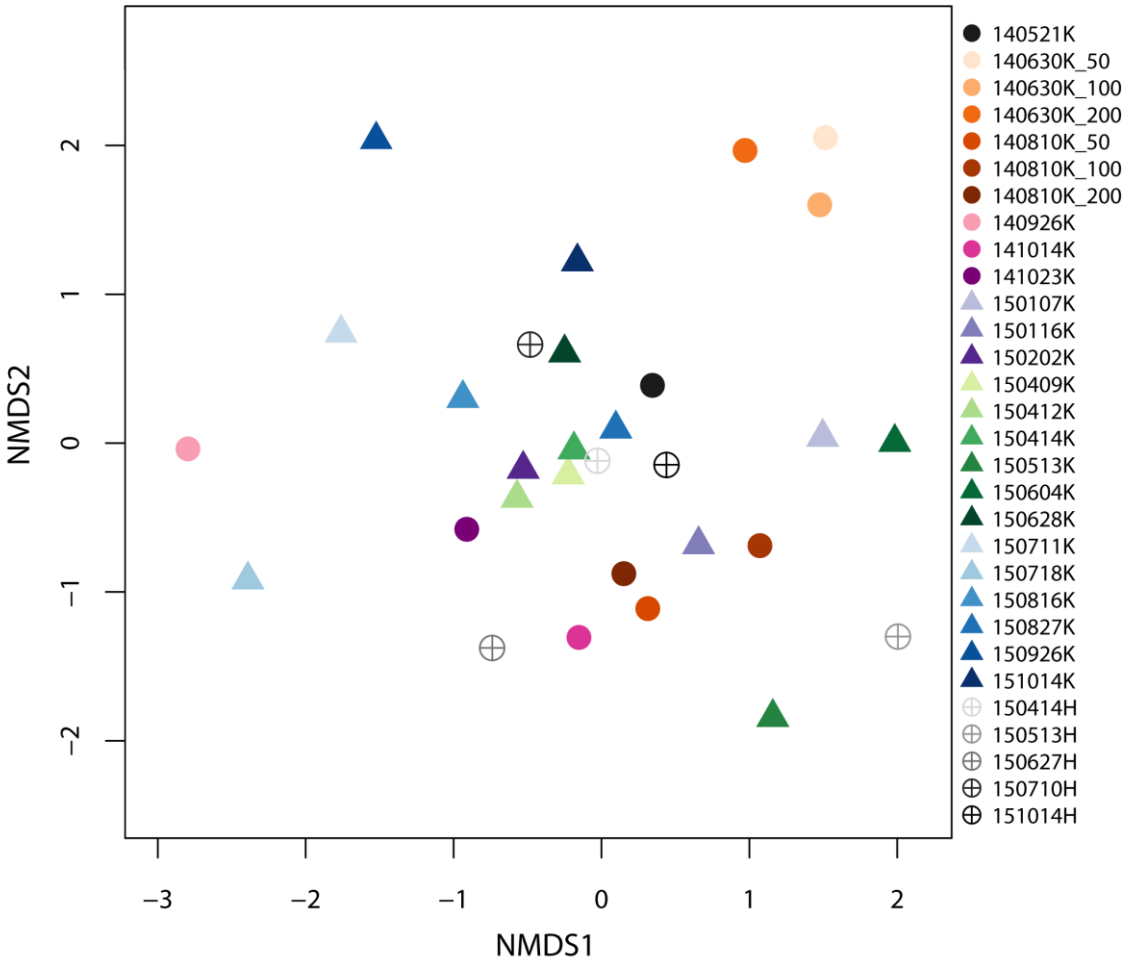

Figure 3-4 | Nonmetric multidimensional scaling plot for OTU compositions. The distance matrix was calculated based on the Bray-Curtis dissimilarity. The stress value of the final configuration was 20.46%.

Taxonomic composition of precipitation microbial communities

Among the 12,089 effective sequences, 11,994 (99.2%) were taxonomically assigned at the phylum level. Almost all sequences were assigned to 24 phyla in the domain Bacteria with the exception of 4 (0.03%) and 219 (1.7%) sequences assigned to Archaea and mitochondria, respectively. This strong bias toward bacterial sequences may reflect the actual composition but may also be attributable to amplification bias introduced by primer specificity. The top three and six most abundant bacterial phyla accounted for >80% and >95%, respectively, of the sequence pool of all precipitation samples (Fig. 3-5A). Proteobacteria was the most abundant phylum (23–88%) across all precipitation samples with the exception of the 140630K, 140926K, and 150116K samples (Firmicutes

(89–94%), Actinobacteria (50%), and Firmicutes (49%) were the most abundant phyla, respectively). A particularly exceptional microbial community dominated by Firmicutes was observed in the 140630K sample. Firmicutes, Bacteroidetes, and Actinobacteria were the other dominant phyla in the total sequence pool. In principle, these results were consistent with those of a previous study in which Proteobacteria, Firmicutes, and Bacteroidetes were the dominant phyla in precipitation samples captured in Seoul, Korea (303), whereas comparatively greater numbers of sequences were assigned to Actinobacteria, Planctomycetes, and Cyanobacteria in this study. At the class level, the abundant groups were Gammaproteobacteria, Betaproteobacteria, and Alphaproteobacteria, followed by Bacilli, Flavobacteriia, Clostridia, Actinobacteria, and Sphingobacteriia (Fig. 3-5B). Notably, the enrichment of these phyla and classes was also reported in previous studies investigating aerosolized (271, 273, 278) and cloud water microbial communities (273, 312).

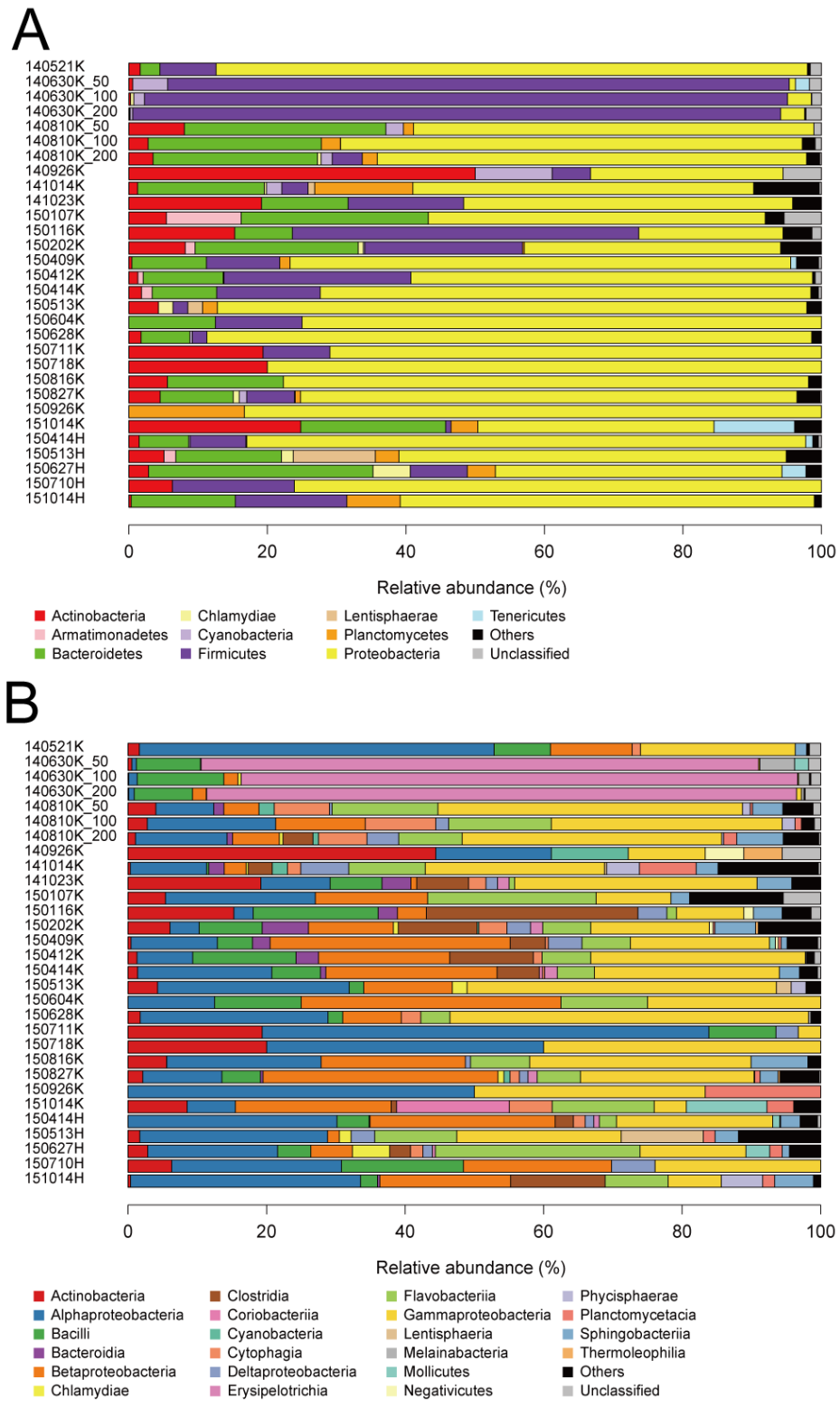

Figure 3-5 | Relative abundances of sequences at the phylum (A) and class (B) levels. Groups demonstrating <5% abundance were summarized as "Others."

Several OTUs were assigned to genera that potentially contain INA bacteria, *i.e.*, *Acinetobacter*, *Bacillus*, *Erwinia*, *Flavobacterium*, *Luteimonas*, *Microbacterium*, *Pseudomonas*, *Psychrobacter*, *Sphingomonas*, and *Stenotrophomonas* (265) (Table 3-4). I also detected several genera containing known pathogens, including typical human pathogens such as *Legionella*, *Streptococcus*, *Arcobacter, Rickettsia* and *Clostridium,* and plant pathogens such as *Erwinia*, although their abundance was low. I did not detect season-specific microbial groups in the typhoon rain, rainy season, and snow samples with statistical significance, probably partly due to small sample sizes.

Table 3-4 | Numbers of sequencing reads of OTUs that assigned to genera including microbial spices that previously reported to have ice nucleation activity (INA).

	140521K	140630K_50	140630K_100	140630K_200	140810K_50	140810K_100	140810K_200	140926K	141014K	141023K	150107K	150116K	150202K	150409K	150412K	150414K	150513K	150604K	150628K	150711K	150718K	150816K	150827K	150926K	151014K	150414H	150513H	150627H	150710H	151014H
Acinetobacter	0	0	0	0	0	9	1	0	2	2	0	0	12	3	10	28	7	0	0	0	0	8	11	0	0	11	0	1	0	1
Bacillus	0	1	0	4	0	0	0	0	0	2	0	0	5	4	7	0	0	0	0	0	2	0	0	0	0	0	4	0	0	0
Erwinia	2	0	0	0	0	0	0	0	0	1	0	0	0	3	0	6	0	0	3	1	4	0	0	0	0	2	6	0	0	0
Flavobacterium	0	0	0	0	0	1	0	0	2	0	0	0	27	14	23	1	14	0	0	0	8	12	0	0	0	23	0	0	0	14
Luteimonas	0	0	0	0	0	7	8	0	1	0	2	0	2	0	0	0	10	0	0	0	0	0	0	0	0	4	12	0	0	0
Microbacterium	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
Pseudomonas	9	1	2	1	2	13	17	0	6	26	0	1	108	59	97	94	98	8	0	0	4	11	6	1	1	6	18	0	3	4
Psychrobacter	0	0	0	0	0	0	0	0	0	0	0	1	13	10	66	16	0	0	0	0	0	0	0	0	0	8	4	0	0	0
Sphingomonas	105	1	0	0	1	3	2	0	4	0	0	0	2	10	7	23	18	2	0	0	2	26	10	2	0	8	23	0	3	5
Stenotrophomonas	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
Total	117	4	2	5	3	33	28	0	15	31	2	2	172	104	210	168	147	10	3	1	20	57	27	3	1	62	80	1	6	24

INA microbes were listed from Després, *et al.* (2012).

Seasonal and meteorological correlations

Taxonomic distribution exhibited seasonal variability (Fig. 3-5). Notably, the abundance of Proteobacteria decreased from summer to winter (p-value < 0.01, Mann-Whitney *U*-test), and a similar trend has consistently been observed in aerosolized microbial communities (276). To more closely investigate the factors underlying changes in the precipitation microbial communities, I performed a correlation analysis between meteorological characteristics and microbial composition (Fig. 3-6). The relative abundance of the order Bacteroidales negatively correlated with temperature (Spearman correlation $\rho = -0.70$, p-value < 0.01 after the Bonferroni correction). Although other correlations were not statistically significant after multiple testing correction, the amount of precipitation, wind speed, and atmospheric pressure showed tendencies of positive correlations with the abundance of the orders Cellvibrionales ($\rho = 0.59$), Cellvibrionales ($\rho = 0.58$), and Pseudomonadales ($\rho = 0.57$), respectively. Notably, the abundance of the order Legionellales, which contains several known pathogens, showed a tendency of a positive

correlation with temperature ($\rho = 0.47$), where aerosolized water is known to facilitate the dispersion of *Legionella* (313) and a warm and wet climate is associated with the incidence of Legionnaires' disease (314, 315). Although cell numbers were not measured except for one sample in this study, I note that seasonal variability in cell numbers would also be important, especially because that of atmospheric samples was reported (275, 302). Similarly, analyses with particulate matter density and O_3 and NO_3 concentrations are also envisioned, because they would substantially affect aerial microbes (273, 285, 310, 316).

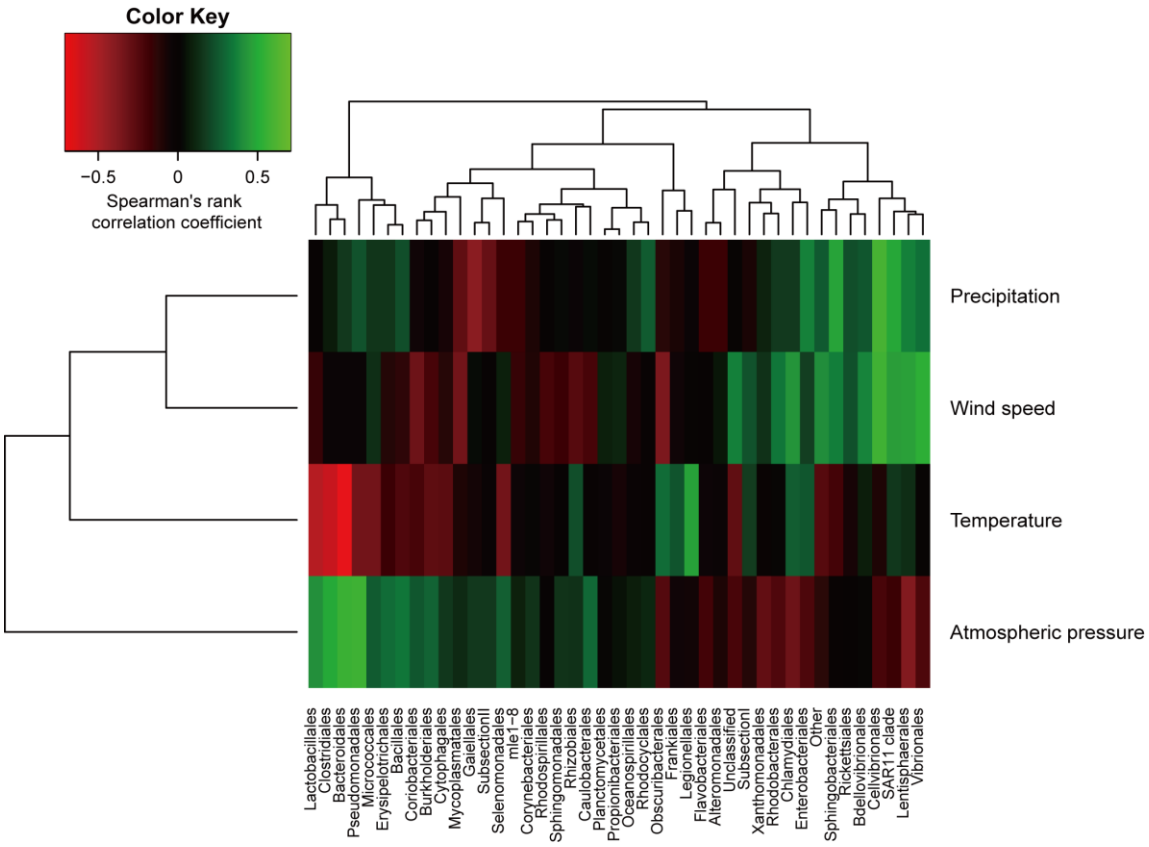

Figure 3-6 | Correlation analysis between relative abundances of sequences at the order level and meteorological data. The color scheme represents Spearman's rank correlation coefficient.

Relationship between ordinary habitats of precipitation microbes and air mass backward trajectories

To estimate the environments from which microbes in precipitation originated, I

performed a microbial habitat index analysis using MetaMetaDB (176), which is a database to estimate the ordinary habitats of microbes based on similarity searches for 16S rRNA gene sequences against amplicon-sequencing and shotgun metagenomic data in public databases. In most samples, animal-associated environments, such as gut microbiota, were estimated to be the most dominant ordinary habitats (52% on average) (Fig. 3-7 and Fig. 3-8), which is consistent with a previous study in which animal feces were the dominant source of airborne bacteria (276). Notably, marine-related environments, such as marine and marine sediment, were estimated to be relatively major ordinary habitats for several samples (*e.g.*, 65.1 and 63.1% in the 140810K and 141014K samples, respectively). Soil-related environments, such as soil and rhizosphere, were also estimated to be major ordinary habitats (11.0% on average). For comparison, I also conducted ordinary habitat analyses using amplicon-sequencing data from aerosol (311) and cloud water (273) samples. The soil-related and animal-associated environments were generally major ordinary habitats as consistent to the present results, whereas marine-related environments were not major possibly because the origins of the microbes or the sampling methods were different from those in this study (Fig. 3-9).

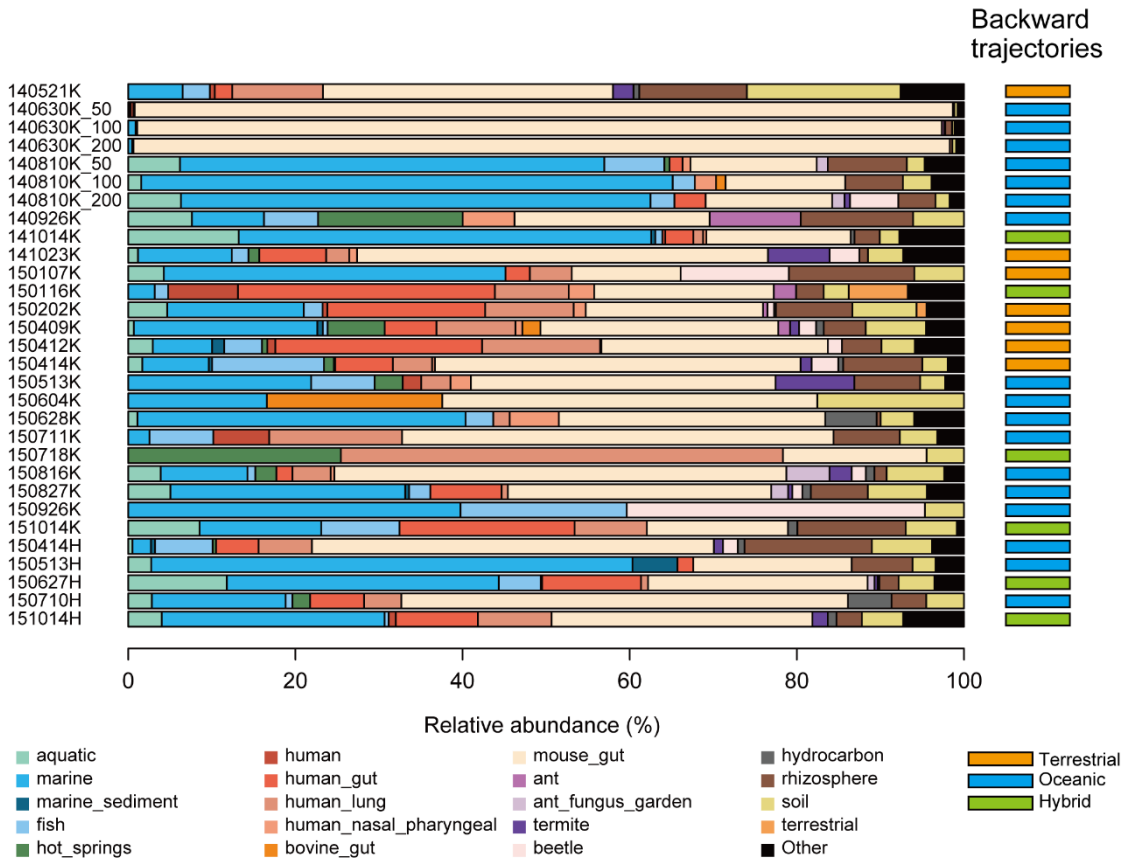

Figure 3-7 | Estimated ordinary habitats of precipitation microbes. Because the ordinary habitat for an individual 16S rRNA sequence cannot be conclusively determined, the microbial habitability index (MHI) was calculated to estimate the probability of an ordinary habitat. Estimated ordinary habitats demonstrating <5% abundance were summarized as "Others." The estimated route of the air mass before each precipitation event is indicated in the right column. The terrestrial, oceanic, and hybrid routes are colored in orange, blue, and green, respectively.

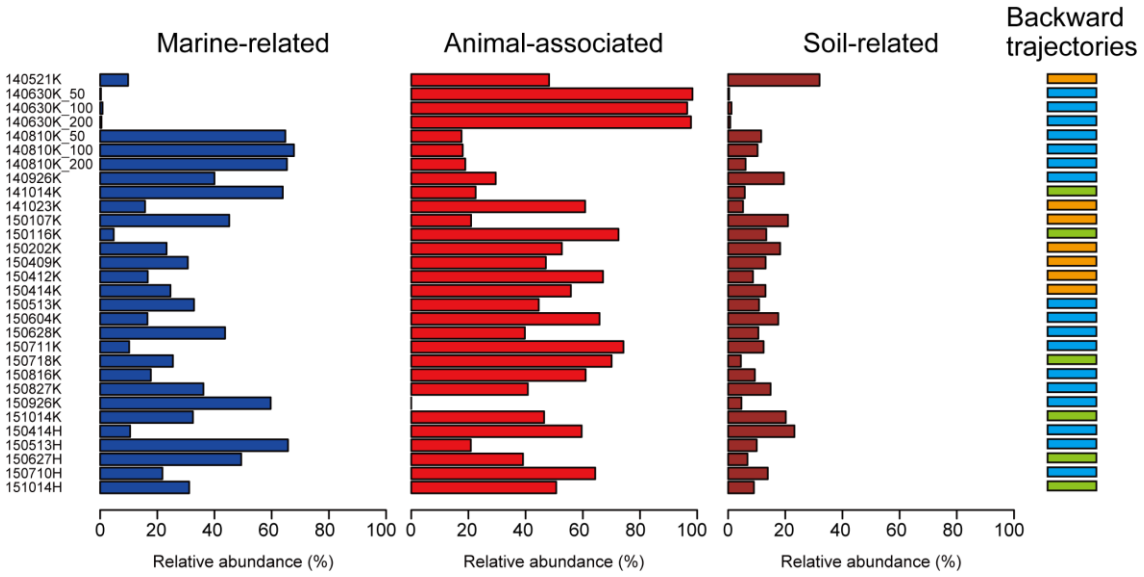

Figure 3-8 | Estimated ordinary habitats of precipitation microbes for three ecosystem groups. The abundance values in each ecosystem group are summation for habitats described below. Marine-related: “aquatic”, “marine”, “marine sediment”, “fish”, and “hot spring”; Animal-associated: “human”, “human gut”, “human lung”, “human nasal pharyngeal”, “bovine gut”, and “mouse gut”; and Soil-related: “hydrocarbon”, “rhizosphere”, “soil”, and “terrestrial.” The estimated route of the air mass before each precipitation event is indicated in the right column.

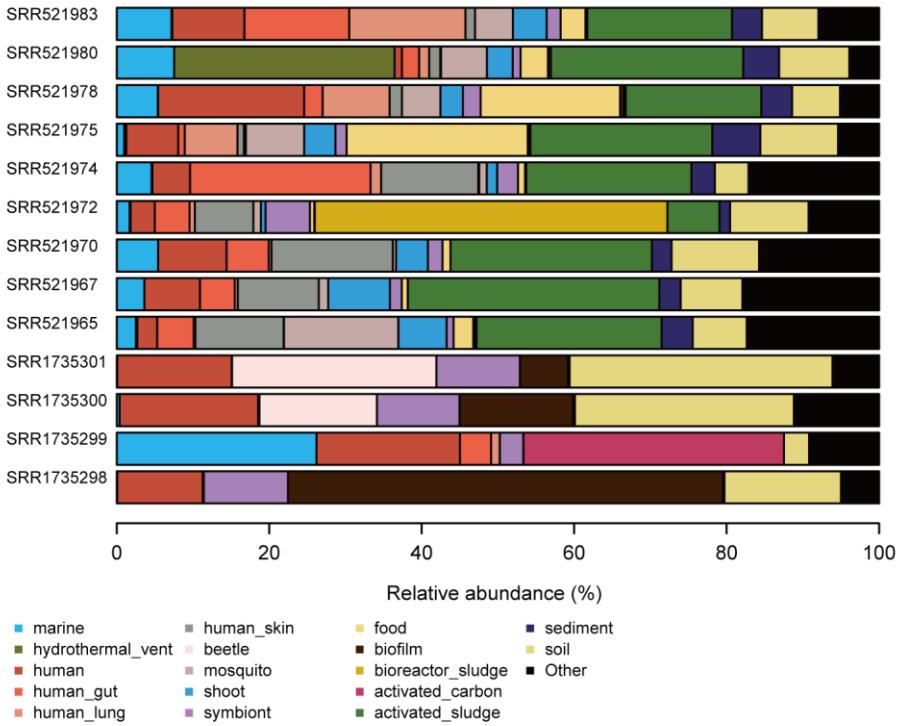

Figure 3-9 | Estimated ordinary habitats of microbes in aerosol and cloud water samples. Estimated ordinary habitats demonstrating <5% abundance were summarized as "Others."

The estimated backward trajectories of air masses that led to the precipitation events at the Kashiwa and Hongo sites were classified as terrestrial, oceanic, and hybrid routes. The terrestrial route typically originated from the middle of the Eurasian continent and passed through the East China Sea, the Yellow Sea, and the Sea of Japan; the oceanic route typically originated from the Pacific Ocean and passed through the East China Sea or the Sea of Okhotsk; and the hybrid route comprised both the terrestrial and oceanic areas. Consistent with the typical pattern of the seasonal winds in Asia, the terrestrial and oceanic routes dominated in winter and summer, respectively (Fig. 3-7 and Fig. 3-8). The estimated ordinary habitats of the precipitation microbes showed agreement with the estimated air mass backward trajectories. For example, Planctomycetes, which contains several aquatic microbes (317), was frequently found when the backward trajectories followed oceanic routes (Fig. 3-5A and Fig. 3-7). PERMANOVA analysis showed a significant relationship between the routes and the estimated composition of ordinary microbial habitats (p-value < 0.05). Notably, the ratios of marine-related environments dominated when the air masses originated from the oceanic route, and animal-related environments dominated when they originated from the terrestrial route. Shannon's diversity indices of

microbes became larger when the air masses originated from the terrestrial route (Shannon's diversity indices were 3.74 ± 0.68, 3.05 ± 1.00, and 3.15 ± 1.36 for the terrestrial, oceanic, and hybrid routes, respectively. The index of each sample is shown in Table 3-1); however, it should be noted that some samples required additional sequences to reach plateaus of rarefaction curves as mentioned already.

