HOKKAIDO UNIVERSITY

Title	Comprehensive analysis of pathogenic bacteria in environmental samples using metagenomic approaches
Author(s)	BORTHONG, Jednipit
Citation	北海道大学. 博士(獣医学) 甲第13294号
Issue Date	2018-09-25
DOI	10.14943/doctoral.k13294
Doc URL	http://hdl.handle.net/2115/76126
Type	theses (doctoral)
File Information	Jednipit_BORTHONG.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Comprehensive analysis of pathogenic bacteria in environmental samples using metagenomic approaches

(メタゲノム解析による環境中の病原性バクテリアの総合解析)

Jednipit Borthong

CONTENTS

Chapter II

Shotgun metagenomic analysis of pathogenic bacteria and antibiotic resistance genes in environmental samples collected from urban and rural areas in Thailand

Abbreviations

AR	Antibiotic resistance
ARB	Antibiotic resistant bacteria
ARGs	Antimicrobial resistance genes
AUC	Area under curve
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CZC	Research Center for Zoonosis Control
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
LCA	Lowest common ancestor
MEGA	Molecular Evolutionary Genetics Analysis
MEGAN	Metagenome Analyzer
NCBI	National Center for Biotechnology Information
NCBI-NT database	NCBI-nucleotide database
NGS	Next Generation Sequencer
Nested PCR	Nested Polymerase Chain Reaction
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
RDP	Ribosomal Database Project
ROC	Receiver operating characteristic curve
rRNA	Ribosomal ribonucleic acid
VBNC	Viable but non-culturable
VFDB	Virulence Factor Gene Database
VFLP	Virulence Factor Gene of *Legionella pneumophila* Database

Preface

Bacteria are abundant in a wide range of environmental reservoirs. They play significant roles in the maintenance of element cycles in ecological systems (Leigh and Dodsworth, 2007;Jones et al., 2009). Most bacteria are harmless in nature, but some are causative agents for infectious diseases threatening human and animal health. In aquatic environments, pathogenic bacteria can be transmitted to humans and animals through direct and indirect contact with contaminated water (Hlavsa et al., 2014). The diseases caused by such pathogens are known as waterborne diseases. These days, waterborne disease is recognized as a global public health concern.

Waterborne diseases have been reported from many countries. In the USA, Legionnaire's disease is a severe infectious disease of the respiratory tract and a total of 3,522 people were sick from this disease from 2000 – 2009 (Centers for Disease Control Prevention, 2011). Cholera causes severe watery diarrhea and 3 – 5 million people are at risk of this disease every year (Mutreja et al., 2011). In Spain, Shigellosis affected more than 180 people after drinking contaminated water from their water supply system (Arias et al., 2006). Leptospirosis is a common disease that is associated with areas of flooding (Thaipadungpanit et al., 2013). The numbers of cases of Campylobacterosis have increased in North America, Europe, and Australia in the past decade (Kaakoush et al., 2015).

Culture-based and Polymerase Chain Reaction (PCR)-based methods are routinely employed to detect pathogenic bacteria in environmental samples. Bacterial cultures have been considered the gold standard method and have been used to detect pathogens in laboratories. However, the method requires selective media and time for bacterial growth. Moreover, its diagnostic sensitivity is low in that it cannot detect bacteria in viable but non-culturable (VBNC) states (Oliver, 2010;Ramamurthy et al., 2014) and uncultured bacteria (Vartoukian et al., 2010). On the other hand, PCR-based methods have been used to detect pathogenic bacteria with higher sensitivity and specificity than cultured-based methods. However, the method requires sequence information of target pathogen in advance to design specific primers (Girones et al., 2010), and thus it is difficult to analyze overall pathogens in environmental samples.

Recently, metagenomic analysis has been established to analyze overall bacterial populations in a given sample. In metagenomic analyses, genetic materials of bacteria in samples are analyzed directly using a next generation sequencer (NGS) (Thomas et al., 2012). In contrast to single gene amplification techniques such as PCR-based assays, metagenomic analysis can detect genomic fragments of thousands of bacteria in a single NGS run (Caporaso et al., 2012). Metagenomic approaches have been used to investigate the bacterial population structure in a variety of samples, including environmental (Daniel, 2005;Sogin et al., 2006;Breitbart et al., 2009), food (Ercolini, 2013), and clinical samples (Cho and Blaser, 2012). The price of NGS platforms and their running costs are decreasing (Lecuit and Eloit, 2014;Muir et al., 2016), increasing the opportunity for application in metagenomic analysis (Garrido-Cardenas and Manzano-Agugliaro, 2017).

There are two major approaches in pathogen detection using metagenomic analysis. The first approach, 16S rRNA metagenomic analysis, uses conserved and variable regions in the bacterial 16S rRNA gene to study the taxonomy of bacteria in samples (Janda and Abbott, 2007). Several studies applied 16S rRNA metagenomic analysis to detect pathogens in aquatic environments, drinking water, and food. Ibekwe et al. (2013) detected potential pathogens from the genera *Aeromonas*, *Clostridium*, *Bacillus*, *Pseudomonas,* and *Treponema* in water samples collected from the Middle Santa Ana River. Ye and Zhang (2011) detected pathogens from wastewater treatment plants in China, USA, Canada, and Singapore; finding all samples contaminated with *Aeromonas* and *Clostridium*. Mukherjee et al. (2016) investigated the bacterial diversity in water supplies from rural areas in Haiti and found human pathogens such as *Aeromonas*, *Bacillus*, *Clostridium*, and *Yersinia* in a high proportion of bacterial communities. Moreover, 16S rRNA metagenomic analysis has been used to check pathogen contamination in drinking water (Shi et al., 2013;Huang et al., 2014;Pinto et al., 2016;Oh et al., 2018) and vegetables (Leonard et al., 2015;Kim et al., 2018). 16S rRNA metagenomic analysis uses PCR for the construction of a DNA library, and the results are affected by the amplification step. Problems due to differences in the copy number of the 16S rRNA gene in a genome of bacteria (Vetrovsky and Baldrian, 2013) and chimeric sequences in PCR products (Haas et al., 2011) may arise. No single hypervariable region can be used to differentiate between bacteria (Chakravorty et al., 2007), and closely related bacteria cannot be differentiated (Weinstock, 2012).

The second approach is shotgun metagenomic analysis. By using random primers, DNA fragments can be captured from any part of the bacterial genomes (Sharpton, 2014), providing narrower sequence coverage than 16S rRNA analysis (Angiuoli et al., 2011). Since the bacterial genome contains sequences specific to bacterial species, there is a possibility of increasing the specificity of pathogen detection. Several studies have applied shotgun metagenomics to the detection of potential pathogens in the environment. Lu et al. (2015) compared bacterial populations in water before and after processing in a sewage treatment system and found that most pathogenic bacteria were eliminated after the treatment. Nordahl Petersen et al. (2015) investigated toilet waste from airplanes using metagenomics and detected *Salmonella enterica* and *Clostridium difficile* from waste after international flights. Shotgun metagenomics have been used to investigate pathogenic bacteria in a variety of water samples, such as wastewater treatment (Cai and Zhang, 2013;Ibarbalz et al., 2016), drinking water and drink water systems (Gomez-Alvarez et al., 2012;Chao et al., 2013;Otten et al., 2016), and freshwater (Van Rossum et al., 2015;Mohiuddin et al., 2017). In addition, shotgun metagenomics are used for the biological evaluation of food safety (Walsh et al., 2017) and the food production chain (Yang et al., 2016). These studies show the potential usefulness of metagenomic analysis in detecting pathogenic bacteria in samples. However, the bacterial diversities analyzed by shotgun metagenomics depend on the method of DNA extraction and/or sequencing protocol (Morgan et al., 2010) and may also capture the host's genetic material (Kuczynski et al., 2011).

An additional advantage of shotgun metagenomics for detecting pathogenic bacteria is the ability to analyze functional genes, such as antibiotic resistant genes (ARGs), in environmental samples. Several studies have applied shotgun metagenomic analysis to study ARGs in a variety of samples, such as lake water (Chen et al., 2016), wastewater (Szczepanowski et al., 2009;Wang et al., 2013), activated sludge (Yang et al., 2013;Zhang et al., 2015), sewage (Yang et al., 2014;Su et al., 2017), soil (Monier et al., 2011;Xiao et al., 2016), drinking water (Shi et al., 2013), animals (Durso et al., 2011;Looft et al., 2012), and humans (Diaz-Torres et al., 2006;Hu et al., 2013).

Antibiotic resistance (AR) in bacteria has become a critical problem in animals and humans worldwide. The misuse and overuse of antibiotics are the main selective

driving force of the emergence and dissemination of antibiotic resistant bacteria (ARB) (Piddock, 1998;Schmieder and Edwards, 2012). Large amounts of antibiotics are used for therapy, prophylaxis, and growth promotion in food animal production, and has become a condition for the selectiveness, spread, and persistence of ARB (Aarestrup, 2005). Bacteria can develop resistance to antibiotics by mutations in their genes and/or acquisition of resistant genes from other species or strains (Munita and Arias, 2016). These genes are called ARGs. Patients infected with ARB are difficult to treat. Recently, human intestinal bacteria have been reported to be a reservoir for ARGs (Salyers et al., 2004;Sommer et al., 2009). The transmission of bacteria carrying ARGs from animals to humans can occur through the food chain (Teale, 2002) or close contact with animals, especially people whose occupations are related to animals (Okeke et al., 1999). Therefore, it is highly possible that ARB spread from animals/humans to the environment and that they are transmitted back to animals/humans from the environment.

Legionella pneumophila is the causative agent of Legionnaire's disease. This pathogenic bacterium is ubiquitous in natural aquatic environments such as ponds, lakes, rivers, and estuaries (Fliermans et al., 1981). Moreover, *L. pneumophila* can also be found in man-made water reservoirs, such as cooling towers (Turetgen et al., 2005), spas (Benkel et al., 2000), and water distribution systems (Stout et al., 1985). Inhalation of water aerosols is the primary cause of transmission to humans, and human-to-human transmission is rare (Correia et al., 2016). In animals, *L. pneumophila* can be found in cattle and calves (Fabbi et al., 1998). Moreover, a prevalence of antibodies against *L. pneumophila* have been reported in sheep, horses, antelopes, buffaloes, and rabbits (Boldur et al., 1987).

The standard methods of detecting *L. pneumophila* in water samples are culture-based and PCR-based. The cultured-based method uses centrifugation, filtration, heat and acid treatments, selective media, and antibiotics (Atlas et al., 1995). This method can be used to show the population of *L. pneumophila* in samples, both quantity and serotypes. However, the cultured-based method takes many days for *L. pneumophila* growth. The nested PCR and real-time PCR are alternative assays for the detection of *L. pneumophila*. There are many researchers who have designed primers for specific detection of *L. pneumophila*. The 5S rRNA (Mahbubani et al., 1990), 16S rRNA (Cloud et al.,

2000;Buchbinder et al., 2002), *dotA* (Yanez et al., 2005), and *mip* (Mahbubani et al., 1990;Catalan et al., 1994) are examples of target genes for the detection of *L. pneumophila*. In recent years, several metagenomic studies were conducted to detect *Legionella* spp. and *L. pneumophila* in water samples (Cai and Zhang, 2013;Delafont et al., 2013;Lu et al., 2015;Mohiuddin et al., 2017). Pereira et al. (2017) conducted 16S rRNA metagenomic analysis and detected six different *Legionella* spp. in freshwater samples. Peabody et al. (2017) investigated *Legionella* spp. in water samples from seven different places throughout a year and found that *L*. pneumophila was the most abundant at all sampling sites.

This thesis consists of two chapters. In Chapter I, shotgun metagenomic and 16S rRNA amplicon analysis were conducted to analyze bacterial communities in water samples collected from aquatic environments in Hokkaido University. By focusing on *Legionella pneumophila*, the different database search methods for detecting *L. pneumophila* were compared, and the detection results were evaluated with those of *L. pneumophila*-specific nested PCR. In Chapter II, shotgun metagenomic analysis was conducted to investigate the potential pathogens and ARGs in samples collected from Phadungkrungkasem Canal and a rice field, located in urban and rural areas of Thailand, respectively. The results of the pathogen population and ARGs profiles were analyzed and compared between these two areas.

Chapter I

Comparison of database search methods for the detection of *Legionella pneumophila* in water samples using metagenomic analysis

Introduction

In metagenomic studies, reference sequences are important for database construction. After constructing a database, an unknown sequence can be aligned against the database to classify it to a bacterial taxonomy. Since the National Center for Biotechnology Information (NCBI) established the taxonomy database in 1991 (Federhen, 2011), complete genomes of bacteria are available to download, and can be used to construct subset databases. In addition to complete genomes, specific marker genes, such as 16S rRNA and virulence factor genes, can also be used to construct databases for classifying bacterial taxonomy and detecting pathogenic bacteria, respectively. At present, these two kinds of reference sequences are commonly used to construct databases for taxonomy classification of bacteria.

Taxonomy classification is a bioinformatics procedure used to infer the population structure of microorganisms based on genomic information obtained from samples, and several computational methods have been developed so far (Lindgreen et al., 2016). The naïve lowest common ancestor (LCA) algorithm implemented in MEGAN assigns sequence reads to taxa on taxonomy trees based on blastn search results of reads against given databases (Huson et al., 2007). Kraken (Wood and Salzberg, 2014), CLARK (Ounit et al., 2015) and One Codex (Minot et al., 2015) use the differences in *k*-mer distributions among taxa to assign reads to nodes in the taxonomy tree. MGmapper uses alignment scores from reference mappings of reads to reference sequences in a database (Petersen et al., 2017). MetaPhlAn2 uses pre-defined sets of clade-specific marker sequences and classifies reads using reference mapping onto marker sequences (Truong et al., 2015). RDP (Cole et al., 2005) and SILVA (Quast et al., 2013) are specialized to analyze 16S rRNA amplicon reads and determine the taxa of reads according to sequence similarity of the 16S rRNA genes.

Despite recent advancements in NGS technologies and classification algorithms,

several studies using metagenomic analyses have exposed important issues associated with sensitivity and specificity. Loman et al. (2013) reported false negative detections of Shiga-Toxigenic *Escherichia coli* O104:H4 in the diagnosis of diarrheal patients using metagenomic analysis. Several groups found that bacterial populations identified by shotgun metagenomics and those by 16S rRNA metagenomics were not always consistent with one another (Shah et al., 2011;Clooney et al., 2016). These results suggest that sensitivity and/or specificity of the two methods are different depending on the bacterial species. It is also known that metagenomic analyses generate different results depending on the taxonomy classification algorithms (Clooney et al., 2016) and reference databases (Miller et al., 2013) used.

Sakushukotoni River, Ohno Pond, and Hyotan Pond are aquatic environments located in the campus of Hokkaido University. These areas are related with several human activities. Some parts of the areas are attractive places for tourists. Some areas are related to agricultural activities, such as experimental fields for plant cultivation and animal rearing. In order to investigate the potential risk of *L. pneumophila* infection, which is one of the most common waterborne diseases in Japan, from the water in these areas, *L. pneumophila* was focused on in the present study.

In this chapter, shotgun metagenomic and 16S rRNA amplicon analyses were conducted to analyze bacterial communities in water samples collected from aquatic environments. Several methods for taxonomy classification of bacteria have been proposed, and the best method for detecting the pathogenic species should be used. By focusing on *L. pneumophila*, the ability of database search methods in the detection of *L. pneumophila* was compared. In order to specify the best method for detecting *L. pneumophila* using metagenomics, the detection results of each database search method were evaluated with those by *L. pneumophila*-specific nested PCR.

Materials and Methods

Water samples

Ten water samples were collected from the Sapporo campus of Hokkaido University on October 16th, 2012. Eight samples were obtained from different points along the Sakushukotoni Stream (HKU_A, HKU_B, HKU_C, HKU_E, HKU_F, HKU_G, HKU_H, and HKU_I), one sample was collected from Ohno Pond (HKU_D), and the other sample was collected from Hyotan Pond (HKU_J) (Figure 1). Two liters of water were collected from the water surface using sterilized containers (Pope and Patel, 2008;Tekera et al., 2011;Silva et al., 2012). The samples were transferred to a laboratory at the Research Center for Zoonosis Control (CZC) in Hokkaido University for further analysis. The process of this study is summarized in Figure 2.

Bacterial concentration and DNA extraction

Bacteria in the water samples were concentrated using a standard membrane filtration technique with four different pore sizes; 100 μm, 10 μm, 5 μm, and 0.22 μm (Millipore, Tokyo, Japan). The filtrates of 0.22 μm-membrane were used to extract DNA using a PowerWater® DNA Isolation Kit (Mo Bio Laboratories, Inc., California, USA). DNA concentration was determined using a Qubit™ fluorometer (Invitrogen, Tokyo, Japan).

Detection of *Legionella* spp. and *L. pneumophila* using nested PCR

Legionella genus-specific nested PCR was conducted amplifying 16S rRNA genes using the outer primers Leg120v and Leg1023r (Buchbinder et al., 2002) and inner primers JFP and JRP (Cloud et al., 2000). *L. pneumophila*-specific nested PCR was conducted amplifying macrophage infectivity potentiator surface protein (*mip*) gene using the outer primers Lmip920 and Lmip1548 (Mahbubani et al., 1990) and inner primers Lmip976 and Lmip1427 (Catalan et al., 1994). All PCR reactions were performed using Tks Gflex DNA Polymerase (TaKaRa Bio Inc., Shiga, Japan). The amplified PCR products were analyzed using agarose gel electrophoresis and visualized with a UV transilluminator. The amplicons of *mip* PCR were subjected to Sanger sequencing analysis. The obtained sequences were aligned using ClustalW (Larkin et al., 2007), and p-distances among sequences were calculated with MEGA6 (Tamura et al., 2013).