Soil, oceanic, and animal-associated microbes are spread in the atmosphere and transported for long distances (292, 318), and precipitation may facilitate this microbial cycle. Sea-living microbes are emitted into the atmosphere via the bursting of bubbles on waves (319), whereas soil-living and animal-associated microbes are transported on soil dust (279, 280, 320, 321). In high-altitude atmospheric environments, microbes may be under substantial selection pressure due to harsh chemical, physical, and nutrient conditions (266, 322, 323). INA microbes play roles in cloud formation (322) and may facilitate the return of aerial microbes to diverse environments. The dispersal of pathogenic microbes causes disease epidemics that threaten public health and agricultural plant and animal health (281–283, 324). Continuous long-term monitoring and large-scale analysis of precipitation microbes is thus envisioned to reveal the full impact of atmospheric microbial transport on microbial ecology, microbial evolution, public health, and climate.

Amplicon sequencing of mock precipitation samples

My collaborators observed a microbial cell density of 8.36×10^3 cell/mL via direct cell counting of a precipitation sample. This number did not differ greatly from densities observed in previous studies of precipitation microbes (289, 300, 325) and was much smaller than the densities of other typical marine (10^6–10^9 cells/mL) (326), soil (10^7–10^{11} cells/g) (327), and human gut samples (10^{11}–10^{12} cells/g) (328). A 50-mL mock precipitation sample was prepared that contained ten bacterial strains at this total density and conducted amplicon-sequencing analysis of their 16S rRNA genes. A sufficient DNA library was produced (10.1 ng/L), and 36.6% of the pyro-sequenced reads were taxonomically assigned (Table 3-5). The existence of seven of the ten strains was verified, but the other three were not. In addition, read number biases were observed as previously reported in amplicon-sequencing studies investigating low-density microbial communities (304, 305, 329–333).

Table 3-5 | Numbers of sequencing reads and assigned taxa retrieved from the artificial mock precipitation sample using this experimental protocol.

Families assignment	Samples			Genus assignment
	C1	C2	C3	
Acanthopleuribacteraceae	8	1	8	***Acanthopleuribacter***
Bacillaceae	73	93	52	***Bacillus***
Staphylococcaceae	3	0	0	***Staphylococcus***
Methylobacteriaceae	0	1	2	*Methylobacterium*
Sphingomonadaceae	1	1	0	*Sphingomonas*
Burkholderiaceae	2	0	0	***Burkholderia***, *Ralstonia*
Pseudoalteromonadaceae	4	0	0	***Pseudoalteromonas***
Enterobacteriaceae	455	176	401	*Buttiauxella, Citrobacter, Enterobacter, Escherichia-Shigella, Kluyvera, Pantoea, Rahnella, Tatumella*
Pseudomonadaceae	11	34	11	***Pseudomonas***
Vibrionaceae	59	34	47	*Aliivibrio, Photobacterium,* ***Vibrio***

The analysis of microbial communities using samples with low cell densities and limited volumes remains a challenge. In addition to precipitation samples, deep-sea sediment core samples below the ocean floor (334), ice core samples in polar regions (335), and hot spring samples (336) also share these characteristics. Notably, microbial cell density in the atmosphere is low (10^4–10^6 cell/m^3) (267, 275, 320), but millions of liters of air can be collected and condensed (337). My results obtained with mock precipitation samples suggest several limitations should be considered when interpreting the amplicon-sequencing analysis of low-density microbial communities; some microbial groups may be missing or over-/under-represented due to DNA extraction and sequencing biases. Another important problem is contamination by experimental reagents during sample preparation and DNA sequencing (338–340). Regardless of my and collaborators efforts to prevent contamination, I observed sequences from several genera that were not part of the mock community, indicating the importance of conducting negative control experiments. Further technical improvements are envisioned to elucidate precipitation and other low-density microbial communities.

Conclusion

Microbes are present nearly everywhere in the Earth, even in precipitation from the sky. Precipitation is supposed to make microbes in the atmosphere finally fall down to the ground surface. In this study, I thoroughly observed microbial communities in precipitation samples that were collected over one year in the Greater Tokyo area, Japan. To my knowledge, this is the first amplicon-sequencing study investigating precipitation microbial communities involving sampling over the duration of a year. Most importantly, the results suggest seasonal variations in the microbial communities in precipitation, and

their community structures were significantly associated with the estimated air mass trajectories. These results highlight importance of precipitation in long-range microbial immigration via the atmosphere, which may answer how tiny microbes can dynamically travel around the globe.

Chapter 4: Culture-independent metagenomic and metaepigenomic analysis of prokaryotes in Lake Biwa, Japan.

Introduction

DNA methylation is a major mechanism of epigenetic modification that is found in the genomes of diverse prokaryotes (341). One of the main roles of prokaryotic DNA methylation is sequence-specific restriction-modification (RM), which protects host cells from invasion by extracellular DNA, such as phage infection (342). Microbes produce several heterogeneous proteins that contain restriction endonucleases (REases) and methyltransferases (MTases). MTases methylate host DNA to protect against digestion by REases, while infected unmethylated DNA is rapidly recognized by REases and degraded. This system for chemical modification of DNA serves a wide variety of biological functions in prokaryotes, including gene expression regulation, chromosome replication, cell cycle regulation, anti-mutagenesis, and mismatch repair (343–347). The sequence specificity of the RM system can easily be altered, and they frequently act as mobilome components, suggesting that the system contributed to genome evolution in prokaryotes (348, 349). Research interest in the field of prokaryotic methylation systems has grown, as has our understanding of fundamental microbiological processes, including microbe adaptability and disease pathogenicity (345, 350). However, community-scale observation of epigenomic characteristics in environmental microbes has been prevented by experimental limitations, although a pioneering study investigated methylation patterns in the sediment community (351).

The recent development of single molecule real-time (SMRT) sequencing technology allows us to obtain DNA methylation information easily. SMRT sequencing can identify the three main types of prokaryotic DNA methylation: N6-methyladenine (m6A), 5-methylcytosine (m5C), and N4-methylcytosine (m4C) (187, 188). Recent studies have reported the pervasive presence of DNA methylation in diverse culturable prokaryotes (189, 352) using this technology. However, little is known about methylation patterns in environmental microbial communities that are typically dominated by unculturable members. Direct observation of methylation patterns in whole microbial communities, known

as metaepigenomic analysis, can provide fundamental knowledge of prokaryotic biology and ecology in the environment (353).

The implementation of SMRT sequencing can also provide ultra-long sequencing reads, up to 60 kb (354). The long read length permits reconstruction of near-complete high-quality microbial genomes from metagenomic shotgun sequencing data, which greatly facilitates genomic analysis, such as gene functional annotation and comparative genomics (355). However, the high rates of base call error (~15%) in SMRT sequencing reads can negatively affect downstream analyses (356, 357). Because the technology is insensitive to various context-specific biases (*e.g.*, GC biases, highly repetitive regions), the circular consensus sequence (CCS), which is an error-corrected consensus read derived from multiple alignment consensuses of subreads belonging to the same single-molecule circular sequence, can improve read accuracy. Recent studies have applied the SMRT CCS technique using full-length 16S rRNA amplicon sequencing for phylogenetic profiling (358) and transcriptome sequencing for determining complete isoform sequences (359). In contrast, expectations for application of long-read sequencing to metagenomic analysis are high (192), few papers have reported applications for reconstructing dominant bacterial genomes (360, 361).

Here, I conducted metagenomic and metaepigenomic analysis using long CCS reads generated using the PacBio Sequel platform to profile the freshwater microbiome of Lake Biwa, the largest lake in Japan. Freshwater habitats are rich in phage-prokaryote interactions, and genetic exchanges are frequent (362). Although many efforts have been made to evaluate the phylogenetic diversity of both phages and prokaryotes (363, 364), little is known about their relationship, for example, variation in RM systems among prokaryotes and their efficacy against phage infections. This proof-of-concept study performed metagenomic analysis using shotgun SMRT sequencing reads to reveal the genetic and epigenomic characteristics of a microbial community that is dominated by unculturable members.

Materials and methods

Sample collection

Water samples were collected at a pelagic site (35°13′09.5″N 135°59′44.7″E) in Lake Biwa, Japan (Fig. 4-1) on December 26, 2016. The sampling site is approximately 3 km from the nearest shore, with a maximum depth of 73 m. The vertical profiles of temperature, dissolved oxygen, and chlorophyll concentration were measured using a

conductivity, temperature, depth (CTD) probe in situ. The lake has a permanently oxygenated hypolimnion and was thermally stratified when the sampling was carried out (Fig. 4-2).

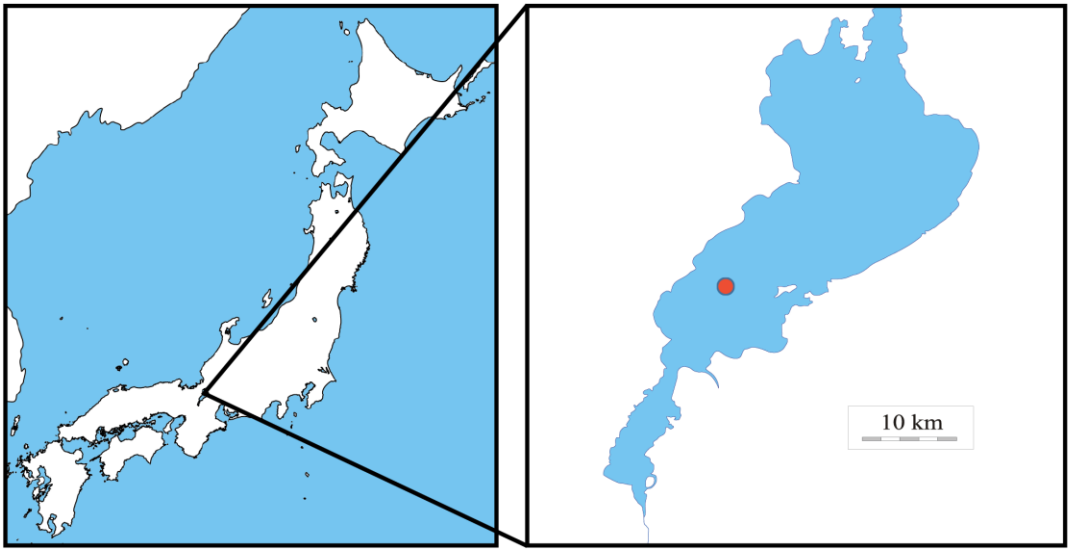

Figure 4-1 | A map of the sampling site. Fresh water was sampled at depths of 5 m and 65 m.

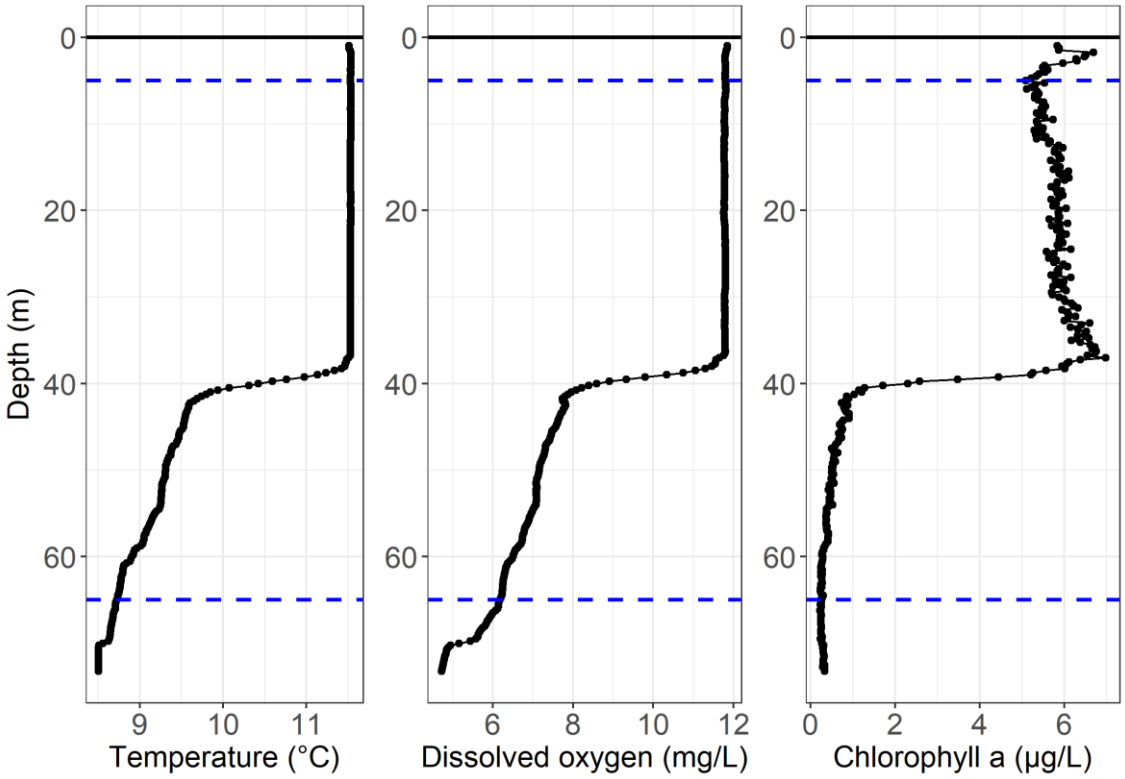

Figure 4-2 | Vertical profiles (water temperature, dissolved oxygen, and chlorophyll a) of the sampling site. The two dotted lines in each figure represent depths where fresh water was collected (5 m and 65 m).

Water sampling was conducted at depths of 5 m and 65 m, immediately above and below the thermally stratified layer, respectively. Water was collected in prewashed 5- L Niskin bottles and then transferred into sterile bottles. I confirmed that the study did not involve endangered or protected species and the sampling required no special permission. All equipment that came in direct contact with the water samples was either sterilized by autoclaving or disinfected with hypochlorous acid solution. The water samples were immediately transported to the laboratory and cell-capture was performed. Approximately 30 L of water samples were prefiltered through 5-µm membrane PC filters (Whatman, UK), and microbial cells were collected using 0.22-µm Sterivex filters (Millipore, USA) and immediately stored at −20°C in a refrigerator until subsequent analysis. The sampling was conducted by the members of the Center of Ecological Research, Kyoto University.

DNA extraction and SMRT sequencing

The microbial DNA captured on the Sterivex filters was retrieved using a PowerSoil DNA Isolation Kit (QIAGEN) according to the supplier's protocol with slight modifications; the filters were removed from the filter container, cut into 3-mm fragments, and directly suspended in the extraction solution in the kit for cell lysis. The bead beating time was extended to 20 minutes in order to yield sufficient quantities of DNA for SMRT sequencing, with reference to Albertsen *et al.* (365). SMRT sequencing was conducted using a PacBio Sequel system (Pacific Biosciences of California, Menlo Park, CA) in two independent runs, as two technical replicates, according to the manufacturer's standard protocols. SMRT libraries were prepared with a 4-kb insertion length and two SMRT cells were used for each sample. All SMRT libraries were prepared and sequenced at the National Institute of Genetics (Mishima, Japan).

Bioinformatic analysis

Subreads that were sequenced in at least three full passes and had >97% minimum basecall accuracy were retained for generating highly accurate consensus sequences for the CCS (referred to as CCS reads) using the standard PacBio SMRT software package. Through this step, mismatches that occurred in only one read, as well as insertions and deletions (which are more common in SMRT sequencing reads), were discarded. Kraken (69) and Kaiju (70) were used for taxonomic assignment of CCS reads. Classification using Kraken (69) was performed using complete prokaryotic genomes from RefSeq (63) for comparison, using the default parameters. Classification using Kaiju (70) was conducted against protein sequences from the NCBI BLAST nr database (366) in "Greedy-5" mode. CCS reads that encoded potential 16S rRNA genes were first extracted using SortMeRNA (226). RNAmmer (217) was used to extract 16S rRNA sequences from CCS reads. The 16S rRNA sequences were taxonomically assigned based on sequence similarity using blastn (213) against the SILVA database (54) and top hit sequences with e-values ≤1E-15 were retrieved.

CCS reads were *de novo* assembled using Canu (367) and Mira (368) individually. Because CCS reads were used for genome assembly, the -pacbio-corrected setting in Canu was used, and the Mira settings for PacBio CCS reads were employed, according to the instructions provided. After exclusion of repeated contigs, the remaining assembled contigs were taxonomically binned using MetaBAT (103) based on genome abundance and tetra-nucleotide frequencies as genomic signatures. The coverage of each genome

was calculated by mapping of CCS reads using BLASR (369). The quality of all genome bins was assessed using CheckM (370). The CheckM is a tool for estimating the completeness and contaminations of each genome bin based on co-located sets of marker genes conserved across wide bacterial and archaeal genome (370). RNAmmer (217) was used to extract 16S rRNA sequences from the contigs. The extracted 16S rRNA sequences were taxonomically annotated as described above. All bins were taxonomically assigned using 16S rRNA similarity when the gene was predicted to be present in the genome bin. Otherwise, each contig was taxonomically assigned using CAT (371) or Kaiju (70) and the most frequently identified linage was regarded as the representative one. Coding sequences (CDSs) were predicted using Prodigal (212) and functional annotations were based on GHOSTZ (372) searches against the eggNOG (215) and Swiss-Prot (214) databases with a cut-off e-value ≤1E-5, and HMMER (373) searches against Pfam (374) with same cut-off e-value. A maximum likelihood (ML) tree of genome bins was constructed on the basis of the set of 400 conserved bacterial marker genes using PhyloPhlAn (222). Prophage sequence regions within genome bins were predicted using PHASTER (375) and sequence alignment of prophages was conducted using LAST (376). CRISPR arrays were predicted using the CRISPR Recognition Tool (377) and *Cas* genes were annotated using 101 known CRISPR-associated genes obtained from TIGRFAM (378) and HMMER (373) searches with an e-value ≤1E-5.

Methylation motif analysis and RM system identification

DNA chemical modification detection and motif analysis were performed using BaseMod, which is an official method for PacBio DNA modification sequence analysis. Briefly, raw sequencing reads were mapped on assembled contigs using BLASR (369) and interpulse duration (IPD) ratios were calculated for DNA modification motif identification. These motifs represent the recognition sequences of active MTase (188). To obtain reliable motifs, possible misidentified motifs were removed; motifs with low presence (<50) or showing a low motif methylation fraction (<1%) in the genome bin were excluded from further analysis.

MTase genes were annotated using Blastp (213) against an experimentally-confirmed gold standard dataset from the Restriction Enzyme Database (REBASE), a comprehensive enzyme database of RM systems (379), with a cut-off e-value ≤1E-15. The sequence specificity of each putative MTase gene was predicted based on significant similarity to reference MTase genes from REBASE. An ML tree of MTases was subjected to

multiple alignment using ClustalW (380) with the default settings. An ML tree was generated by MEGA 7 (381) using the LG substitution model that incorporates the gamma distribution of the dataset (LG+G), the AIC-selected model, and 100 bootstrap replicates.

Data deposition

The sequence data were deposited in the DDBJ Sequenced Read Archive under the accession numbers DRX114265-114268. All data were registered under BioProject ID PRJDB6656.

Results and discussion

CCS reads quality assessment

PacBio Sequel returned a total of 9.6 Gbp (2.6 million subreads) and 6.4 Gbp (2.0 million subreads) of sequencing subreads from the biwa_5m and biwa_65m samples, respectively (Table 4-1). Following the circular consensus analysis, 168,599 and 117,802 CCS reads remained, respectively. The average length of the CCS reads from the two samples was 4,474 ± 931 and 4,394 ± 587 bp, respectively (Fig. 4-2). Although base quality score declined slightly with read base position, the read quality scores exceeded a 20 Phred quality score of >90% at each position (Fig. 4-3), suggesting a low rate of base call errors in the CCS reads. The resulting high-accuracy long CCS reads were therefore used for subsequent downstream analyses.

Table 4-1 | General features of the metagenomic sequences.

Sample	biwa_5m	biwa_65m
Sequenced reads	850,494	688,436
Total base pairs (bp)	9,570,723,004	6,419,717,083
CCS reads	168,599	117,802
Read length (bp)	4,474 ± 931	4,394 ± 587
Total base (bp)	754,416,328	517,663,806
16S rRNA	170	106
Length (bp)	1,491 ± 64	1,468 ± 104

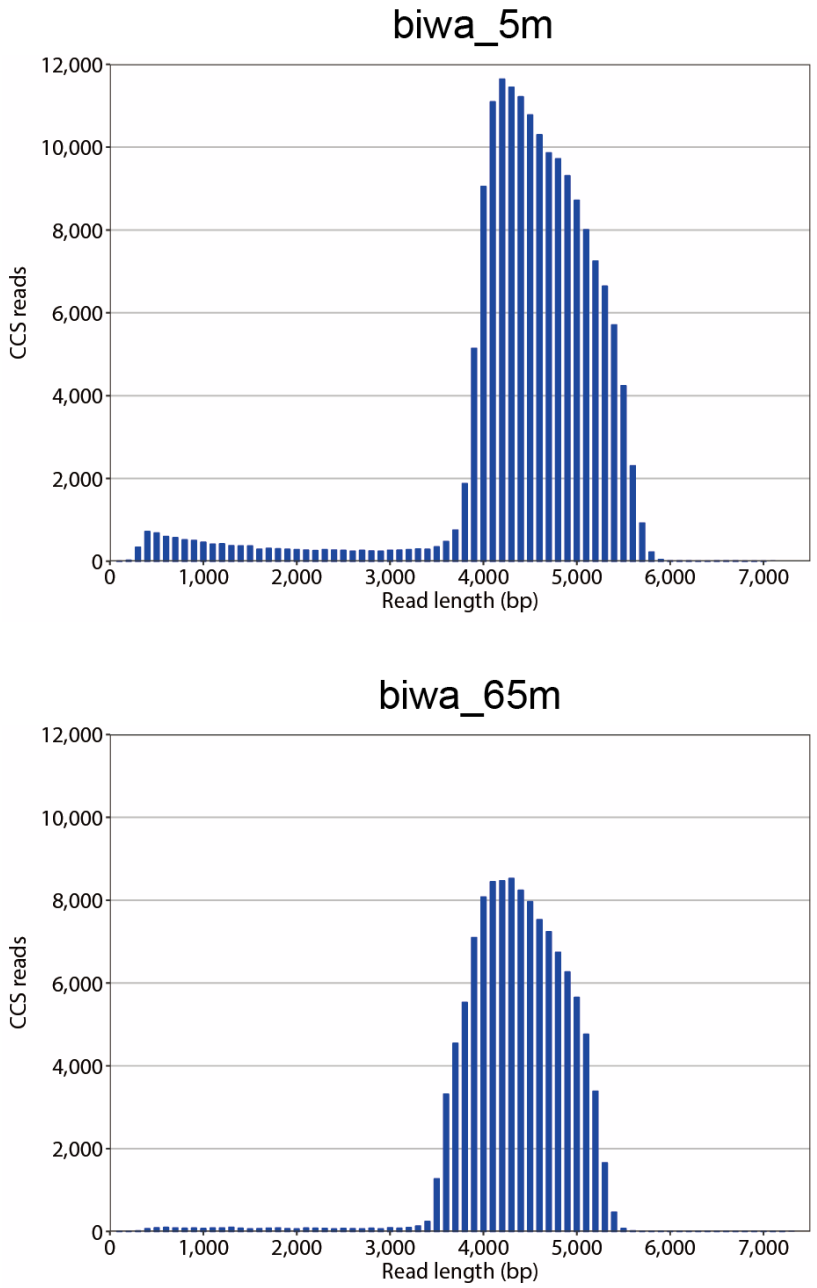

Figure 4-3 | Frequency distribution of the length of the circular consensus sequence (CCS) reads with at least 3-fold coverage and a quality score >97%. The SMRT libraries were size-selected for a 4-kbp length. CCS sequence counts were binned in 100-bp increments.

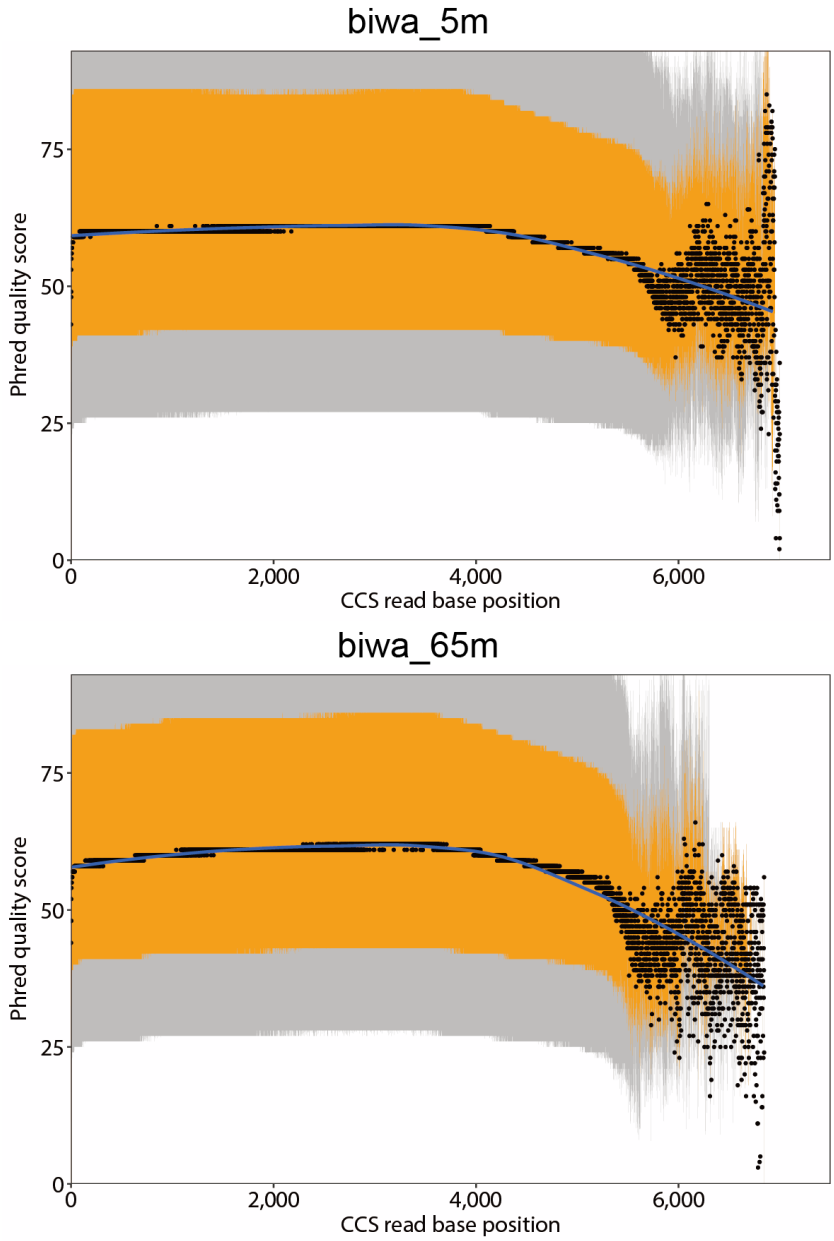

Figure 4-4 | Base quality scores of CCS reads. Outer whiskers (gray region) represent the 10th to 90th percentiles of position quality scores. Inner whiskers (orange region) represent the 25th to 75th percentiles. Dots are the mean quality score at each base position. The lines show a fitted LOESS model.