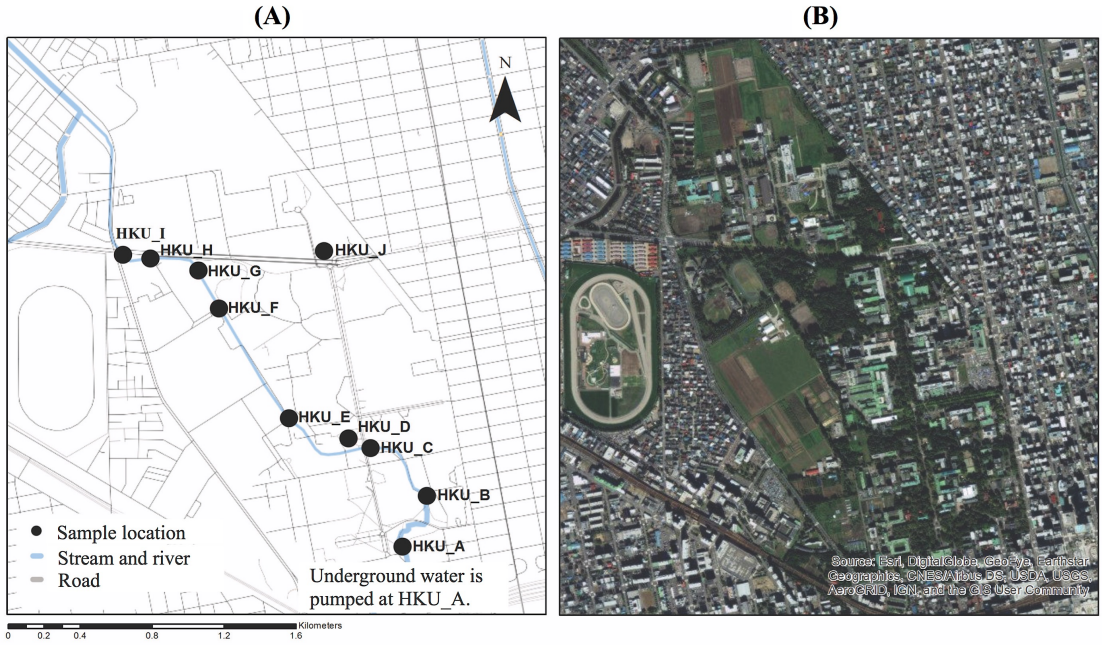

Figure 1. Locations of sample collection. Water samples were collected from Sakushukotoni Stream (HKU_A, HKU_B, HKU_C, HKU_E, HKU_F, HKU_G, HKU_H, and HKU_I), Ohno Pond (HKU_D), and Hyotan Pond (HKU_J). (A) A map from OpenStreetmap. (B) A satellite image from Google Earth**.**

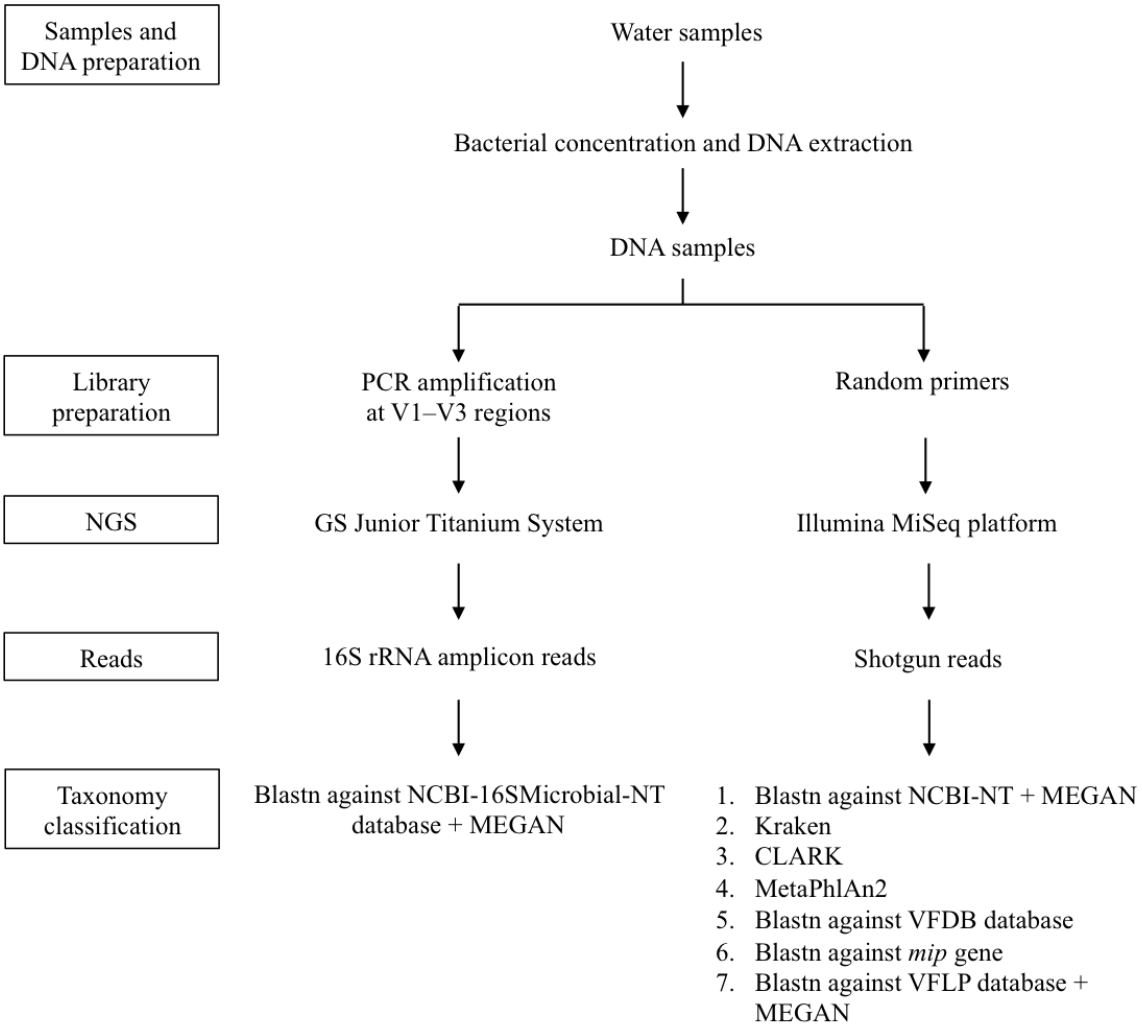

Figure 2. Flow chart of metagenomic analysis and different database search methods for the detection of *L. pneumophila*. A total of 2 liters of water samples were collected from Sakushukotoni Stream, Ohno Pond, and Hyotan Pond. Bacteria in water samples were concentrated using a filtration technique, and DNA was directly extracted from a 0.22 μm-membrane using a commercial kit. After DNA libraries were prepared, GS Junior Titanium system and Illumina MiSeq were used to analyze the nucleotide sequences. 16S rRNA amplicons were classified to bacterial genera using the blastn search against the NCBI-16SMicrobial-NT. Shotgun reads were classified to bacterial genera and species using a blastn search against NCBI-NT, VFDB, *mip*, and VFLP databases, and Kraken, CLARK, and MetaPhlAn2.

Illumina sequencing for shotgun metagenomic analysis

The Illumina MiSeq platform was used for shotgun metagenomic analysis. The sequencing libraries were prepared with a Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Libraries from each sample were tagged with multiplexing barcodes for analysis in one run. The final concentration of the purified libraries was normalized to 4 nM and the pooled libraries were sequenced with a MiSeq Reagent Kit v3 (Illumina). The resulting sequence data were made available at the DNA Data Bank of Japan (DDBJ) with an accession number of DRA006698. The barcoding sequences were removed using CLC Genomic Workbench software 8.0 (CLC bio, Tokyo, Japan). The resulting clean reads were used as shotgun reads for further analysis.

GS Junior sequencing for 16S rRNA amplicon analysis

The GS Junior Titanium System (Roche, Basel, Switzerland) was used for 16S rRNA amplicon analysis. The V1 – V3 hypervariable regions of the 16S rRNA gene was amplified with primers 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 518R (5'-GTATTACCGCGGCTGCTG-3'). The 27F primer was tagged with ten bases, which were used as barcodes for identifying reads from pooled samples in a single run. A total of 50 μl PCR mixture was prepared from 10 ng of DNA template, 1X PCR buffer, 2 mM MgCl2, 400 μM dNTP, 1 μM of each primer, and 2.5U Taq DNA polymerase (Life Technologies, Tokyo, Japan). The PCR condition was comprised of pre-denaturation at 94°C for 5 min, followed by 30 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 1 min, and finished with a final extension step at 72°C for 5 min. The amplicon products were checked on 2% (w/v) agarose gel using gel electrophoresis. The amplicon products were purified using a Wizard® SV Gel and PCR Clean-Up System (Promega, Tokyo, Japan) and sequenced using a GS Junior Titanium System. The resulting sequence data were made available at the DDBJ with an accession number of DRA006697. Barcoding sequences were removed as described above and reads shorter than 250 bp were also removed using the CLC Genomic Workbench software. Potential chimera sequences were removed using Chimera.Slayer (Haas et al., 2011), and the clean reads were used for further analysis.

Taxonomy classification of reads from shotgun metagenomic and 16S rRNA amplicon analyses by MEGAN

A blastn search (Altschul et al., 1990) and MEGAN (Huson et al., 2016) were used for the taxonomy classification of reads. Briefly, the NCBI-NT and NCBI-16SMicrobial-NT databases were downloaded from NCBI. Shotgun reads were used to align against the NCBI-NT database using a blastn search. The command line for blastn is:

blastn –db NCBI-NT –evalue 1e-04 –max_target_seqs 1 –num_threads 8 –outfmt "6 std staxids scomnames stitle salltitles" –query Sample.fasta –out Output.txt,

where Sample.fasta is a file name of shotgun reads and Output.txt is a result file name. Then, the blastn results were analyzed using the naïve LCA algorithm in MEGAN with parameters of min score = 50.0, max expected = 0.01, top percent = 10.0, min support percent = 0.001, and min support = 1. The proportions of a bacterial genus (or species) were calculated using the numbers of reads classified to the genus (or species) divided by the number of reads classified as bacteria. Numbers of reads mapped to each bacterial genus in each sample were subjected to principal component analysis (PCA) using the "prcomp" command in R. The number of reads identified as *L. pneumophila* was collected after taxonomy classification. Separately, the reads generated from the 454 GS Junior Titanium System were aligned against the NCBI-16SMicrobial-NT database using a blastn search. Taxonomy classification and downstream analysis were conducted as mentioned above. The proportions of *Legionella* spp. were calculated by dividing the number of reads identified as *Legionella* spp. by the number of reads identified as bacteria.

Detection of *L. pneumophila* using Kraken and CLARK

Two *k*-mer based taxonomy classification algorithms, Kraken (Wood and Salzberg, 2014) and CLARK (Ounit et al., 2015), were used to detect *L. pneumophila* from shotgun metagenomic analysis. For the Kraken analysis, the reference sequences (RefSeq) of bacteria, archaea, and viruses were downloaded from the Kraken webpage, and a standard Kraken database was constructed. Shotgun reads were aligned and classified to the bacterial taxonomy using Kraken v1.0 with default parameters. For CLARK, only the RefSeq of bacteria were obtained from the CLARK webpage, and they were used to construct a bacterial database. Shotgun reads of each sample were aligned and classified to the bacterial taxonomy using CLARK v1.2.3.2 with default parameters.

After taxonomy classification, the numbers of reads identified as *L. pneumophila* were collected and the proportions of reads identified as *L. pneumophila* were calculated.

Detection of *L. pneumophila* using MetaPhlAn2

In addition to the database search using complete genomes, MetaPhlAn2 (Truong et al., 2015) was used to detect *L. pneumophila*. The database of unique clade-specific marker genes from bacteria, archaea, viruses, and eukaryotes were downloaded from the MetaPhlAn2 webpage. Each sample of shotgun reads was aligned against the database using MetaPhlAn2 with the default parameters. The numbers of reads identified as *L. pneumophila* were collected and the proportions of reads identified as *L. pneumophila* were calculated.

Detection of *L. pneumophila* using blastn against VFDB database

Nucleotide sequences of virulence factor genes were downloaded from the Virulence Factor Gene database (Chen et al., 2005). All reference sequences were used to construct the VFDB blast database using the "makeblastdb" command from the blast package. The command for database construction is:

makeblastdb –in Sequence.fasta –input_type fasta –dbtype nucl –out VFDB,

where Sequence.fasta is a file of nucleotide sequences downloaded from the Virulence Factor Gene database and VFDB is the name of the constructed database. Shotgun reads were aligned against the database using blastn with the parameters as described above. Blastn results with multiple hits from the same query to different regions of the same reference sequence were removed, except one. The proportions of *L. pneumophila* hits were calculated by dividing the number of reads classified as *L. pneumophila* by the number of reads classified as bacteria.

Detection of *L. pneumophila* using blastn against *mip* database

A nucleotide sequence of the *mip* gene from *L. pneumophila* subsp. *philadelphia* str. Philadelphia 1 (NC_002942.5) was downloaded from NCBI. A *mip* blast database was constructed using the "makeblastdb" command. Shotgun reads from each sample

were aligned to this database using a blastn search, and the number of hit reads was collected.

Detection of *L. pneumophila* using blastn against virulence factor genes of *L. pneumophila* (VFLP) database

Based on the results of a blastn search of shotgun reads against the VFDB blast database, *L. pneumophila* reads were identified from nineteen virulence factor genes (n = 19). All reads identified as *L. pneumophila* were extracted from sequence files and categorized based on the information of gene names from the sequence id. These reads were aligned again against the NCBI-NT database using a blastn search for confirmation of *L. pneumophila* identification. Reads identified as *L. pneumophila* were included for further analysis, whereas reads identified as non-*L. pneumophila* were excluded. Finally, the reads of *L. pneumophila* origin were identified from nine virulence factor genes.

For each virulence factor gene (n = 9), the protein sequences of *L. pneumophila* subsp. *philadelphia* str. Philadelphia 1 were downloaded from NCBI. These protein sequences are CcmC (YP_094893.1), CcmF (YP_094896.1), DotA (YP_096691.1), IcmO (YP_094490.1), KatB (YP_096397.1), LvhB10 (YP_095278.1), PilT (YP_096029.1), GTP pyrophosphokinase (YP_095486.1), and superoxide dismutase (YP_096960.1). Nucleotide sequences encoding these nine proteins were collected using the tblastn search at NCBI (4,267, 4,526, 483, 707, 2,686, 5,506, 5,000, 5,000, 5,000 sequences were obtained for *ccmC*, *ccmF*, *dotA*, *icmO*, *lvhB10*, *katB*, *pilT*, *relA*, and *sodB*, respectively) and VFLP blast databases of each gene were constructed. A blastn search of shotgun reads against each of a VFLP database was performed, and the number of reads identified as *L. pneumophila* was obtained using the naïve LCA algorithm in MEGAN.

Phylogenetic tree analysis of nucleotide sequences encoding catalase peroxidase from *L. pneumophila* and other bacteria

Based on the blast results of shotgun reads against each of nine VFLP databases, the blastn search of shotgun reads against the database using *katB* demonstrated the best agreement in detecting *L. pneumophila* out of those using other VFLP databases, compared with *L. pneumophila*-specific nested PCR. It is known that *katB* encodes

catalase-peroxidase. However, this protein can also be encoded by *katG* from *L. pneumophila* and other bacteria. In order to confirm the differences of nucleotide sequences encoding the catalase-peroxidase, nucleotide sequences encoding this protein from *L. pneumophila* and other bacteria were downloaded from NCBI. All sequences were aligned using clustalW, and the tree was constructed using Maximum likelihood in MEGA6.

Comparison of database search methods for *L. pneumophila* detection

The area under the curve (AUC) of the receiver operating characteristic curve (ROC curve) was used to compare the results of different database search methods in the detection of *L. pneumophila*. The results of *L. pneumophila*-specific nested PCR were considered as correct. The true positive rate and false positive rate (1 – specificity) of each database search method were calculated. The area under curves were determined using the AUC package (Ballings and Van den Poel, 2013) in R.

Results

Detection of *Legionella* spp. and *L. pneumophila* using nested PCR

Legionella spp. was detected in all samples by *Legionella* genus-specific nested PCR (Figure 3A). The amplification of the *mip* gene by *L. pneumophila*-specific nested PCR was observed in only three samples; HKU_G, HKU_H, and HKU_I (Figure 3B). These results suggest that seven samples, i.e. all except HKU_G, HKU_H, and HKU_I, contained *Legionella* spp. not classified as *L. pneumophila*. The pairwise distances among the *mip* gene sequences from amplified samples and positive control were within a range of 0.018 – 0.030, indicating that there was no cross contamination from the positive control during the PCR process. Therefore, the samples HKU_G, HKU_H, and HKU_I were contaminated with *L. pneumophila*.

Next generation sequencing

Next generation sequencing was conducted using Illumina MiSeq and GS Junior Titanium System, from which a total of 51,162,136 and 353,913 reads were obtained, respectively (Table 1). The average lengths of bacterial reads obtained from 16S rRNA amplicon analysis were within a range of 453.3 – 473.4 bp, whereas the average length of bacterial reads obtained from MiSeq were within a range of 288.8 – 293.1 bp.

Bacterial communities inferred from 16S rRNA amplicon and shotgun metagenomic analyses

16S rRNA amplicon and shotgun sequence reads were subjected to a blastn search against the NCBI-16SMicrobial-NT and NCBI-NT databases, respectively. The proportions of bacterial genera inferred using the naïve LCA algorithm of MEGAN are shown in Figures 4A and 4B for 16S rRNA amplicon and shotgun reads, respectively. More than 75% of reads generated by the GS Junior Titanium System were identified as having bacterial origins, whereas 19.9% – 45.0% of the Illumina reads were identified as having bacterial origins. A total of 977 bacterial genera were detected from the 16S rRNA amplicon analysis, while a total of 897 bacterial genera were found in the shotgun metagenomic analysis. The PCA suggested that the bacterial communities in the samples were divided into three groups in both the 16S rRNA amplicon and shotgun metagenomic analyses (Figures 4C and 4D).

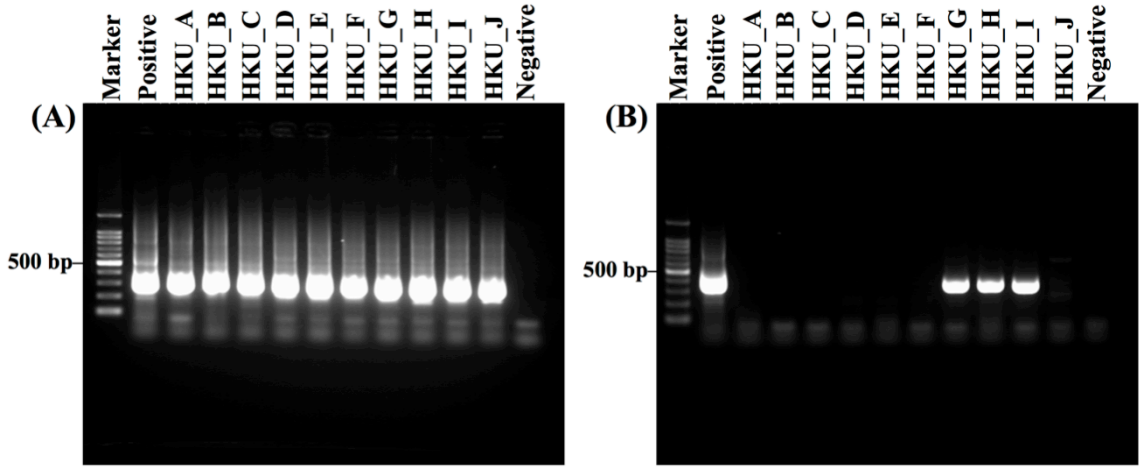

Figure 3. Gel electrophoresis of DNA amplification by *Legionella* genus-specific and *Legionella pneumophila*-specific nested PCRs. (A) Amplification results of *Legionella* genus-specific and (B) *L. pneumophila*-specific nested PCRs are shown. Lane Marker : 100 bp DNA marker; lane Positive : *L. pneumophila*; lanes HKU_A – HKU_C and HKU_E – HKU_I : DNA from water samples of Sakushokotoni Stream; lane HKU_D : DNA from water samples of Ohno Pond; lane HKU_J : DNA from water samples of Hyotan Pond; and lane Negative : distilled water (no DNA).

Table 1. Summary of next generation sequencing reads.

Methods	Samples	Number of raw reads	Number of passed-QC reads	Number of reads hit with database	Number of reads identified as bacteria by MEGAN	Average length of bacterial reads
16S rRNA analysis	HKU_A	46,968	39,245	39,166[a]	35,404	455.7
	HKU_B	29,684	25,183	25,093[a]	24,711	453.3
	HKU_C	39,167	32,628	32,534[a]	32,256	450.5
	HKU_D	35,360	28,564	28,504[a]	28,409	461.5
	HKU_E	32,936	25,826	25,735[a]	25,654	464.9
	HKU_F	29,649	24,215	24,143[a]	24,063	465.4
	HKU_G	43,416	33,846	33,692[a]	33,636	466.8
	HKU_H	28,235	21,948	21,852[a]	21,735	462.8
	HKU_I	38,581	30,646	30,554[a]	30,510	468.5
	HKU_J	29,917	25,313	25,232[a]	25,065	473.4
Shotgun analysis	HKU_A	1,554,614	N/A	318,064[b]	309,063	288.8
	HKU_B	5,291,304	N/A	1,628,823[b]	1,600,198	293.1
	HKU_C	7,078,858	N/A	2,354,608[b]	2,323,879	289.7
	HKU_D	5,430,216	N/A	1,891,874[b]	1,873,938	284.5
	HKU_E	6,046,758	N/A	2,283,330[b]	2,264,076	285.4
	HKU_F	6,350,502	N/A	2,024,966[b]	2,006,002	284.7
	HKU_G	4,992,354	N/A	1,769,319[b]	1,752,738	285.8
	HKU_H	6,039,572	N/A	1,896,294[b]	1,872,777	286.3
	HKU_I	4,581,078	N/A	1,729,619[b]	1,714,434	285.5
	HKU_J	3,796,880	N/A	1,719,066[b]	1,710,319	289.7

[a] Number of reads hit with the NCBI-16SMicrobial-NT database

[b] Number of reads hit with the NCBI-NT database

N/A means these reads do not check quality, and all of them were used for an analysis.

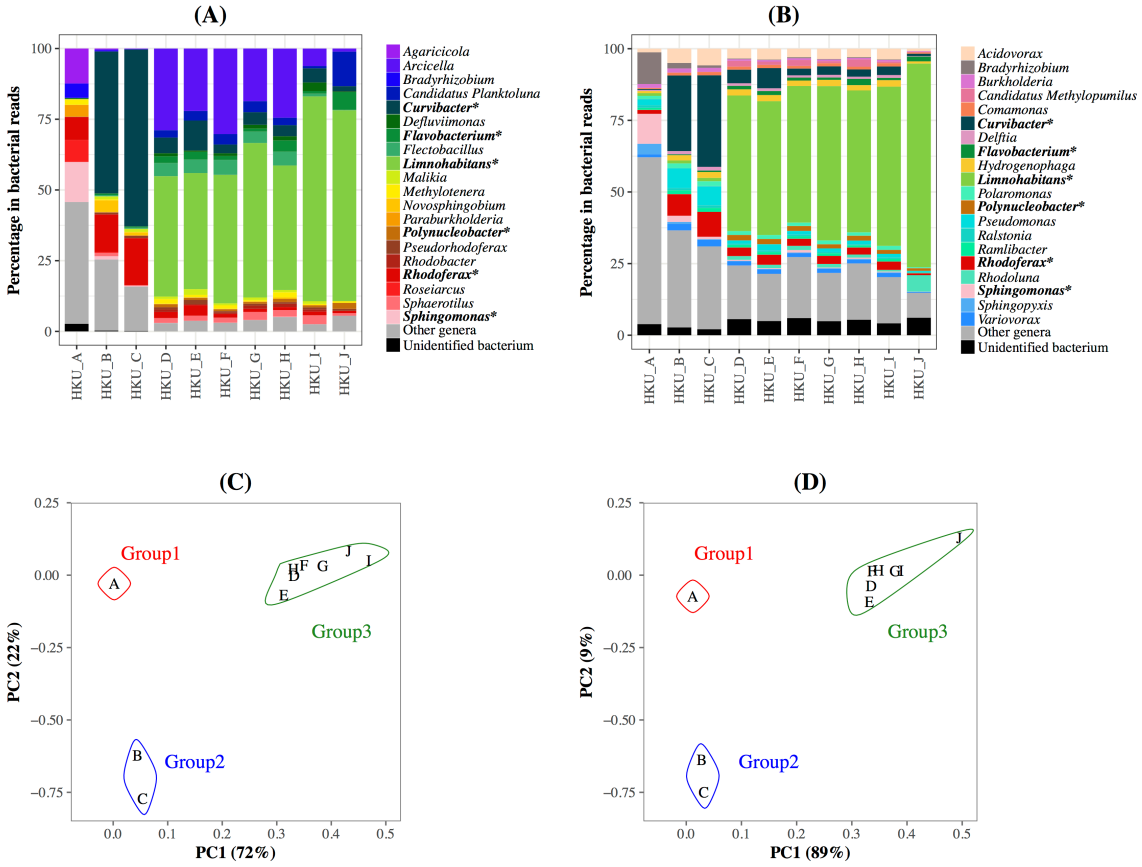

Figure 4. Bacterial communities at the genus level in water samples determined by the naïve LCA algorithm in MEGAN. (A) Bacterial communities based on the results of a blastn search of 16S rRNA amplicon reads against nucleotide sequences from the NCBI-16SMicrobial-NT database. (B) Bacterial communities based on the results of the blastn search of shotgun reads against nucleotide sequences from the NCBI-NT database. Color bars represent the top 20 abundant genera in all samples. Reads from other minor genera are represented in gray, and the reads with unidentified genera are represented in black. The genera ranked in top 20 in both 16S rRNA amplicon and shotgun metagenomic analyses are indicated with asterisks. Unclassified reads at genus level were presented as unidentified bacterium. The results of principal component analysis of the bacterial communities using 16S rRNA reads and shotgun reads are shown in panel (C) and (D), respectively.

Some genera showed similar proportions of reads between 16S rRNA amplicon and shotgun metagenomic analyses, whereas others did not. For example, more than 10% of reads were identified as *Sphingomonas* in both the 16S rRNA amplicon and shotgun metagenomic analyses (14.1% and 10.4%, respectively) in group 1 (HKU_A). Shotgun metagenomic analysis identified *Pseudomonas* (6.6% – 7.0%) in group 2, but this genus was not found in the top 20 genera in the 16S rRNA amplicon analysis. The highest portion of a bacterium in group 3 was *Limnohabitans* in both the 16S rRNA amplicon and shotgun metagenomic analyses (41.0% – 72.4% and 47.3% – 71.1%, respectively). In contrast, the 16S rRNA amplicon analysis identified a moderate number of reads from *Arcicella* (1.1% – 30.2%) in group 3, but this abundant genus was not listed in the top 20 genera of the shotgun metagenomic analysis.

Detection of *L. pneumophila* using MEGAN, Kraken, CLARK, MetaPhlAn2, VFDB, and *mip* database

To investigate the diagnostic ability of different database search methods in the detection of *L. pneumophila*, the results of each method were compared with that of *L. pneumophila*-specific nested PCR (Table 2). Although the nested PCR amplified sequences of the *mip* gene of *L. pneumophila* in three samples, blastn searches of shotgun reads did not detect any reads encoding the *mip* gene (Table 2, Figure 5F). Similarly, MetaPhlAn2 detected no read of *L. pneumophila* in all samples (Table 2, Figure 5E). In contrast, MEGAN with the NCBI-NT database, Kraken and CLARK with the RefSeq database detected a moderate number of *L. pneumophila* sequences in all samples (Table 2). MEGAN, Kraken, and CLARK identified the highest proportion of *L. pneumophila* reads in HKU_A (Figures 5A, 5B, and 5C) even though HKU_A was negative by *L. pneumophila*-specific nested PCR assay. On the other hand, the use of VFDB detected no *L. pneumophila* read in HKU_A, and a relatively higher proportion of *L. pneumophila* reads in HKU_G, HKU_H, and HKU_I (Figure 5D), which were positive by nested PCR (Figure 3B). VFDB hits contained 19 virulence factor genes (Table 3). Blastn searches of detected sequences against NCBI-NT indicated that 10 virulence factor genes were derived from other bacterial species. Finally, 9 virulence factor genes (*ccmC*, *ccmF*, *dotA*, *icmO*, *lvhB10*, *katB*, *pilT*, *relA*, and *sodB*) were identified as *L. pneumophila* origin (Table 4).

Table 2. Numbers of shotgun reads identified as *Legionella pneumophila* by MEGAN, Kraken, CLARK, MetaPhlAn2, VFDB, and *mip* database.

Sample	*L. pneumophila*-specific nested PCR	Number of shotgun reads identified as *Legionella pneumophila*					
		MEGAN + NCBI-NT	Kraken	CLARK	VFDB	MetaPhlAn2	*mip*
HKU_A	Negative	130	90	99	0	0	0
HKU_B	Negative	220	136	125	19	0	0
HKU_C	Negative	200	134	117	22	0	0
HKU_D	Negative	135	63	63	27	0	0
HKU_E	Negative	129	45	45	59	0	0
HKU_F	Negative	159	71	83	28	0	0
HKU_G	**Positive**	100	27	42	28	0	0
HKU_H	**Positive**	178	75	79	24	0	0
HKU_I	**Positive**	81	30	40	18	0	0
HKU_J	Negative	104	86	34	3	0	0

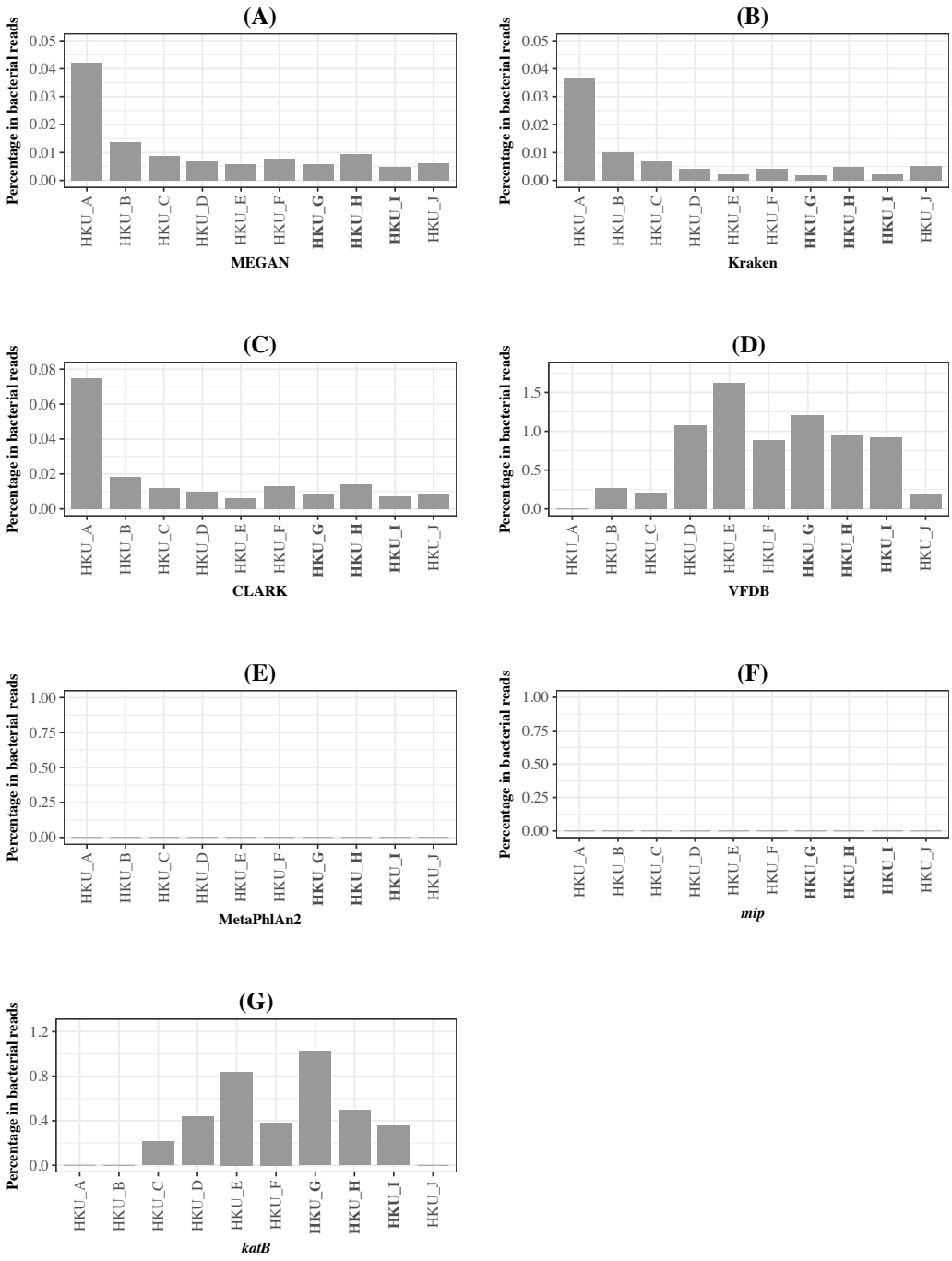

Figure 5. Percentage of shotgun reads identified as *Legionella pneumophila* sequences using different database search methods. (A) Blastn search against NCBI-NT and MEGAN, (B) Kraken, (C) CLARK, (D) blastn search against VFDB, (E) MetaPhlAn2, (F) blastn search against *mip* database, and (G) blastn search against *katB* database. The bold labels indicate *L. pneumophila* positive samples by *L. pneumophila* nested PCR.

Table 3. Number of shotgun reads identified as *Legionella pneumophila* sequences categorized by virulence factor genes.

Gene	Sample									
	HKU_A (865)*	**HKU_B** (7,163)*	**HKU_C** (10,682)*	**HKU_D** (2,520)*	**HKU_E** (3,645)*	**HKU_F** (3,147)*	**HKU_G** (2,320)*	**HKU_H** (2,552)*	**HKU_I** (1,949)*	**HKU_J** (1,490)*
ccmC		3	6			1	1			
ccmF			1		1					
dotA		2								
fleQ			2							
htpB		4	1	7	12	10	4	7	4	3
icm0			1							
katA		1								
katB		1	2	16	36	9	16	15	13	
legK3			1							
letS		2	2	4	8	3	2	1		
lpg0773						1				
lvhB10		1								
motB		1	4		2	3	5			
pgi									1	
pilT		2								
pilZ								1		
relA		1								
sdcB			2							
sodB		1				1				

* Numbers of shotgun reads identified as bacterial reads.

Table 4. Number of shotgun reads identified as *Legionella pneumophila* using the blastn search against VFDB database and NCBI-NT database.

Gene	Species	Sample									
		HKU_A	HKU_B	HKU_C	HKU_D	HKU_E	HKU_F	HKU_G	HKU_H	HKU_I	HKU_J
ccmC	*Achromobacter xylosoxidans*						1				
	Legionella pneumophila		2	3							
	Oceanisphaera profunda		1								
	Ralstonia pickettii							1			
	Sulfuricella denitrificans			2							
ccmF	*Achromobacter xylosoxidans*						1				
	Legionella pneumophila		2	3							
dotA	***Legionella pneumophila***		2								
fleQ	*Candidatus Midichloria mitochondrii*			2							
htpB	*Acinetobacter johnsonii*			1							
	Beta proteobacterium				1				1		
	Bdellovibrio bacteriovorus						1				
	Bdellovibrio exovorus				2						
	Bacillus sp.		1								
	Cellvibrio sp.		1				1				
	Legionella hackeliae								1		
	Methylotenera mobilis		1								
	Polynucleobacter duraquae					4	1	1	1	1	
	Polynucleobacter necessarius				4	8	7	3	4	3	3
icmO	***Legionella pneumophila***			1							
katA	*Dyella japonica*		1								
katB	*Aeromonas salmonicida*				4	6	1	5	4	2	
	Bacillus lehensis								1		
	Dechlorosoma suillum			1						2	
	Flavobacterium sp.						1				
	Hydrogenophaga sp.					2					
	Lacinutrix sp.								1		
	Laribacter hongkongensis					2					
	Legionella pneumophila			1	2	3	3	4	2	2	
	Marinovum algicola				3	2				2	
	Pseudomonas syringae					5		1			
	Vibrio owensii		1								
	Vitreoscilla filiformis				7	16	4	6	7	5	
legK3	No bacterium was identified										
letS	*Caldilinea aerophila*						1				
	Cellvibrio sp.		1								
	Curvibacter sp.		1		2	2			1		
	Geitlerinema sp.			1							
	Limnohabitans sp.				1	5	2	2			
	Pseudomonas sp.					1					
	Rhodoferax saidenbachensis				1						
	Vitreoscilla filiformis			1							
lpg0773	No bacterium was identified										
lvhB10	***Legionella pneumophila***		1								
motB	*Curvibacter* sp.			1		1		1			
	Curvibacter sp.		1	3		1	3	4			
pgi	*Polynucleobacter necessarius*									1	
pilT	***Legionella pneumophila***		2								
pilZ	No bacteria was identified										
relA	***Legionella pneumophila***		1								
sdcB	*Legionella longbeachae*			2							
sodB	*Glaciecola psychrophila*		1								
	Legionella pneumophila						1				

Detection of *L. pneumophila* using VFLP databases

Table 5 shows the number of shotgun reads identified as virulence factor genes associated with *L. pneumophila*. Among the nine genes, a blastn search of shotgun reads against the *katB* gene showed the best agreement with the results from *L. pneumophila*-specific nested PCR. Moreover, the *katB* genes indicated the difference in nucleotide sequences among bacteria and were different from *katG* (Figure 6).

Diagnostic ability of *L. pneumophila* using a *katB* gene

Figure 5 presents the percentage of *L. pneumophila*-associated reads identified by 7 different database search methods. Among the 7 database search methods, the blastn search against the *katB* gene showed the best agreement with the results from *L. pneumophila*-specific nested PCR. The highest percentage of shotgun reads identified as *L. pneumophila* origin was observed in HKU_G (Figure 5G). None of the reads identified as *L. pneumophila* were found in HKU_A, HKU_B, and HKU_J. Separately, the non-bacterial reads were classified as archaea, fungi, and metazoan reads.

The AUC of database search methods demonstrated that the detection of *L. pneumophila* using the *katB* gene had the highest AUC at 0.8095 (Figure 7E). Other database search methods such as MEGAN with NCBI-NT, Kraken and CLARK with RefSeq database had AUC values with a range between 0.2142 and 0.3095; lower than that using the *katB* gene (Figures 7A, 7B, and 7C). The database search method using the VFDB database had an AUC value of 0.7619 (Figure 7D). These results indicate that the blastn search against the *katB* gene database had higher diagnostic capability than searches against databases containing complete genome sequences of *L. pneumophila.*

Table 5. Numbers of shotgun reads identified as *Legionella pneumophila* using blastn search against nine databases of virulence factor genes.