Diversity of microbial taxonomy

Taxonomic assignment of CCS reads was performed using two different tools; Kraken (69) with complete prokaryotic genomes from RefSeq (63) (Fig. 4-5); and Kaiju (70) with the NCBI BLAST nr protein sequence database (366) (Fig. 4-6). Moreover, 16S

rRNA sequences were used for a blastn similarity search against the SILVA database (382) (Fig. 4-7). The assignment ratio was >88% at the phylum level and >56% at the genus level when using Kaiju. In contrast, only 29% of CCS reads could be assigned to phyla when using Kraken (Table 4-1), likely due to the lack of genomic data for freshwater microbes in RefSeq. The ratios of assigned reads were higher than previously reported (70), perhaps due to the long read length. The protein-based taxonomic assignment analysis showed that the two samples had similar compositions (Fig. 4-5) and analyses using two other strategies showed similar tendencies (Fig. 4-6, 4-7). The results of Kaiju analysis and 16S rRNA sequence analysis showed similar compositions (Fig. 4-5, 4-7), likely indicating high assignment accuracy. Thus, the Kaiju results were used for subsequent analyses.

At the phylum level, Proteobacteria were dominant in both samples, followed by Actinobacteria, Firmicutes, and Bacteroidetes. Nitrospirae, Chloroflexi, and Thaumarchaeota were abundant in the deep water sample, consistent with previous findings (383, 384). The number of archaea was negligible in both samples (0.6 and 6.9% in the biwa_5m and biwa_65m samples, respectively), but archaea were more abundant in the deep water sample than the shallow water sample. Thaumarchaeota was the most abundant archaeal phylum in the deep water sample (6.1%), followed by Euryarchaeota (0.6%), Candidatus Pacearchaeota (0.04%), and Crenarchaeota (0.02%). The overwhelming abundance of Thaumarchaeota in the hypolimnion was consistent with a previous study (384). Although viruses and eukaryotes were more abundant in surface water (Fig. 4-5), they were relatively less abundant than bacteria, because the filter size range (5–0.2 μm) was not suitable for most viruses and eukaryotic cells. The dominant eukaryotic phylum was Opisthokonta (2.68 and 0.92%), followed by Alveolata (1.67 and 0.45%) and Stramenopiles (1.45 and 0.15%). Among viruses, Caudovirales and Phycodnaviridae were the most abundant families in both samples. Phycodnaviridae mainly infect eukaryotic algae, while Caudovirales are known to be bacteriophages. The third most abundant viral family was Mimiviridae, eukaryophages that are known as “Megavirales” owing to their large genome size (0.6–1.3 Mbp) (385, 386). Phages without double-stranded DNA (*i.e.*, single-stranded DNA and RNA phages) were not included in this study owing to the experimental method used. In general, the results were consistent with those of previous studies that reported the microbial component in freshwater lake environments. Notably, previous studies reported that community compositions predicted using SMRT sequencing reads show good concordance at the genus level with those using current short read technologies, such as Illumina MiSeq and HiSeq (387, 388).

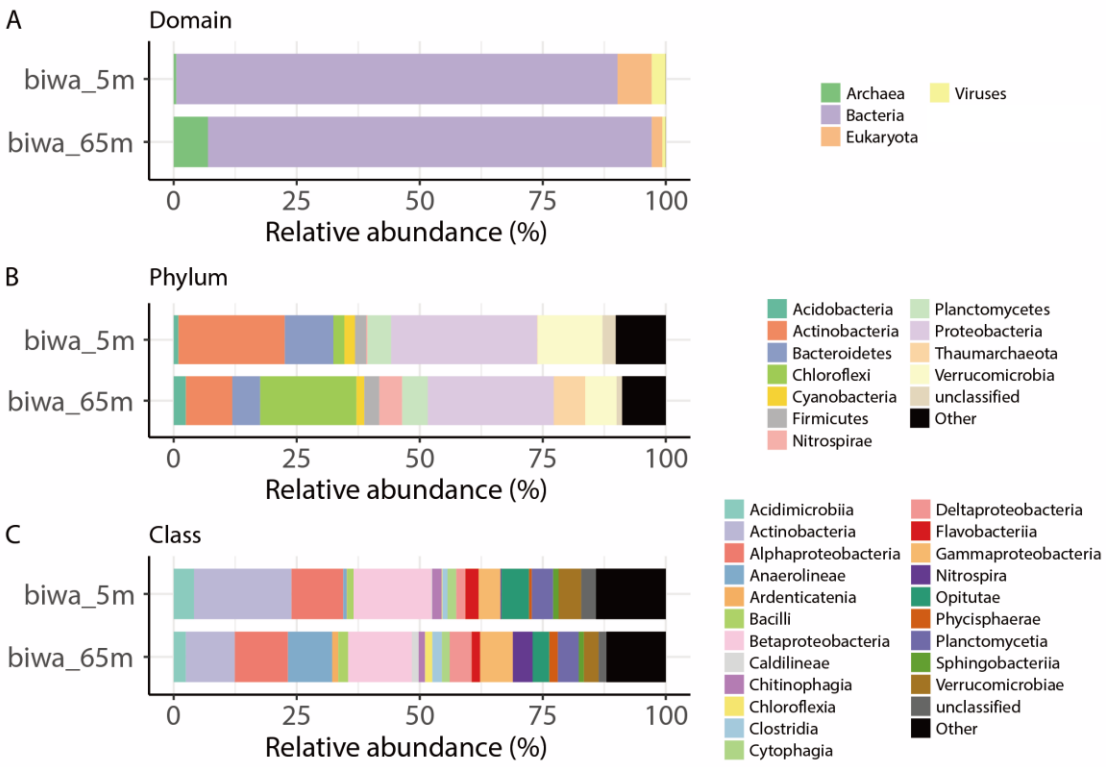

Figure 4-5 | Relative abundances of sequences at the domain (A), phylum (B), and class (C) levels predicted using Kaiju with the NR database. The ratio of bacterial and archaeal taxa are shown in (B) and (C). Groups with <1% abundance were grouped as "Others" in (B) and (C).

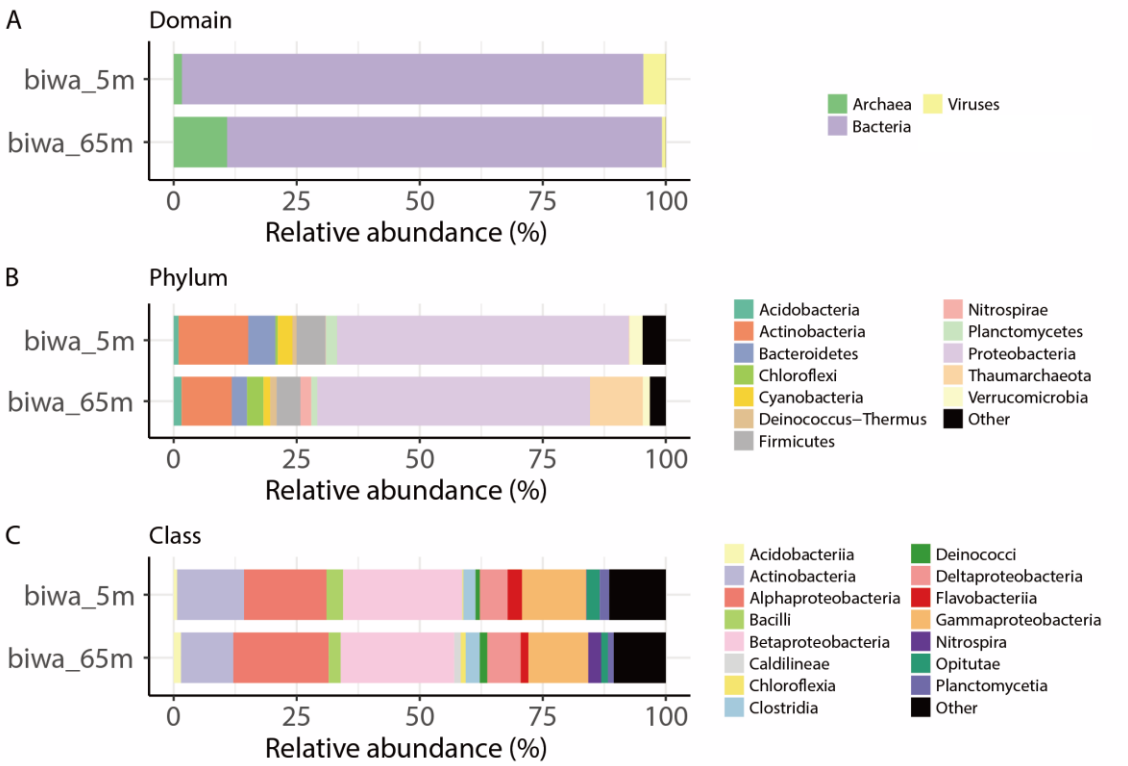

Figure 4-6 | Relative abundances of sequences at the domain (A), phylum (B), and class (C) levels predicted using Kraken with reference to archaea and bacteria in the RefSeq database.

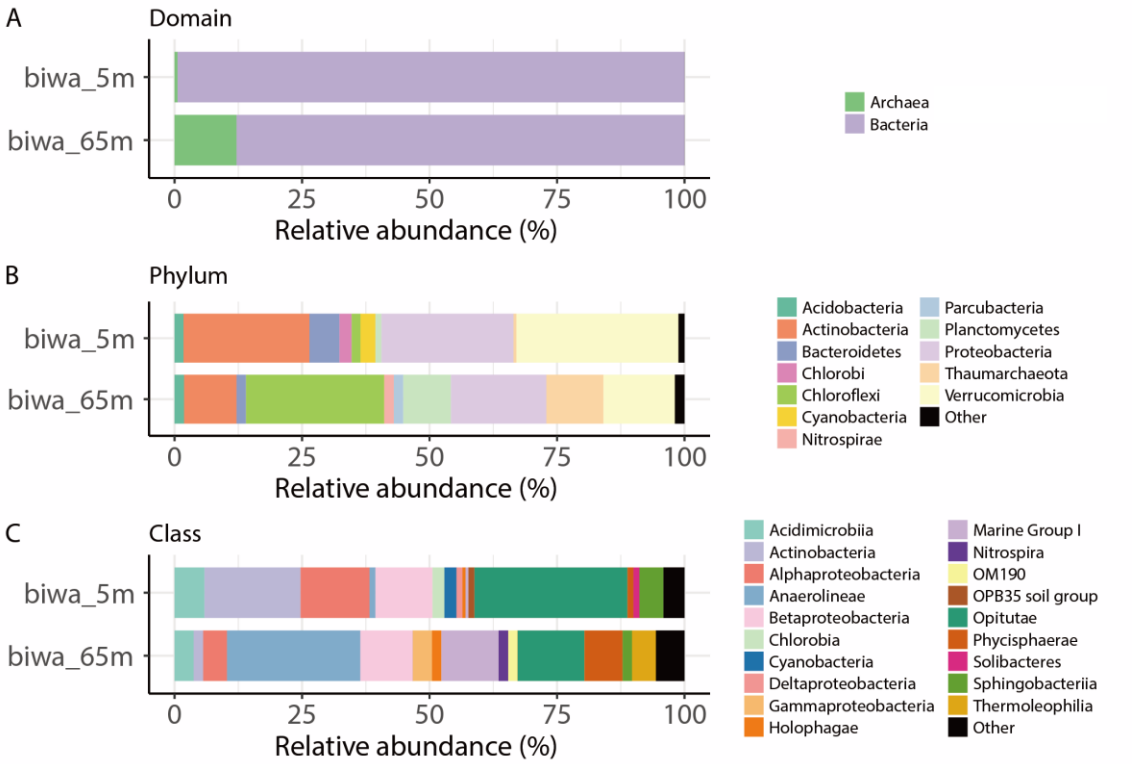

Figure 4-7 | Relative abundances of 16S rRNA sequences at the domain (A), phylum (B), and class (C) levels. The 16S rRNA sequences were extracted from CCS reads and taxonomically assigned using blastn sequence similarity search against the Silva database.

Metagenomic assembly and genome binning

Using Canu (367), the subreads from the biwa_5m and biwa_65m samples were assembled into 511 and 323 contigs, respectively (Table 4-2, Fig. 4-8). The N50 values were 101 and 83 kb, and the longest contigs from the biwa_5m and biwa_65m samples reached 549 and 740 kbp, respectively. The contigs were much longer than those previously reported from an active sludge microbial community (361). Although I also used Mira (368) for metagenomic assembly, per the methods published in a previous study (361), the longest contigs (148 and 151 kbp, respectively) and N50 values (19 and 18 kbp, respectively) obtained using this method were lower. Therefore, the contigs assembled using Canu were used for subsequent analyses.

Table 4-2 | Statistical analysis of metagenomic assembly and genome binning.

	Total length (bp)	Contigs	Longest length (bp)	Average length (bp)	N50 (bp)	Bins
Biwa_5m	22,609,702	554	481,299	40,812	83,238	16
Biwa_65m	10,687,383	345	739,933	30,978	75,701	6

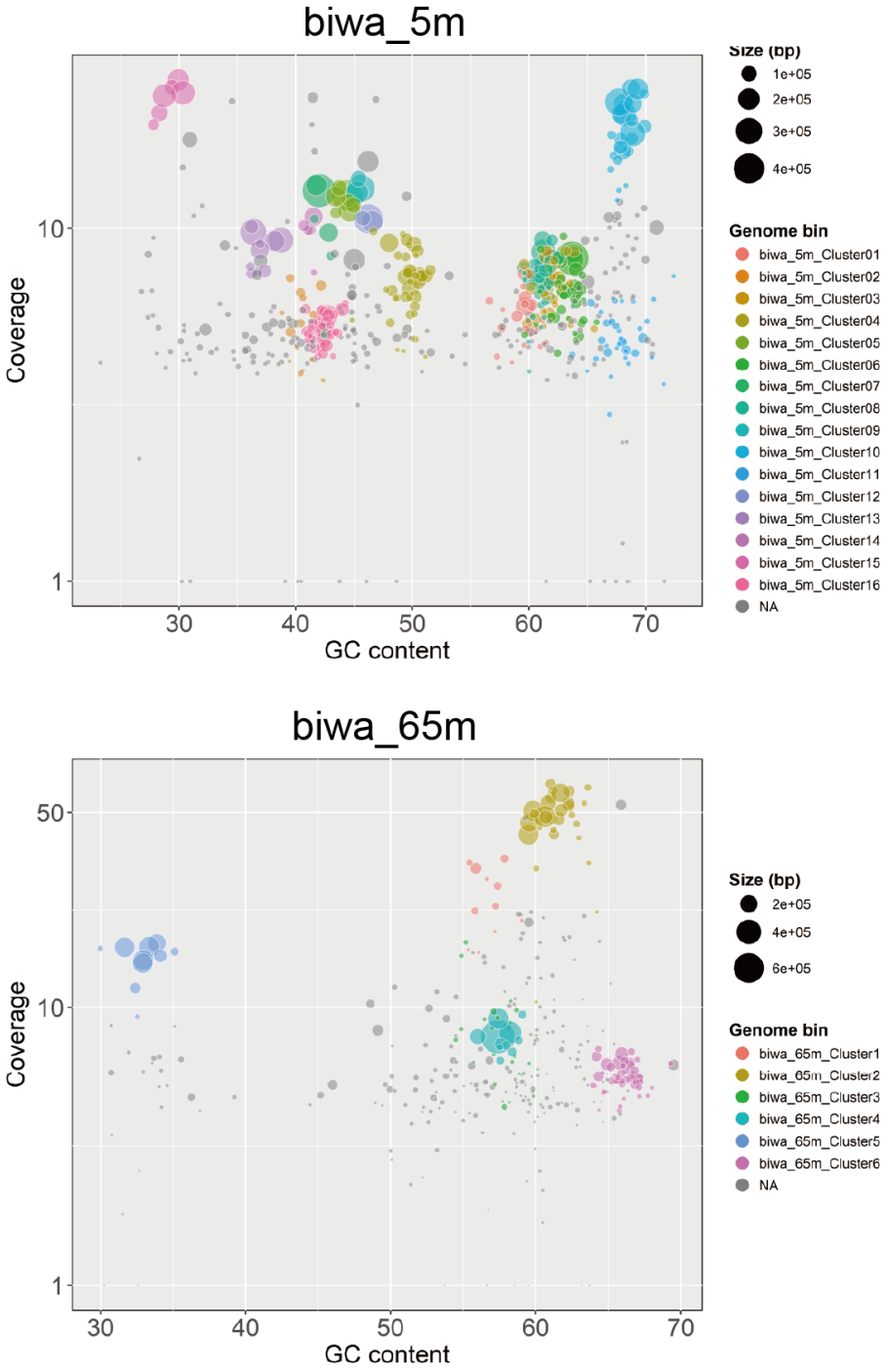

Figure 4-8 | Visualization of GC%, coverage, and size of assembled contigs generated from CCS reads. Contigs are colored based on genome bins. Contigs which not assigned any bins were grouped as “NA”.

Discrete genome bins were reconstructed using MetaBAT, a reference-independent binning tool based on genome abundance and tetranucleotide frequency. This binning analysis generated fifteen and four bins from the biwa_5m and biwa_65m samples, respectively (Table 4-3, Fig. 4-8). Although the coverage depths of each genome bin were generally low (11.3× on average), I obtained high-quality draft genomes for which the estimated completeness ranged from 16–99% (67% on average), and contamination was found to be extremely low (<3%), suggesting low heterogeneity within each genome bin. Despite the phylogenetic diversity of the microbial community (Fig. 4-5), few individual genomes were successfully reconstructed, likely due to the low numbers of CCS reads. From a total of 899 contigs, 425 were assigned to bins, representing an integration rate of 47.3%. The estimated genome size, based on total bin size and estimated genome completeness, ranged from 1.0–5.6 Mbp. The GC content ranged from 29–68% and the average N50 was 24 kbp, with a maximum of 1.67 Mbp.

Table 4-3 | Summary of contig bins.

Bin	Lineage	Estimated genome size (Mb)	Contigs	N50 (bp)	GC content (%)	Completeness (%)	Contamination (%)	16S rRNA	CDSs	Coverage	Prophage	Toxin-antitoxin system
biwa_5m_Cluster1	Bacteria;Chloroflexi[1]	2.24	21	64,528	59.5	30.6	0.0	0	751	5.79	0	0
biwa_5m_Cluster2	Bacteria;Actinobacteria[1]	1.57	13	28,617	40.6	16.9	0.0	0	363	5.13	1	0
biwa_5m_Cluster3	Bacteria;Chloroflexi;Anaerolineae;Anaerolineales;Anaerolineaceae; uncultured;uncultured Crater Lake bacterium CL500-11	3.35	36	58,996	61.8	49.1	0.0	1	1,646	6.91	0	0
biwa_5m_Cluster4	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;Acidimicrobiaceae;CL500-29 marine group;	2.31	40	61,750	49.8	76.8	1.3	1	2,066	6.67	2	1
biwa_5m_Cluster5	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Sporichthyaceae; hgcI clade; uncultured Clavibacter sp.	1.51	8	190,417	44.2	71.6	0.0	1	1,209	10.02	2	0
biwa_5m_Cluster6	Bacteria;Verrucomicrobia;Opitutae;Opitutae vadinHA64; uncultured bacterium	2.27	37	100,045	63.4	89.2	0.7	1	1889	6.85	1	0
biwa_5m_Cluster7	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Sporichthyaceae; hgcI clade;uncultured Candidatus Planktophila sp.	1.49	6	470,028	42.1	58.4	0.6	1	948	9.26	0	0
biwa_5m_Cluster8	Bacteria;Verrucomicrobia[2]	2.71	34	102,020	61.2	82.5	2.0	0	2,121	7.34	0	0
biwa_5m_Cluster9	Bacteria;Actinobacteria[2]	1.65	3	315,861	45.5	37.6	0.0	0	677	12.09	3	0
biwa_5m_Cluster10	Bacteria;Verrucomicrobia;Opitutae;Opitutae vadinHA64;uncultured bacterium	2.55	24	1,672,582	68.4	95.9	2.7	1	2,165	17.93	2	1
biwa_5m_Cluster12	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Sporichthyaceae; hgcI clade; uncultured actinobacterium	1.03	3	365,154	46.3	62.1	0.0	1	675	10.28	0	0
biwa_5m_Cluster13	Bacteria;Proteobacteria;Betaproteobacteria;Methylophilales;Methylophilaceae;Candidatus Methylopumilus;uncultured bacterium	1.40	10	169,468	37.3	80.7	0.4	1	1,289	8.37	1	0
biwa_5m_Cluster14	Bacteria;Actinobacteria;Actinobacteria;[1]	1.49	5	47,968	41.3	19.0	0.0	0	351	7.56	1	1
biwa_5m_Cluster15	Proteobacteria;Alphaproteobacteria;Pelagibacterales;[1]	1.02	6	222,441	29.4	88.6	0.0	0	1,075	20.45	1	0
biwa_5m_Cluster16	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;Filimonas;uncultured bacterium	4.08	44	45,979	42.4	43.1	0.1	1	1,908	5.57	0	2
biwa_65m_Cluster2	Bacteria;Chloroflexi;Anaerolineae;Anaerolineales;Anaerolineaceae; uncultured;	2.89	30	157,947	60.9	90.9	0.9	1	2,429	45.74	0	0
biwa_65m_Cluster4	Bacteria;Nitrospirae	1.92	11	313,929	57.6	93.9	0.9	0	1,890	8.01	2	1
biwa_65m_Cluster5	Archaea;Thaumarchaeota;Marine Group I;Unknown Order;Unknown Family;Candidatus Nitrosoarchaeum;	1.48	10	250,506	33.0	98.5	1.9	1	1,869	13.93	0	1
biwa_65m_Cluster6	Verrucomicrobia	2.09	49	46,663	65.9	81.5	0.7	0	1,705	5.98	0	0

[1] Estimated using CAT

[2] Estimated using Kaiju

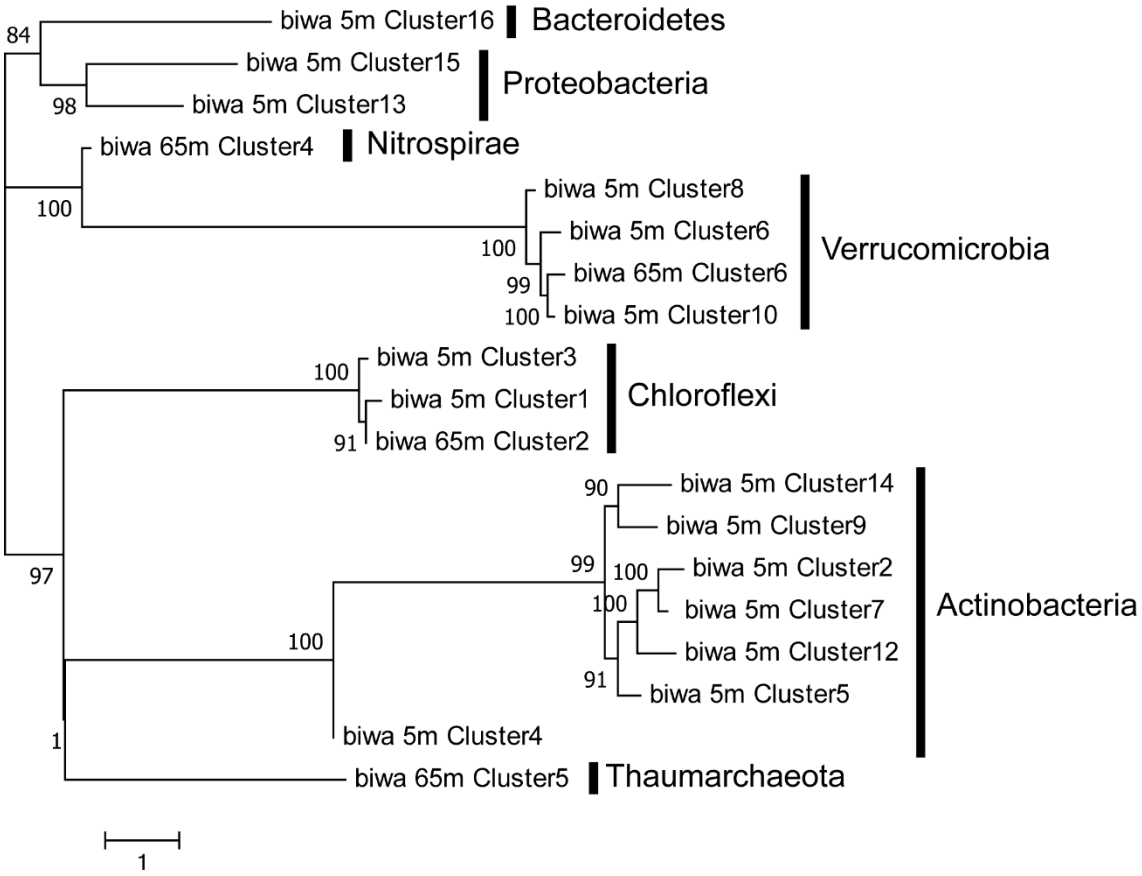

Figure 4-9 | Phylogenetic tree of genome bins. The phylogenetic tree was reconstructed using a set of up to 400 conserved bacterial marker genes with 1,000 bootstrap replicates by the maximum-likelihood method.

The genome bins belonged to seven phyla (Table 4-3). The archaeal genome bin only obtained from the biwa_65m sample, likely reflecting the microbial composition of the sampling locations (Fig. 4-5). All eleven 16S rRNA genes obtained from the genome bins showed greatest similarity to organisms belonging to uncultured clades, suggesting that the genome bins primarily contain unculturable microbes. Most of the genome bins were predicted to contain sequences belonging to phylum Actinobacteria, one of the dominant linages (Fig. 4-5). The genome bins biwa_5m_Cluster1, biwa_5m_Cluster3, and biwa_65m_Cluster2 likely represent members of the CL500-11 group belonging to the Chloroflexi phylum, one of the dominant clades in the hypolimnion of Lake Biwa (383) and frequently found world-wide in deep oligotrophic freshwater environments (389). Supporting this, the highest read coverage (45×) was obtained in biwa_65m_Cluster2,

from the deep lake water sample. Other taxa, such as CL500-29, hgcI, and Opitutae vadinHA64, are also typical members of freshwater habitats (363, 390), although the characteristics of these unculturable organisms are largely unknown. Overall, the phylogeny of the genome bins likely reflects the major dominant lineages.

Methylation patterns of reconstructed genomes

Using the SMRT analysis modification and motif detection tools, a total of 29 DNA methylation motifs were identified (Table 4-4). Although the methylation level is generally bistable at 0 and 100% (189), the ratio of motif methylation ranged from 21–93%. The low ratios (Table 4-4, items marked with '*') possibly reflect low modification detection power, due to the low coverage and level of genome completeness (391), and/or existence of unmethylated sites that frequently detected in culturable strains (391, 392). The motif sets were composed of 21 unique motifs, including 14 motifs that could not be matched to existing recognition sequences in the REBASE repository. Interestingly, no motifs were identified in 6 of the genome bins, including all Actinobacteria bins. These results are inconsistent with those of a previous study, which reported a number of methylation motifs and corresponding MTases in genomes of Actinobacteria (189). In contrast, all three genome bins containing members of the Chloroflexi phylum shared the same motif set (G**A**NTC, TT**A**A, and G**C**WGC). Sequences in the four genome bins assigned to the phylum Verrucomicrobia showed different motif sets. Three similar motifs predicted from the biwa_5m_Cluster13 genome bin could be attributed to an overlapping methylation motif; *HCAG**C**TKC* and B*GMAG**C**TGD* methylation motifs could be attributed to methylation activity on the palindromic *GMAG**C**TKC* motif. In support of this explanation, the methylation ratios of the above two motifs were lower (42.3 and 42.2%) than the last comparable motif (76.8%). The incongruent specificity of terminal bases (left H in HCAG**C**TKC, and right D in BGMAG**C**TGD) possibly be explain as noncanonical recognition which known as star activity in many REases (393). Among the 11 methylation motifs detected in the biwa_5m_Cluster16 genome bin, one contained a palindromic sequence (*i.e.*, G**A**ANNNNTTC) and the other 3 pairs showed exactly complementary sequences (*e.g.*, a pair of **A**GCNNNNNNCAT and **A**TGNNNNNNGCT), suggesting that on the genome, these motif sequences contained methylated bases on both strands. The pair *CA**A***NNNNNNNN*CTTG* and *CA**A**G*NNNNNNND*TTG*, and the pair *GYT**A***NNNNNNN*TTRG* and *CYA**A***NNNNNNN*TAVC*H may also be a complementary motif set. I did not observe any characteristics of the observed motifs that could be used to distinguish the shallow and deep water samples. In summary, these results suggested

that DNA methylation patterns were varied and widespread across prokaryotes in the freshwater environment, and a large number of novel methylation motifs remain undetected in environmental microbes, including unculturable members. I note that some identified motifs show low mean motif coverages (<50x, items marked with '#'), indicating the ambiguity still remains and further experimental validation is required.