Samples	Number of reads identified as *Legionella pneumophila* / Number of reads identified as bacterial sequences								
	ccmC[a]	***ccmF***[a]	***dotA***[a]	***icmO***[a]	***lvhB10***[a]	***katB***[a]	***pilT***[a]	***relA***[a]	***sodB***[a]
HKU_A	0 / 123	0 / 193	0 / 0	0 / 0	1 / 120	0 / 226	0 / 176	0 / 71	0 / 77
HKU_B	2 / 172	0 / 317	2 / 4	0 / 18	1 / 113	0 / 733	2/ 1423	1 / 480	0 / 370
HKU_C	3 / 182	0 / 373	0 / 5	1 / 6	0 / 67	2 / 942	0 / 2440	0 / 742	0 / 529
HKU_D	0 / 279	0 / 595	0 / 0	0 / 0	0 / 10	2 / 458	0 / 637	0 / 452	0 / 603
HKU_E	1 / 386	0 / 704	0 / 0	0 / 0	0 / 15	5 / 596	0 / 1019	0 / 507	0 / 603
HKU_F	0 / 356	0 / 672	0 / 1	0 / 0	0 / 19	2 / 520	0 / 793	0 / 494	1 / 665
HKU_G[b]	0 / 300	0 / 456	0 / 0	0 / 0	0 / 7	4 / 390	0 / 512	0 / 446	0 / 523
HKU_H[b]	0 / 346	0 / 559	0 / 0	0 / 0	0 / 7	2 / 405	0 / 615	0 / 460	0 / 652
HKU_I[b]	0 / 299	0 / 583	0 / 0	0 / 0	0 / 7	1 / 281	0 / 500	0 / 423	0 / 533
HKU_J	0 / 374	0 / 795	0 / 0	0 / 0	0 / 3	0 / 197	0 / 196	0 / 577	0 / 380

[a] Nucleotide lengths of *ccmC*, *ccmF*, *dotA*, *icmO*, *lvhB10*, *katB*, *pilT*, *relA*, and *sodB* are 789 bp, 1,950 bp, 3,144 bp, 2,349 bp, 1,089 bp, 2,193 bp, 1,032 bp, 2,202 bp, and 588 bp, respectively.

[b] These samples were positive for *L. pneumophila* detection by *L. pneumophila*-specific nested PCR.

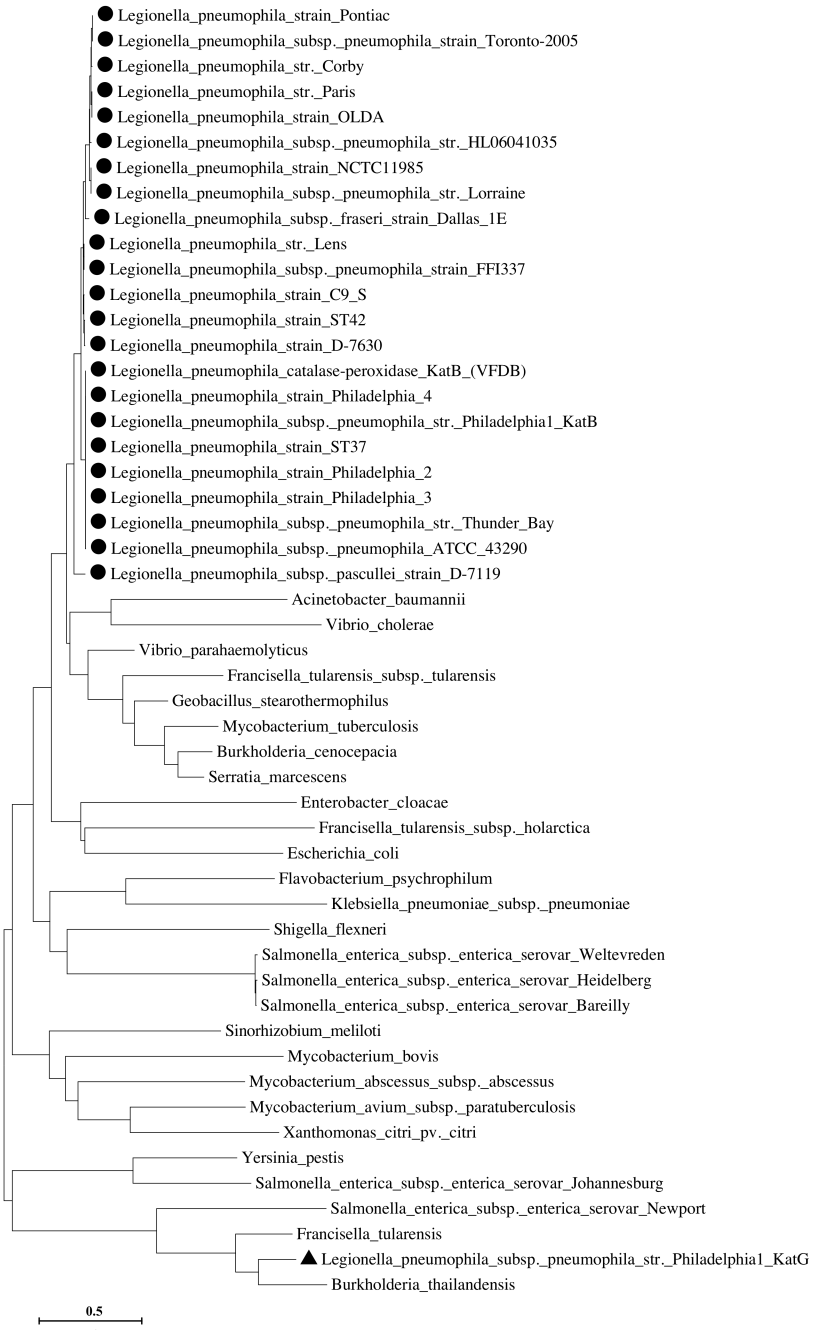

Figure 6. Phylogenetic tree of genes coding catalase-peroxidase from *Legionella pneumophila* and other bacteria. The tree was constructed using maximum likelihood in MEGA based on the nucleotide sequences of genes. The bootstrap confidence values were generated using 1,000 permutations. Different symbols indicate different genes from *L. pneumophila*: circles represent *katB* sequences and triangle represents *katG* sequences from *L. pneumophila*.

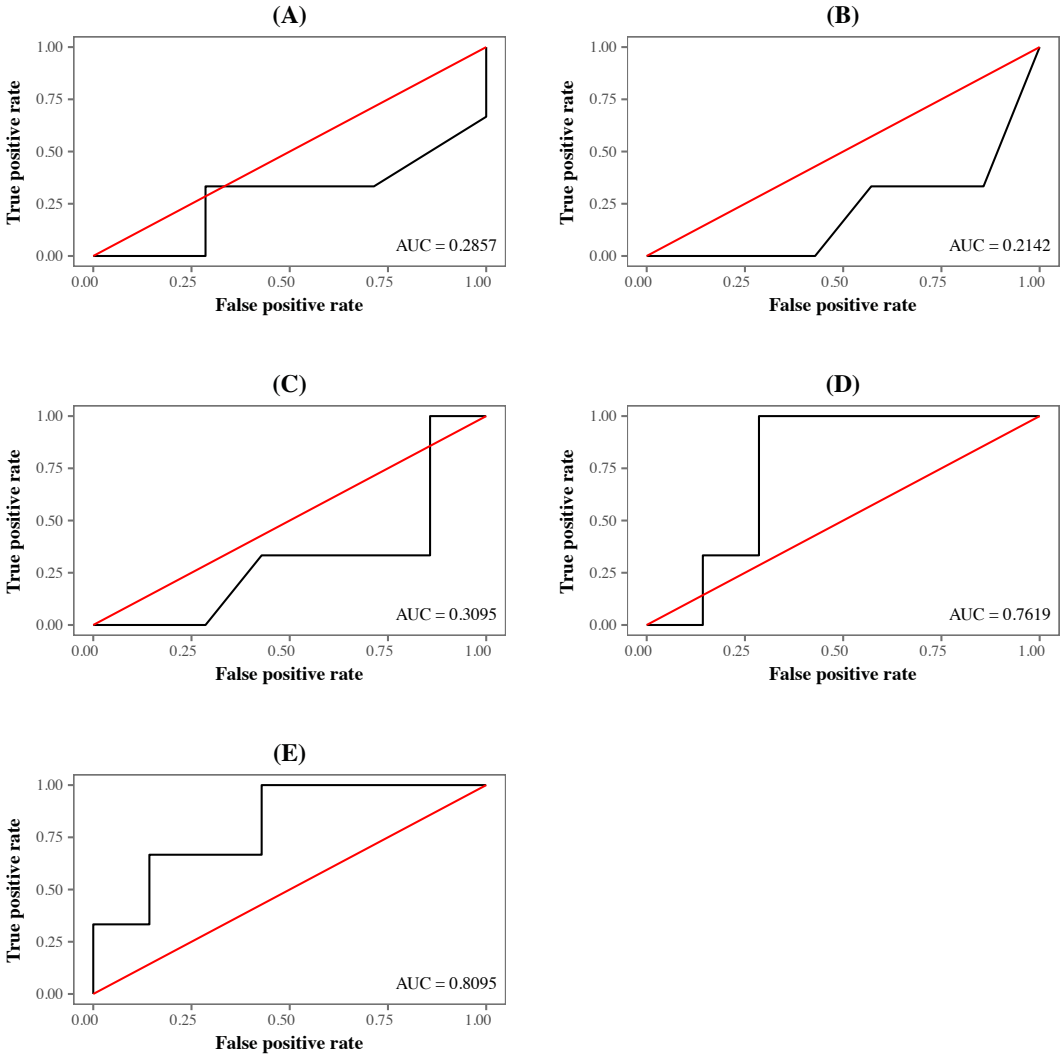

Figure 7. The ROC curves for different database search methods. (A) MEGAN with blastn search against NCBI-NT database, (B) Kraken with RefSeq archaea, bacteria, and viruses, (C) CLARK with RefSeq archaea and bacteria, (D) blastn search against VFDB database, and (E) blastn search against *kat*B database. The red line is the reference line indicating the test without diagnostic benefit, i.e., random diagnosis.

Discussion

In this study, metagenomic analyses were conducted to analyze bacterial populations in water samples collected from a stream and ponds in the campus of Hokkaido University. By focusing on *L. pneumophila*, different database search methods were evaluated to detect a specific bacterium in water samples by validating the results with those of nested PCR assay. A blastn search of shotgun reads against the NCBI-NT database showed false positive detection and had a potential problem in specificity. The results indicated that a blastn search against the genes of species-specific virulence factors had better agreement with the results of *L. pneumophila*-specific nested PCR.

The population structures inferred by 16S rRNA amplicon analysis and those by shotgun metagenomic analysis showed different bacterial communities even at the genus level (Figures 4A and 4B). However, PCA using 16S rRNA amplicon and shotgun metagenomic analyses clustered the samples in a similar way (Figures 4C and 4D). These results indicated that both 16S rRNA amplicon and shotgun metagenomic analyses captured the similarity in population structures among samples, but sensitivity and/or specificity of the two methods were different depending on bacterial genera.

The nested PCR assay detected *L. pneumophila* DNA in only three out of ten water samples (Figure 3B). In contrast, MEGAN with NCBI-NT database and Kraken and CLARK with the RefSeq database detected a moderate number of *L. pneumophila* sequences in the shotgun reads from all samples (Table 2). Furthermore, MEGAN with NCBI-NT, Kraken, and CLARK with RefSeq database detected a larger number of *L. pneumophila* sequences in PCR-negative samples such as HKU_B and HKU_C than in PCR-positive samples including HKU_G, HKU_H, and HKU_I (Table 2). Since the sensitivity of nested PCR assay with the employed primer sets is known to be 10 fg or 10 CFU per ml (Nintasen et al., 2007), the inconsistency is probably attributed to false positive detections due to the low specificity of these database search methods in detecting *L. pneumophila*.

The NCBI-NT and RefSeq databases contain complete genome sequences of *L. pneumophila*. The sequences of some of the bacterial genomic regions, for example the loci encoding housekeeping genes, are conserved among closely related bacterial species.

Wrong assignment of the reads from such conserved genomic loci may be a possible cause of the false positive detection with MEGAN with NCBI-NT, Kraken and CLARK with RefSeq databases. A large fraction of reads assigned to *L. pneumophila* in HKU_A may be attributed to wrong assignment of reads from other abundant species in genus *Legionella* (Figures 5A, 5B, and 5C).

The ROC plot analysis showed that detection using the *katB* gene had the largest area under the curve, indicating that the method was the best among the database search methods (Figure 7). The *katB* gene can be found in several bacterial species, but nucleotide sequences of *katB* are divergent among different bacterial species (Figure 6). This would be the reason for the high diagnostic ability of the method using the *katB* gene. The *mip* gene is a genetic marker for detecting *L. pneumophila* using PCR-based assay (Cianciotto et al., 1989). However, the shotgun reads did not contain a DNA fragment of the *mip* gene (Table 2 and Figure 4F). The nucleotide length of the *mip* gene is 702 bp, while the length of a *katB* gene is 2,163 bp. The read depth of certain genes in shotgun metagenomic sequencing is proportional to the length of the gene. It was speculated that the length of the *mip* gene might affect the absence of the gene in the metagenomic sequencing data. Despite *dotA* (3,144 bp) having more nucleotides than the *katB* gene, the number of reads identified as *L. pneumophila* using the *dotA* gene is smaller than that using the *katB* gene (Table 5). It is known that *dotA* determines the serogroup of *L. pneumophila* (Ko et al., 2003). There is a possibility that the *L. pneumophila* present in our samples belong to different serogroups from *L. pneumophila* subsp. *philadelphia* str. Philadelphia 1, which is used as the reference sequence for the tblastn search to collect nucleotide sequences.

Nested PCR using specific primers to amplify a *mip* gene detected *L. pneumophila* in only three samples; HKU_G, HKU_H, and HKU_I (Figure 2B). *L. pneumophila* can be found in natural water supplies (Mahbubani et al., 1990), and there is no report of outbreaks of *L. pneumophila* on the university campus. Since the sampling sites of HKU_G, HKU_H, and HKU_I are near a primeval forest conserved by the university, the pathogen has probably existed naturally and is not associated with the emergence of Legionnaires' disease.

Although the detection of *L. pneumophila* using PCR-based methods is relatively rapid and sensitive, it is necessary to know the sequences of the target bacteria in advance. Conversely, a shotgun metagenomic approach does not require sequence information and thus is potentially useful in the detection of new and/or unexpected organisms. High throughput is another advantage of the metagenomic approach in that the method can detect multiple organisms in a single run. In fact, several studies have demonstrated the usefulness of metagenomic analysis in water science. Gomez-Alvarez et al. (2012) used metagenomics to investigate microbial populations in drinking water and found that *Legionella*-like genes were abundant in free-chlorine-treated drinking water. Metagenomic analysis showed the potential risk of *Mycobacterium tuberculosis*-like in water samples from wastewater treatment plants (Cai and Zhang, 2013). Several studies have detected bacterial genes related to antibiotic resistance in water samples (Zhang et al., 2011;Durso et al., 2012;Wang et al., 2013). Pereira et al. (2017) proposed a novel approach to increase the sensitivity of *Legionella* detection in metagenomics. These studies are examples of possible directions for future application of metagenomics in detecting pathogens in water.

This study has a limitation due to a lack of information for *L. pneumophila* in water samples. The conventional method could be used to enumerate the number of *L. pneumophila* in the water samples. Based on the sensitivity of the *L. pneumophila*-specific nested PCR (Nintasen et al., 2007), the number of *L. pneumophila* were estimated to be at least 10 CFU/ml. Another limitation of this study was the number of reads generated by MiSeq. HiSeq can produce a larger number of sequence reads with deeper coverage. In this sense, the sensitivity of detection of *L. pneumophila* might be increased by HiSeq. At the same time, however, the length of reads from HiSeq are 100 – 150 bp shorter than those of MiSeq, which produces 300 bp reads. In this sense, the specificity of detection might be decreased by HiSeq. The number and the length of sequence reads are a tradeoff as well as sensitivity and specificity. These tradeoffs should be considered when conducting shotgun metagenomic analysis to detect pathogens in water samples. One possible future work is the evaluation of the detection limit of *L. pneumophila* in water samples using metagenomic analysis. Comparison of results among culture-based methods, quantitative RT PCR, and metagenomic analysis can be used to discuss the detection limit of *L. pneumophila* in water samples.

Summary

In metagenomic studies, the taxonomy classification of bacteria requires an appropriate database search method for bacterial identification at the genus and/or species levels. In this Chapter, the ability of different database search methods for detecting *L. pneumophila* using metagenomic analyses was compared with the detection result of *L. pneumophila*-specific nested PCR. The database search methods using complete genomes or RefSeq sequences led to false positive results. A blastn search against the catalase-peroxidase (*katB*) gene sequences indicates the highest AUC of *L. pneumophila* detection among the tested search methods. This study suggests that sequence searches targeting a long gene specifically associated with a bacterial species of interest have better potential diagnosis using current NGS technologies.

Chapter II

Shotgun metagenomic analysis of pathogenic bacteria and antibiotic resistance genes in environmental samples collected from urban and rural areas in Thailand

Introduction

In Thailand, ARB have been a concern in humans, animals, and food products from animals for decades. Several studies have demonstrated evidence of bacteria with antibiotic resistance in various kinds of samples. Hoge et al. (1998) examined antibiotic resistance in enteric bacteria from indigenous persons and travelers, and found that 97% – 100% of *Shigella dysenteriae* resisted ciprofloxacin. Petersen and Dalsgaard (2003) reported that 87% and 95% of *Enterococcus* isolated from fish farms resisted erythromycin and oxytetracycline, respectively. Angkititrakul et al. (2005) found that antimicrobial resistant *Salmonella* were widely spread among pork, chicken meat, and healthy workers in food plants. Padungtod et al. (2006) indicated that all isolates of *Campylobacter* from animals, healthy farm workers, and children with diarrhea resisted fluoroquinolones and tetracycline. Agerso and Petersen (2007) found that *tet*(39) and *suIII* are common resistance genes for tetracycline and sulphonamide resistances in *Acinetobacter* spp. isolated from fish farms. Mootsikapun et al. (2009) reported the increasing trend of AR in *Streptococcus pneumoniae* from clinical specimens in hospitals from 2000 – 2005. Changkaew et al. (2014) documented that the resistance to tetracycline, ampicillin, and trimethoprim were obviously found in *Escherichia coli* isolated from shrimp and their environments. Boonyasiri et al. (2014) reported a high prevalence of ESBL producing *Escherichia coli* in stool samples from healthy workers and rectal swabs from healthy pigs. Ngamwongsatit et al. (2016) presented high resistance rates of antibiotic resistance in *Clostridium perfringens* isolated from neonatal pig diarrhea in swine farms. Lim et al. (2016) studied the mortality rate of patients in hospitals from Northeast Thailand. They found that 43% of deaths from hospital-acquired infection were caused by infection of multidrug-resistant bacteria. Although several studies have reported evidence of pathogenic bacteria related with antibiotic resistance in Thailand, their studies focused on the analysis of antibiotic resistance from specific bacterial pathogens isolated from clinical, animal, and food samples; whereas information of pathogenic bacteria and ARGs in the environment is very rare.

In this chapter, environmental samples were collected from Phadungkrungkasem Canal and a rice field in Thailand. Phadungkrungkasem Canal is an aquatic reservoir located near household areas and various kinds of markets, such as aquatic animal markets, plant markets, and fresh markets. This canal receives wastewater from around 1,200 houses near the canal and markets along the canal via drainpipes. Some water in this canal is used for feeding crops. On the other hand, Si Nawa is an area of agriculture; most of the land in this area is used for rice growth with a few small households. Thus, these aquatic environmental are related to human activities.

In order to prevent waterborne diseases from these areas, shotgun metagenomic analysis was employed to analyze the pathogenic bacteria and ARB. As demonstrated in Chapter I, database search methods using complete genome or RefSeq as reference sequences had false positives. Thus, in this chapter, the virulence factor genes of bacteria were used to construct the database for analyzing potential pathogens in the samples. Moreover, ARGs were analyzed using database searches containing resistance genes. The results of the pathogen populations and ARG profiles in samples collected from urban and rural areas were compared.

Materials and Methods

Locations and sample collection

A total of fifteen samples were collected from urban and rural areas in Thailand. The urban samples were collected from Phadungkrungkasem Canal, Dusit District, Bangkok, and the rural samples were collected from a rice field at Si Nawa, Mueang Nakhon Nayok, Nakhon Nayok (Figure 8). Ten urban samples were obtained from five different points along the Phadungkrungkasem Canal in Bangkok at two different times. The samples collected on January 19th, 2015 were named as BKK_**X**1, while the samples collected from the same points on November 23rd, 2015 were named as BKK_**X**2. The denoted "**X**" represents the locations where the samples were collected. Two liters of each sample were collected from the water surface at the junction of the canal with the Chao Phraya River (BKK_A1 and BKK_A2), dam (BKK_B1 and BKK_B2), plant and fresh markets (BKK_C1 and BKK_C2), and household areas and street food stalls at night time (BKK_D1, BKK_D2, BKK_E1, and BKK_E2). The bacterial concentration and DNA extraction were performed at the Faculty of Science, Mahidol University, Bangkok, Thailand. DNA samples were transported to the CZC for further analysis

A total of five samples were collected at the rice field in the rural area on November 18th, 2015. One sample was collected from the water surface in the rice field (NYK_A) and the other sample was collected from a marsh (NYK_B) located a 5m-distance from the rice field (NYK_A). Both of these aquatic reservoirs are stand-alone, with no connection to other aquatic environments. For the soil samples, one hundred grams each of three samples were collected from the ground surface, located 5 m (NYK_C), 10 m (NYK_D), and 15 m (NYK_E) distances from the rice field (NYK_A). All samples were transferred to the Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand for bacterial concentration and DNA extraction. DNA samples were then sequenced at the CZC.

Bacterial concentration and DNA extraction

Bacteria in the water samples were concentrated using standard filtration, and DNA was extracted using the PowerWater® DNA Isolation Kit as described in Chapter I. For the soil samples, a total of 0.25 g of the sample was directly used to extract DNA using a PowerSoil® DNA Isolation Kit (Mo Bio Laboratories, Inc., California, USA). The

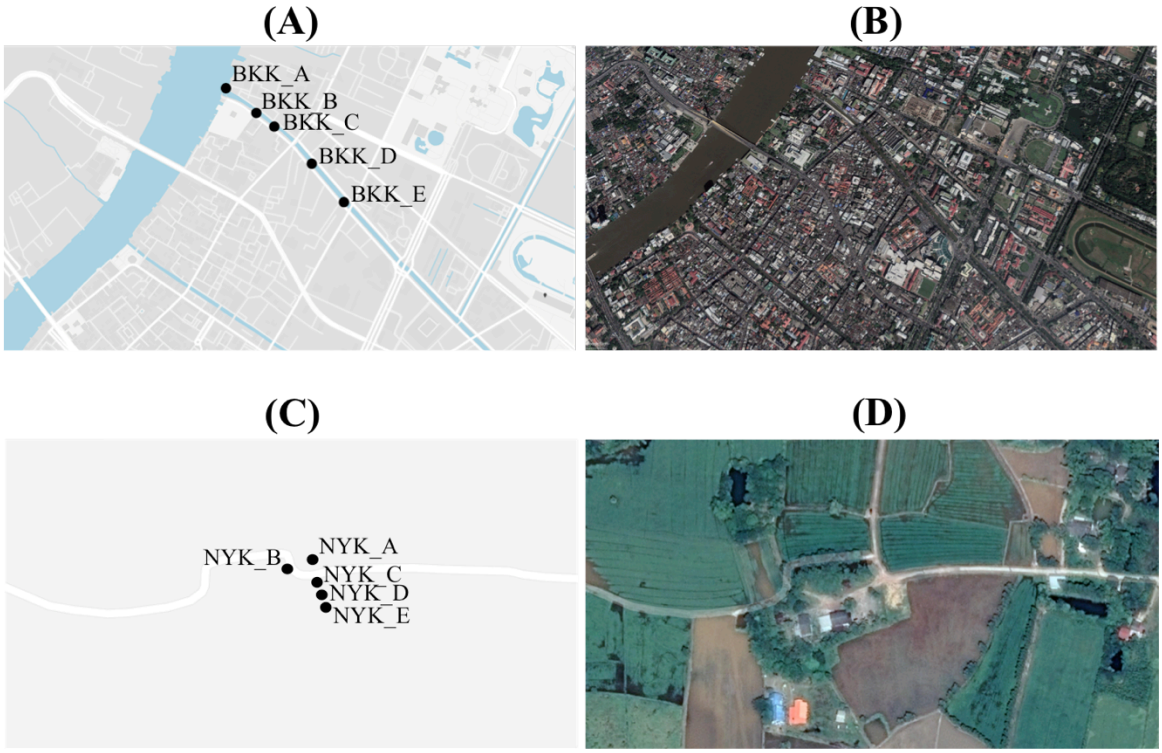

Figure 8. Locations of sample collection from Phadungkrungkasem Canal and a rice field. (A) A map of sample locations for the Phadungkrungkasem Canal, Dusit District, Bangkok from Google Map. (B) Satellite image of Phadungkrungkasem Canal from Google Earth. (C) A map of sample locations for the rice field at Si Nawa, Nakhon Nayok from Google Map. (D) Satellite map of the rice field at Si Nawa, Nakhon Nayok from Google Earth.

DNA concentration was determined using a Qubit™ fluorometer.

Illumina sequencing for metagenomic analysis

The Illumina MiSeq Platform was used for shotgun metagenomic analysis in this study. The methods for DNA library construction and sequencing were indicated in Chapter I. The sequence data of BKK and NYK samples were deposited in the DDBJ with an accession number of DRA006774 and DRA006775, respectively. The barcoding sequences were removed using CLC Genomic Workbench software 8.0 (CLC bio, Tokyo, Japan). The resulting clean reads were used for the downstream analysis.

Taxonomy classification of bacteria

A blastn search and MEGAN were used for the taxonomic classification of bacteria. For each sample, reads were aligned against the NT-NCBI database using blastn with an e-value of 1e-04. The blast results were then analyzed using the naïve LCA algorithm in MEGAN with parameters of min score = 50, max expect = 0.001, top percent = 10, min support percent = 0.001, and min support = 1. The proportions of bacterial species were calculated using the number of reads identified as bacterial species divided by the number of bacterial reads.

Detection of pathogens using blastn against VFDB database

In order to identify shotgun reads for pathogen origin, a blastn search against the VFDB blast database was conducted as described in Chapter I. After reads were identified as pathogens, the abundance of pathogens was calculated using the number of reads classified as pathogenic species divided by the total number of reads. The number of reads identified as pathogenic species in each sample was subjected to PCA using the "prcomp" command in R. In addition, the similarity of detected pathogenic species (beta diversity) in pair-compared samples was determined using the Sørensen–Dice coefficient.

Detection of ARGs using blastn against ARG database

Non-mutated nucleotide sequences of ARGs were downloaded from the Comprehensive Antibiotic Resistance Database (Jia et al., 2017). Each ARG sequence contained information on the ontology number (ARO), antibiotic class, and gene name in its sequence id. Redundant sequences were removed using a custom Python script, and

then the ARG blast database was constructed. Reads were aligned against the ARG database using a blastn search. Positive identification of ARGs was obtained from read hits with ≥ 90% identity over at least 150 bp sections, with a cut off e-value < 1e-05 (McCall and Xagoraraki, 2018). The positive ARG reads were grouped to antibiotic classes and genes. The proportions of antibiotic classes and ARGs in a sample were calculated using the number of reads classified as antibiotic class or ARG divided by the total number of reads. The number of reads identified as antibiotic class and ARGs in each sample were subjected to PCA using the "prcomp" command in R.

Results

NGS reads and read origin

The summary of reads and number of identified reads as bacteria, pathogens, and ARGs are shown in Table 6. An Illumina MiSeq sequencer produced a total of 51,493,800 reads from all samples. Of these reads, approximately 7.3% – 20.3% and 0.03% – 0.14% were identified as having bacterial origins and pathogenic bacteria origins, respectively. A few reads (>0.001% – 0.02%) were identified as having ARG origins. No ARG reads were found in one sample, NYK_B, which was a water sample collected in the rural area.

Bacterial communities in samples

The bacterial communities in each sample were analyzed using a blastn search against the NT-NCBI database and the naïve LCA algorithm in MEGAN. The bacterial proportions in each sample are shown in Figure 9. A total of 3,256 bacterial species were detected from all samples. Approximately 11% – 45% of bacterial-origin samples were assigned as high proportion and ranged in the top 30 species, while 55% – 81% of those were assigned to a minor population and grouped as other bacteria. The remaining proportions were identified as uncultured bacteria and unidentified bacteria at species level.

The populations of bacterial species were different between samples collected from BKK (Phadungkrungkasem Canal, urban area) and NYK (rice field at Si Nawa, Nakhon Nayok, rural area). For instance, *Limnohabitans* sp. was found in high proportions in BKK samples (2.8% – 8.2%) but was low in NYK samples (0.02% – 1.8%). Similarly, the proportions of *Dechloromonas aromatic* and *Arcobacter* sp. in BKK samples (1.6% – 2.4% and 0.05% – 0.8%, respectively) were higher than those in NYK samples (0.04% – 0.1% and 0% – 0.01%, respectively). Although, *Aeromonas hydrophila* was found in both BKK and NYK samples, the proportions of this species in BKK samples (0.1% – 2.0%) were higher than those in NYK samples (0.05% – 0.09%). Conversely, the highest proportion of *Ralstonia solanacearum* was found in one sample, NYK_B (13.3%), while proportions of this bacterium were low in the BKK samples (0.3% – 1.9%). The proportions of *Candidatus* Koribacter versatilis and *Candidatus* Solibacter usitatus in the NYK samples (0.01% – 2.6% and 0.03% – 2.5%, respectively)

Table 6. Summary of NGS reads and number reads identified as bacteria, pathogens, and ARG origin

Location	Samples	Total reads	Number of classified reads (Percentage in total reads)		
			Bacteria[a]	Pathogens[b]	Antibiotics and ARGs[c]
Phadungkrungkasem Canal (urban area)	BKK_A1	3,345,700	397,749 (11.9%)	2,144 (0.06%)	240 (0.007%)
	BKK_B1	3,448,830	380,037 (11.0%)	2,195 (0.06%)	146 (0.004%)
	BKK_C1	3,054,372	619,714 (20.3%)	4,350 (0.14%)	719 (0.024%)
	BKK_D1	3,053,704	510,349 (16.7%%)	2,628 (0.09%)	323 (0.011%)
	BKK_E1	3,366,900	562,355 (16.7%)	3,056 (0.09%)	507 (0.015%)
	BKK_A2	861,078	62,550 (7.3%)	235 (0.03%)	26 (0.003%)
	BKK_B2	3,338,072	360,304 (10.8%)	1,577 (0.05%)	62 (0.002%)
	BKK_C2	3,581,078	463,771 (13.0%)	3,258 (0.09%)	393 (0.011%)
	BKK_D2	3,569,074	351,945 (9.9%)	1,397 (0.04%)	216 (0.006%)
	BKK_E2	3,781,236	536,737 (14.2%)	2,123 (0.06%)	249 (0.007%)
Rice fields (rural area)	NYK_A	3,622,552	267,921 (7.4%)	1,357 (0.04%)	6 (< 0.001%)
	NYK_B	3,683,094	53,304 (1.5%)	172 (>0.04%)	0 (0%)
	NYK_C	4,003,642	236,955 (5.9%)	1,573 (0.04%)	1 (< 0.001%)
	NYK_D	4,977,156	299,374 (6.0%)	2,118 (0.04%)	1 (< 0.001%)
	NYK_E	3,807,312	197,232 (5.2%)	1,359 (0.04%)	4 (< 0.001%)

[a] These reads were classified as bacteria using a blastn search against the NT-NCBI and MEGAN.

[b] These reads were classified as potential pathogens using a blastn search against the VFDB blast database.

[c] These reads were classified as ARGs using a blastn search against the ARG blast database.

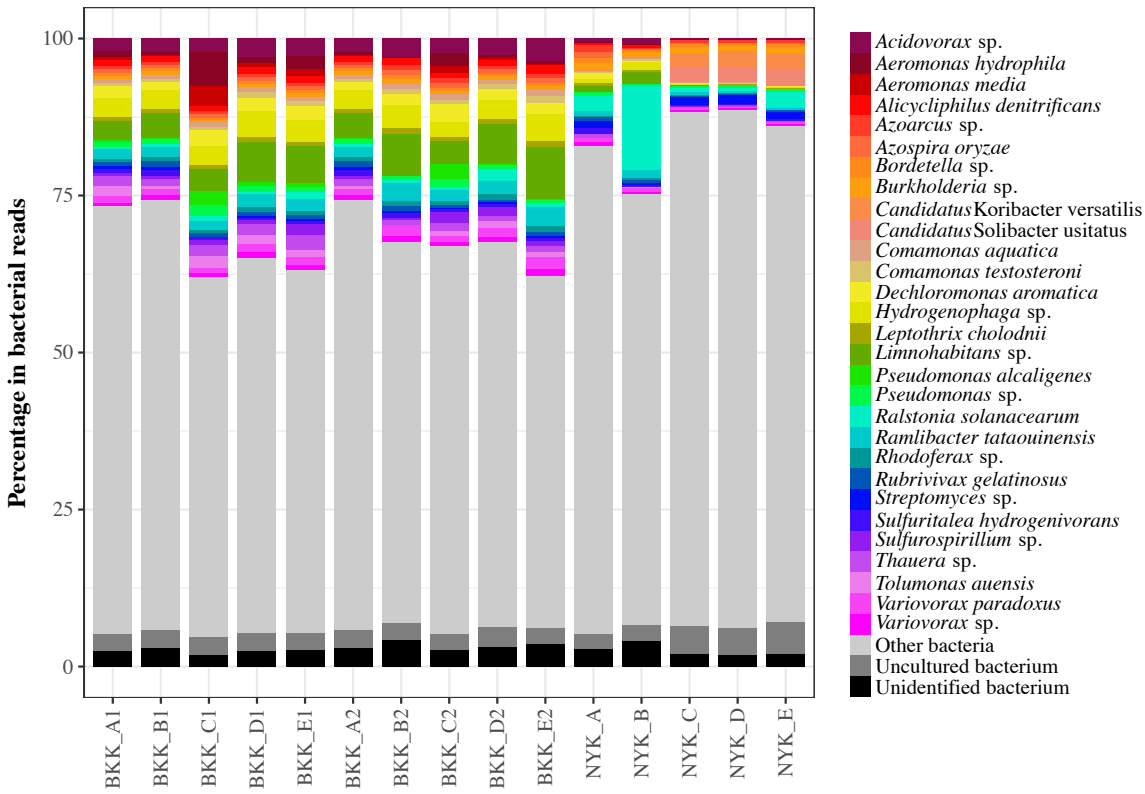

Figure 9. Bacterial communities in samples collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). The urban samples collected from Phadungkrungkasem Canal were named BKK, while rural samples collected from a rice field were named NYK. The colored bars represent abundant species in all samples. Reads from minor species are categorized as other bacteria and represented in light gray. Reads hit to sequences of uncultured bacterium from NCBI-NT database are presented in dark gray. The reads unclassified to bacterial species level are represented in black.

were also higher than those in the BKK samples (<0.01% – 0.01% and 0.01% – 0.03%, respectively).

Abundance of pathogenic bacteria in samples

The read abundance of pathogenic bacteria identified using a blastn search against the VFDB database is shown in Figure 10. A total of 140 pathogenic species were detected from all samples. Of these, 25 pathogenic species including *Aeromonas hydrophila*, *Bordetella bronchiseptica*, *B. pertussis*, *Burkholderia cenocepacia*, *B. pseudomallei*, *B. thailandensis*, *Mycobacterium* sp., *M. abscessus*, *M. avium*, *M. canettii*, *M. gilvum*, *M. intracellulare*, *M. liflandii*, *M. marinum*, *M. smegmatis*, *M. tuberculosis*, *M. vanbaalenii*, *M. yongonense*, *Pseudomonas aeruginosa*, *P. entomophila*, *P. fluorescens*, *P. mendocina*, *P. putida*, *P. syringae*, and *Ralstonia solanacearum* were detected in each sample with a read abundance higher than -4.00 on a $\log_{10}$ scale. Most species (72% of detected species; 102 species) indicated their read abundance lower than -4.00 on a $\log_{10}$ scale in any samples.

Although 25 pathogenic species with high abundance were detected in all samples, some species, for instance *A. hydrophila*, *B. bronchiseptica*, *B. pertussis*, *M. gilvum*, and *P. aeruginosa*, presented higher abundance in the BKK samples than in the NYK samples. On the other hand, some species demonstrated very similar abundance in all samples. These included *B. cenocepaci B. pseudomallei*, and *B. thailandensis*. From these results, it can be assumed that these 25 species from 6 genera were common pathogens in the samples collected from urban and rural areas.

In addition, some pathogenic species, for instance *Escherichia coli, Legionella pneumophila*, and *Mycoplasma penetrans*, presented high abundance in only some samples. These species indicated high abundance only in the BKK samples but not in the NYK samples. In addition, most of the pathogens detected in the BKK samples showed higher abundance than in the NYK samples.