Table 4-4 | Detected DNA methylation motifs.

Genome bin	Motif[1]	Modification Type	Motif in REBASE	Number of methylations	Number of motifs in contigs	motif methylation (%)	Mean modification QV	Mean motif coverage
biwa_5m_Cluster1	G**A**NTC	m6A	Yes	1,813	2,070	87.6%	58.0	35.2 #
	TTA**A**	m6A	Yes	1,264	1,522	83.0%	55.5	34.1 #
	G**C**WGC	m4C	Yes	3,026	15,948	19.0% *	38.4	40.6 #
biwa_5m_Cluster3	**G**ANTC	m6A	Yes	3,724	4,014	92.8%	66.1	41.3 #
	TTA**A**	m6A	Yes	3,036	3,338	91.0%	62.4	40.4 #
	G**C**WGC	m4C	Yes	13,821	54,026	25.6% *	39.5	46.4 #
biwa_5m_Cluster8	**A**GGNNNNNRTTT	m6A	No	80	276	29.0% *	39.6	65.8
biwa_5m_Cluster10	ACG**A**G	m6A	No	1,986	7,185	27.6% *	45.0	171.4
biwa_5m_Cluster13	GMAG**C**TKC	m4C	No	169	220	76.8%	50.9	83.5
	(HCAG**C**TKC)[2]	m4C	No	124	293	42.3% *	46.8	79.0
	(BGMAG**C**TGD)[2]	m4C	No	78	185	42.2% *	46.3	76.3
biwa_5m_Cluster15	G**A**NTC	m6A	Yes	2,856	2,880	99.2%	190.6	166.9
biwa_5m_Cluster16	G**A**ANNNNTTC	m6A	Yes	1,309	1,472	88.9%	55.6	30.9 #
	AGCNNNNNNCAT	m6A	No	642	726	88.4%	56.0	29.4 #
	ATGNNNNNNGCT	m6A	No	619	726	85.3%	52.0	29.8 #
	AGCNNNNNNGTG	m6A	No	311	349	89.1%	56.9	30.4 #
	C**A**CNNNNNNGCT	m6A	No	293	349	84.0%	53.3	30.9 #
	CA**A**NNNNNNNNCTTG	m6A	No	205	256	80.1%	49.4	29.1 #
	CA**A**GNNNNNNNDTTG	m6A	No	164	214	76.6%	48.7	28.7 #
	TT**A**GNNNNNCCT	m6A	No	87	99	87.9%	51.3	29.8 #
	AGGNNNNNCTAA	m6A	No	77	99	77.8%	49.4	29.7 #
	GYT**A**NNNNNNNTTRG	m6A	No	76	89	85.4%	56.0	31.3 #
	CYA**A**NNNNNNNTAVCH	m6A	No	59	127	46.5% *	53.5	32.6 #
biwa_65m_Cluster2	G**C**WGC	m4C	Yes	72,730	77,932	93.3%	140.2	297.3
	G**A**NTC	m6A	Yes	6,754	6,844	98.7%	346.3	281.7
	TTA**A**	m6A	Yes	5,475	5,564	98.4%	325.3	270.9
biwa_65m_Cluster4	TANGG**A**B	m6A	No	1,276	1,367	93.3%	64.4	48.5 #
biwa_65m_Cluster5	G**A**TC	m6A	Yes	9,446	9,618	98.2%	122.1	93.7
	AG**C**T	m4C	Yes	5,974	6,224	96.0%	84.0	92.1

Recognition sequences representations use the standard abbreviations to represent ambiguity:
R= G/A, Y= T/C, M= A/C, K= G/T, S= G/C, W= A/T, H= A/C/T, B= G/T/C, V= G/C/A, D= G/A/T, N= G/A/T/C

[1] Methylated base is underlined

[2] Methylated motifs partly matched with GMAG**C**TKC

* Low motif methylation rate (<75%)

Low mean motif coverage (<50x)

Characterization of methyltransferases

To identify MTases corresponding to the detected methylation motifs, systematic annotation of MTase genes was performed. The sequence specificities of each candidate MTase were estimated based on significant similarity to reference MTase genes with experimentally confirmed specificity. In total, 20 candidate MTase genes were identified from 9 genome bins, with a wide range of identity scores against reference MTase genes (23–71%) (Table 4-5). The bins contained up to 6 candidate MTases, and none were detected in 10 of the genome bins, partly due incomplete coverage. Type II RM were the most abundant candidate MTases, followed by Type I and Type III MTases, consistent with previous phylogeny-wide surveys (189, 394). Surprisingly, only eight predicted recognition sequences exactly matched the motifs detected by SMRT sequencing, and target motifs matching the other 13 candidate MTases were not identified. This result strongly suggested that similarity-based specificity prediction frequently leads to misannotation of the targeted motifs, including novel motifs.

Table 4-5 | Candidate MTases and REases that showed homology with reference genes in REBASE.

		Bioinformatic predictions						
	CDS ID	Gene type	REBASE candidate RM gene name	Identity (%)	Motif	Modification type	RM type	Motif identification
biwa_5m_3	tig00001319_41	M	M.SstE37II	58.9	G**A**NTC	m6A	II	Yes
	tig00001774_10	M	M.Sth20745I	71.4	TTA**A**	m6A	II	Yes
	tig00002121_20	M	M1.BceSIII	22.9	A**C**GGC	m4C	II	No
biwa_5m_6	tig00001209_84	M	M.SinI	57.0	GGW**C**C	m5C	II	No
biwa_5m_8	tig00001263_77	R	DvuI	36.3	?	-	I	-
	tig00001263_79	S	S.PveNS15I	32.4	?	-	I	-
	tig00001263_80	M	M.RbaNRL2II	55.6	ACG**A**NNNNNNGRTC	m6A	I	No
biwa_5m_10	tig00021821_21	RM	CjeFIII	23.7	GCA**A**GG	m6A	II	No
biwa_5m_15	tig00068316_171	M	M.Bsp460I	56.7	G**A**NTC	m6A	II	Yes
biwa_5m_16	tig00001681_20	M	M.Bli37I	56.6	G**A**YNNNNNRTC	m6A	I	No
	tig00001681_21	M	M.EcoNIH1III	59.2	G**A**TGNNNNNNTAC	m6A	I	No
	tig00001681_24	S	S.PveNS15I	47.2	?	-	I	-
	tig00001681_30	R	DvuI	38.4	?	-	I	-
	tig00001708_7	M	M.EcoGI	25.8	non-specific	m6A	II	?
	tig00001708_8	R	XmnI	34.0	GAANNNNTTC	-	II	Yes
	tig00001749_11	R	GmeII	33.8	TCCAGG	-	III	-
	tig00001749_15	M	M.FpsJII	53.4	CGC**A**G	m6A	III	No
	tig00001763_52	M	M.FnuDI	59.8	GGCC[1]	m4C	II	No
	tig00001763_54	R	BhaII	45.6	GGCC	-	II	-
	tig00002196_32	M	M.Mva1261III	37.1	CT**A**NNNNNNRTTC	m6A	I	No
biwa_65m_2	tig00012391_19	M	M.Sth20745I	71.0	TTA**A**	m6A	II	Yes
	tig00012395_52	M	M1.BceSIII	22.9	A**C**GGC	m4C	II	No
	tig00012461_58	M	M.SstE37II	58.9	G**A**NTC	m6A	II	Yes
biwa_65m_4	tig00000921_69	M	M.HgiDII	55.0	GTCGAC[1]	m5C	II	No
	tig00000921_72	RM	AquIV	28.5	GRGGA**A**G	m6A	II	No
	tig00000921_73	R	LpnPI	56.3	CCDG			
biwa_65m_5	tig00000166_68	M	M.Mma5219II	45.9	AG**C**T	m4C	II	Yes
	tig00000166_197	M	M.AvaVI	50.3	G**A**TC	m6A	II	Yes

[1] Modified base undetermined

Sequences identified in both the methylation analysis and MTase specificity prediction are likely to be active. For example, the genome bin biwa_65m_Cluster5 cont

ained two candidate MTases with AG**C**T and G**A**TC specificity, congruent with detected methylation motifs (Table 4-4, 4-5). This motif pattern exactly matched that of a closely related genus, *Candidatus Nitrosomarinus catalina,* that was investigated using enrichment culture (395).

The genome bins biwa_5m_Cluster3 and biwa_65m_Cluster2 contained the

same MTase specificity set (Table 4-5), and both genome bins shared the same methylation motifs (Table 4-4) and were identified as Chloroflexi (Table 4-3). These results suggested that microbes in the two genome bins share similar methylation levels. However, although two of three detected MTases were predicted to have G**A**NTC and TT**A**A specificity, matching the identified methylation motifs, other candidate MTase genes (tig00002121_20 and tig00012395_52 in biwa_5m_Cluster3 and biwa_65m_Cluster2 bins, respectively) showed highest sequence similarity to Type II M1.BceSIII, which was experimentally confirmed to have A**C**GGC target specificity, different from the G**C**WGC motif predicted in the methylation analysis. Because the CDS and the reference gene showed weak sequence similarity (22.9%), I hypothesized that the candidate MTases show G**C**WGC specificity although the protein sequence showed greater similarity to sequences with A**C**GGC specificity. Consistent with this hypothesis, phylogenetic analysis placed the candidate MTases between two clades with A**C**GGC and G**C**WGC specificity (Fig. 4-10). Although further experimental analysis will be required to verify this hypothesis, the proposed novel MTase clade will create a stir in the field of DNA methylation because G**C**WGC specific MTases have been deeply investigated since they were first discovered in 1975 (396).

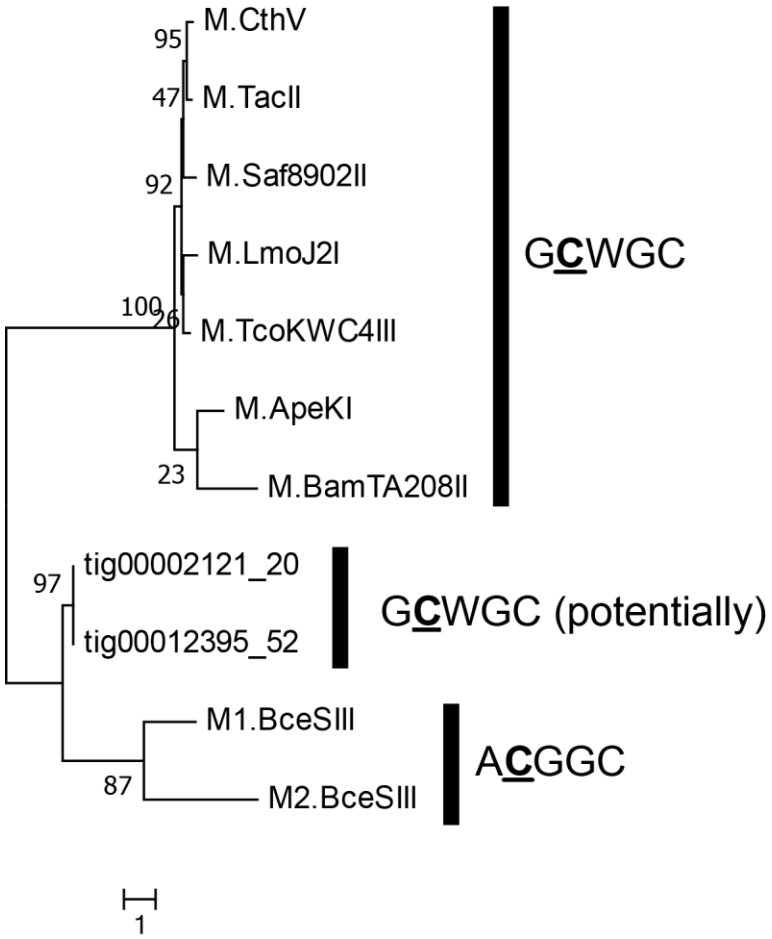

Figure 4-10 | Phylogenetic tree of MTases, including all of those with G**C**WGC and A**C**-GGC sequence specificity in the gold standard REBASE resource, and those identified from two genome bins belonging to Chloroflexi. Numbers adjacent to branch points are bootstrap percentages (100 replicates). Sequences adjusted with node MTases represent recognition sequences.

In the biwa_5m_Cluster16 genome bin, two Type I, two Type II, and one Type I potential RM systems were identified (Table 4-5), although at least 6 methylation motif pairs were expected (Table 4-4). The tig00001708_8 gene was assigned to Type II XmnI REase that are known to recognize G**A**ANNNNTTC motifs, which exactly matched the detected motif on the genome. The REase was adjusted to MTase (tig00001708_7) genes, as compatible with typical gene structure of Type II RM systems, suggesting the genes compose one RM system and targeted G**A**ANNNNTTC motifs. Another one Type III MTase (tig00001749_15) and REase (tig00001749_11) also placed closely, implying the genes are active and have specificity matched to ones identified in methylation analysis. The two Type I MTases (tig00001681_21 and tig00002196_32) have different specificity

from motifs observed in the methylation analysis. The type I RM system contains large pentameric proteins encoded by three separate restriction (R), methylation (M), and DNA sequence-recognition (S) subunits (397). The tig00002196_32–36 and tig00001681_17–30 genes show cluster structures composed of the three subunits (Fig. 4-11). Thus, these candidate RM systems are likely to be active and to show specificity for one of the identified methylation motifs. I would like to note that a purely bioinformatic approach is not sufficient to resolve which system recognizes which sequence. One of the candidate Type I RM system genes (tig00001681_20–30) were adjusted with transposase (tig00001681_17–18), suggesting that the gene cluster may behave as a mobile element, as described previously (348). On contrary, the methylation motifs complementary to the one Type II (tig00001763_52) MTases, which modify cytosine to m4C, were not found; this MTases are likely to be inactive.

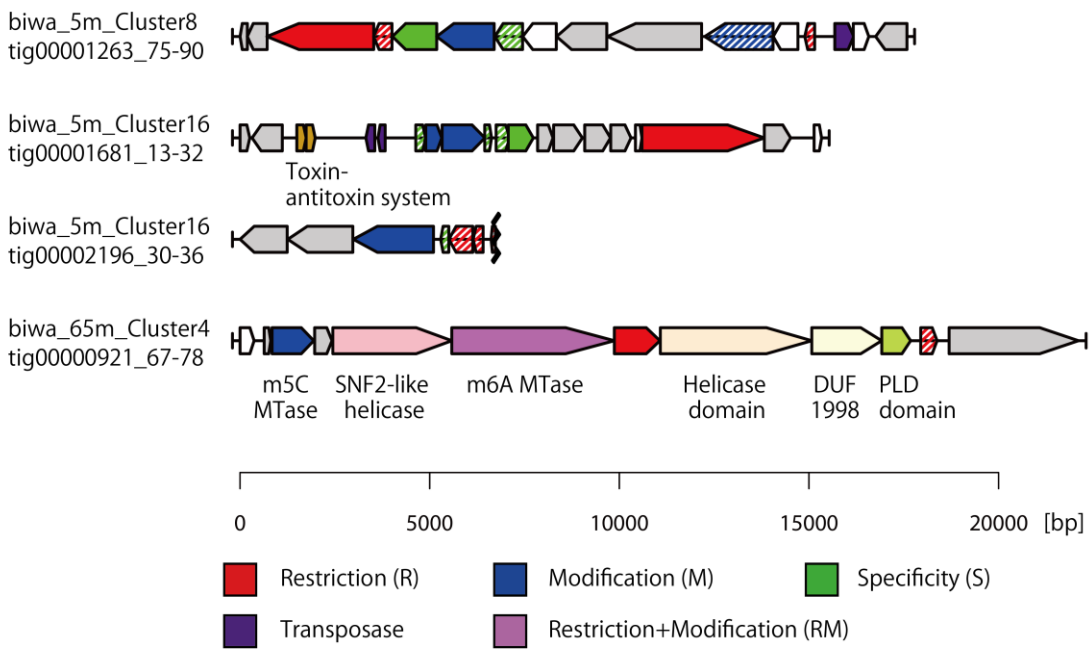

Figure 4-11 | Gene structure map of the region containing the restriction modification system genes. Arrow length is proportional to gene length. Restriction (R), methylation (M), and DNA sequence-recognition (S) subunits, and other functional genes are shown in different colors according to their annotation. The R, M, and S genes which show similarity to those in REBASE are shown in filled color, and others, which annotated using other gene databases, are shown in slanting lines. Other functionally annotated genes are shown in gray. Zigzag lines indicate contig terminals.

The biwa_5m_Cluster8 genome bin contained an incongruence between the identified methylation motif (**A**GGNNNNNRTTT) and the predicted MTase specificity (ACG**A**NNNNNNGRTC) (Table 4-4, 4-5). The MTase likely act as an RM system cooperating with neighboring M, R, and S proteins (Fig. 4-11). Similar, candidate MTase that show the greatest similarity to those contain a GCA**A**GG motif may recognize the ACG**A**G motif in the biwa_5m_Cluster10 genome.

In biwa_65m_Cluster4, a candidate Type II RM protein was predicted with low sequence similarity to sequences with GRGGA**A**G specificity. Considering the high degree of completeness of the reconstructed genome and the results of methylation analysis, the MTase is likely to have TANGG**A**B specificity, unlike the other candidate MTase that is predicted to have m5C modification ability. The two MTases were located in close proximity to three neighboring genes in the following order: m5C MTase, SNF2 family helicase, m6A REase and MTase, putative helicase, function unknown DUF1998, and phospholipase D-nuclease (PLD) domain proteins (Fig. 4-11). This gene placement is very similar to the recently proposed DISARM phage-infection defense system (398). The system was categorized into two groups: Class1, m6A MTases; and Class2, m5C MTases. Considering that the m6A methylation motif was identified, the system is likely Class1. Unfortunately, because m5C modification was difficult to identify in this study, due to its weak signature in SMRT sequencing, I was unable to determine whether the m5C MTase was active or not, although a previous study experimentally verified C**C**WGG motif methylation in one *Bacillus* strain with an active Class2 DISARM system (398). Therefore, it remains a possibility that the two MTases were synergistically active in the same DISARM system. To my knowledge, this is the first time a candidate DISARM system has been identified in phylum Nitrospirae, and further examination of this linage should be performed.

The biwa_5m_Cluster6 contained one candidate MTase but no methylation motif, suggesting the MTase is inactive. Although a methylation motif was identified in biwa_5m_Cluster13, no MTase was identified, possibly because the corresponding MTase has little similarity with known MTases. In all, I proposed six candidate novel MTases; the target motifs are difficult to estimate using only a similarity-based search against the REBASE repository. Although further experimental validation is required to confirm those hypotheses, these findings highlight the large potential of metaepigenomic analysis using SMRT sequencing technology for identification of novel RM systems in unculturable microbes.

Genome bins that potentially lack RM systems

As expected from the methylation analysis (Table 4-4), no MTase genes were identified in any of the 7 genome bins belonging to Actinobacteria (Table 4-3, 4-5), suggesting that the dominant Actinobacteria in Lake Biwa lack RM systems. This finding was incompatible with the results of a previous study, which reported that many Actinobacteria strains possess DNA methylation ability and contain corresponding MTases (189, 394). Moreover, bacterial and archaeal genomes rarely lack DNA methylation motifs, and the organisms in question shared no obvious characteristics (189). Although I cannot exclude the possibility that some novel MTases were not detected using the similarity-search screening strategy due to sequence diversity, these results suggested an underlying biological explanation.

Considering that RM systems play a crucial role in preventing uptake of exogenous DNA (399, 400), lack of an RM system can facilitate DNA exchange with bacteriophages and present opportunities to gain novel abilities (346). In addition, the high frequency of phage–prokaryote encounters in freshwater environments can place microbes under huge infection pressure. To investigate potential genetic exchange in the genomes, systematic prophage prediction was carried out (Table 4-6). A prophage is a phage genome integrated into the host genome that remains latent until conditions favor its reactivation (401). On the contrary, in some cases, bacterial and archaeal host cells can acquire phage genes that can increase host fitness, such as those for antibiotic production, toxin secretion, and biofilm formation, through prophage integration (401). However, in most cases, phage infection is virulent and prophages can be molecular time bombs that kill the host cell following their eventual induction (402). Thus, protection against phage infection is fundamental for survival in phage-rich environments such as fresh water.

In silico prediction showed that more than one prophage was present in 10 of the genome bins, 7 of which belonged to organisms in which no methylation motifs were identified (Table 4-3). Although I cannot draw conclusions from these data, due to the incomplete reconstructions of the genomes and difficulty of prophage identification, this unevenness likely cannot be explained by chance. The prophages showed little sequence similarity, except for two pairs (Fig. 4-12), suggesting that most of the detected prophages were not a result of vertical inheritance but independent infection and integration events. The size of prophages ranged from 4.3–10.8 kbp, smaller than typical prophages detected

in genome-wide analysis (403, 404). The prophages contained genes encoding tail-associated, chaperonin, protease, glycosyltransferase, ATP-dependent serine proteases, terminase, integrase, and tRNAs, that are frequently found in prophages (403, 404). The prophages also contained RNA polymerase sigma factor subunits that likely facilitate direct transcription of phage genes (405, 406). In contrast, functional annotation showed that some contigs lacked some of these essential genes (403), likely due to removal from the host genome. These data indicate that phage infection and prophage integration frequently occurred in microbes lacking RM systems.

Table 4-6 | Genomic features of prophages in genome bins.

Genome bin	Prophages	Contig	Coordinates	Size (kbp)
biwa_5m_Cluster2	phiB0502_1	tig00002506	10670-17702	7.0
biwa_5m_Cluster4	phiB0504_1	tig00001057	109914-117124	7.2
	phiB0504_2	tig00001402	61630-71371	9.7
biwa_5m_Cluster5	phiB0505_1	tig00000702	1808-9066	7.3
	phiB0505_2	tig00001168	78416-87062	8.6
biwa_5m_Cluster6	phiB0506_1	tig00001369	62143-68903	6.8
biwa_5m_Cluster9	phiB0509_1	tig00000318	22428-33243	10.8
	phiB0509_2	tig00000318	210817-215076	4.3
	phiB0509_3	tig00000914	58326-65877	7.6
biwa_5m_Cluster10	phiB0510_1	tig00021807	121169-127892	6.7
	phiB0510_2	tig00021821	246252-253779	7.5
biwa_5m_Cluster12	phiB0512_1	tig00001281	18969-24445	5.5
biwa_5m_Cluster13	phiB0513_1	tig00001534	3340-12017	8.7
biwa_5m_Cluster15	phiB0515_1	tig00000489	166958-172559	5.6
biwa_65m_Cluster1	phiB6501_1	tig00012480	21159-26551	5.4
biwa_65m_Cluster4	phiB6504_1	tig00000357	22052-29797	7.7
	phiB6504_2	tig00001590	874-10534	9.7

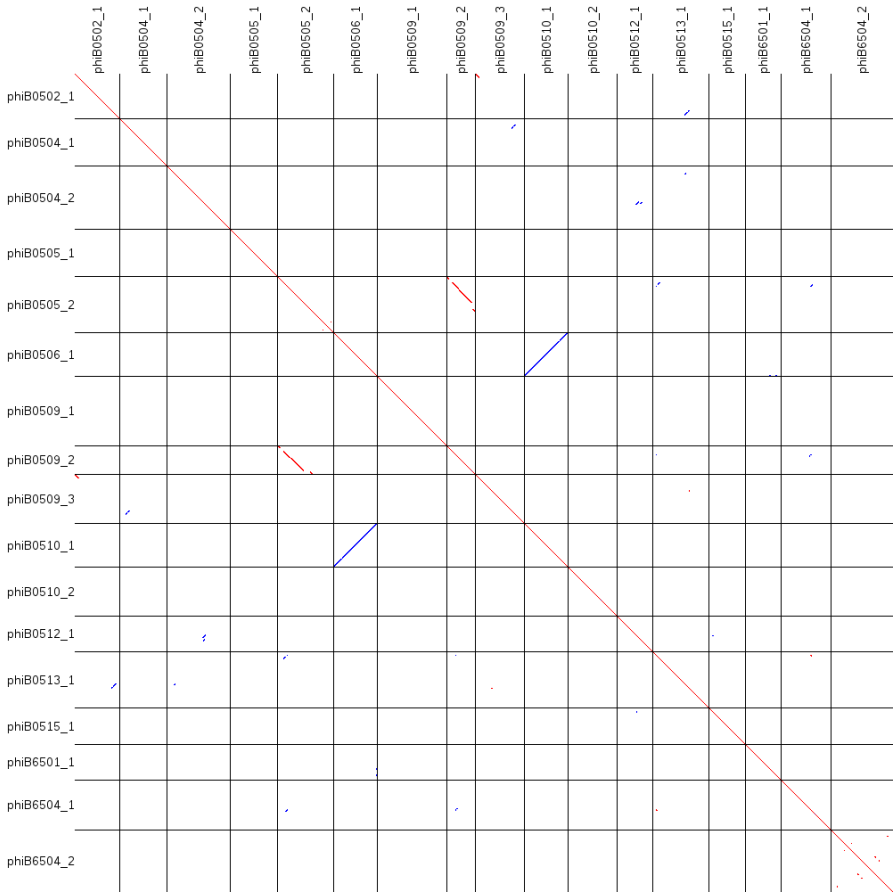

Figure 4-12 | Comparative analyses of prophage sequences.

Abortive infection and CRISPR/Cas system

Next, I investigated defense systems other than RM systems. Interestingly, I identified genes encoding serine threonine protein kinase, which are associated with abortive infection (Abi) systems, in phiB0504_1 prophage (407, 408) (Table 4-5). The system inhibits phage replication and proliferation through programmed cell death induced when phages infect host cells (409). The system is often found in prophages and probably favors their dissemination (410). Thus, the presence of an Abi system in genomes that lack RM systems probably acts as a repressor of further superinfection by phages and a deterrent to prophage expansion, substituting for the RM system.

Toxin-antitoxin (TA) systems are an antiphage mechanism composed of a stable toxin and an unstable antitoxin that cause cell death when phage infection occurs, similar to Abi systems (342, 411, 412). Type II TA system loci were identified from six genome bins, and no relationship was detected between RM system and prophage presence (Table 4-3). One TA system gene was adjacent to RM system genes in biwa_5m_Cluster16 (Fig.

4-11), likely indicating that the individual TA and RM systems comprised a defense island for coordinated defense against phage infection.

The CRISPR/Cas system is another system for protection from phage infection, similar to RM systems (409, 413, 414); it has been detected in the majority of bacterial and archaea genomes investigated to date (172). This system is typically composed of a *Cas* locus, a leader region, and a CRISPR array arranged in tandem, with nearly identical repeats and unique fragments acquired from phage genomes (spacers). I identified three possible CRISPR arrays in the biwa_5m_Cluster3, biwa_5m_Cluster8, and biwa_65m_Cluster5 genome bins. Within the three full-length arrays, two contained small numbers of repeat sequences (7 and 4 repeats in biwa_5m_Cluster3 and biwa_5m_Cluster8, respectively) and the other contained 50 repeats. However, no genes on the same contigs as the CRISPR array were assigned to *Cas* genes in the first two genome bins, and the array from the last genome bin showed tandem repeats, suggesting it might be a 'false-CRISPR,' as proposed in a recent study (415). I cannot conclude whether the arrays were mistaken sequence repeats, remnants of ancient functional CRISPR/Cas systems, or active CRISPR/Cas systems with remote *Cas* genes and/or novel Cas-like proteins that could not be assigned in this analysis. However, no prophage was predicted from any of the three genome bins that possessed a CRISPR repeat array (Table 4-3), which may reflect the efficiency of the CRISPR/Cas system to inhibit phage infection and prophage integration similar to the RM system, as previously described (404, 416, 417).