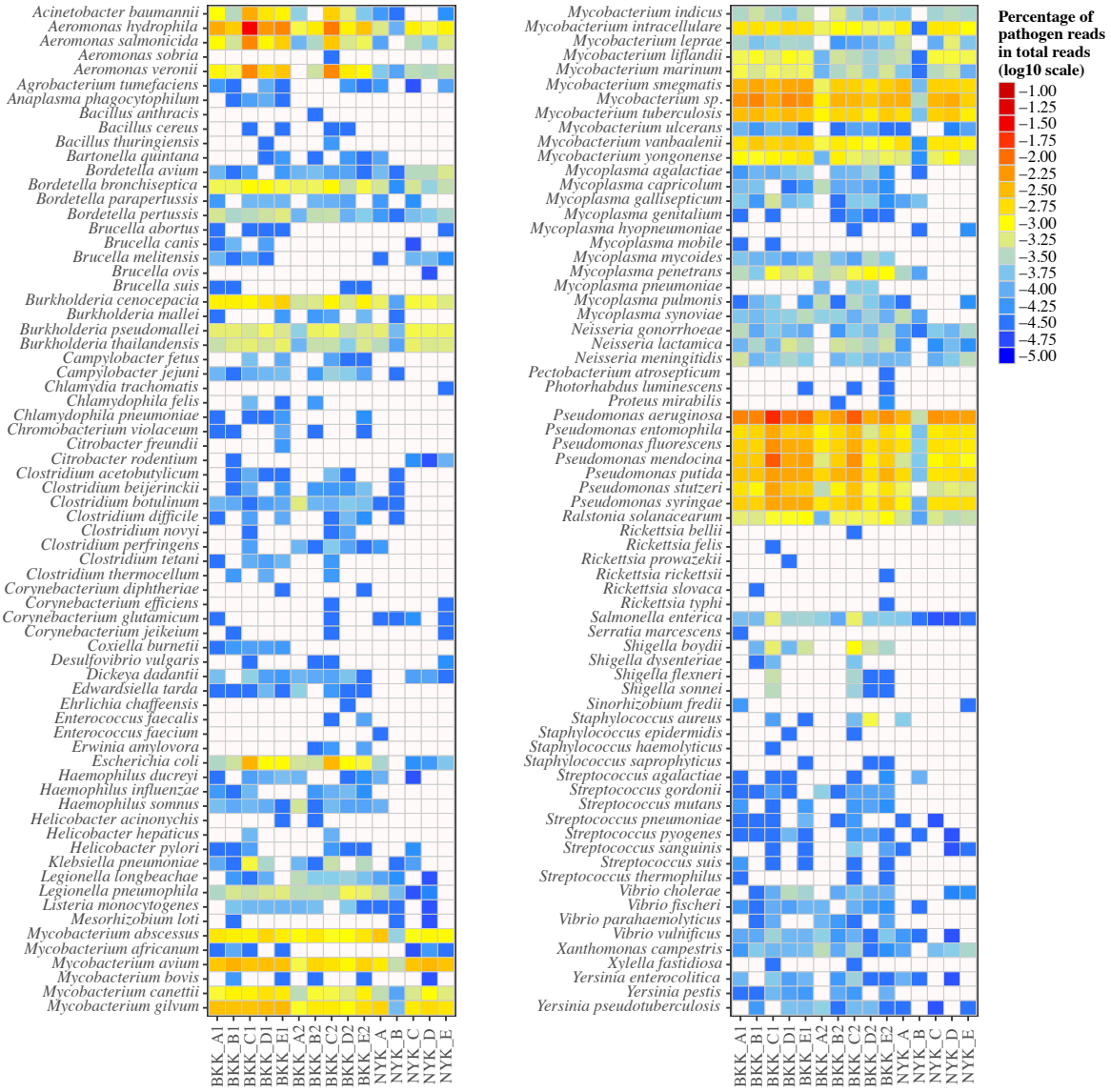

Figure 10. Potential pathogens in samples from Phadungkrungkasem Canal and a rice field detected by a blastn search against the VFDB database. BKK and NYK samples were collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). The proportion of potential pathogens was calculated using the number of reads identified as potential bacteria divided by the total number of reads generated by NGS. Color squares represent the abundance of reads identified as pathogenic bacteria ranging from -1.00 to -5.00 on a $\log_{10}$ scale. White squares represent no read detected for each pathogen.

Four species in genera *Burkholderia* were detected from the samples collected from both areas (Figure 10). Of these four species, *B. cenocepacia*, *B. pseudomalli*, and *B. thailandensis* were detected in all samples with read abundance higher than -4.00 on a $\log_{10}$ scale. On the other hand, *B. mallei* was detected in six samples: BKK_A1, BKK_E1, BKK_B2, BKK_C2, BKK_E2, and NYK_B with low read abundance.

Figure 11A represents the pathogen populations in samples analyzed by PCA. It clearly indicates that the pathogen populations in samples BKK_C1 and BKK_C2 are different from those in the other samples. The pathogen populations in four samples collected in January (BKK_A1, BKK_B1, BKK_D1, and BKK_E1) presented the difference in pathogen population among these samples, whereas pathogen populations in samples collected in November (BKK_A2, BKK_B2, BKK_D2, and BKK_E2) shared similar populations. Moreover, the pathogen populations in the BKK samples collected from January were different from those collected in November, except BKK_A1. On the other hand, the pathogen populations collected from the rural area showed only a small difference in samples NYK_A, NYK_C, NYK_D, and NYK_E. Only the pathogen population in NYK_B indicated a different population among the NYK samples. Although the pathogen population showed difference in some samples, each sample shared a similarity of detected pathogens within 58% – 86%, analyzed by the Sørensen–Dice coefficient (Figure 11B). The highest similarity of detected pathogenic species, at 86%, was found from the comparison of samples BKK_C1 and BKK_C2. In the rural area, the similarity of detected species in samples NYK_A, NYK_C, NYK_D, and NYK_E was within a range of 70% – 82%.

Abundance of ARGs classified as antibiotic classes and gene level

The read abundance of ARGs in each sample analyzed by the blastn search against the ARG database is shown in Figure 12. A total of 2,893 reads were identified as ARG origin from all samples, and those genes could be categorized into 16 antibiotic classes (Figure 12A). Of these 16 antibiotic classes, aminoglycoside, beta-lactam, and multidrug resistance (MDR) indicted high abundance of more than -4.00 on a $\log_{10}$ scale in the BKK samples, but were very low or disappeared in the NYK samples (less than -4.00 on a $\log_{10}$ scale). In the BKK samples, each sample contained at least 7 classes of

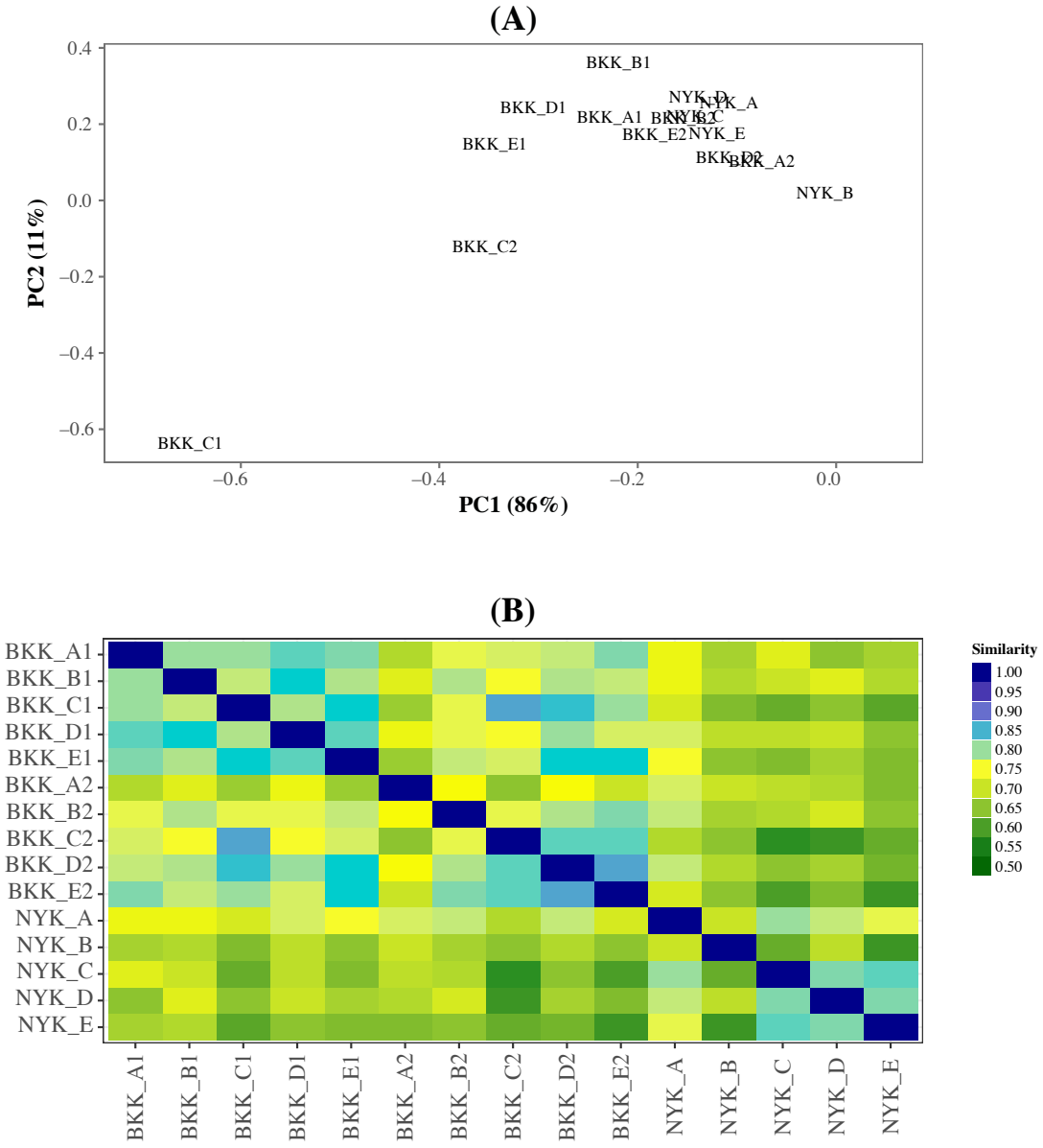

Figure 11. Comparison analysis of pathogen populations in each sample collected from Phadungkrungkasem Canal and a rice field. BKK and NYK samples were collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). (A) Population of pathogens analyzed by principle component analysis. (B) Beta diversity of pathogens in samples analyzed by the Sørensen–Dice coefficient.

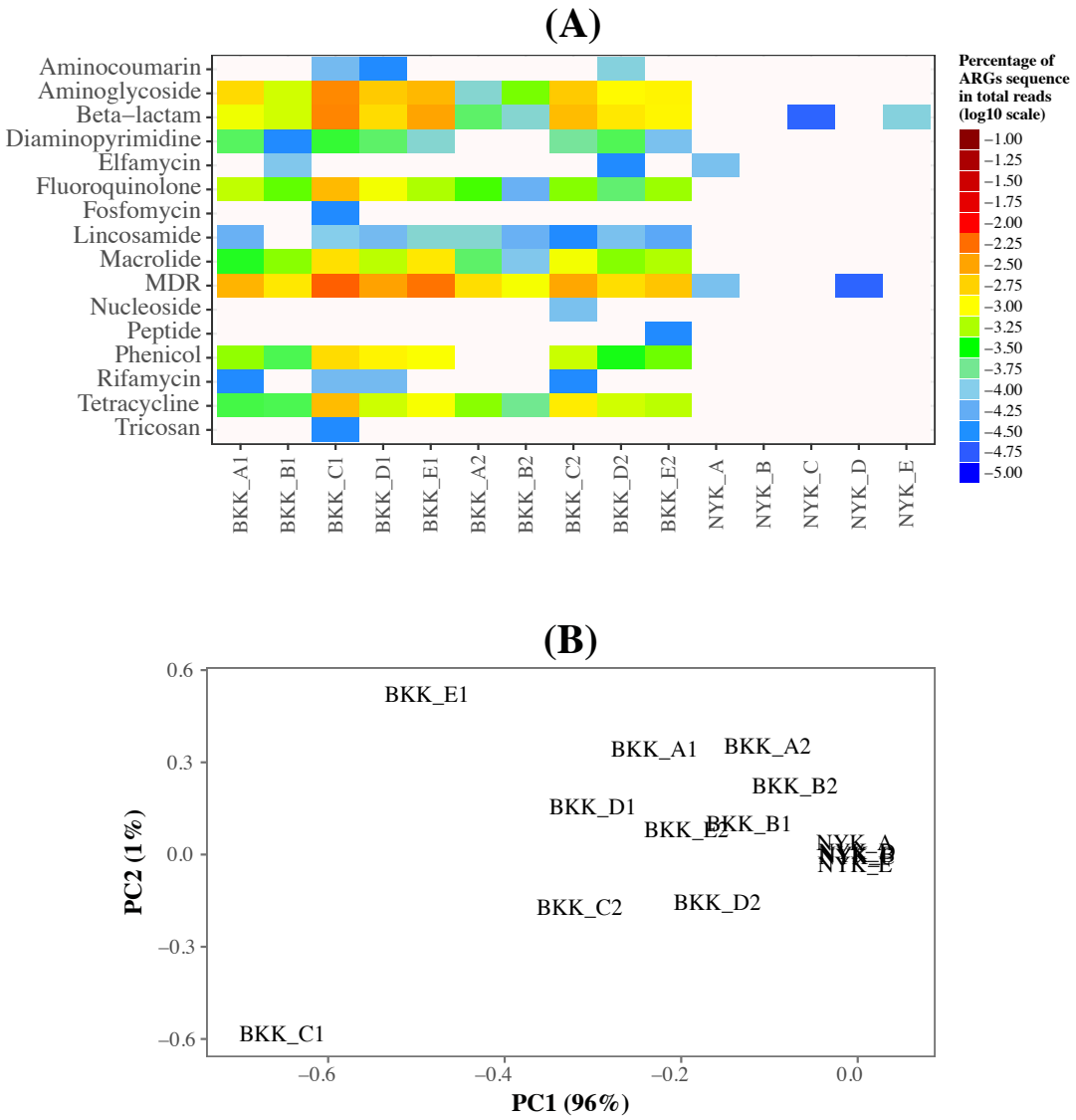

Figure 12. The abundance of ARGs classified by drug class in samples collected from Phadungkrungkasem Canal and a rice field. BKK and NYK samples were collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). The ARGs were identified using a blastn search against a CARD database. (A) Abundance of ARGs classified by antibiotic class. (B) PCA of ARGs classified by drug class. BKK samples were collected from an urban area, while NYK samples were collected from a rural area.

antibiotic resistance, and MDR was in the highest abundance of antibiotic classes in each sample. Moreover, ARGs showing resistance to nucleoside, peptide, and tricosan were found in the samples, BKK_C2, BKK_E2, and BKK_C1, respectively. For the NYK samples, only 12 ARG reads were detected, and those reads were categorized as beta-lactam, elfamycin, and MDR (Table 1). PCA indicated that the antibiotic classes in the samples BKK_C1 and BKK_E1 were different from the rest of the samples and that all the samples from NYK were clustered together (Figure 12B).

At the gene level, a total of 212 ARGs were grouped from 2,893 reads of ARG origin (Figure 13). It was clearly shown that the ARG abundance in the BKK samples were higher than those in the NYK samples. In the BKK samples, only four genes were represented in abundance higher than -3.00 on a $\log_{10}$ scale in any samples. Those genes were *QnrS2* (0.00006% – 0.0016% of total reads), *sul1* (0.00032% – 0.00184% of total reads), *sul2* (0.00034% – 0.00124% of total reads), and *cmlA* (0.00014% – 0.00134% of total reads), which referred to the resistance to fluoroquinolone, MDR, MDR, and phenicol, respectively. In the NYK samples, a total of four ARGs were detected. They were *TEM-171*, *TEM-4*, *MuxB*, and *rpoB2*, which correspond to the resistance to beta-lactam, beta-lactam, MDR, and MDR, respectively.

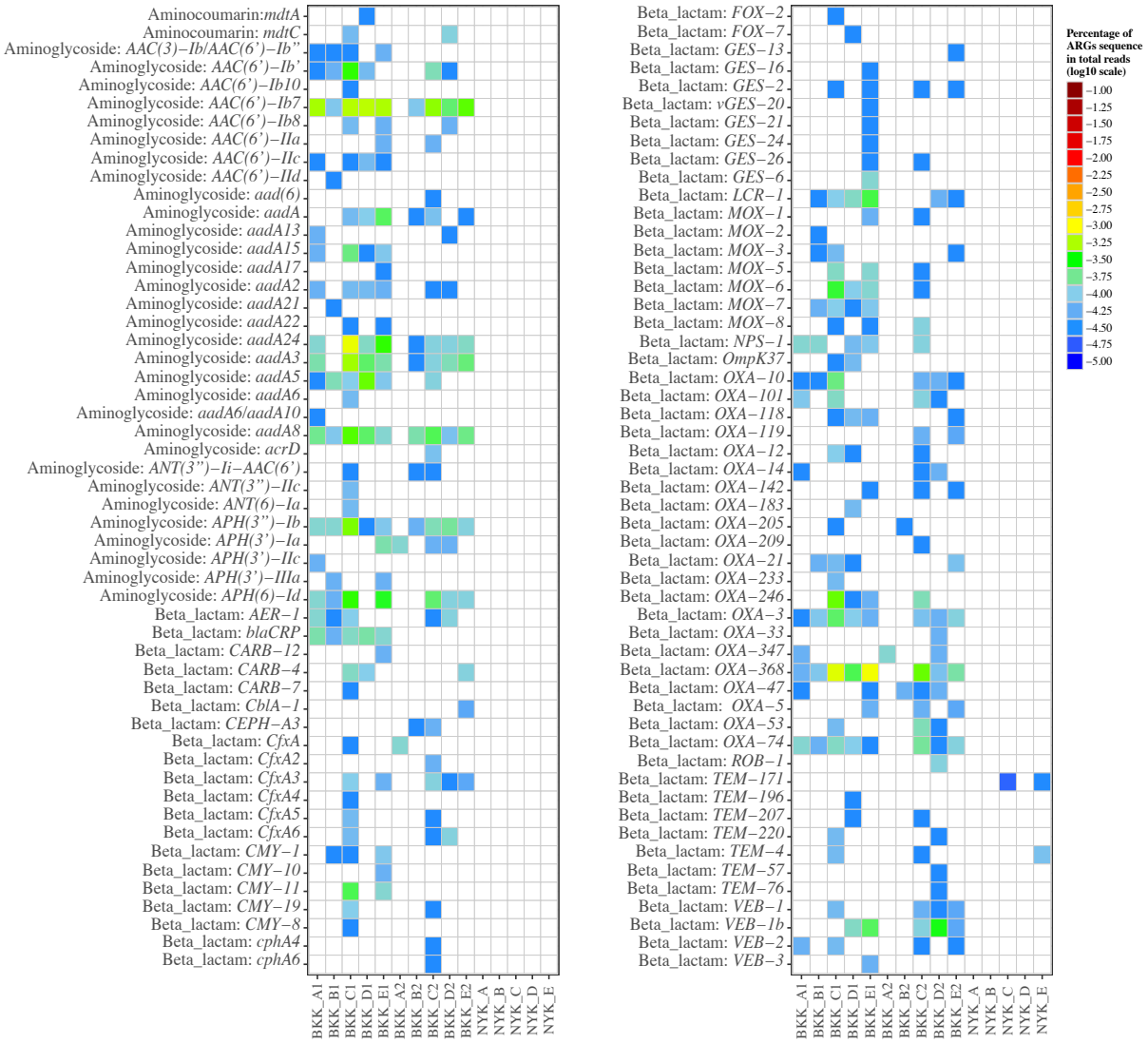

Figure 13. The abundance of ARGs classified by gene levels in samples collected from Phadungkrungkasem Canal and a rice field. BKK and NYK samples were collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). The ARGs were identified using a blastn search against the ARG database. The ARGs were grouped by the same gene and represented by colored squares. White squares represent no read of ARGs is detected.

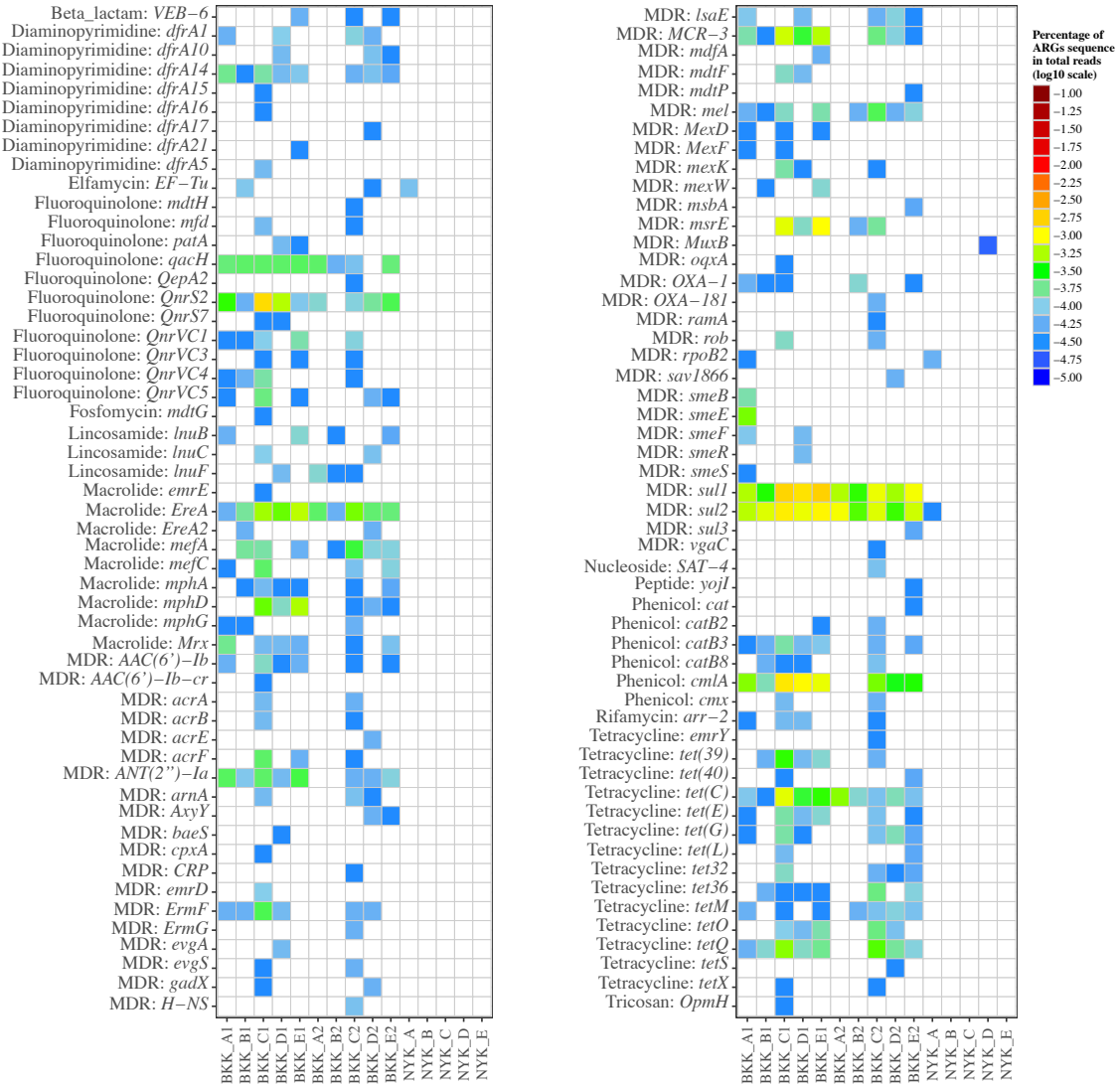

Figure 13. (continued) The abundance of ARGs classified by gene levels in samples collected from Phadungkrungkasem Canal and a rice field. BKK and NYK samples were collected from Phadungkrungkasem Canal (urban area) and a rice field (rural area). The ARGs were identified using a blastn search against the ARG database. The ARGs were grouped by the same gene and represented by colored squares. White squares represent no read of ARGs is detected.

Discussion

In this study, a metagenomic analysis was conducted to comprehensively investigate pathogenic bacteria and ARGs in samples collected from Phadungkrungkasem Canal (BKK sample) and a rice field (NYK sample), located in urban and rural areas in Thailand, respectively. Twenty-five pathogenic species from genera *Aeromonas*, *Bordetella*, *Burkholderia*, *Mycobacterium*, *Pseudomonas*, and *Ralstonia* were detected in all samples with high abundance. Some pathogenic species such as *E. coli, L. pneumophila*, and *M. penetrans* indicated high abundance in all samples collected from the urban area. Moreover, a total of 212 ARGs related to the resistance to 16 antibiotic classes were detected from all samples. However, the abundance of ARGs in the BKK samples was higher than those in the NYK samples. From these results, it suggests that the samples collected from the urban area contained more divergent pathogens and ARGs than those in the rural area.

The bacterial communities analyzed by blastn against the NCBI-NT database and naïve LCA in MEGAN demonstrated the difference in bacterial proportions in the BKK and NYK samples (Figure 9). The patterns of bacterial species in the top 30 species were the same in all samples. However, some species indicated high proportions in all samples. *A. hydrophila* indicated a high percentage of 5.3% in one sample BKK_C1, while the proportion of this species in other BKK samples was within a range of 0.4% – 2.0%. Similarly, *Ralstonia solanacearum* showed the highest abundance of 13.3% in sample NYK_B, but had a low percentage (0.5% – 2.4%) in other samples from the rural area. These results may suggest that a high proportion of some bacterial species may be related to the area of sample collection.

Shotgun metagenomic analysis detected *B. mallei* in six samples: BKK_A1, BKK_E1, BKK_B2, BKK_C2, BKK_E2, and NYK_B (Figure 10). Generally, *B. mallei* is an etiological agent of glanders which is an infectious disease of equines (Galyov et al., 2010). Horses are the primary host of this bacterium and responsible for transmission to healthy animals and humans (Whitlock et al., 2007). Interestingly, there is no report for the detection of *B. mallei* in Thailand. However, in this study, genetic material of *B. mallei* was detected in six samples by shotgun metagenomic analysis. This indicates an advantage of metagenomics in detecting unexpected pathogens in environmental samples.

PCA indicated that the pathogen populations in samples BKK_C1 and BKK_C2 were different from other samples (Figure 11A). In fact, these two samples were collected at the same point at different times. The numbers of detected pathogens in these samples were 98 and 103 species for BKK_C1 and BKK_C2, respectively, which are higher than those in other samples (53 – 96 species) collected from the urban area. This result indicates that this location affected the number of pathogenic species. In addition, sample BKK_C1 presented similar results of PCA as analyzed by pathogens (Figure 11A) and ARGs (Figure 12B). Generally, ARGs are related to the bacteria in the sample because ARGs are genetic material presented on mobile genetic elements, e.g. plasmids, transposons, and integrons in bacteria. With this result it can be assumed that the sample BKK_C1 may serve as an important reservoir for the circulation of antibiotic resistant bacteria in this canal. Interestingly, the location of BKK_C1 is near fresh and plant markets. There is high potential that some pathogens or ARB may travel from these places to the canal through the water drains. Thus, an investigation of the pathogen population inside the markets should be analyzed to verify the original source of pathogens or ARB. However, this analysis could not be performed due to economic reasons. In an alternative investigation, a time series collection of water can be performed to monitor pathogens and ARB for the prevention of infectious disease in this area.

The PCA demonstrated that the ARG compositions were different between the BKK and NYK samples collected from urban and rural areas (Figure 12B). When considering the locations of sampling, the NYK samples were collected from the water surface of a rice field (NYK_A) and a marsh (NYK_B). These were stand-alone aquatic reservoirs which have no connection to other aquatic reservoirs or drainpipes from residents. In contrast, BKK samples were collected from water in a canal that receives wastewater from drainpipes along the canal. Thus, it has a high possibility for bacteria carrying ARGs travelling from these locations to the water in the canal through drainpipes, especially at the locations BKK_C, BKK_D, and BKK_E, which showed high abundance of *sul1*, *sul2* and *cmlA* (Figure 13). The results indicate that the connection of aquatic reservoirs with wastewater drainpipes plays an important role for the high abundance of ARGs in urban areas.

A few reads were identified as ARGs in NYK_A though no read was identified as ARGs in NYK_B (Table 6); these results indicate that these samples contain a low abundance of ARGs, which is different from the water samples collected from urban areas. One possible reason for the low abundance of ARGs in samples NYK_A and NYK_B is that the area nearby these locations is affected by the ARG populations. To clarify the question, three soil samples (NYK_C, NYK_D, and NYK_E) were collected, and ARGs were analyzed using a shotgun metagenomic approach. However, the PCA results indicated that the composition of ARGs in all NYK samples collected from water and soil were similar (Figure 12B). These results suggest that there is no difference in the populations of bacteria carrying ARGs in water and soil in this area.

Metagenomic analysis is a powerful technique and helpful for comprehensive analysis of ARGs in a community. Although metagenomic analysis can analyze the overall number of ARGs in an environmental sample, this approach cannot specifically clarify which bacterial species is an original source of ARGs. Basically, metagenomic analysis directly analyzes all genomic DNA extracted from a given sample. Thus, it provides information from the overall DNA in a community (Thomas et al., 2012). Researchers should keep in mind that this is a classical limitation of metagenomic analysis.

Summary

The biggest obstacle to understanding the circulation of pathogenic bacteria and ARGs is a lack of comprehensive information of all the pathogens and ARGs in the environment. In this study, shotgun metagenomic analysis was conducted for a comprehensive analysis of pathogens and ARGs in environmental samples collected from Phadungkrungkasem Canal and a rice field. The results revealed that 25 species with high abundance were common pathogens in samples. Some species, such as *E. coli, L. pneumophila*, and *M. penetrans,* were found in high abundance in samples collected from urban areas; indicating that these species were associated with the location of sample collections. In addition, high numbers of ARGs were found in the samples collected from Phadungkrungkasem Canal. These results indicated that Phadungkrungkasem Canal was more likely to be contaminated with pathogenic and antibiotic resistant bacteria than the rice field.

Conclusion

Metagenomic analysis has become a powerful tool for analyzing genetic materials in environmental samples. By this analysis, genomic fragments of thousands of bacteria in a sample are analyzed using NGS technologies. This tool is helpful to comprehensively investigate the bacterial communities and functional genes in a given sample.

In Chapter I, different database search methods were compared to detect *L. pneumophila*. The results showed that the current search method using complete genome or RefSeq as reference sequences led to false positives. The search method using the *mip* gene, which is a species-specific marker gene for *L. pneumophila*, showed false negatives. The use of a species-specific long gene, e.g. *katB,* provided the best agreement of detection with the results of *L. pneumophila*-specific nested PCR. This result suggests that a database search method using a long gene specifically associated with a bacterial species has better potential diagnosis than using the current search methods.

In Chapter II, shotgun metagenomic analysis was conducted to comprehensively analyze the pathogens and ARGs in samples collected from urban and rural areas in Thailand. By using the database search method with virulence factor genes for analyzing pathogens, 25 pathogenic species with high abundance were detected in samples collected from both areas; indicating these species were common pathogens in these samples. In addition, some species, such as *E. coli, L. pneumophila*, and *M. penetrans,* were detected in high abundance in samples collected from some urban locations; indicating that these species may be shared from other nearby locations. Moreover, a high abundance of ARGs were also detected in samples collected from Phadungkrungkasem Canal, while a few reads were detected from samples collected from a rice field. Our results revealed that the Phadungkrungkasem Canal might serve as a reservoir for the circulation of pathogens and antibiotic resistant bacteria in the environment.

Through the metagenomic analysis of environmental samples, the ability of current database search methods using the complete genome as reference sequences has a problem in the detection of specific-species, whereas a long gene for a specific marker gene showed better diagnostic detection. In order to analyze pathogens in a given sample

from the environment, virulence factor genes can be used as the database search method. From the results, this idea for a database search method will be helpful for the analysis of pathogens in environmental samples using a metagenomic approach.

Acknowledgements

First of all, I am exceptionally grateful to my supervisor, Professor Kimihito Ito (Division of Bioinformatics, CZC, Hokkaido University, Sapporo, Japan) who is a great teacher, for his guidance, active discussions, and helping me to conduct this research. I am grateful to Assistant Professor Ryosuke Omori (Division of Bioinformatics, CZC) for his technical assistance, active discussion, and his excellent support. I am also grateful to the advisors who gave academic comments, and advice for my research: Professor Chihiro Sugimoto (Division of Collaboration and Education, CZC), Associate Professor Chie Nakajima (Division of Bioresources, CZC), Associate Professor Kanako Koyanagi (Division of Bioengineering and Bioinformatics, Graduate School of Information Science and Technology, Hokkaido University), and Associate Professor Ryo Nakao (Laboratory of Parasitology, Department of Disease Control, Faculty of Veterinary Medicine, Hokkaido University).

I would like to express my gratitude to my thesis committees: Professor Yasuhiko Suzuki (Division of Bioresources, CZC), Professor Kazuhiko Ohashi (Laboratory of Infectious Disease, Department of Disease Control, Faculty of Veterinary Medicine, Hokkaido University), Associate Professor Chie Nakajima, and Associate Professor Ryo Nakao for their kind discussions and suggestions for my thesis.

I would like to express my gratitude to Professor Emeritus Orasa Sutheinkul (Faculty of Public Health, Thammasat University, Thailand), Associate Professor Apinya Assavanig (Department of Biotechnology, Faculty of Science, Mahidol University), and Professor Sunee Krobseatsiri (Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University) for their suggestions, discussions, and providing laboratory space and facilities for my experiments when I collected samples in Thailand.

I am thankful to Dr. Heidi Lynn Tessmer who helped correct my English grammar for both my manuscript and thesis, and provided suggestions for writing computer programs. I would like to thank to all members of the Division of Bioinformatics: Dr. Gabriel Gonzalez, Dr. Kiyeon Kim, Dr. Nipawit Karnbunchob, Dr. Heidi Lynn Tessmer, Ms. Wessam Mohamed Ahmed, Mr. Teiji Murakami, and Ms. Sayaka Iida for their

sincere friendship, support, and valuable suggestions, and for filling my five years in Japan with unforgettable experiences. Also, I would like to thank all members of the CZC, the Thai students, and Thai people in Sapporo for their kind help and support during my stay in Japan.

Finally, I would like to express my deepest appreciation to my family for their moral support, encouragement, and understanding throughout my studies in Japan.

Abstract

Waterborne diseases caused by pathogenic bacteria are a concern for public health worldwide. The lack of information of overall pathogens in aquatic environments makes it difficult to establish a strategy for the prevention of pathogen infections from water. Recently, metagenomic analysis combined with the ability of next generation sequencing technology can be used for the direct detection of pathogenic bacteria in water samples. One important process in metagenomic studies is the taxonomy classification of bacteria. Several database search methods have been used for taxonomy classification of bacteria using bioinformatics algorithms, such as blastn searches. In this thesis, the different database search methods for taxonomy classification of bacteria were compared for their diagnostic ability in detecting pathogens. By focusing on a specific bacterium species, *L. pneumophila*, the detection results of each search method were compared with the results from *L. pneumophila*-specific nested PCR. The results demonstrate that database search methods using complete genomes as reference sequences showed numerous false positives and had a potential problem with specificity. On the other hand, database search methods using long specific marker genes as reference sequences provided the best agreement of detection with *L. pneumophila*-specific PCR. Subsequently, database search methods using specific marker genes were selected to analyze pathogenic bacteria in environmental samples collected from urban and rural areas in Thailand. The results indicated that pathogenic species from the genera *Aeromonas*, *Bordetella*, *Burkholderia*, *Mycobacterium*, *Pseudomonas*, and *Ralstonia* were common in these samples; they were detected in all samples with high abundance. In addition, the profiles of antibiotic resistance genes (ARGs) in each sample group were analyzed using a blastn search against the ARG database which contains reference sequences of ARGs. The results showed that the abundance of reads identified as ARGs in samples collected from urban areas were higher than those in samples collected from rural areas; indicating that urban environments may serve as reservoirs for the circulation of antibiotic resistant bacteria in the environment. In conclusion, metagenomic analysis could be a useful tool to establish a strategy for the prevention of waterborne diseases.

References

Aarestrup, F.M. (2005). Veterinary drug usage and antimicrobial resistance in bacteria of animal origin. *Basic Clin Pharmacol Toxicol*. 96, 271-281.

Agerso, Y., Petersen, A. (2007). The tetracycline resistance determinant Tet39 and the sulphonamide resistance gene *sulII* are common among resistant *Acinetobacter* spp. isolated from integrated fish farms in Thailand. *J Antimicrob Chemother*. 59, 23-27.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol*. 215, 403-410.

Angiuoli, S.V., White, J.R., Matalka, M., White, O., Fricke, W.F. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS One*. 6, e26624.

Angkititrakul, S., Chomvarin, C., Chaita, T., Kanistanon, K., Waethewutajarn, S. (2005). Epidemiology of antimicrobial resistance in *Salmonella* isolated from pork, chicken meat and humans in Thailand. *Southeast Asian J Trop Med Public Health*. 36, 1510-1515.

Arias, C., Sala, M.R., Dominguez, A., Bartolome, R., Benavente, A., Veciana, P., et al. (2006). Waterborne epidemic outbreak of *Shigella sonnei* gastroenteritis in Santa Maria de Palautordera, Catalonia, Spain. *Epidemiol Infect*. 134, 598-604.

Atlas, R.M., Williams, J.F., Huntington, M.K. (1995). *Legionella* contamination of dental-unit waters. *Appl Environ Microbiol*. 61, 1208-1213.

Ballings, M., Van Den Poel, D. (2013). Threshold independent performance mesures for probabilistic classifiers. Available at: *http://cran.r-project.org/web/packages/AUC/AUC.pdf*

Benkel, D.H., Mcclure, E.M., Woolard, D., Rullan, J.V., Miller, G.B., Jr., Jenkins, S.R., et al. (2000). Outbreak of Legionnaires' disease associated with a display whirlpool spa. *Int J Epidemiol*. 29, 1092-1098.

Boldur, I., Cohen, A., Tamarin-Landau, R., Sompolinsky, D. (1987). Isolation of *Legionella pneumophila* from calves and the prevalence of antibodies in cattle, sheep, horses, antelopes, buffaloes and rabbits. *Vet Microbiol*. 13, 313-320.

Boonyasiri, A., Tangkoskul, T., Seenama, C., Saiyarin, J., Tiengrim, S., Thamlikitkul, V. (2014). Prevalence of antibiotic resistant bacteria in healthy adults, foods, food

animals, and the environment in selected areas in Thailand. *Pathog Glob Health*. 108, 235-245.

Breitbart, M., Hoare, A., Nitti, A., Siefert, J., Haynes, M., Dinsdale, E., et al. (2009). Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environ Microbiol*. 11, 16-34.

Buchbinder, S., Trebesius, K., Heesemann, J. (2002). Evaluation of detection of *Legionella* spp. in water samples by fluorescence in situ hybridization, PCR amplification and bacterial culture. *Int J Med Microbiol*. 292, 241-245.

Cai, L., Zhang, T. (2013). Detecting human bacterial pathogens in wastewater treatment plants by a high-throughput shotgun sequencing technique. *Environ Sci Technol*. 47, 5433-5441.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. 6, 1621-1624.

Catalan, V., Moreno, C., Dasi, M.A., Munoz, C., Apraiz, D. (1994). Nested polymerase chain reaction for detection of *Legionella pneumophila* in water. *Res Microbiol*. 145, 603-610.

Centers for Disease Control Prevention (2011). Legionellosis --- United States, 2000-2009. *MMWR Morb Mortal Wkly Rep*. 60, 1083-1086.

Chakravorty, S., Helb, D., Burday, M., Connell, N., Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*. 69, 330-339.