Conclusions

The present study demonstrated the effectiveness of SMRT circular consensus sequencing for metagenomic and metaepigenomic analyses, with obvious advantages over short-read sequencing and experimental methylation detection approaches. The high ratio of taxonomical read classification indicates that the combination of long accurate CCS reads and protein-based taxonomic assignment methods will be a suitable strategy to assess entire sequencing reads without ambiguity. The CCS reads also facilitate the metagenomic assembly and binning processes to reconstruct high-quality draft genomes, most of which are from dominant unculturable microbes. Most importantly, the analysis disclosed a number of DNA methylation motifs and candidate corresponding MTases, including novel single motifs as well as pairs. The presence of prophage and methylation motif loci were mutually exclusive, consistent with past experimental observations that

RM systems inhibit phage-mediated genetic exchange. Although further experimental validation is required, the results provide new insight into microbial ecology and phage-prokaryote interactions in freshwater environments.

Unfortunately, the current low throughput of SMRT sequencing made it difficult to obtain sufficient sequencing reads to capture the entire microbial community, including 'rare' species (typically with <1% relative abundance). Moreover, because deep sequencing coverage (>25× in subreads) was required for reliable detection of DNA methylation, the metagenomic setting easily leads underestimation of methylation identification. Although metagenomic analysis using long reads remains challenging, sequencing throughput and read length may not be the primary concern for long-read metagenomics when considering the continuing advances in sequencing technology. For example, Moleculo technology can provide long-linked synthetic reads and has already been applied to metagenomic research (418, 419), although use of PCR in library preparation cause biases that affect downstream analysis, in contrast to PCR-free SMRT sequencing. The 10x Chromium system can also generate long linked reads derived from the same single cell and similar to single-cell sequencing (420), likely have potential for metagenomic applications. Also, in the near future DNA methylation data will be able to be obtained using not only SMRT sequencing but also the Oxford Nanopore Technology (421, 422). Although detectable types of DNA modification are limited (*i.e.*, m4C, m5C, and m6A) using current SMRT sequencing technology, many DNA chemical modifications frequently occur in nature and potentially play significant biological roles (423). In all, advances in sequencing technology, modification of measuring schemes, and enhancements in bioinformatic tool development should be key for reliable and further detailed metagenomic and metaepigenomic analyses of environmental microbes.

This study provides an example of metagenomic and metaepigenomic analysis using SMRT sequencing technology of the environmental microbial community. Importantly, the present method is available not only for fresh water but also various environmental samples. Therefore, it provides insight into microbial ecology, such as diversity of phage-infection defense systems and phage–prokaryote interactions. Because the biological significance of most methylation motifs is not yet clear, further study is needed to assess how chemical modification of DNA contributes to microbial ecology. As this is the first attempt to characterize DNA methylation in unculturable microbes, it is expected that further studies performed under different sampling conditions and environments will broaden the potential of metaepigenomic analysis.

Chapter 5: Concluding remarks

In conclusion of this thesis, I describe summaries of the researches presented in this thesis and discuss the future works based on these researches.

In Chapter 2, I found that microbes impacted by the tsunami and resulting floods had adapted to an environment high in iron. Whole-genome sequencing of four of the isolated *Arthrobacter* strains revealed independent losses of siderophore-synthesis genes from their genomes. Indeed, chemical analysis confirmed the investigated soil samples to be rich in iron, and culture experiments confirmed weak cultivability of some of these strains in iron-limited media. Furthermore, metagenomic analyses demonstrated over-representation of denitrification-related genes in the tsunami-affected soil sample, as well as the presence of pathogenic and marine-living genera and genes related to salt-tolerance. The present results would provide an example of microbial characteristics of soil disturbed by the tsunami, which may give an insight into microbial adaptation to drastic environmental changes. Further analyses on microbial ecology after a tsunami are envisioned to develop a deeper understanding of the recovery processes of terrestrial microbial ecosystems.

In Chapter 3, I provided evidences supporting a precipitation-mediated microbial cycle model in which soil, oceanic, and animal-associated microbes are spread in the atmosphere, transported for long distances, and deposited via precipitation. The community-wide and seasonal analyses show the precipitation microbial communities were dominated by Proteobacteria, Firmicutes, Bacteroidetes, and Actinobacteria and were overall consistent with those previously reported in atmospheric aerosols and cloud water. Seasonal variations in composition were observed; specifically, Proteobacteria abundance significantly decreased from summer to winter. Notably, estimated ordinary habitats of precipitation microbes were dominated by animal-associated, soil-related, and marine-related environments, and reasonably consistent with estimated air mass backward trajectories. To my knowledge, this is the first amplicon-sequencing study investigating precipitation microbial communities involving sampling over the duration of a year.

In Chapter 4, I investigated genomic and metaepigenomic characteristics of environmental prokaryotic community, which dominated by unculturable members. Total of 19 phylogenetically-diverse draft genomes were obtained from two freshwater samples. Metaepigenomic analysis identified numbers of DNA methylation motifs including novel

ones and corresponding MTases were estimated from the assembled genome bins. The past hypothesis that the RM system acts as an inhibitor of genetic exchange, such as prophage mediated by phage infection, was supported in my analysis. To my knowledge, this is the first report that demonstrated the effectiveness of SMRT sequencing in metagenomic and metaepigenomic analysis against environmental prokaryotic community. Despite the ambiguity of DNA methylation identification due to the technical limitations, further experimental efforts will give new insights sinto microbial ecology especially in phage-prokaryote interactions.

Genomic, metagenomic, and bioinformatic approaches have already been common in microbial ecology and have been used to investigate whole communities containing many types of culturable and unculturable microbes. However, to date, most analyses have depended on straightforward sequence similarity searches against reference databases. This situation may not be satisfactory because microbial genomes should be the fundamental basis for microbial ecology and evolution. Enrichment of reference sequences (for both microbial taxa and functional genes) is one of the fundamental issues to promote various kinds of analyses.

Today, it has continued to accumulate in large quantities of publicly-available sequencing dataset. Importantly, the accumulation speed is higher in Whole-Genome Shotgun (WGS) data than simple genomes of individual organism registered in GenBank (https://www.ncbi.nlm.nih.gov/genbank/statistics/). Large efforts have been made to reveal weak relationship within microbes and between environmental characteristics. However, knowledge obtained from meta-analysis was still little used reductively for microbial studies likely due to a lack of useful knowledge resources and bioinformatic tools. Platforms that enable meta-analysis of diverse metagenomic datasets will allow us to discover hidden laws of the microbial ecosystem from publicly available data.

Recent advancement of sequencing technology allows us to obtain ultra-long sequencing reads and another information such as DNA chemical modification. Although current technical limitations restrict study design, the technology opens a new world especially in epigenomics, which is unable to reach when using only current short-read sequencing technology and experimental methodology. In order to go beyond current microbial genomic and metagenomic analyses, more powerful bioinformatic methods for analyzing data from diverse perspectives is required.

Acknowledgments

My research would not have been possible without help of people around me. I am using this opportunity to express my gratitude to everyone who supported me throughout the Ph.D. course.

Firstly, I would like to show my respect and gratitude to my supervisor, Dr. Wataru Iwasaki, for providing me with an opportunity to try this new and exciting research area. His insightful advice based on his broader perspective often helped me.

I thank all co-authors of our paper published in BMC Genomics. Dr. Shotaro Hirase, Dr. Asako Machiyama, Dr. Kenshiro Oshima, and Dr. Masahira Hattori kindly provided the experimental and sequencing data. Dr. Minoru Ijichi, Dr. Kentaro Inoue, Dr. Susumu Yoshizawa, and Dr. Kazuhiro Kogure made continuous discussion on the analyses. Their comments helped me a lot to conduct the research in depth.

I would like to thank all co-authors of our review paper published in Microbes and Environments. Dr. Ching-chia Yang revised my manuscripts and corrected my English descriptions. Her broad and profound knowledge on metagenomics and microbiology helped me to deepen insights into microbial ecology and methodology of microbial genomic analysis.

I would like to thank all co-authors of our paper published in Frontiers in Microbiology. I would also like to thank Mr. Masaya Miyahara, Mr. Kazushi Fujii, and Dr. Asako Machiyama for providing experimental data and processing of DNA sequencing.

Finally, I would like to show my respect and gratitude to all members in the Iwasaki laboratory of the University of Tokyo.

January 31, 2018

Satoshi Hiraoka

References

1. **Iwasaki W**, **Takagi T**. 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics **23**:i230–i239.
2. **Snel B**, **Bork P**, **Huynen MA**. 2002. Genomes in flux: The evolution of Archaeal and Proteobacterial gene content. Genome Res **12**:17–25.
3. **Rappé MS**, **Giovannoni SJ**. 2003. The uncultured microbial majority. Annu Rev Microbiol **57**:369–394.
4. **Mori K**, **Kamagata Y**. 2014. The challenges of studying the anaerobic microbial world. Microbes Environ **29**:335–337.
5. **Narihiro T**, **Kamagata Y**. 2013. Cultivating yet-to-be cultivated microbes: The challenge continues. Microbes Environ **28**:163–165.
6. **Puspita ID**, **Kamagata Y**, **Tanaka M**, **Asano K**, **Nakatsu CH**. 2012. Are uncultivated bacteria really uncultivable? Microbes Environ **27**:356–66.
7. **Zoetendal EG**, **Vaughan EE**, **De Vos WM**. 2006. A microbial world within us. Mol Microbiol **59**:1639–1650.
8. **Akkermans ADL**, **Mirza MS**, **Harmsen HJM**, **Blok HJ**, **Herron PR**, **Sessitsch A**, **Akkermans WM**. 1994. Molecular ecology of microbes: A review of promises, pitfalls, and true progress. FEMS Microbiol Rev **15**:185–194.
9. **Su C**, **Lei L**, **Duan Y**, **Zhang KQ**, **Yang J**. 2012. Culture-independent methods for studying environmental microorganisms: Methods, application, and perspective. Appl Microbiol Biotechnol **93**:993–1003.
10. **Béjà O**, **Suzuki MT**, **Koonin E V.**, **Aravind L**, **Hadd A**, **Nguyen LP**, **Villacorta R**, **Garrigues C**, **Jovanovich SB**, **Feldman RA**, **DeLong EF**. 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol **2**:516–529.
11. **Felczykowska A**, **Bloch SK**, **Nejman-Faleńczyk B**, **Barańska S**. 2012. Metagenomic approach in the investigation of new bioactive compounds in the marine environment. Acta Biochim Pol **59**:501–505.
12. **Coughlan LM**, **Cotter PD**, **Hill C**, **Alvarez-Ordóñez A**. 2015. Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. Front Microbiol **6**:672.
13. **Jackson SA**, **Borchert E**, **O'Gara F**, **Dobson ADW**. 2015. Metagenomics for the discovery of novel biosurfactants of environmental interest from marine ecosystems. Curr Opin Biotechnol **33**:176–182.
14. **Gillespie DE**, **Brady SF**, **Bettermann AD**, **Cianciotto NP**, **Liles MR**, **Rondon MR**, **Clardy J**, **Goodman RM**, **Handelsman J**. 2002. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol **68**:4301–4306.
15. **Banik JJ**, **Brady SF**. 2008. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc Natl Acad Sci U S A **105**:17273–17277.

16. **Wichmann F**, **Udikovic-Kolic N**, **Andrew S**, **Handelsman J**. 2014. Diverse antibiotic resistance genes in dairy cow manure. MBio **5**:e01017-13.

17. **Culligan EP**, **Marchesi JR**, **Hill C**, **Sleator RD**. 2012. Mining the human gut microbiome for novel stress resistance genes. Gut Microbes **3**:394–397.

18. **Guazzaroni ME**, **Morgante V**, **Mirete S**, **González-Pastor JE**. 2013. Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. Environ Microbiol **15**:1088–1102.

19. **Lewin A**, **Wentzel A**, **Valla S**. 2013. Metagenomics of microbial life in extreme temperature environments. Curr Opin Biotechnol **24**:516–525.

20. **Cowan DA**, **Ramond JB**, **Makhalanyane TP**, **De Maayer P**. 2015. Metagenomics of extreme environments. Curr Opin Microbiol **25**:97–102.

21. **Chen Y**, **Wu L**, **Boden R**, **Hillebrand A**, **Kumaresan D**, **Moussard H**, **Baciu M**, **Lu Y**, **Colin Murrell J**. 2009. Life without light: Microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave. ISME J **3**:1093–1104.

22. **Gill SR**, **Pop M**, **Deboy RT**, **Eckburg PB**, **Turnbaugh PJ**, **Samuel BS**, **Gordon JI**, **Relman DA**, **Fraser-Liggett CM**, **Nelson KE**. 2006. Metagenomic analysis of the human distal gut microbiome. Science **312**:1355–1359.

23. **Schwabe RF**, **Jobin C**. 2013. The microbiome and cancer. Nat Rev Cancer **13**:800–812.

24. **Kostic AD**, **Gevers D**, **Pedamallu CS**, **Michaud M**, **Duke F**, **Earl AM**, **Ojesina AI**, **Jung J**, **Bass AJ**, **Tabernero J**, **Baselga J**, **Liu C**, **Shivdasani RA**, **Ogino S**, **Birren BW**, **Huttenhower C**, **Garrett WS**, **Meyerson M**. 2012. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. Genome Res **22**:292–298.

25. **Greenblum S**, **Turnbaugh PJ**, **Borenstein E**. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A **109**:594–599.

26. **Qin J**, **Li Y**, **Cai Z**, **Li S**, **Zhu J**, **Zhang F**, **Liang S**, **Zhang W**, **Guan Y**, **Shen D**, **Peng Y**, **Zhang D**, **Jie Z**, **Wu W**, **Qin Y**, **Xue W**, **Li J**, **Han L**, **Lu D**, **Wu P**, **Dai Y**, **Sun X**, **Li Z**, **Tang A**, **Zhong S**, **Li X**, **Chen W**, **Xu R**, **Wang M**, **Feng Q**, **Gong M**, **Yu J**, **Zhang Y**, **Zhang M**, **Hansen T**, **Sanchez G**, **Raes J**, **Falony G**, **Okuda S**, **Almeida M**, **LeChatelier E**, **Renault P**, **Pons N**, **Batto J-M**, **Zhang Z**, **Chen H**, **Yang R**, **Zheng W**, **Li S**, **Yang H**, **Wang J**, **Ehrlich SD**, **Nielsen R**, **Pedersen O**, **Kristiansen K**, **Wang J**. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490**:55–60.

27. **Manichanh C**, **Rigottier-Gois L**, **Bonnaud E**, **Gloux K**, **Pelletier E**, **Frangeul L**, **Nalin R**, **Jarrin C**, **Chardon P**, **Marteau P**, **Roca J**, **Dore J**. 2006. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut **55**:205–11.

28. **Erickson AR**, **Cantarel BL**, **Lamendella R**, **Darzi Y**, **Mongodin EF**, **Pan C**, **Shah M**, **Halfvarson J**, **Tysk C**, **Henrissat B**, **Raes J**, **Verberkmoes NC**, **Fraser CM**, **Hettich RL**, **Jansson JK**. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One

7:e49138.

29. **Belda-Ferre P**, **Alcaraz LD**, **Cabrera-Rubio R**, **Romero H**, **Simón-Soro A**, **Pignatelli M**, **Mira A**. 2012. The oral metagenome in health and disease. ISME J **6**:46–56.

30. **Kong HH**, **Oh J**, **Deming C**, **Conlan S**, **Grice EA**, **Beatson MA**, **Nomicos E**, **Polley EC**, **Komarow HD**, **Mullikin J**, **Thomas J**, **Blakesley R**, **Young A**, **Chu G**, **Ramsahoye C**, **Lovett S**, **Han J**, **Legaspi R**, **Sison C**, **Montemayor C**, **Gregory M**, **Hargrove A**, **Johnson T**, **Riebow N**, **Schmidt B**, **Novotny B**, **Gupta J**, **Benjamin B**, **Brooks S**, **Coleman H**, **Ho SL**, **Schandler K**, **Stantripop M**, **Maduro Q**, **Bouffard G**, **Dekhtyar M**, **Guan X**, **Masiello C**, **Maskeri B**, **McDowell J**, **Park M**, **Vemulapalli M**, **Murray PR**, **Turner ML**, **Segre JA**. 2012. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome Res **22**:850–859.

31. **Warnecke F**, **Luginbühl P**, **Ivanova N**, **Ghassemian M**, **Richardson TH**, **Stege JT**, **Cayouette M**, **McHardy AC**, **Djordjevic G**, **Aboushadi N**, **Sorek R**, **Tringe SG**, **Podar M**, **Martin HG**, **Kunin V**, **Dalevi D**, **Madejska J**, **Kirton E**, **Platt D**, **Szeto E**, **Salamov A**, **Barry K**, **Mikhailova N**, **Kyrpides NC**, **Matson EG**, **Ottesen E a**, **Zhang X**, **Hernández M**, **Murillo C**, **Acosta LG**, **Rigoutsos I**, **Tamayo G**, **Green BD**, **Chang C**, **Rubin EM**, **Mathur EJ**, **Robertson DE**, **Hugenholtz P**, **Leadbetter JR**. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature **450**:560–565.

32. **Hongoh Y**. 2010. Diversity and genomes of uncultured microbial symbionts in the termite gut. Biosci Biotechnol Biochem **74**:1145–1151.

33. **Woyke T**, **Teeling H**, **Ivanova NN**, **Huntemann M**, **Richter M**, **Gloeckner FO**, **Boffelli D**, **Anderson IJ**, **Barry KW**, **Shapiro HJ**, **Szeto E**, **Kyrpides NC**, **Mussmann M**, **Amann R**, **Bergin C**, **Rubin EM**, **Dubilier N**. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443**:950–955.

34. **Li RW**. 2010. Metagenomics and its applications in agriculture, biomedicine, and environmental studies. Nova Science Publisher's, New York, United States.

35. **Yang Y**, **Xie B**, **Yan J**. 2014. Application of next-generation sequencing technology in forensic science. Genomics Proteomics Bioinformatics **12**:190–197.

36. **Khodakova AS**, **Smith RJ**, **Burgoyne L**, **Abarno D**, **Linacre A**. 2014. Random whole metagenomic sequencing for forensic discrimination of soils. PLoS One **9**:e104996.

37. **Fierer N**, **Lauber CCL**, **Zhou N**, **McDonald D**, **Costello EK**, **Knight R**. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A **107**:6477–6481.

38. **Kakizaki E**, **Ogura Y**, **Kozawa S**, **Nishida S**, **Uchiyama T**, **Hayashi T**, **Yukawa N**. 2012. Detection of diverse aquatic microbes in blood and organs of drowning victims: First metagenomic approach using high-throughput 454-pyrosequencing. Forensic Sci Int **220**:135–146.

39. **Turnbaugh PJ**, **Ley RE**, **Hamady M**, **Fraser-Liggett CM**, **Knight R**, **Gordon JI**. 2007. The human microbiome project. Nature **449**:804–810.

40. **Nelson KE**. 2011. Metagenomics of the human body. Springer New York, New York, United States.

41. **Gilbert JA**, **Jansson JK**, **Knight R**. 2014. The Earth microbiome project: Successes and aspirations. BMC Biol **12**:69.

42. **Vogel TM**, **Simonet P**, **Jansson JK**, **Hirsch PR**, **Tiedje JM**, **Elsas V**, **Dirk J**, **Bailey MJ**, **Nalin R**, **Philippot L**. 2009. TerraGenome: A consortium for the sequencing of a soil metagenome. Nat Rev Microbiol **7**:252.

43. **Rusch DB**, **Halpern AL**, **Sutton G**, **Heidelberg KB**, **Williamson S**, **Yooseph S**, **Wu D**, **Eisen JA**, **Hoffman JM**, **Remington K**, **Beeson K**, **Tran B**, **Smith H**, **Baden-Tillson H**, **Stewart C**, **Thorpe J**, **Freeman J**, **Andrews-Pfannkoch C**, **Venter JE**, **Li K**, **Kravitz S**, **Heidelberg JF**, **Utterback T**, **Rogers YH**, **Falcón LI**, **Souza V**, **Bonilla-Rosso G**, **Eguiarte LE**, **Karl DM**, **Sathyendranath S**, **Platt T**, **Bermingham E**, **Gallardo V**, **Tamayo-Castillo G**, **Ferrari MR**, **Strausberg RL**, **Nealson K**, **Friedman R**, **Frazier M**, **Venter JC**. 2007. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol **5**:0398–0431.

44. **Bork P**, **Bowler C**, **Vargas C de**, **Gorsky G**, **Karsenti E**, **Wincker P**. 2015. Tara Oceans studies plankton at planetary scale. Science **348**:873.

45. **Caporaso JG**, **Lauber CL**, **Costello EK**, **Berg-Lyons D**, **Gonzalez A**, **Stombaugh J**, **Knights D**, **Gajer P**, **Ravel J**, **Fierer N**, **Gordon JI**, **Knight R**. 2011. Moving pictures of the human microbiome. Genome Biol **12**:R50.

46. **Voigt AY**, **Costea PI**, **Kultima JR**, **Li SS**, **Zeller G**, **Sunagawa S**, **Bork P**. 2015. Temporal and technical variability of human gut metagenomes. Genome Biol **16**:73.

47. **Jung JY**, **Lee SH**, **Kim JM**, **Park MS**, **Bae JW**, **Hahn Y**, **Madsen EL**, **Jeon CO**. 2011. Metagenomic analysis of kimchi, a Traditional Korean fermented food. Appl Environ Microbiol **77**:2264–2274.

48. **Chaillou S**, **Chaulot-Talmon A**, **Caekebeke H**, **Cardinal M**, **Christieans S**, **Denis C**, **Hélène Desmonts M**, **Dousset X**, **Feurer C**, **Hamon E**, **Joffraud J-J**, **La Carbona S**, **Leroi F**, **Leroy S**, **Lorre S**, **Macé S**, **Pilet M-F**, **Prévost H**, **Rivollier M**, **Roux D**, **Talon R**, **Zagorec M**, **Champomier-Vergès M-C**. 2015. Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J **9**:1105–1118.

49. **Neelakanta G**, **Sultana H**. 2013. The use of metagenomic approaches to analyze changes in microbial communities. Microbiol insights **6**:37–48.

50. **Aziz RK**, **Bartels D**, **Best AA**, **DeJongh M**, **Disz T**, **Edwards RA**, **Formsma K**, **Gerdes S**, **Glass EM**, **Kubal M**, **Meyer F**, **Olsen GJ**, **Olson R**, **Osterman AL**, **Overbeek RA**, **McNeil LK**, **Paarmann D**, **Paczian T**, **Parrello B**, **Pusch GD**, **Reich C**, **Stevens R**, **Vassieva O**, **Vonstein V**, **Wilke A**, **Zagnitko O**. 2008. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics **9**:75.

51. **Meyer F**, **Paarmann D**, **D'Souza M**, **Olson R**, **Glass E**, **Kubal M**, **Paczian T**, **Rodriguez A**, **Stevens R**, **Wilke A**, **Wilkening J**, **Edwards R**. 2008. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**:386.

52. **Markowitz VM**, **Chen IMA**, **Chu K**, **Szeto E**, **Palaniappan K**, **Pillay M**, **Ratner A**, **Huang J**, **Pagani I**, **Tringe S**, **Huntemann M**, **Billis K**, **Varghese N**, **Tennessen K**, **Mavromatis K**, **Pati A**, **Ivanova NN**, **Kyrpides NC**. 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res **42**:D568–D573.

53. **Hunter S**, **Corbett M**, **Denise H**, **Fraser M**, **Gonzalez-Beltran A**, **Hunter C**, **Jones P**, **Leinonen R**, **McAnulla C**, **Maguire E**, **Maslen J**, **Mitchell A**, **Nuka G**, **Oisel A**, **Pesseat S**, **Radhakrishnan R**, **Rocca-Serra P**, **Scheremetjew M**, **Sterk P**, **Vaughan D**, **Cochrane G**, **Field D**, **Sansone SA**. 2014. EBI metagenomics-A new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res **42**:D600–D606.

54. **Quast C**, **Pruesse E**, **Yilmaz P**, **Gerken J**, **Schweer T**, **Yarza P**, **Peplies J**, **Glöckner FO**. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res **41**:D590–D596.

55. **Huson DH**, **Weber N**. 2013. Microbial community analysis using MEGAN. Methods Enzymol **531**:465–485.

56. **Chen HM**, **Lifschitz CH**. 1989. Preparation of fecal samples for assay of volatile fatty acids by gas-liquid chromatography and high-performance liquid chromatography. Clin Chem **35**:74–76.

57. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75**:7537–7541.

58. **Scholz MB**, **Lo CC**, **Chain PSG**. 2012. Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. Curr Opin Biotechnol **23**:9–15.

59. **Kunin V**, **Copeland A**, **Lapidus A**, **Mavromatis K**, **Hugenholtz P**. 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev **72**:557–578.

60. **Di Bella JM**, **Bao Y**, **Gloor GB**, **Burton JP**, **Reid G**. 2013. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods **95**:401–414.

61. **Mendoza MLZ**, **Sicheritz-Pontén T**, **Thomas Gilbert MP**. 2014. Environmental genes and genomes: Understanding the differences and challenges in the approaches and software for their analyses. Brief Bioinform **16**:745–758.

62. **Oulas A**, **Pavloudi C**, **Polymenakou P**, **Pavlopoulos GA**, **Papanikolaou N**, **Kotoulas G**, **Arvanitidis C**, **Iliopoulos I**. 2015. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights **9**:75–88.

63. **Tatusova T**, **Ciufo S**, **Fedorov B**, **O'Neill K**, **Tolstoy I**. 2014. RefSeq microbial genomes database: New representation and annotation strategy. Nucleic Acids Res **42**:D553–D559.

64. **McDonald D**, **Price MN**, **Goodrich J**, **Nawrocki EP**, **DeSantis TZ**, **Probst A**, **Andersen GL**, **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.

65. **Cole JR**, **Wang Q**, **Fish JA**, **Chai B**, **McGarrell DM**, **Sun Y**, **Brown CT**, **Porras-Alfaro A**, **Kuske CR**, **Tiedje JM**. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res **42**:D633–D642.

66. **Chun J**, **Lee J-H**, **Jung Y**, **Kim M**, **Kim S**, **Kim BK**, **Lim Y-W**. 2007. EzTaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int J Syst Evol Microbiol **57**:2259–2261.

67. **Alonso-Alemany D**, **Barré A**, **Beretta S**, **Bonizzoni P**, **Nikolski M**, **Valiente G**. 2014. Further steps in TANGO: Improved taxonomic assignment in metagenomics. Bioinformatics **30**:17–23.

68. **Segata N**, **Waldron L**, **Ballarini A**, **Narasimhan V**, **Jousson O**, **Huttenhower C**. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods **9**:811–814.

69. **Wood DE**, **Salzberg SL**. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15**:R46.

70. **Menzel P**, **Ng KL**, **Krogh A**. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun **7**:11257.

71. **Pignatelli M**, **Aparicio G**, **Blanquer I**, **Hernández V**, **Moya A**, **Tamames J**. 2008. Metagenomics reveals our incomplete knowledge of global diversity. Bioinformatics **24**:2124–2125.

72. **Grigoriev I V**, **Nikitin R**, **Haridas S**, **Kuo A**, **Ohm R**, **Otillar R**, **Riley R**, **Salamov A**, **Zhao X**, **Korzeniewski F**, **Smirnova T**, **Nordberg H**, **Dubchak I**, **Shabalov I**. 2014. MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res **42**:D699–D704.

73. **Yamazaki Y**, **Akashi R**, **Banno Y**, **Endo T**, **Ezura H**, **Fukami-Kobayashi K**, **Inaba K**, **Isa T**, **Kamei K**, **Kasai F**, **Kobayashi M**, **Kurata N**, **Kusaba M**, **Matuzawa T**, **Mitani S**, **Nakamura T**, **Nakamura Y**, **Nakatsuji N**, **Naruse K**, **Niki H**, **Nitasaka E**, **Obata Y**, **Okamoto H**, **Okuma M**, **Sato K**, **Serikawa T**, **Shiroishi T**, **Sugawara H**, **Urushibara H**, **Yamamoto M**, **Yaoita Y**, **Yoshiki A**, **Kohara Y**. 2010. NBRP databases: Databases of biological resources in Japan. Nucleic Acids Res **38**:D26–D32.

74. **Haas BJ**, **Gevers D**, **Earl AM**, **Feldgarden M**, **Ward D V**, **Giannoukos G**, **Ciulla D**, **Tabbaa D**, **Highlander SK**, **Sodergren E**, **Methé B**, **DeSantis TZ**, **Consortium THM**, **Petrosino JF**, **Knight R**, **Birren BW**. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res **21**:494–504.

75. **Quince C**, **Lanzen A**, **Davenport RJ**, **Turnbaugh PJ**. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics **12**:38.

76. **Edgar RC**, **Haas BJ**, **Clemente JC**, **Quince C**, **Knight R**. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**:2194–2200.

77. **Fu L**, **Niu B**, **Zhu Z**, **Wu S**, **Li W**. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics **28**:3150–3152.

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

2461.

79. **Edgar RC**. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nat Methods **10**:996–998.

80. **Sunagawa S**, **Mende DR**, **Zeller G**, **Izquierdo-Carrasco F**, **Berger S a**, **Kultima JR**, **Coelho LP**, **Arumugam M**, **Tap J**, **Nielsen HB**, **Rasmussen S**, **Brunak S**, **Pedersen O**, **Guarner F**, **de Vos WM**, **Wang J**, **Li J**, **Dore J**, **Ehrlich SD**, **Stamatakis A**, **Bork P**. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods **10**:1196–1199.

81. **Sharpton TJ**, **Riesenfeld SJ**, **Kembel SW**, **Ladau J**, **O'Dwyer JP**, **Green JL**, **Eisen JA**, **Pollard KS**. 2011. PhyLOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput Biol **7**:e1001061.

82. **Matsen F a**, **Kodner RB**, **Armbrust EV**. 2010. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics **11**:538.

83. **Darling AE**, **Jospin G**, **Lowe E**, **Matsen FA**, **Bik HM**, **Eisen JA**. 2014. PhyloSift: Phylogenetic analysis of genomes and metagenomes. PeerJ **2**:e243.

84. **Prosser JI**. 2015. Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology. Nat Rev Microbiol **13**:439–446.

85. **Luo C**, **Tsementzi D**, **Kyrpides NC**, **Konstantinidis KT**. 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J **6**:898–901.

86. **Albertsen M**, **Hugenholtz P**, **Skarshewski A**, **Nielsen KL**, **Tyson GW**, **Nielsen PH**. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol **31**:533–538.

87. **Tyson GW**, **Chapman J**, **Hugenholtz P**, **Allen EE**, **Ram RJ**, **Richardson PM**, **Solovyev V V**, **Rubin EM**, **Rokhsar DS**, **Banfield JF**. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**:37–43.

88. **Brown CT**, **Hug LA**, **Thomas BC**, **Sharon I**, **Castelle CJ**, **Singh A**, **Wilkins MJ**, **Wrighton KC**, **Williams KH**, **Banfield JF**. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature **523**:208–211.

89. **Narasingarao P**, **Podell S**, **Ugalde JA**, **Brochier-Armanet C**, **Emerson JB**, **Brocks JJ**, **Heidelberg KB**, **Banfield JF**, **Allen EE**. 2012. *De novo* metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J **6**:81–93.

90. **Evans PN**, **Parks DH**, **Chadwick GL**, **Robbins SJ**, **Orphan VJ**, **Golding SD**, **Tyson GW**. 2015. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. Science **350**:434–438.

91. **Rodrigue S**, **Malmstrom RR**, **Berlin AM**, **Birren BW**, **Henn MR**, **Chisholm SW**. 2009. Whole genome amplification and *de novo* assembly of single bacterial cells. PLoS One **4**:e6864.

92. **Lasken RS**. 2012. Genomic sequencing of uncultured microorganisms from single cells. Nat Rev Microbiol

10:631–640.

93. **Eloe-Fadrosh EA**, **Paez-Espino D**, **Jarett J**, **Dunfield PF**, **Hedlund BP**, **Dekas AE**, **Grasby SE**, **Brady AL**, **Dong H**, **Briggs BR**, **Li W-J**, **Goudeau D**, **Malmstrom R**, **Pati A**, **Pett-Ridge J**, **Rubin EM**, **Woyke T**, **Kyrpides NC**, **Ivanova NN**. 2016. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat Commun **7**:10476.

94. **Mende DR**, **Aylward FO**, **Eppley JM**, **Nielsen TN**, **DeLong EF**. 2016. Improved environmental genomes via integration of metagenomic and single-cell assemblies. Front Microbiol **7**:143.

95. **Pop M**, **Phillippy A**, **Delcher AL**. 2004. Comparative genome assembly. Bioinformatics **5**:237–248.

96. **Afiahayati**, **Sato K**, **Sakakibara Y**. 2015. MetaVelvet-SL: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res **22**:69–77.

97. **Nurk S**, **Bankevich A**, **Antipov D**, **Gurevich AA**, **Korobeynikov A**, **Lapidus A**, **Prjibelski AD**, **Pyshkin A**, **Sirotkin A**, **Sirotkin Y**, **Stepanauskas R**, **Clingenpeel SR**, **Woyke T**, **McLean JS**, **Lasken R**, **Tesler G**, **Alekseyev MA**, **Pevzner PA**. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol **20**:714–737.

98. **Peng Y**, **Leung HCM**, **Yiu SM**, **Chin FYL**. 2012. IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28**:1420–1428.

99. **Mavromatis K**, **Ivanova N**, **Barry KW**, **Shapiro HJ**, **Goltsman E**, **McHardy AC**, **Rigoutsos I**, **Salamov A**, **Korzeniewski F**, **Land M**, **Lapidus A**, **Grigoriev I**, **Richardson P**, **Hugenholtz P**, **Kyrpides N**. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods **4**:495–500.

100. **Pignatelli M**, **Moya A**. 2011. Evaluating the fidelity of *de novo* short read metagenomic assembly using simulated data. PLoS One **6**:e19984.

101. **Vázquez-Castellanos JF**, **García-López R**, **Pérez-Brocal V**, **Pignatelli M**, **Moya A**. 2014. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. BMC Genomics **15**:37.

102. **Alneberg J**, **Bjarnason BS**, **de Bruijn I**, **Schirmer M**, **Quick J**, **Ijaz UZ**, **Lahti L**, **Loman NJ**, **Andersson AF**, **Quince C**. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods **11**:1144–1146.

103. **Kang DD**, **Froula J**, **Egan R**, **Wang Z**. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3**:e1165.

104. **Wu Y-W**, **Tang Y-H**, **Tringe SG**, **Simmons B a**, **Singer SW**. 2014. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**:26.

105. **Cleary B**, **Brito IL**, **Huang K**, **Gevers D**, **Shea T**, **Young S**, **Alm EJ**. 2015. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol **33**:1053–1060.

106. **Nielsen HB**, **Almeida M**, **Juncker AS**, **Rasmussen S**, **Li J**, **Sunagawa S**, **Plichta DR**, **Gautier L**,

Pedersen AG, **Le Chatelier E**, **Pelletier E**, **Bonde I**, **Nielsen T**, **Manichanh C**, **Arumugam M**, **Batto J-M**, **Quintanilha Dos Santos MB**, **Blom N**, **Borruel N**, **Burgdorf KS**, **Boumezbeur F**, **Casellas F**, **Doré J**, **Dworzynski P**, **Guarner F**, **Hansen T**, **Hildebrand F**, **Kaas RS**, **Kennedy S**, **Kristiansen K**, **Kultima JR**, **Léonard P**, **Levenez F**, **Lund O**, **Moumen B**, **Le Paslier D**, **Pons N**, **Pedersen O**, **Prifti E**, **Qin J**, **Raes J**, **Sørensen S**, **Tap J**, **Tims S**, **Ussery DW**, **Yamada T**, **Renault P**, **Sicheritz-Ponten T**, **Bork P**, **Wang J**, **Brunak S**, **Ehrlich SD**. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol **32**:822–828.

107. **Dostie J**, **Dekker J**. 2007. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc **2**:988–1002.

108. **Belton JM**, **McCord RP**, **Gibcus JH**, **Naumova N**, **Zhan Y**, **Dekker J**. 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods **58**:268–276.

109. **Burton JN**, **Liachko I**, **Dunham MJ**, **Shendure J**. 2014. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 **4**:1339–1346.

110. **Hastie AR**, **Dong L**, **Smith A**, **Finklestein J**, **Lam ET**, **Huo N**, **Cao H**, **Kwok PY**, **Deal KR**, **Dvorak J**, **Luo MC**, **Gu Y**, **Xiao M**. 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. PLoS One **8**:e55864.

111. **Kanehisa M**, **Goto S**, **Sato Y**, **Furumichi M**, **Tanabe M**. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res **40**:D109–D114.

112. **Caspi R**, **Altman T**, **Billington R**, **Dreher K**, **Foerster H**, **Fulcher CA**, **Holland TA**, **Keseler IM**, **Kothari A**, **Kubo A**, **Krummenacker M**, **Latendresse M**, **Mueller LA**, **Ong Q**, **Paley S**, **Subhraveti P**, **Weaver DS**, **Weerasinghe D**, **Zhang P**, **Karp PD**. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res **42**:D459–D471.

113. **Overbeek R**, **Begley T**, **Butler RM**, **Choudhuri J V.**, **Chuang HY**, **Cohoon M**, **de Crécy-Lagard V**, **Diaz N**, **Disz T**, **Edwards R**, **Fonstein M**, **Frank ED**, **Gerdes S**, **Glass EM**, **Goesmann A**, **Hanson A**, **Iwata-Reuyl D**, **Jensen R**, **Jamshidi N**, **Krause L**, **Kubal M**, **Larsen N**, **Linke B**, **McHardy AC**, **Meyer F**, **Neuweger H**, **Olsen G**, **Olson R**, **Osterman A**, **Portnoy V**, **Pusch GD**, **Rodionov DA**, **Rül;ckert C**, **Steiner J**, **Stevens R**, **Thiele I**, **Vassieva O**, **Ye Y**, **Zagnitko O**, **Vonstein V**. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res **33**:5691–5702.

114. **Takami H**, **Taniguchi T**, **Moriya Y**, **Kuwahara T**, **Kanehisa M**, **Goto S**. 2012. Evaluation method for the potential functionome harbored in the genome and metagenome. BMC Genomics **13**:699.

115. **Ye Y**, **Doak TG**. 2011. A parsimony approach to biological pathway reconstruction/inference for metagenomes, p. 453–460. *In* Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., Hoboken, NJ, USA.

116. **Jiao D**, **Ye Y**, **Tang H**. 2013. Probabilistic inference of biochemical reactions in microbial communities from metagenomic sequences. PLoS Comput Biol **9**:e1002981.

117. **Abubucker S**, **Segata N**, **Goll J**, **Schubert AM**, **Izard J**, **Cantarel BL**, **Rodriguez-Mueller B**, **Zucker J**, **Thiagarajan M**, **Henrissat B**, **White O**, **Kelley ST**, **Methé B**, **Schloss PD**, **Gevers D**, **Mitreva M**, **Huttenhower C**. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol **8**:e1002358.

118. **Langille MG**, **Zaneveld J**, **Caporaso JG**, **McDonald D**, **Knights D**, **Reyes JA**, **Clemente JC**, **Burkepile DE**, **Vega Thurber RL**, **Knight R**, **Beiko RG**, **Huttenhower C**. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol **31**:814–821.

119. **Okuda S**, **Tsuchiya Y**, **Kiriyama C**, **Itoh M**, **Morisaki H**. 2012. Virtual metagenome reconstruction from 16S rRNA gene sequences. Nat Commun **3**:1203.

120. **Nijkamp JF**, **Pop M**, **Reinders MJT**, **De Ridder D**. 2013. Exploring variation-aware contig graphs for (comparative) metagenomics using MARYGOLD. Bioinformatics **29**:2826–2834.

121. **Kagan J**, **Sharon I**, **Beja O**, **Kuhn JC**. 2008. The tryptophan pathway genes of the Sargasso Sea metagenome: new operon structures and the prevalence of non-operon organization. Genome Biol **9**:R20.

122. **Suenaga H**, **Koyama Y**, **Miyakoshi M**, **Miyazaki R**, **Yano H**, **Sota M**, **Ohtsubo Y**, **Tsuda M**, **Miyazaki K**. 2009. Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. ISME J **3**:1335–1348.

123. **Gatte-Picchi D**, **Weiz A**, **Ishida K**, **Hertweck C**, **Dittmann E**. 2014. Functional analysis of environmental DNA-derived microviridins provides new insights into the diversity of the tricyclic peptide family. Appl Environ Microbiol **80**:1380–1387.

124. **Korem T**, **Zeevi D**, **Suez J**, **Weinberger A**, **Avnit-Sagi T**, **Pompan-Lotan M**, **Matot E**, **Jona G**, **Harmelin A**, **Cohen N**, **Sirota-Madi A**, **Thaiss CA**, **Pevsner-Fischer M**, **Sorek R**, **Xavier R**, **Elinav E**, **Segal E**. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science **349**:1101–1106.

125. **Ochman H**, **Lawrence JG**, **Groisman E a**. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

126. **Soucy SM**, **Huang J**, **Gogarten JP**. 2015. Horizontal gene transfer: Building the web of life. Nat Rev Genet **16**:472–482.

127. **Roberts AP**, **Kreth J**. 2014. The impact of horizontal gene transfer on the adaptive ability of the human oral microbiome. Front Cell Infect Microbiol **4**:124.

128. **Chen J**, **Novick RP**. 2009. Transfer of toxin genes. Science **323**:139–141.

129. **Smillie CS**, **Smith MB**, **Friedman J**, **Cordero OX**, **David LA**, **Alm EJ**. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature **480**:241–244.

130. **Iwasaki W**, **Takagi T**. 2009. Rapid pathway evolution facilitated by horizontal gene transfers across prokaryotic lineages. PLoS Genet **5**:e1000402.

131. **Tasse L**, **Bercovici J**, **Pizzut-Serin S**, **Robe P**, **Tap J**, **Klopp C**, **Cantarel BL**, **Coutinho PM**, **Henrissat B**, **Leclerc M**, **Doré J**, **Monsan P**, **Remaud-Simeon M**, **Potocki-Veronese G**. 2010. Functional

metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. Genome Res **20**:1605–1612.

132. **Palenik B**, **Ren Q**, **Tai V**, **Paulsen IT**. 2009. Coastal *Synechococcus* metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. Environ Microbiol **11**:349–359.

133. **Guo J**, **Wang Q**, **Wang X**, **Wang F**, **Yao J**, **Zhu H**. 2015. Horizontal gene transfer in an acid mine drainage microbial community. BMC Genomics **16**:496.

134. **Tamames J**, **Moya A**. 2008. Estimating the extent of horizontal gene transfer in metagenomic sequences. BMC Genomics **9**:136.

135. **Silver S**, **Misra TK**. 1988. Plasmid-mediated heavy metal resistances. Annu Rev Microbiol **42**:717–743.

136. **Smalla K**, **Sobecky PA**. 2002. The prevalence and diversity of mobile genetic elements in bacterial communities of different environmental habitats: Insights gained from different methodological approaches. FEMS Microbiol Ecol **42**:165–175.

137. **Top EM**, **Holben WE**, **Forney LJ**. 1995. Characterization of diverse 2,4-dichlorophenoxyacetic acid-degradative plasmids isolated from soil by complementation. Appl Environ Microbiol **61**:1691–8.

138. **Bennett PM**. 2008. Plasmid encoded antibiotic resistance: Acquisition and transfer of antibiotic resistance genes in bacteria. Br J Pharmacol **153**:S347–S357.

139. **Walker A**. 2012. Welcome to the plasmidome. Nat Rev Microbiol **467**:379.

140. **Dib JR**, **Wagenknecht M**, **Farías ME**, **Meinhardt F**. 2015. Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements? Front Microbiol **6**:463.

141. **Brown Kav A**, **Sasson G**, **Jami E**, **Doron-Faigenboim A**, **Benhar I**, **Mizrahi I**. 2012. Insights into the bovine rumen plasmidome. Proc Natl Acad Sci U S A **109**:5452–5457.

142. **Jørgensen TS**, **Xu Z**, **Hansen MA**, **Sørensen SJ**, **Hansen LH**. 2014. Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. PLoS One **9**:e87924.

143. **Sentchilo V**, **Mayer AP**, **Guy L**, **Miyazaki R**, **Green Tringe S**, **Barry K**, **Malfatti S**, **Goessmann A**, **Robinson-Rechavi M**, **van der Meer JR**. 2013. Community-wide plasmid gene mobilization and selection. ISME J **7**:1173–1186.

144. **Matus-Garcia M**, **Nijveen H**, **Van Passel MWJ**. 2012. Promoter propagation in prokaryotes. Nucleic Acids Res **40**:10032–10040.

145. **Oren Y**, **Smith MB**, **Johns NI**, **Kaplan Zeevi M**, **Biran D**, **Ron EZ**, **Corander J**, **Wang HH**, **Alm EJ**, **Pupko T**. 2014. Transfer of noncoding DNA drives regulatory rewiring in bacteria. Proc Natl Acad Sci U S A **111**:16112–16117.

146. **Fernandez L**, **Mercader JM**, **Planas-Fèlix M**, **Torrents D**. 2014. Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities. BMC Genomics **15**:877.

147. **Boon E**, **Meehan CJ**, **Whidden C**, **Wong DHJ**, **Langille MGI**, **Beiko RG**. 2014. Interactions in the microbiome: Communities of organisms and communities of genes. FEMS Microbiol Rev **38**:90–118.

148. **Zhang Y**, **Lun CY**, **Tsui SKW**. 2015. Metagenomics: A new way to illustrate the crosstalk between

infectious diseases and host microbiome. Int J Mol Sci **16**:26263–26279.

149. **Faust K**, **Raes J**. 2012. Microbial interactions: From networks to models. Nat Rev Microbiol **10**:538–550.

150. **Toju H**, **Guimarães PR**, **Olesen JM**, **Thompson JN**. 2014. Assembly of complex plant–fungus networks. Nat Commun **5**:5273.

151. **Chaffron S**, **Rehrauer H**, **Pernthaler J**, **Mering C**. 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res **2010**:947–959.

152. **Beman JM**, **Steele JA**, **Fuhrman JA**. 2011. Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. ISME J **5**:1077–1085.

153. **Gilbert JA**, **Steele JA**, **Caporaso JG**, **Steinbrück L**, **Reeder J**, **Temperton B**, **Huse S**, **McHardy AC**, **Knight R**, **Joint I**, **Somerfield P**, **Fuhrman JA**, **Field D**. 2012. Defining seasonal marine microbial community dynamics. ISME J **6**:298–308.

154. **Chow C-ET**, **Kim DY**, **Sachdeva R**, **Caron DA**, **Fuhrman JA**. 2014. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. ISME J **8**:816–829.

155. **Lozupone C**, **Faust K**, **Raes J**, **Faith JJ**, **Frank DN**, **Zaneveld J**, **Gordon JI**, **Knight R**. 2012. Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. Genome Res **22**:1974–1984.

156. **Navarrete AA**, **Tsai SM**, **Mendes LW**, **Faust K**, **De Hollander M**, **Cassman NA**, **Raes J**, **Van Veen JA**, **Kuramae EE**. 2015. Soil microbiome responses to the short-term effects of Amazonian deforestation. Mol Ecol **24**:2433–2448.

157. **Soffer N**, **Zaneveld J**, **Vega Thurber R**. 2015. Phage-bacteria network analysis and its implication for the understanding of coral disease. Environ Microbiol **17**:1203–1218.

158. **Miya M**, **Sato Y**, **Fukunaga T**, **Sado T**, **Poulsen JY**, **Sato K**, **Minamoto T**, **Yamamoto S**, **Yamanaka H**, **Araki H**, **Kondoh M**, **Iwasaki W**. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. R Soc Open Sci **2**:150088.

159. **Santos F**, **Yarza P**, **Parro V**, **Meseguer.I.Inmaculada**, **Rosselló-Móra R**, **Antón J**. 2012. Culture-independent approaches for studying viruses from hypersaline environments. Appl Environ Microbiol **78**:1635–1643.

160. **Santos F**, **Yarza P**, **Parro V**, **Briones C**, **Antón J**. 2010. The metavirome of a hypersaline environment. Environ Microbiol **12**:2965–2976.

161. **Minot S**, **Sinha R**, **Chen J**, **Li H**, **Keilbaugh SA**, **Wu GD**, **Lewis JD**, **Bushman FD**. 2011. The human gut virome: Inter-individual variation and dynamic response to diet. Genome Res **21**:1616–1625.

162. **Modi SR**, **Lee HH**, **Spina CS**, **Collins JJ**. 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature **499**:219–22.

163. **Dutilh BE**, **Cassman N**, **McNair K**, **Sanchez SE**, **Silva GGZ**, **Boling L**, **Barr JJ**, **Speth DR**, **Seguritan V**,

Aziz RK, **Felts B**, **Dinsdale EA**, **Mokili JL**, **Edwards RA**. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun **5**:4498.

164. **Urayama S**, **Yoshida-Takashima Y**, **Yoshida M**, **Tomaru Y**, **Moriyama H**, **Takai K**, **Nunoura T**. 2015. A new fractionation and recovery method of viral genomes based on nucleic acid composition and structure using tandem column chromatography. Microbes Environ **30**:199–203.

165. **Wylie TN**, **Wylie KM**, **Herter BN**, **Storch GA**, **Wylie TN**. 2015. Enhanced virome sequencing using targeted sequence capture. Genome **25**:1910–1920.

166. **Rampelli S**, **Soverini M**, **Turroni S**, **Quercia S**, **Biagi E**, **Brigidi P**, **Candela M**. 2016. ViromeScan: A new tool for metagenomic viral community profiling. BMC Genomics **17**:165.

167. **Roux S**, **Tournayre J**, **Mahul A**, **Debroas D**, **Enault F**. 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics **15**:76.

168. **Bolotin A**, **Quinquis B**, **Sorokin A**, **Dusko Ehrlich S**. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151**:2551–2561.

169. **Pourcel C**, **Salvignol G**, **Vergnaud G**. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151**:653–663.

170. **Mojica FJM**, **Díez-Villaseñor C**, **García-Martínez J**, **Soria E**. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol **60**:174–182.

171. **Grissa I**, **Vergnaud G**, **Pourcel C**. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformaticss **8**:172.

172. **Makarova KS**, **Wolf YI**, **Alkhnbashi OS**, **Costa F**, **Shah SA**, **Saunders SJ**, **Barrangou R**, **Brouns SJJ**, **Charpentier E**, **Haft DH**, **Horvath P**, **Moineau S**, **Mojica FJM**, **Terns RM**, **Terns MP**, **White MF**, **Yakunin AF**, **Garrett RA**, **van der Oost J**, **Backofen R**, **Koonin E V**. 2015. An updated evolutionary classification of CRISPR–Cas systems. Nat Rev Microbiol **13**:722.

173. **Kodama Y**, **Shumway M**, **Leinonen R**. 2012. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res **40**:D54–D56.

174. **Hildebrand F**, **Meyer A**, **Eyre-walker A**. 2010. Evidence of selection on genomic GC content in bacteria. PLoS Genet **6**:e1001107.

175. **Bentkowski P**, **Van Oosterhout C**, **Mock T**. 2015. A model of genome size evolution for prokaryotes in stable and fluctuating environments. Genome Biol Evol **7**:2344–2351.

176. **Yang C-C**, **Iwasaki W**. 2014. MetaMetaDB: A database and analytic system for investigating microbial habitability. PLoS One **9**:e87126.

177. **Gianoulis TA**, **Raes J**, **Patel P V**, **Bjornson R**, **Korbel JO**, **Letunic I**, **Yamada T**, **Paccanaro A**, **Jensen LJ**, **Snyder M**, **Bork P**, **Gerstein MB**. 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A **106**:1374–1379.

178. **Grzymski JJ**, **Murray AE**, **Campbell BJ**, **Kaplarevic M**, **Gao GR**, **Lee C**, **Daniel R**, **Ghadiri A**,

Feldman RA, **Cary SC**. 2008. Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. Proc Natl Acad Sci U S A **105**:17516–17521.

179. **Mirete S**, **De Figueras CG**, **González-Pastor JE**. 2007. Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. Appl Environ Microbiol **73**:6001–6011.

180. **Hemme CL**, **Deng Y**, **Gentry TJ**, **Fields MW**, **Wu L**, **Barua S**, **Barry K**, **Tringe SG**, **Watson DB**, **He Z**, **Hazen TC**, **Tiedje JM**, **Rubin EM**, **Zhou J**. 2010. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME J **4**:660–672.

181. **Trabelsi D**, **Mengoni A**, **Aouani ME**, **Bazzicalupo M**, **Mhamdi R**. 2010. Genetic diversity and salt tolerance of sinorhizobium populations from two tunisian soils. Ann Microbiol **60**:541–547.

182. **Patel P V.**, **Gianoulis TA**, **Bjornson RD**, **Yip KY**, **Engelman DM**, **Gerstein MB**. 2010. Analysis of membrane proteins in metagenomics: Networks of correlated environmental features and protein families. Genome Res **20**:960–971.

183. **Lax S**, **Smith DP**, **Hampton-Marcell J**, **Owens SM**, **Handley KM**, **Scott NM**, **Gibbons SM**, **Larsen P**, **Shogan BD**, **Weiss S**, **Metcalf JL**, **Ursell LK**, **Vazquez-Baeza Y**, **Van Treuren W**, **Hasan NA**, **Gibson MK**, **Colwell R**, **Dantas G**, **Knight R**, **Gilbert JA**. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science **345**:1048–1052.

184. **Hiraoka S**, **Machiyama A**, **Ijichi M**, **Inoue K**, **Oshima K**, **Hattori M**, **Yoshizawa S**, **Kogure K**, **Iwasaki W**. 2016. Genomic and metagenomic analysis of microbes in a soil environment affected by the 2011 Great East Japan Earthquake tsunami. BMC Genomics **17**:53.

185. **Alquezar-Planas DE**, **Mourier T**, **Bruhn CAW**, **Hansen AJ**, **Vitcetz SN**, **Mørk S**, **Gorodkin J**, **Nielsen HA**, **Guo Y**, **Sethuraman A**, **Paxinos EE**, **Shan T**, **Delwart EL**, **Nielsen LP**. 2013. Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing. Sci Rep **3**:2468.

186. **Ikuta T**, **Takaki Y**, **Nagai Y**, **Shimamura S**, **Tsuda M**, **Kawagucci S**, **Aoki Y**, **Inoue K**, **Teruya M**, **Satou K**, **Teruya K**, **Shimoji M**, **Tamotsu H**, **Hirano T**, **Maruyama T**, **Yoshida T**. 2016. Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J **10**:990–1001.