Changkaew, K., Utrarachkij, F., Siripanichgon, K., Nakajima, C., Suthienkul, O., Suzuki, Y. (2014). Characterization of antibiotic resistance in *Escherichia coli* isolated from shrimps and their environment. *J Food Prot*. 77, 1394-1401.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., et al. (2013). Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep*. 3, 3550.

Chen, B., Yuan, K., Chen, X., Yang, Y., Zhang, T., Wang, Y., et al. (2016). Metagenomic analysis revealing antibiotic resistance genes (ARGs) and their genetic compartments in the Tibetan environment. *Environ Sci Technol*. 50, 6670-6679.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., et al. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 33, D325-328.

Cho, I., Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 13, 260-270.

Cianciotto, N.P., Eisenstein, B.I., Mody, C.H., Toews, G.B., Engleberg, N.C. (1989). A *Legionella pneumophila* gene encoding a species-specific surface protein potentiates initiation of intracellular infection. *Infect Immun*. 57, 1255-1262.

Clooney, A.G., Fouhy, F., Sleator, R.D., A, O.D., Stanton, C., Cotter, P.D., et al. (2016). Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. *PLoS One*. 11, e0148028.

Cloud, J.L., Carroll, K.C., Pixton, P., Erali, M., Hillyard, D.R. (2000). Detection of *Legionella* species in respiratory specimens using PCR with sequencing confirmation. *J Clin Microbiol*. 38, 1709-1712.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., Mcgarrell, D.M., et al. (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*. 33, D294-296.

Correia, A.M., Ferreira, J.S., Borges, V., Nunes, A., Gomes, B., Capucho, R., et al. (2016). Probable person-to-person transmission of Legionnaires' disease. *N Engl J Med*. 374, 497-498.

Daniel, R. (2005). The metagenomics of soil. *Nat Rev Microbiol*. 3, 470-478.

Delafont, V., Brouke, A., Bouchon, D., Moulin, L., Hechard, Y. (2013). Microbiome of free-living amoebae isolated from drinking water. *Water Res*. 47, 6958-6965.

Diaz-Torres, M.L., Villedieu, A., Hunt, N., Mcnab, R., Spratt, D.A., Allan, E., et al. (2006). Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiol Lett*. 258, 257-262.

Durso, L.M., Harhay, G.P., Bono, J.L., Smith, T.P. (2011). Virulence-associated and antibiotic resistance genes of microbial populations in cattle feces analyzed using a metagenomic approach. *J Microbiol Methods*. 84, 278-282.

Durso, L.M., Miller, D.N., Wienhold, B.J. (2012). Distribution and quantification of antibiotic resistant genes and bacteria across agricultural and non-agricultural metagenomes. *PLoS One*. 7, e48325.

Ercolini, D. (2013). High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology. *Appl Environ Microbiol*. 79, 3148-3155.

Fabbi, M., Pastoris, M.C., Scanziani, E., Magnino, S., Di Matteo, L. (1998). Epidemiological and environmental investigations of *Legionella pneumophila* infection in cattle and case report of fatal pneumonia in a calf. *J Clin Microbiol*. 36, 1942-1947.

Federhen, S. (2011). The NCBI Taxonomy database. *Nucleic Acids Res*. 40, D136-143.

Fliermans, C.B., Cherry, W.B., Orrison, L.H., Smith, S.J., Tison, D.L., Pope, D.H. (1981). Ecological distribution of *Legionella pneumophila*. *Appl Environ Microbiol*. 41, 9-16.

Galyov, E.E., Brett, P.J., Deshazer, D. (2010). Molecular insights into Burkholderia pseudomallei and *Burkholderia mallei* pathogenesis. *Annu Rev Microbiol*. 64, 495-517.

Garrido-Cardenas, J.A., Manzano-Agugliaro, F. (2017). The metagenomics worldwide research. *Curr Genet*. 63, 819-829.

Girones, R., Ferrus, M.A., Alonso, J.L., Rodriguez-Manzano, J., Calgua, B., Correa Ade, A., et al. (2010). Molecular detection of pathogens in water--the pros and cons of molecular techniques. *Water Res*. 44, 4325-4339.

Gomez-Alvarez, V., Revetta, R.P., Santo Domingo, J.W. (2012). Metagenomic analyses of drinking water receiving different disinfection treatments. *Appl Environ Microbiol*. 78, 6095-6102.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 21, 494-504.

Hlavsa, M.C., Roberts, V.A., Kahler, A.M., Hilborn, E.D., Wade, T.J., Backer, L.C., et al. (2014). Recreational water-associated disease outbreaks--United States, 2009-2010. *MMWR Morb Mortal Wkly Rep*. 63, 6-10.

Hoge, C.W., Gambel, J.M., Srijan, A., Pitarangsi, C., Echeverria, P. (1998). Trends in antibiotic resistance among diarrheal pathogens isolated in Thailand over 15 years. *Clin Infect Dis*. 26, 341-345.

Hu, Y., Yang, X., Qin, J., Lu, N., Cheng, G., Wu, N., et al. (2013). Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun*. 4, 2151.

Huang, K., Zhang, X.X., Shi, P., Wu, B., Ren, H. (2014). A comprehensive insight into bacterial virulence in drinking water using 454 pyrosequencing and Illumina high-throughput sequencing. *Ecotoxicol Environ Saf*. 109, 15-21.

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Res*. 17, 377-386.

Huson, D.H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN community edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*. 12, e1004957.

Ibarbalz, F.M., Orellana, E., Figuerola, E.L., Erijman, L. (2016). Shotgun metagenomic profiles have a high capacity to discriminate samples of activated sludge according to wastewater type. *Appl Environ Microbiol*. 82, 5186-5196.

Ibekwe, A.M., Leddy, M., Murinda, S.E. (2013). Potential human pathogenic bacteria in a mixed urban watershed as revealed by pyrosequencing. *PLoS One*. 8, e79490.

Janda, J.M., Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol*. 45, 2761-2764.

Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 45, D566-D573.

Jones, S.E., Newton, R.J., Mcmahon, K.D. (2009). Evidence for structuring of bacterial community composition by organic carbon source in temperate lakes. *Environ Microbiol*. 11, 2463-2472.

Kaakoush, N.O., Castano-Rodriguez, N., Mitchell, H.M., Man, S.M. (2015). Global Epidemiology of *Campylobacter* Infection. *Clin Microbiol Rev*. 28, 687-720.

Kim, D., Hong, S., Kim, Y.T., Ryu, S., Kim, H.B., Lee, J.H. (2018). Metagenomic approach to identifying foodborne pathogens on Chinese cabbage. *J Microbiol Biotechnol*. 28, 227-235.

Ko, K.S., Hong, S.K., Lee, H.K., Park, M.Y., Kook, Y.H. (2003). Molecular evolution of the *dotA* gene in *Legionella pneumophila*. *J Bacteriol*. 185, 6269-6277.

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet*. 13, 47-58.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., Mcwilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*. 23, 2947-2948.

Lecuit, M., Eloit, M. (2014). The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Front Cell Infect Microbiol*. 4, 25.

Leigh, J.A., Dodsworth, J.A. (2007). Nitrogen regulation in bacteria and archaea. *Annu Rev Microbiol*. 61, 349-377.

Leonard, S.R., Mammel, M.K., Lacher, D.W., Elkins, C.A. (2015). Application of metagenomic sequencing to food safety: detection of Shiga toxin-producing *Escherichia coli* on fresh bagged spinach. *Appl Environ Microbiol*. 81, 8183-8191.

Lim, C., Takahashi, E., Hongsuwan, M., Wuthiekanun, V., Thamlikitkul, V., Hinjoy, S., et al. (2016). Epidemiology and burden of multidrug-resistant bacterial infection in a developing country. 6, e18082.

Lindgreen, S., Adair, K.L., Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep*. 6, 19233.

Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z., Quick, J., et al. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA*. 309, 1502-1510.

Looft, T., Johnson, T.A., Allen, H.K., Bayles, D.O., Alt, D.P., Stedtfeld, R.D., et al. (2012). In-feed antibiotic effects on the swine intestinal microbiome. *Proc Natl Acad Sci USA*. 109, 1691-1696.

Lu, X., Zhang, X.X., Wang, Z., Huang, K., Wang, Y., Liang, W., et al. (2015). Bacterial pathogens and community composition in advanced sewage treatment systems revealed by metagenomics analysis based on high-throughput sequencing. *PLoS One*. 10, e0125549.

Mahbubani, M.H., Bej, A.K., Miller, R., Haff, L., Dicesare, J., Atlas, R.M. (1990). Detection of *Legionella* with polymerase chain reaction and gene probe methods. *Mol Cell Probes*. 4, 175-187.

Mccall, C., Xagoraraki, I. (2018). Comparative study of sequence aligners for detecting antibiotic resistance in bacterial metagenomes. *Lett Appl Microbiol*. 66, 162-168.

Miller, R.R., Montoya, V., Gardy, J.L., Patrick, D.M., Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med*. 5, 81.

Minot, S.S., Krumm, N., Greenfield, N.B. (2015). *One Codex: A sensitive and accurate data platform for genomic microbial identification*. Available: *http://www.onecodex.com/*.

Mohiuddin, M.M., Salama, Y., Schellhorn, H.E., Golding, G.B. (2017). Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res*. 115, 360-369.

Monier, J.M., Demaneche, S., Delmont, T.O., Mathieu, A., Vogel, T.M., Simonet, P. (2011). Metagenomic exploration of antibiotic resistance in soil. *Curr Opin Microbiol*. 14, 229-235.

Mootsikapun, P., Trakulsomboon, S., Sawanpanyalert, P., Aswapokee, N., Suankratay, C. (2009). An overview of antimicrobial susceptibility patterns of Gram-postive bacteria from National Antimicrobial Resistance Surveillance Thailand (NARST) program from 2000 to 2005. *J Med Assoc Thai*. 92, S87-90.

Morgan, J.L., Darling, A.E., Eisen, J.A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One*. 5, e10209.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol*. 17, 53.

Mukherjee, N., Bartelli, D., Patra, C., Chauhan, B.V., Dowd, S.E., Banerjee, P. (2016). Microbial diversity of source and point-of-use water in rural Haiti - A pyrosequencing-based metagenomic survey. *PLoS One*. 11, e0167353.

Munita, J.M., Arias, C.A. (2016). Mechanisms of antibiotic resistance. *Microbiol Spectr*. 4. doi:10.1128/microbiolspec.VMBF-0016-2015.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*. 477, 462-465.

Ngamwongsatit, B., Tanomsridachchai, W., Suthienkul, O., Urairong, S., Navasakuljinda, W., Janvilisri, T. (2016). Multidrug resistance in *Clostridium perfringens* isolated from diarrheal neonatal piglets in Thailand. *Anaerobe*. 38, 88-93.

Nintasen, R., Utrarachkij, F., Siripanichgon, K., Bhumiratana, A., Suzuki, Y., Suthienkul, O. (2007). Enhancement of *Legionella pneumophila* culture isolation from

microenvironments by macrophage infectivity potentiator (mip) gene-specific nested polymerase chain reaction. *Microbiol Immunol*. 51, 777-785.

Nordahl Petersen, T., Rasmussen, S., Hasman, H., Caroe, C., Baelum, J., Schultz, A.C., et al. (2015). Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci Rep*. 5, 11444.

Oh, S., Hammes, F., Liu, W.T. (2018). Metagenomic characterization of biofilter microbial communities in a full-scale drinking water treatment plant. *Water Res*. 128, 278-285.

Okeke, I.N., Lamikanra, A., Edelman, R. (1999). Socioeconomic and behavioral factors leading to acquired bacterial resistance to antibiotics in developing countries. *Emerg Infect Dis*. 5, 18-27.

Oliver, J.D. (2010). Recent findings on the viable but nonculturable state in pathogenic bacteria. *FEMS Microbiol Rev*. 34, 415-425.

Otten, T.G., Graham, J.L., Harris, T.D., Dreher, T.W. (2016). Elucidation of taste- and odor-producing bacteria and toxigenic cyanobacteria in a midwestern drinking water supply reservoir by shotgun metagenomic analysis. *Appl Environ Microbiol*. 82, 5410-5420.

Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k-mers*. *BMC Genomics*. 16, 236.

Padungtod, P., Kaneene, J.B., Hanson, R., Morita, Y., Boonmar, S. (2006). Antimicrobial resistance in *Campylobacter* isolated from food animals and humans in northern Thailand. *FEMS Immunol Med Microbiol*. 47, 217-225.

Peabody, M.A., Caravas, J.A., Morrison, S.S., Mercante, J.W., Prystajecky, N.A., Raphael, B.H., et al. (2017). Characterization of *Legionella* species from watersheds in British Columbia, Canada. *mSphere*. 2, 00246-17.

Pereira, R.P., Peplies, J., Brettar, I., Hofle, M.G. (2017). Development of a genus-specific next generation sequencing approach for sensitive and quantitative determination of the *Legionella* microbiome in freshwater systems. *BMC Microbiol*. 17, 79.

Petersen, A., Dalsgaard, A. (2003). Species composition and antimicrobial resistance genes of *Enterococcus* spp, isolated from integrated and traditional fish farms in Thailand. *Environ Microbiol*. 5, 395-402.

Petersen, T.N., Lukjancenko, O., Thomsen, M.C.F., Maddalena Sperotto, M., Lund, O., Moller Aarestrup, F., et al. (2017). MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. *PLoS One*. 12, e0176469.

Piddock, L.J. (1998). Fluoroquinolone resistance: overuse of fluoroquinolones in human and veterinary medicine can breed resistance. *BMJ*. 317, 1029-1030.

Pinto, A.J., Marcus, D.N., Ijaz, U.Z., Bautista-De Lose Santos, Q.M., Dick, G.J., Raskin, L. (2016). Metagenomic evidence for the presence of Comammox *Nitrospira*-like bacteria in a drinking water system. *mSphere*. 1, e00054-15.

Pope, P.B., Patel, B.K. (2008). Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol Ecol*. 64, 9-27.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 41, D590-596.

Ramamurthy, T., Ghosh, A., Pazhani, G.P., Shinoda, S. (2014). Current perspectives on viable but non-culturable (VBNC) pathogenic bacteria. *Front Public Health*. 2, 103.

Salyers, A.A., Gupta, A., Wang, Y. (2004). Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends Microbiol*. 12, 412-416.

Schmieder, R., Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiol*. 7, 73-89.

Shah, N., Tang, H., Doak, T.G., Ye, Y. (2011). Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac Symp Biocomput*. 165-176.

Sharpton, T.J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci*. 5, 209.

Shi, P., Jia, S., Zhang, X.X., Zhang, T., Cheng, S., Li, A. (2013). Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. *Water Res*. 47, 111-120.

Silva, C.C., Hayden, H., Sawbridge, T., Mele, P., Kruger, R.H., Rodrigues, M.V., et al. (2012). Phylogenetic and functional diversity of metagenomic libraries of phenol degrading sludge from petroleum refinery wastewater treatment system. *AMB Express*. 2, 18.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U SA*. 103, 12115-12120.

Sommer, M.O.A., Dantas, G., Church, G.M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*. 325, 1128-1131.

Stout, J.E., Yu, V.L., Best, M.G. (1985). Ecology of *Legionella pneumophila* within water distribution systems. *Appl Environ Microbiol*. 49, 221-228.

Su, J.Q., An, X.L., Li, B., Chen, Q.L., Gillings, M.R., Chen, H., et al. (2017). Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome*. 5, 84.

Szczepanowski, R., Linke, B., Krahn, I., Gartemann, K.H., Gutzkow, T., Eichler, W., et al. (2009). Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology*. 155, 2306-2319.

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.

Teale, C.J. (2002). Antimicrobial resistance and the food chain. *Symp Ser Soc Appl Microbiol*. 85S-89S.

Tekera, M., Lotter, A., Oliver, J., Jonker, N., Venter, S. (2011). Metagenomic analysis of bacterial diveristy of Siloam hot water spring, Limpopo, South Africa. *Afr J Biotechnol*. 10, 18005-18012.

Thaipadungpanit, J., Wuthiekanun, V., Chantratita, N., Yimsamran, S., Amornchai, P., Boonsilp, S., et al. (2013). *Leptospira* species in floodwater during the 2011 floods in the Bangkok Metropolitan Region, Thailand. *Am J Trop Med Hyg*. 89, 794-796.

Thomas, T., Gilbert, J., Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp*. 2, 3.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 12, 902-903.

Turetgen, I., Sungur, E.I., Cotuk, A. (2005). Enumeration of *Legionella pneumophila* in cooling tower water systems. *Environ Monit Assess*. 100, 53-58.

Van Rossum, T., Peabody, M.A., Uyaguari-Diaz, M.I., Cronin, K.I., Chan, M., Slobodan, J.R., et al. (2015). Year-long metagenomic study of river microbiomes across land use and water quality. *Front Microbiol*. 6, 1405.

Vartoukian, S.R., Palmer, R.M., Wade, W.G. (2010). Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol Lett*. 309, 1-7.

Vetrovsky, T., Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One*. 8, e57923.

Walsh, A.M., Crispie, F., Daari, K., O'sullivan, O., Martin, J.C., Arthur, C.T., et al. (2017). Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks. *Appl Environ Microbiol*. 83, e01144-17.

Wang, Z., Zhang, X.X., Huang, K., Miao, Y., Shi, P., Liu, B., et al. (2013). Metagenomic profiling of antibiotic resistance genes and mobile genetic elements in a tannery wastewater treatment plant. *PLoS One*. 8, e76079.

Weinstock, G.M. (2012). Genomic approaches to studying the human microbiota. *Nature*. 489, 250-256.

Whitlock, G.C., Estes, D.M., Torres, A.G. (2007). Glanders: off to the races with *Burkholderia mallei*. *FEMS Microbiol Lett*. 277, 115-122.

Wood, D.E., Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 15, R46.

Xiao, K.Q., Li, B., Ma, L., Bao, P., Zhou, X., Zhang, T., et al. (2016). Metagenomic profiles of antibiotic resistance genes in paddy soils from South China. *FEMS Microbiol Ecol*. 92.

Yanez, M.A., Carrasco-Serrano, C., Barbera, V.M., Catalan, V. (2005). Quantitative detection of *Legionella pneumophila* in water samples by immunomagnetic purification and real-time PCR amplification of the *dotA* gene. *Appl Environ Microbiol*. 71, 3433-3441.

Yang, X., Noyes, N.R., Doster, E., Martin, J.N., Linke, L.M., Magnuson, R.J., et al. (2016). Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. *Appl Environ Microbiol*. 82, 2433-2443.

Yang, Y., Li, B., Ju, F., Zhang, T. (2013). Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environ Sci Technol*. 47, 10197-10205.

Yang, Y., Li, B., Zou, S., Fang, H.H., Zhang, T. (2014). Fate of antibiotic resistance genes in sewage treatment plant revealed by metagenomic approach. *Water Res*. 62, 97-106.

Ye, L., Zhang, T. (2011). Pathogenic bacteria in sewage treatment plants as revealed by 454 pyrosequencing. *Environ Sci Technol*. 45, 7173-7179.

Zhang, T., Yang, Y., Pruden, A. (2015). Effect of temperature on removal of antibiotic resistance genes by anaerobic digestion of activated sludge revealed by metagenomic approach. *Appl Microbiol Biotechnol*. 99, 7771-7779.

Zhang, T., Zhang, X.X., Ye, L. (2011). Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLoS One*. 6, e26041.