187. **Flusberg BA**, **Webster DR**, **Lee JH**, **Travers KJ**, **Olivares EC**, **Clark TA**, **Korlach J**, **Turner SW**. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods **7**:461–465.

188. **Clark TA**, **Murray IA**, **Morgan RD**, **Kislyuk AO**, **Spittle KE**, **Boitano M**, **Fomenkov A**, **Roberts RJ**, **Korlach J**. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res **40**:e29.

189. **Blow MJ**, **Clark TA**, **Daum CG**, **Deutschbauer AM**, **Fomenkov A**, **Fries R**, **Froula J**, **Kang DD**, **Malmstrom RR**, **Morgan RD**, **Posfai J**, **Singh K**, **Visel A**, **Wetmore K**, **Zhao Z**, **Rubin EM**, **Korlach J**, **Pennacchio LA**, **Roberts RJ**. 2016. The Epigenomic Landscape of Prokaryotes. PLoS Genet **12**:e1005854.

190. **Kim SW**, **Suda W**, **Kim S**, **Oshima K**, **Fukuda S**, **Ohno H**, **Morita H**, **Hattori M**. 2013. Robustness of

gut microbiota of healthy adults in response to probiotic intervention revealed by high-throughput pyrosequencing. DNA Res **20**:241–253.

191. **Morita H**, **Kuwahara T**, **Ohshima K**, **Sasamoto H**, **Itoh K**, **Hattori M**, **Hayashi T**, **Takami H**. 2007. An improved DNA isolation method for metagenomic analysis of the microbial flora of the human intestine. Microbes Environ **22**:214–222.

192. **Hiraoka S**, **Yang C**, **Iwasaki W**. 2016. Metagenomics and bioinformatics in microbial ecology: Current status and beyond. Microbes Environ **31**:204–212.

193. **Hiraoka S**, **Miyahara M**, **Fujii K**, **Machiyama A**, **Iwasaki W**. 2017. Seasonal Analysis of Microbial Communities in Precipitation in the Greater Tokyo Area, Japan. Front Microbiol **8**:1506.

194. **Simons M**, **Minson SE**, **Sladen A**, **Ortega F**, **Jiang J**, **Owen SE**, **Meng L**, **Ampuero J-P**, **Wei S**, **Chu R**, **Helmberger D V**, **Kanamori H**, **Hetland E**, **Moore AW**, **Webb FH**. 2011. The 2011 Magnitude 9.0 Tohoku-Oki Earthquake: Mosaicking the Megathrust from Seconds to Centuries. Science **332**:1421–1425.

195. **McLeod MK**, **Slavich PG**, **Irhas Y**, **Moore N**, **Rachman A**, **Ali N**, **Iskandar T**, **Hunt C**, **Caniago C**. 2010. Soil salinity in Aceh after the December 2004 Indian Ocean tsunami. Agric Water Manag **97**:605–613.

196. **Rengalakshmi R**, **Senthilkumar R**, **Selvarasu T**, **Thamizoli P**. 2007. Reclamation and status of tsunami damaged soil in Nagappattinam District, Tamil Nadu. Curr Sci **92**:1221–1223.

197. **Ranjan RK**, **Ramanathan AL**, **Singh G**. 2008. Evaluation of geochemical impact of tsunami on Pichavaram mangrove ecosystem, southeast coast of India. Environ Geol **55**:687–697.

198. **Szczuciński W**, **Niedzielski P**, **Rachlewicz G**, **Sobczyński T**, **Zioła A**, **Kowalski A**, **Lorenc S**, **Siepak J**. 2005. Contamination of tsunami sediments in a coastal zone inundated by the 26 December 2004 tsunami in Thailand. Environ Geol **49**:321–331.

199. **Srinivasalu S**, **Thangadurai N**, **Jonathan MP**, **Armstrong-Altrin JS**, **Ayyamperumal T**, **Ram-Mohan V**. 2008. Evaluation of trace-metal enrichments from the 26 December 2004 tsunami sediments along the Southeast coast of India. Environ Geol **53**:1711–1721.

200. **Prasath P**, **Khan TH**. 2008. Impact of tsunami on the heavy metal accumulation in water, Sediments and fish at Poompuhar coast, Southeast Coast of India. J Chem **5**:16–22.

201. **Chandrasekharan H**, **Sarangi A**, **Nagarajan M**, **Singh VP**, **Rao DUM**, **Stalin P**, **Natarajan K**, **Chandrasekaran B**, **Anbazhagan S**. 2008. Variability of soil-water quality due to Tsunami-2004 in the coastal belt of Nagapattinam district, Tamilnadu. J Environ Manage **89**:63–72.

202. **Szczucinski W**, **Chaimanee N**, **Niedzielski P**, **Rachlewicz G**, **Saisuttichai D**, **Tepsuwan T**, **Lorenc S**, **Siepak J**. 2006. Environmental and geological impacts of the 26 December 2004 tsunami in coastal zone of Thailand-overview of short and long-term effects. Polish J Environ Stud **15**:793–810.

203. **Curran PJ**, **Dash J**, **Llewellyn GM**. 2007. Indian Ocean tsunami: The use of MERIS (MTCI) data to infer salt stress in coastal vegetation. Int J Remote Sens **28**:729–735.

204. **Hayasaka D**, **Shimada N**, **Konno H**, **Sudayama H**, **Kawanishi M**, **Uchida T**, **Goka K**. 2012. Floristic variation of beach vegetation caused by the 2011 Tohoku-oki tsunami in northern Tohoku, Japan. Ecol Eng

44:227–232.

205. **Somboonna N**, **Wilantho A**, **Jankaew K**, **Assawamakin A**, **Sangsrakru D**, **Tangphatsornruang S**, **Tongsima S**. 2014. Microbial ecology of Thailand tsunami and non-tsunami affected terrestrials. PLoS One **9**:e94236.

206. **Wada K**, **Fukuda K**, **Yoshikawa T**, **Hirose T**, **Ikeno T**, **Umata T**, **Irokawa T**, **Taniguchi H**, **Aizawa Y**. 2012. Bacterial hazards of sludge brought ashore by the tsunami after the great East Japan earthquake of 2011. J Occup Health **54**:255–262.

207. **Crosa JH**. 1989. Genetics and molecular biology of siderophore-mediated iron transport in bacteria. Microbiol Rev **53**:517–530.

208. **Neilands JB**. 1995. Siderophores: Structure and function of microbial iron transport compounds. J Biol Chem **270**:26723–26726.

209. **Hider RC**, **Kong X**. 2010. Chemistry and biology of siderophores. Nat Prod Rep **27**:637–657.

210. **Lane DJ**. 1991. 16S/23S rRNA sequencing, p. 125–175. *In* Stackebrandt, E, Goodfellow, M (eds.), Nucleic acid techniques in bacterial systematics. John Wiley and Sons, Chichester, United Kingdom.

211. **Margulies M**, **Egholm M**, **Altman WE**, **Attiya S**, **Bader JS**, **Bemben LA**, **Berka J**, **Braverman MS**, **Chen Y-J**, **Chen Z**, **Dewell SB**, **Du L**, **Fierro JM**, **Gomes X V.**, **Godwin BC**, **He W**, **Helgesen S**, **Begley RF**, **Rothberg JM**. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376–380.

212. **Hyatt D**, **Chen G-L**, **LoCascio P**, **Land M**, **Larimer F**, **Hauser L**. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**:119.

213. **Camacho C**, **Coulouris G**, **Avagyan V**, **Ma N**, **Papadopoulos J**, **Bealer K**, **Madden TL**. 2009. BLAST+: Architecture and applications. BMC Bioinformatics **10**:1–9.

214. **UniProt Consortium**. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res **41**:D43–D47.

215. **Powell S**, **Forslund K**, **Szklarczyk D**, **Trachana K**, **Roth A**, **Huerta-Cepas J**, **Gabaldón T**, **Rattei T**, **Creevey C**, **Kuhn M**, **Jensen LJ**, **Von Mering C**, **Bork P**. 2014. EggNOG v4.0: Nested orthology inference across 3686 organisms. Nucleic Acids Res **42**:D231–D239.

216. **Lowe TM**, **Eddy SR**. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25**:955–964.

217. **Lagesen K**, **Hallin P**, **Rødland EA**, **Stærfeldt H-H**, **Rognes T**, **Ussery DW**. 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res **35**:3100–3108.

218. **Benson DA**, **Clark K**, **Karsch-Mizrachi I**, **Lipman DJ**, **Ostell J**, **Sayers EW**. 2014. GenBank. Nucleic Acids Res **42**:D32–D37.

219. **Yao Y**, **Tang H**, **Su F**, **Xu P**. 2015. Comparative genome analysis reveals the molecular basis of nicotine degradation and survival capacities of *Arthrobacter*. Sci Rep **5**:8642.

220. **Edgar RC**. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic

Acids Res **32**:1792–1797.

221. **Tamura K**, **Stecher G**, **Peterson D**, **Filipski A**, **Kumar S**. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol **30**:2725–2729.

222. **Segata N**, **Börnigen D**, **Morgan XC**, **Huttenhower C**. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun **4**:2304.

223. **Payne SM**. 1994. Detection, isolation, and characterization of siderophores. Methods Enzymol **235**:329–344.

224. **Niu B**, **Fu L**, **Sun S**, **Li W**. 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics **11**:187.

225. **Hyatt D**, **LoCascio PF**, **Hauser LJ**, **Uberbacher EC**. 2012. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics **28**:2223–2230.

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

227. **Jones D**, **Keddie RM**. 2006. The genus *Arthrobacter*, p. 945–960. *In* The Prokaryotes. Springer New York, New York, United States.

228. **Cacciari I**, **Lippi D**. 1987. Arthrobacters: Successful arid soil bacteria: A review. Arid L Res Manag **1**:1–30.

229. **Crocker FH**, **Fredrickson JK**, **White DC**, **Ringelberg DB**, **Balkwill DL**. 2000. Phylogenetic and physiological diversity of *Arthrobacter* strains isolated from unconsolidated subsurface sediments. Microbiology **146**:1295–1310.

230. **Bowman JP**, **McCammon SA**, **Brown M V**, **Nichols DS**, **McMeekin TA**. 1997. Diversity and association of psychrophilic bacteria in Antarctic sea ice. Appl Environ Microbiol **63**:3068–3078.

231. **Fredrickson JK**, **Zachara JM**, **Balkwill DL**, **Kennedy D**, **Shu-mei WL**, **Kostandarithes HM**, **Daly MJ**, **Romine MF**, **Brockman FJ**. 2004. Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford Site, Washington State. Appl Environ Microbiol **70**:4230–4241.

232. **Trajanovska S**, **Britz ML**, **Bhave M**. 1997. Detection of heavy metal ion resistance genes in Gram-positive and Gram-negative bacteria isolated from a lead-contaminated site. Biodegradation **8**:113–124.

233. **Boylen CW**. 1973. Survival of *Arthrobacter crystallopoietes* during prolonged periods of extreme desiccation. J Bacteriol **113**:33–37.

234. **Ensign JC**. 1970. Long-term starvation survival of rod and spherical cells of *Arthrobacter crystallopoietes*. J Bacteriol **103**:569–577.

235. **Margesin R**, **Schinner F**. 1997. Heavy metal resistant *Arthrobacter* sp. - A tool for studying conjugational plasmid transfer between Gram-negative and Gram-positive bacteria. J Basic Microbiol **37**:217–227.

236. **Bafana A**, **Krishnamurthi K**, **Patil M**, **Chakrabarti T**. 2010. Heavy metal resistance in *Arthrobacter ramosus* strain G2 isolated from mercuric salt-contaminated soil. J Hazard Mater **177**:481–486.

237. **Suzuki Y**, **Banfield JF**. 2004. Resistance to, and accumulation of, uranium by bacteria from a uranium-

contaminated site. Geomicrobiol J **21**:113–121.

238. **Guerinot M Lou**. 1994. Microbial iron transport. Annu Rev Microbiol **48**:743–772.

239. **Sugiura Y**, **Nomoto K**. 1984. Phytosiderophores structures and properties of mugineic acids and their metal complexes, p. 107–135. *In* Siderophores from Microorganisms and Plants. Springer Berlin Heidelberg, Berlin, Heidelberg.

240. **Jurkevitch E**, **Hadar Y**, **Chen Y**. 1992. Differential siderophore utilization and iron uptake by soil and rhizosphere bacteria. Appl Environ Microbiol **58**:119–124.

241. **Barona-Gómez F**, **Wong U**, **Giannakopulos AE**, **Derrick PJ**, **Challis GL**. 2004. Identification of a cluster of genes that directs desferrioxamine biosynthesis in *Streptomyces coelicolor* M145. J Am Chem Soc **126**:16282–16283.

242. **Günter K**, **Toupet C**, **Schupp T**. 1993. Characterization of an iron-regulated promoter involved in desferrioxamine B synthesis in *Streptomyces pilosus*: Repressor-binding site and homology to the diphtheria toxin gene promoter. J Bacteriol **175**:3295–3302.

243. **Schwyn B**, **Neilands JB**. 1987. Universal chemical assay for the detection and determination of siderophores. Anal Biochem **160**:47–56.

244. **Shenker M**, **Chen Y**. 2005. Increasing iron availability to crops: Fertilizers, organo-fertilizers, and biological approaches. Soil Sci Plant Nutr **51**:1–17.

245. **Radzki W**, **Gutierrez Mañero FJ**, **Algar E**, **Lucas García JA**, **García-Villaraco A**, **Ramos Solano B**. 2013. Bacterial siderophores efficiently provide iron to iron-starved tomato plants in hydroponics culture. Antonie Van Leeuwenhoek **104**:321–330.

246. **Duhan JS**, **Dudeja SS**, **Khurana AL**. 1998. Siderophore production in relation to N2 fixation and iron uptake in pigeon pea-*Rhizobium* symbiosis. Folia Microbiol (Praha) **43**:421–426.

247. **Jflrgensen BB**. 1977. The sulfur cycle of a coastal marine sediment (Limfjorden, Denmark). Limnol Oceanogr **22**:814–832.

248. **Schippers A**, **Jørgensen BB**. 2002. Biogeochemistry of pyrite and iron sulfide oxidation in marine sediments. Geochim Cosmochim Acta **66**:85–92.

249. **Garcia-Gil LJ**, **Golterman HL**. 1993. Kinetics of FeS-mediated denitrification in sediments from the Camargue (Rhone delta, southern France). FEMS Microbiol Ecol **13**:85–91.

250. **Straub KL**, **Benz M**, **Schink B**, **Widdel F**. 1996. Anaerobic, nitrate-dependent microbial oxidation of ferrous iron. Appl Environ Microbiol **62**:1458–1460.

251. **Hauck S**, **Benz M**, **Brune A**, **Schink B**. 2001. Ferrous iron oxidation by denitrifying bacteria in profundal sediments of a deep lake (Lake Constance). FEMS Microbiol Ecol **37**:127–134.

252. **Haaijer SCM**, **Lamers LPM**, **Smolders AJP**, **Jetten MSM**, **den Camp HJM**. 2007. Iron sulfide and pyrite as potential electron donors for microbial nitrate reduction in freshwater wetlands. Geomicrobiol J **24**:391–401.

253. **Szczuciński W**, **Niedzielski P**, **Kozak L**, **Frankowski M**, **Zioła A**, **Lorenc S**. 2007. Effects of rainy season

on mobilization of contaminants from tsunami deposits left in a coastal zone of Thailand by the 26 December 2004 tsunami. Environ Geol **53**:253–264.

254. **Ogawa Y**, **Ooka T**, **Shi F**, **Ogura Y**, **Nakayama K**, **Hayashi T**, **Shimoji Y**. 2011. The Genome of Erysipelothrix rhusiopathiae, the Causative Agent of Swine Erysipelas, Reveals New Insights into the Evolution of Firmicutes and the Organism's Intracellular Adaptations. J Bacteriol **193**:2959–2971.

255. **Veraldi S**, **Girgenti V**, **Dassoni F**, **Gianotti R**. 2009. Erysipeloid: A review. Clin Exp Dermatol **34**:859–862.

256. **Makita K**, **Inoshita K**, **Kayano T**, **Uenoyama K**, **Hagiwara K**, **Asakawa M**, **Ogawa K**, **Kawamura S**, **Noda J**, **Sera K**, **Sasaki H**, **Nakatani N**, **Higuchi H**, **Ishikawa N**, **Iwano H**, **Tamura Y**. 2013. Temporal changes in environmental health risks and socio-psychological status in areas affected by the 2011 tsunami in Ishinomaki, Japan. Environ Pollut **3**:p1.

257. **Cho J-C**, **Giovannoni SJ**. 2003. *Croceibacter atlanticus* gen. nov., sp. nov., a novel marine bacterium in the family Flavobacteriaceae. Syst Appl Microbiol **26**:76–83.

258. **Prieur D**, **Marteinsson VT**, **Alain K**, **Bonch-Osmolovskaya EA**, **Miroshnichenko ML**, **Birrien J-L**. 2002. *Marinitoga piezophila* sp. nov., a rod-shaped, thermo-piezophilic bacterium isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. Int J Syst Evol Microbiol **52**:1331–1339.

259. **Fiala G**, **Stetter KO**. 1986. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch Microbiol **145**:56–61.

260. **Erauso G**, **Reysenbach A-L**, **Godfroy A**, **Meunier J-R**, **Crump B**, **Partensky F**, **Baross JA**, **Marteinsson V**, **Barbier G**, **Pace NR**, **Prieur D**. 1993. *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Arch Microbiol **160**:338–349.

261. **Zeng X**, **Birrien J-L**, **Fouquet Y**, **Cherkashov G**, **Jebbar M**, **Querellou J**, **Oger P**, **Cambon-Bonavita M-A**, **Xiao X**, **Prieur D**. 2009. Pyrococcus CH1, an obligate piezophilic hyperthermophile: Extending the upper pressure-temperature limits for life. ISME J **3**:873–876.

262. **Oren A**, **Heldal M**, **Norland S**, **Galinski EA**. 2002. Intracellular ion and organic solute concentrations of the extremely halophilic bacterium *Salinibacter ruber*. Extremophiles **6**:491–498.

263. **Kosono S**, **Haga K**, **Tomizawa R**, **Kajiyama Y**, **Hatano K**, **Takeda S**, **Wakai Y**, **Hino M**, **Kudo T**. 2005. Characterization of a multigene-encoded sodium/hydrogen antiporter (Sha) from *Pseudomonas aeruginosa*: Its involvement in pathogenesis. J Bacteriol **187**:5242–5248.

264. **Sugawara D**, **Goto K**, **Imamura F**, **Matsumoto H**, **Minoura K**. 2012. Assessing the magnitude of the 869 Jogan tsunami using sedimentary deposits: Prediction and consequence of the 2011 Tohoku-oki tsunami. Sediment Geol **282**:14–26.

265. **Després V**, **Huffman JA**, **Burrows SM**, **Hoose C**, **Safatov A**, **Buryak G**, **Fröhlich-Nowoisky J**, **Elbert W**, **Andreae M**, **Pöschl U**, **Jaenicke R**. 2012. Primary biological aerosol particles in the atmosphere: A review. Tellus B Chem Phys Meteorol **64**:15598.

266. **Smith DJ**. 2013. Microbes in the upper atmosphere and unique opportunities for astrobiology research.

Astrobiology **13**:981–990.

267. **Burrows SM**, **Elbert W**, **Lawrence MG**, **Pöschl U**. 2009. Bacteria in the global atmosphere—Part 1: Review and synthesis of literature data for different ecosystems. Atmos Chem Phys **9**:9263–9280.

268. **Kellogg CA**, **Griffin DW**. 2006. Aerobiology and the global transport of desert dust. Trends Ecol Evol **21**:638–644.

269. **Fröhlich-Nowoisky J**, **Kampf CJ**, **Weber B**, **Huffman JA**, **Pöhlker C**, **Andreae MO**, **Lang-Yona N**, **Burrows SM**, **Gunthe SS**, **Elbert W**, **Su H**, **Hoor P**, **Thines E**, **Hoffmann T**, **Després VR**, **Pöschl U**. 2016. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. Atmos Res **182**:346–376.

270. **Zweifel UL**, **Hagström Å**, **Holmfeldt K**, **Thyrhaug R**, **Geels C**, **Frohn LM**, **Skjøth CA**, **Karlson UG**. 2012. High bacterial 16S rRNA gene diversity above the atmospheric boundary layer. Aerobiologia (Bologna) **28**:481–498.

271. **Bowers RM**, **McLetchie S**, **Knight R**, **Fierer N**. 2011. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. ISME J **5**:601–612.

272. **Bowers RM**, **Clements N**, **Emerson JB**, **Wiedinmyer C**, **Hannigan MP**, **Fierer N**. 2013. Seasonal variability in bacterial and fungal diversity of the near-surface atmosphere. Environ Sci Technol **47**:12097–12106.

273. **DeLeon-Rodriguez N**, **Lathem TL**, **Rodriguez-R LM**, **Barazesh JM**, **Anderson BE**, **Beyersdorf AJ**, **Ziemba LD**, **Bergin M**, **Nenes A**, **Konstantinidis KT**. 2013. Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. Proc Natl Acad Sci U S A **110**:2575–2580.

274. **Woo AC**, **Brar MS**, **Chan Y**, **Lau MCY**, **Leung FCC**, **Scott JA**, **Vrijmoed LLP**, **Zawar-Reza P**, **Pointing SB**. 2013. Temporal variation in airborne microbial populations and microbially-derived allergens in a tropical urban landscape. Atmos Environ **74**:291–300.

275. **Dong L**, **Qi J**, **Shao C**, **Zhong X**, **Gao D**, **Cao W**, **Gao J**, **Bai R**, **Long G**, **Chu C**. 2016. Concentration and size distribution of total airborne microbes in hazy and foggy weather. Sci Total Environ **541**:1011–1018.

276. **Bowers RM**, **Sullivan AP**, **Costello EK**, **Collett JL**, **Knight R**, **Fierer N**. 2011. Sources of bacteria in outdoor air across cities in the midwestern United States. Appl Environ Microbiol **77**:6350–6356.

277. **Vaïtilingom M**, **Attard E**, **Gaiani N**, **Sancelme M**, **Deguillaume L**, **Flossmann AI**, **Amato P**, **Delort A-M**. 2012. Long-term features of cloud microbiology at the puy de Dôme (France). Atmos Environ **56**:88–100.

278. **Bowers RM**, **Lauber CL**, **Wiedinmyer C**, **Hamady M**, **Hallar AG**, **Fall R**, **Knight R**, **Fierer N**. 2009. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. Appl Environ Microbiol **75**:5121–5130.

279. **Maki T**, **Aoki K**, **Kobayashi F**, **Kakikawa M**, **Tobo Y**, **Matsuki A**, **Hasegawa H**, **Iwasaka Y**. 2011.

Characterization of halotolerant and oligotrophic bacterial communities in Asian desert dust (KOSA) bioaerosol accumulated in layers of snow on Mount Tateyama, Central Japan. Aerobiologia (Bologna) **27**:277–290.

280. **Echigo A**, **Hino M**, **Fukushima T**, **Mizuki T**, **Kamekura M**, **Usami R**. 2005. Endospores of halophilic bacteria of the family Bacillaceae isolated from non-saline Japanese soil may be transported by Kosa event (Asian dust storm). Saline Systems **1**:8.

281. **Rodó X**, **Ballester J**, **Cayan D**, **Melish ME**, **Nakamura Y**, **Uehara R**, **Burns JC**. 2011. Association of Kawasaki disease with tropospheric wind patterns. Sci Rep **1**:152.

282. **Rodó X**, **Curcoll R**, **Robinson M**, **Ballester J**, **Burns JC**, **Cayan DR**, **Lipkin WI**, **Williams BL**, **Couto-Rodriguez M**, **Nakamura Y**, **Uehara R**, **Tanimoto H**, **Morguí J-A**. 2014. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proc Natl Acad Sci U S A **111**:7952–7957.

283. **Brown JKM**, **Hovmøller MS**. 2002. Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. Science **297**:537–541.

284. **Fitt BDL**, **McCartney HA**, **Walklate PJ**. 1989. The role of rain in dispersal of pathogen inoculum. Annu Rev Phytopathol **27**:241–270.

285. **Kaushik R**, **Balasubramanian R**, **de la Cruz AA**. 2012. Influence of air quality on the composition of microbial pathogens in fresh rainwater. Appl Environ Microbiol **78**:2813–2818.

286. **Hervàs A**, **Camarero L**, **Reche I**, **Casamayor EO**. 2009. Viability and potential for immigration of airborne bacteria from Africa that reach high mountain lakes in Europe. Environ Microbiol **11**:1612–1623.

287. **Peter H**, **Hörtnagl P**, **Reche I**, **Sommaruga R**. 2014. Bacterial diversity and composition during rain events with and without Saharan dust influence reaching a high mountain lake in the Alps. Environ Microbiol Rep **6**:618–624.

288. **Hamilton WD**, **Lenton TM**. 1998. Spora and Gaia: How microbes fly with their clouds. Ethol Ecol Evol **10**:1–16.

289. **Christner BC**, **Morris CE**, **Foreman CM**, **Cai R**, **Sands DC**. 2008. Ubiquity of biological ice nucleators in snowfall. Science **319**:1214–1214.

290. **Hara K**, **Maki T**, **Kobayashi F**, **Kakikawa M**, **Wada M**, **Matsuki A**. 2016. Variations of ice nuclei concentration induced by rain and snowfall within a local forested site in Japan. Atmos Environ **127**:1–5.

291. **Konstantinidis KT**. 2014. Do airborne microbes matter for atmospheric chemistry and cloud formation? Environ Microbiol **16**:1482–1484.

292. **Morris CE**, **Conen F**, **Huffman JA**, **Phillips V**, **Pöschl U**, **Sands DC**. 2014. Bioprecipitation: A feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. Glob Chang Biol **20**:341–351.

293. **Stopelli E**, **Conen F**, **Morris CE**, **Herrmann E**, **Bukowiecki N**, **Alewell C**. 2015. Ice nucleation active particles are efficiently removed by precipitating clouds. Sci Rep **5**:16433.

294. **Hoose C**, **Möhler O**. 2012. Heterogeneous ice nucleation on atmospheric aerosols: A review of results from laboratory experiments. Atmos Chem Phys **12**:9817–9854.

295. **Joly M**, **Attard E**, **Sancelme M**, **Deguillaume L**, **Guilbaud C**, **Morris CE**, **Amato P**, **Delort A-M**. 2013. Ice nucleation activity of bacteria isolated from cloud water. Atmos Environ **70**:392–400.

296. **Mortazavi R**, **Hayes CT**, **Ariya PA**. 2008. Ice nucleation activity of bacteria isolated from snow compared with organic and inorganic substrates. Environ Chem **5**:373.

297. **Amato P**, **Demeer F**, **Melaouhi A**, **Fontanella S**, **Martin-Biesse A-S**, **Sancelme M**, **Laj P**, **Delort A-M**. 2007. A fate for organic acids, formaldehyde and methanol in cloud water: Their biotransformation by micro-organisms. Atmos Chem Phys **7**:4159–4169.

298. **Vaitilingom M**, **Deguillaume L**, **Vinatier V**, **Sancelme M**, **Amato P**, **Chaumerliac N**, **Delort A-M**. 2013. Potential impact of microbial activity on the oxidant capacity and organic carbon budget in clouds. Proc Natl Acad Sci U S A **110**:559–564.

299. **Hill KA**, **Shepson PB**, **Galbavy ES**, **Anastasio C**, **Kourtev PS**, **Konopka A**, **Stirm BH**. 2007. Processing of atmospheric nitrogen by clouds above a forest environment. J Geophys Res **112**:D11301.

300. **Šantl-Temkiv T**, **Finster K**, **Dittmar T**, **Hansen BM**, **Thyrhaug R**, **Nielsen NW**, **Karlson UG**. 2013. Hailstones: A window into the microbial and chemical inventory of a storm cloud. PLoS One **8**:e53550.

301. **Ahern HE**, **Walsh KA**, **Hill TCJ**, **Moffett BF**. 2007. Fluorescent Pseudomonads isolated from Hebridean cloud and rain water produce biosurfactants but do not cause ice nucleation. Biogeosciences **4**:115–124.

302. **Kaushik R**, **Balasubramanian R**, **Dunstan H**. 2014. Microbial quality and phylogenetic diversity of fresh rainwater and tropical freshwater reservoir. PLoS One **9**:e100737.

303. **Cho BC**, **Jang G Il**. 2014. Active and diverse rainwater bacteria collected at an inland site in spring and summer 2011. Atmos Environ **94**:409–416.

304. **Wang Y**, **Qian P-Y**. 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One **4**:e7401.

305. **Claesson MJ**, **Wang Q**, **O'Sullivan O**, **Greene-Diniz R**, **Cole JR**, **Ross RP**, **O'Toole PW**. 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res **38**:e200.

306. **Cox MP**, **Peterson DA**, **Biggs PJ**. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11**:1–6.

307. **Stein AF**, **Draxler RR**, **Rolph GD**, **Stunder BJB**, **Cohen MD**, **Ngan F**. 2015. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bull Am Meteorol Soc **96**:2059–2077.

308. **Kobayashi F**, **Maki T**, **Kakikawa M**, **Yamada M**, **Puspitasari F**, **Iwasaka Y**. 2015. Bioprocess of Kosa bioaerosols: Effect of ultraviolet radiation on airborne bacteria within Kosa (Asian dust). J Biosci Bioeng **119**:570–579.

309. **Smith DJ**, **Timonen HJ**, **Jaffe DA**, **Griffin DW**, **Birmele MN**, **Perry KD**, **Ward PD**, **Roberts MS**. 2013. Intercontinental dispersal of bacteria and archaea by transpacific winds. Appl Environ Microbiol **79**:1134–

1139.

310. **Xu C**, **Wei M**, **Chen J**, **Sui X**, **Zhu C**, **Li J**, **Zheng L**, **Sui G**, **Li W**, **Wang W**, **Zhang Q**, **Mellouki A**. 2017. Investigation of diverse bacteria in cloud water at Mt. Tai, China. Sci Total Environ **580**:258–265.

311. **Xia X**, **Wang J**, **Ji J**, **Zhang J**, **Chen L**, **Zhang R**. 2015. Bacterial communities in marine aerosols revealed by 454 pyrosequencing of the 16S rRNA gene. J Atmos Sci **72**:2997–3008.

312. **Kourtev PS**, **Hill KA**, **Shepson PB**, **Konopka A**. 2011. Atmospheric cloud water contains a diverse bacterial community. Atmos Environ **45**:5399–5405.

313. **Nguyen TMN**, **Ilef D**, **Jarraud S**, **Rouil L**, **Campese C**, **Che D**, **Haeghebaert S**, **Ganiayre F**, **Marcel F**, **Etienne J**, **Desenclos J-C**. 2006. A community-wide outbreak of legionnaires disease linked to industrial cooling towers-how far can contaminated aerosols spread? J Infect Dis **193**:102–111.

314. **Fisman DN**, **Lim S**, **Wellenius GA**, **Johnson C**, **Britz P**, **Gaskins M**, **Maher J**, **Mittleman MA**, **Victor Spain C**, **Haas CN**, **Newbern C**. 2005. It's not the heat, It's the humidity: Wet weather increases legionellosis risk in the Greater Philadelphia Metropolitan Area. J Infect Dis **192**:2066–2073.

315. **Fisman DN**. 2007. Seasonality of infectious diseases. Annu Rev Public Health **28**:127–143.

316. **Wei M**, **Xu C**, **Chen J**, **Zhu C**, **Li J**, **Lv G**. 2016. Characteristics of bacterial community in fog water at Mt. Tai: Similarity and disparity under polluted and non-polluted fog episodes. Atmos Chem Phys Discuss **2016**:1–30.

317. **Fuerst JA**. 1995. The Palnctomycetes: Emerging models for microbial ecology evolution and cell biology. Microbiology **141**:1493–1506.

318. **Smets W**, **Moretti S**, **Denys S**, **Lebeer S**. 2016. Airborne bacteria in the atmosphere: Presence, purpose, and potential. Atmos Environ **139**:214–221.

319. **Fahlgren C**, **Gómez-Consarnau L**, **Zábori J**, **Lindh M V**, **Krejci R**, **Mårtensson EM**, **Nilsson D**, **Pinhassi J**. 2015. Seawater mesocosm experiments in the arctic uncover differential transfer of marine bacteria to aerosols. Environ Microbiol Rep **7**:460–470.

320. **Prospero JM**, **Blades E**, **Mathison G**, **Naidu R**. 2005. Interhemispheric transport of viable fungi and bacteria from Africa to the Caribbean with soil dust. Aerobiologia (Bologna) **21**:1–19.

321. **Yamaguchi N**, **Ichijo T**, **Baba T**, **Nasu M**. 2014. Long-range transportation of bacterial cells by asian dust. Genes Environ **36**:145–151.

322. **Morris CE**, **Monteil CL**, **Berge O**. 2013. The life history of </i>Pseudomonas syringae</i>: Linking agriculture to earth system processes. Annu Rev Phytopathol **51**:85–104.

323. **Delort A-M**, **Vaïtilingom M**, **Amato P**, **Sancelme M**, **Parazols M**, **Mailhot G**, **Laj P**, **Deguillaume L**. 2010. A short overview of the microbial population in clouds: Potential roles in atmospheric chemistry and nucleation processes. Atmos Res **98**:249–260.

324. **Cao C**, **Jiang W**, **Wang B**, **Fang J**, **Lang J**, **Tian G**, **Jiang J**, **Zhu TF**. 2014. Inhalable microorganisms in Beijing's PM2.5 and PM10 pollutants during a severe smog event. Environ Sci Technol **48**:1499–1507.

325. **Bauer H**, **Kasper-Giebl A**, **Löflund M**, **Giebl H**, **Hitzenberger R**, **Zibuschka F**, **Puxbaum H**. 2002. The

contribution of bacteria and fungal spores to the organic carbon content of cloud water, precipitation and aerosols. Atmos Res **64**:109–119.

326. **Kennedy J**, **Marchesi JR**, **Dobson ADW**. 2008. Marine metagenomics: Strategies for the discovery of novel enzymes with biotechnological applications from marine environments. Microb Cell Fact **7**:1–8.

327. **Raynaud X**, **Nunan N**. 2014. Spatial ecology of bacteria at the microscale in soil. PLoS One **9**:e87217.

328. **Ley RE**, **Peterson DA**, **Gordon JI**. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell **124**:837–848.

329. **Klindworth A**, **Pruesse E**, **Schweer T**, **Peplies J**, **Quast C**, **Horn M**, **Glöckner FO**. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res **41**:e1.

330. **Hong S**, **Bunge J**, **Leslin C**, **Jeon S**, **Epstein SS**. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J **3**:1365–1373.

331. **Kumar PS**, **Brooker MR**, **Dowd SE**, **Camerlengo T**. 2011. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. PLoS One **6**:e20956.

332. **Feinstein LM**, **Woo JS**, **Blackwood CB**. 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Appl Environ Microbiol **75**:5428–5433.

333. **Morgan JL**, **Darling AE**, **Eisen JA**. 2010. Metagenomic sequencing of an *in vitro*-simulated microbial community. PLoS One **5**:e10209.

334. **Inagaki F**, **Kubo Y**, **Bowles MW**, **Heuer VB**, **Ijiri A**, **Imachi H**, **Ito M**, **Kaneko M**, **Lever MA**, **Morita S**, **Morono Y**, **Tanikawa W**, **Bihan M**, **Bowden SA**, **Elvert M**, **Glombitza C**, **Gross D**, **Harrington GJ**, **Hori T**, **Li K**, **Limmer D**, **Murayama M**, **Ohkouchi N**, **Ono S**, **Purkey M**, **Sanada Y**, **Sauvage J**, **Snyder G**, **Takano Y**, **Tasumi E**, **Terada T**, **Tomaru H**, **Wang DT**, **Yamada Y**. 2015. Exploring deep microbial life in coal-bearing sediment down to ~2.5 km below the ocean floor. Science **349**:420–424.

335. **Karl DM**. 1999. Microorganisms in the accreted ice of lake Vostok, Antarctica. Science **286**:2144–2147.

336. **Belkova NL**, **Tazaki K**, **Zakharova JR**, **Parfenova V V**. 2007. Activity of bacteria in water of hot springs from southern and central Kamchatskaya geothermal provinces, Kamchatka Peninsula, Russia. Microbiol Res **162**:99–107.

337. **Yooseph S**, **Andrews-Pfannkoch C**, **Tenney A**, **McQuaid J**, **Williamson S**, **Thiagarajan M**, **Brami D**, **Zeigler-Allen L**, **Hoffman J**, **Goll JB**, **Fadrosh D**, **Glass J**, **Adams MD**, **Friedman R**, **Venter JC**. 2013. A metagenomic framework for the study of airborne microbial communities. PLoS One **8**:e81862.

338. **Salter SJ**, **Cox MJ**, **Turek EM**, **Calus ST**, **Cookson WO**, **Moffatt MF**, **Turner P**, **Parkhill J**, **Loman NJ**, **Walker AW**. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **12**:87.

339. **Mohammadi T**, **Reesink HW**, **Vandenbroucke-Grauls CMJE**, **Savelkoul PHM**. 2005. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. J Microbiol Methods **61**:285–288.

340. **Grahn N**, **Olofsson M**, **Ellnebo-Svedlund K**, **Monstein HJ**, **Jonasson J**. 2003. Identification of mixed

bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. FEMS Microbiol Lett **219**:87–91.

341. **Kumar R**, **Rao DN**. 2013. Role of DNA methyltransferases in epigenetic regulation in bacteria, p. 81–102. *In* Kundu, TK (ed.), Subcellular Biochemistry. Springer Netherlands, Dordrecht.

342. **Labrie SJ**, **Samson JE**, **Moineau S**. 2010. Bacteriophage resistance mechanisms. Nat Rev Microbiol **8**:317.

343. **Wion D**, **Casadesús J**. 2006. N^6-methyl-adenine: An epigenetic signal for DNA–protein interactions. Nat Rev Microbiol **4**:183–192.

344. **Low DA**, **Casadesús J**. 2008. Clocks and switches: Bacterial gene regulation by DNA adenine methylation. Curr Opin Microbiol **11**:106–112.

345. **Casadesus J**, **Low D**. 2006. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev **70**:830–856.

346. **Vasu K**, **Nagaraja V**. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev **77**:53–72.

347. **Kozdon JB**, **Melfi MD**, **Luong K**, **Clark TA**, **Boitano M**, **Wang S**, **Zhou B**, **Gonzalez D**, **Collier J**, **Turner SW**, **Korlach J**, **Shapiro L**, **McAdams HH**. 2013. Global methylation state at base-pair resolution of the *Caulobacter* genome throughout the cell cycle. Proc Natl Acad Sci U S A **110**:E4658–E4667.

348. **Kobayashi I**. 2001. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res **29**:3742–3756.

349. **Makarova KS**, **Wolf YI**, **Snir S**, **Koonin E V.** 2011. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol **193**:6039–6056.

350. **Srikhanta YN**, **Fox KL**, **Jennings MP**. 2010. The phasevarion: Phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. Nat Rev Microbiol **8**:196.

351. **Rambo IM**, **Marsh A**, **Biddle JF**. 2017. Cytosine methylation within marine sediment microbial communities: Potential epigenetic adaptation to the environment. bioRxiv 167189.

352. **Murray IA**, **Clark TA**, **Morgan RD**, **Boitano M**, **Anton BP**, **Luong K**, **Fomenkov A**, **Turner SW**, **Korlach J**, **Roberts RJ**. 2012. The methylomes of six bacteria. Nucleic Acids Res **40**:11450–11462.

353. **Allen White R**, **Borkum MI**, **Rivas-Ubach A**, **Bilbao A**, **Wendler JP**, **Colby SM**, **Köberl M**, **Jansson C**. 2017. From data to knowledge: The future of multi-omics data analysis for the rhizosphere. Rhizosphere **3**:222–229.

354. **Rhoads A**, **Au KF**. 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics **13**:278–289.

355. **Koren S**, **Phillippy AM**. 2015. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol **23**:110–120.

356. **Eid J**, **Fehr A**, **Gray J**, **Luong K**, **Lyle J**, **Otto G**, **Peluso P**, **Rank D**, **Baybayan P**, **Bettman B**, **Bibillo A**, **Bjornson K**, **Chaudhuri B**, **Christians F**, **Cicero R**, **Clark S**, **Dalal R**, **DeWinter A**, **Dixon J**, **Foquet M**, **Gaertner A**, **Hardenbol P**, **Heiner C**, **Hester K**, **Holden D**, **Kearns G**, **Kong X**, **Kuse R**, **Lacroix Y**, **Lin**

S, **Lundquist P**, **Ma C**, **Marks P**, **Maxham M**, **Murphy D**, **Park I**, **Pham T**, **Phillips M**, **Roy J**, **Sebra R**, **Shen G**, **Sorenson J**, **Tomaney A**, **Travers K**, **Trulson M**, **Vieceli J**, **Wegener J**, **Wu D**, **Yang A**, **Zaccarin D**, **Zhao P**, **Zhong F**, **Korlach J**, **Turner S**. 2009. Real-time DNA sequencing from single polymerase molecules. Science **323**:133–138.

357. **Koren S**, **Schatz MC**, **Walenz BP**, **Martin J**, **Howard J**, **Ganapathy G**, **Wang Z**, **Rasko DA**, **McCombie WR**, **Jarvis ED**, **Phillippy AM**. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol **30**:693–700.

358. **Fichot EB**, **Norman RS**. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome **1**:10.

359. **Gao S**, **Ren Y**, **Sun Y**, **Wu Z**, **Ruan J**, **He B**, **Zhang T**, **Yu X**, **Tian X**, **Bu W**. 2016. PacBio full-length transcriptome profiling of insect mitochondrial gene expression. RNA Biol **13**:820–825.

360. **Bossé JT**, **Chaudhuri RR**, **Li Y**, **Leanse LG**, **Fernandez Crespo R**, **Coupland P**, **Holden MTG**, **Bazzolli DM**, **Maskell DJ**, **Tucker AW**, **Wren BW**, **Rycroft AN**, **Langford PR**. 2016. Complete genome sequence of MIDG2331, a genetically tractable serovar 8 clinical isolate of *Actinobacillus pleuropneumoniae*. Genome Announc **4**:e01667-15.

361. **Frank JA**, **Pan Y**, **Tooming-Klunderud A**, **Eijsink VGH**, **McHardy AC**, **Nederbragt AJ**, **Pope PB**. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep **6**:25373.

362. **Kenzaka T**, **Tani K**, **Nasu M**. 2010. High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. ISME J **4**:648–659.

363. **Newton RJ**, **Jones SE**, **Eiler A**, **McMahon KD**, **Bertilsson S**. 2011. A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev **75**:14–49.

364. **Paez-Espino D**, **Eloe-Fadrosh EA**, **Pavlopoulos GA**, **Thomas AD**, **Huntemann M**, **Mikhailova N**, **Rubin E**, **Ivanova NN**, **Kyrpides NC**. 2016. Uncovering Earth's virome. Nature **536**:425.

365. **Albertsen M**, **Karst SM**, **Ziegler AS**, **Kirkegaard RH**, **Nielsen PH**. 2015. Back to basics—The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One **10**:e0132783.

366. **Coordinators NR**. 2017. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **45**:D12–D17.

367. **Koren S**, **Walenz BP**, **Berlin K**, **Miller JR**, **Bergman NH**, **Phillippy AM**. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res **27**:722–736.

368. **Chevreux B**, **Wetter T**, **Suhai S**. 1999. Genome sequence assembly using trace signals and additional sequence information, p. 45–56. *In* Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99. Hanover, Germany.

369. **Chaisson MJ**, **Tesler G**. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. BMC Bioinformatics **13**:238.

370. **Parks DH**, **Imelfort M**, **Skennerton CT**, **Hugenholtz P**, **Tyson GW**. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res **25**:1043–1055.

371. **Cambuy DD**, **Coutinho FH**, **Dutilh BE**. 2016. Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. bioRxiv 72868.

372. **Suzuki S**, **Kakuta M**, **Ishida T**, **Akiyama Y**. 2015. Faster sequence homology searches by clustering subsequences. Bioinformatics **31**:1183–1190.

373. **Mistry J**, **Finn RD**, **Eddy SR**, **Bateman A**, **Punta M**. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res **41**:e121.

374. **Finn RD**, **Coggill P**, **Eberhardt RY**, **Eddy SR**, **Mistry J**, **Mitchell AL**, **Potter SC**, **Punta M**, **Qureshi M**, **Sangrador-Vegas A**, **Salazar GA**, **Tate J**, **Bateman A**. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res **44**:D279–D285.

375. **Arndt D**, **Grant JR**, **Marcu A**, **Sajed T**, **Pon A**, **Liang Y**, **Wishart DS**. 2016. PHASTER: A better, faster version of the PHAST phage search tool. Nucleic Acids Res **44**:W16–W21.

376. **Kielbasa SM**, **Wan R**, **Sato K**, **Horton P**, **Frith MC**. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res **21**:487–493.

377. **Bland C**, **Ramsey TL**, **Sabree F**, **Lowe M**, **Brown K**, **Kyrpides NC**, **Hugenholtz P**. 2007. CRISPR recognition tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics **8**:209.

378. **Haft DH**, **Selengut JD**, **Richter RA**, **Harkins D**, **Basu MK**, **Beck E**. 2013. TIGRFAMs and genome properties in 2013. Nucleic Acids Res **41**:D387–D395.

379. **Roberts RJ**, **Vincze T**, **Posfai J**, **Macelis D**. 2010. REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. Nucleic Acids Res **38**:D234–D236.

380. **Larkin MA**, **Blackshields G**, **Brown NP**, **Chenna R**, **McGettigan PA**, **McWilliam H**, **Valentin F**, **Wallace IM**, **Wilm A**, **Lopez R**, **Thompson JD**, **Gibson TJ**, **Higgins DG**. 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23**:2947–2948.

381. **Kumar S**, **Stecher G**, **Tamura K**. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol **33**:1870–1874.

382. **Yilmaz P**, **Parfrey LW**, **Yarza P**, **Gerken J**, **Pruesse E**, **Quast C**, **Schweer T**, **Peplies J**, **Ludwig W**, **Glöckner FO**. 2014. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Res **42**:D643-8.

383. **Okazaki Y**, **Nakano S-I**. 2016. Vertical partitioning of freshwater bacterioplankton community in a deep mesotrophic lake with a fully oxygenated hypolimnion (Lake Biwa, Japan). Environ Microbiol Rep **8**:780–788.

384. **Okazaki Y**, **Fujinaga S**, **Tanaka A**, **Kohzu A**, **Oyagi H**, **Nakano S**. 2017. Ubiquity and quantitative significance of bacterioplankton lineages inhabiting the oxygenated hypolimnion of deep freshwater lakes. ISME J **11**:2279–2293.

385. **Colson P**, **De Lamballerie X**, **Yutin N**, **Asgari S**, **Bigot Y**, **Bideshi DK**, **Cheng X-W**, **Federici BA**, **Van Etten JL**, **Koonin E V**, **La Scola B**, **Raoult D**. 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol **158**:2517–2521.

386. **Claverie J-M**, **Grzela R**, **Lartigue A**, **Bernadac A**, **Nitsche S**, **Vacelet J**, **Ogata H**, **Abergel C**. 2009. Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol **101**:172–180.

387. **Tsai Y-C**, **Conlan S**, **Deming C**, **Program NCS**, **Segre JA**, **Kong HH**, **Korlach J**, **Oh J**. 2016. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio **7**:e01948-15.

388. **Singer E**, **Andreopoulos B**, **Bowers RM**, **Lee J**, **Deshpande S**, **Chiniquy J**, **Ciobanu D**, **Klenk H-P**, **Zane M**, **Daum C**, **Clum A**, **Cheng J-F**, **Copeland A**, **Woyke T**. 2016. Next generation sequencing data of a defined microbial mock community. Sci Data **3**:160081.

389. **Urbach E**, **Vergin KL**, **Young L**, **Morse A**, **Larson GL**, **Giovannoni SJ**. 2001. Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. Limnol Oceanogr **46**:557–572.

390. **Rodrigues JLM**, **Isanapong J**. 2014. The family Opitutaceae, p. 751–756. *In* Rosenberg, E, DeLong, EF, Lory, S, Stackebrandt, E, Thompson, F (eds.), The Prokaryotes. Springer Berlin Heidelberg, Berlin, Heidelberg.

391. **Fang G**, **Munera D**, **Friedman DI**, **Mandlik A**, **Chao MC**, **Banerjee O**, **Feng Z**, **Losic B**, **Mahajan MC**, **Jabado OJ**, **Deikus G**, **Clark TA**, **Luong K**, **Murray IA**, **Davis BM**, **Keren-Paz A**, **Chess A**, **Roberts RJ**, **Korlach J**, **Turner SW**, **Kumar V**, **Waldor MK**, **Schadt EE**. 2012. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat Biotechnol **30**:1232–1239.

392. **Zhu L**, **Zhong J**, **Jia X**, **Liu G**, **Kang Y**, **Dong M**, **Zhang X**, **Li Q**, **Yue L**, **Li C**, **Fu J**, **Xiao J**, **Yan J**, **Zhang B**, **Lei M**, **Chen S**, **Lv L**, **Zhu B**, **Huang H**, **Chen F**. 2016. Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. Nucleic Acids Res **44**:730–743.

393. **Wei H**, **Therrien C**, **Blanchard A**, **Guan S**, **Zhu Z**. 2008. The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases. Nucleic Acids Res **36**:e50.

394. **Oliveira PH**, **Touchon M**, **Rocha EPC**. 2014. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res **42**:10618–10631.

395. **Ahlgren NA**, **Chen Y**, **Needham DM**, **Parada AE**, **Sachdeva R**, **Trinh V**, **Chen T**, **Fuhrman JA**. 2017. Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. Environ Microbiol **19**:2434–2452.

396. **Vanyushin BF**, **Dobritsa AP**. 1975. On the nature of the cytosine-methylated sequence in DNA of Bacillus brevis var. G.-B. Biochim Biophys Acta **407**:61–72.

397. **Loenen WAM**, **Dryden DTF**, **Raleigh EA**, **Wilson GG**. 2014. Type I restriction enzymes and their relatives. Nucleic Acids Res **42**:20–44.

398. **Ofir G**, **Melamed S**, **Sberro H**, **Mukamel Z**, **Silverman S**, **Yaakov G**, **Doron S**, **Sorek R**. 2018. DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat Microbiol **3**:90–98.

399. **Ando T**, **Xu Q**, **Torres M**, **Kusugami K**, **Israel DA**, **Blaser MJ**. 2000. Restriction–modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. Mol Microbiol **37**:1052–1065.

400. **Budroni S**, **Siena E**, **Hotopp JCD**, **Seib KL**, **Serruto D**, **Nofroni C**, **Comanducci M**, **Riley DR**, **Daugherty SC**, **Angiuoli S V**, **Covacci A**, **Pizza M**, **Rappuoli R**, **Moxon ER**, **Tettelin H**, **Medini D**. 2011. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A **108**:4494–4499.

401. **Fortier L-C**, **Sekulovic O**. 2013. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence **4**:354–365.

402. **Canchaya C**, **Proux C**, **Fournous G**, **Bruttin A**, **Brussow H**. 2003. Prophage genomics. Microbiol Mol Biol Rev **67**:238–276.

403. **Kang HS**, **McNair K**, **Cuevas D**, **Bailey B**, **Segall A**, **Edwards RA**. 2017. Prophage genomics reveals patterns in phage genome organization and replication. bioRxiv 114819.

404. **Touchon M**, **Bernheim A**, **Rocha EPC**. 2016. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J **10**:2744–2754.

405. **Malik S**, **Zalenskaya K**, **Goldfarb A**. 1987. Competition between sigma factors for core RNA polymerase. Nucleic Acids Res **15**:8521–8530.

406. **Kolesky S**, **Ouhammouch M**, **Brody EN**, **Geiduschek EP**. 1999. Sigma competition: The contest between bacteriophage T4 middle and late transcription. J Mol Biol **291**:267–281.

407. **Depardieu F**, **Didier J-P**, **Bernheim A**, **Sherlock A**, **Molina H**, **Duclos B**, **Bikard D**. 2016. A eukaryotic-like serine/threonine kinase protects Staphylococci against phages. Cell Host Microbe **20**:471–481.

408. **Friedman DI**, **Mozola CC**, **Beeri K**, **Ko C-C**, **Reynolds JL**. 2011. Activation of a prophage-encoded tyrosine kinase by a heterologous infecting phage results in a self-inflicted abortive infection. Mol Microbiol **82**:567–577.

409. **Seed KD**. 2015. Battling phages: How bacteria defend against viral attack. PLoS Pathog **11**:e1004847.

410. **Samson JE**, **Magadán AH**, **Sabri M**, **Moineau S**. 2013. Revenge of the phages: Defeating bacterial defences. Nat Rev Microbiol **11**:675–687.

411. **Stern A**, **Sorek R**. 2011. The phage-host arms race: Shaping the evolution of microbes. BioEssays **33**:43–51.

412. **Mruk I**, **Kobayashi I**. 2014. To be or not to be: Regulation of restriction–modification systemas and other toxin–antitoxin systems. Nucleic Acids Res **42**:70–86.

413. **Rath D**, **Amlinger L**, **Rath A**, **Lundgren M**. 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie **117**:119–128.

414. **Jore MM**, **Brouns SJJ**, **van der Oost J**. 2012. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. Cold Spring Harb Perspect Biol **4**:a003657.

415. **Zhang Q**, **Ye Y**. 2017. Not all predicted CRISPR–Cas systems are equal: Isolated cas genes and classes of CRISPR like elements. BMC Bioinformatics **18**:92.

416. **Nozawa T**, **Furukawa N**, **Aikawa C**, **Watanabe T**, **Haobam B**, **Kurokawa K**, **Maruyama F**, **Nakagawa I**. 2011. CRISPR inhibition of prophage acquisition in *Streptococcus pyogenes*. PLoS One **6**:e19543.

417. **Edgar R**, **Qimron U**. 2010. The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. J Bacteriol **192**:6291–6294.

418. **White RA**, **Bottos EM**, **Roy Chowdhury T**, **Zucker JD**, **Brislawn CJ**, **Nicora CD**, **Fansler SJ**, **Glaesemann KR**, **Glass K**, **Jansson JK**. 2016. Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems **1**:e00045-16.

419. **Sharon I**, **Kertesz M**, **Hug LA**, **Pushkarev D**, **Blauwkamp TA**, **Castelle CJ**, **Amirebrahimi M**, **Thomas BC**, **Burstein D**, **Tringe SG**, **Williams KH**, **Banfield JF**. 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res **25**:534–543.

420. **Danko DC**, **Meleshko D**, **Bezdan D**, **Mason C**, **Hajirasouliha I**. 2017. Minerva: An alignment and reference free approach to deconvolve linked-reads for metagenomics. bioRxiv 217869.

421. **Rand AC**, **Jain M**, **Eizenga JM**, **Musselman-Brown A**, **Olsen HE**, **Akeson M**, **Paten B**. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods **14**:411–413.

422. **Stoiber MH**, **Quick J**, **Egan R**, **Lee JE**, **Celniker SE**, **Neely R**, **Loman N**, **Pennacchio L**, **Brown JB**. 2016. *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. bioRxiv 94672.

423. **Davis BM**, **Chao MC**, **Waldor MK**. 2013. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr Opin Microbiol **16**:192–198.