이학박사 학위논문

원핵미생물의 분류체계에 기반한 16S rRNA 유전자 및 유전체 데이터베이스의 개발

Development of prokaryotic taxonomy-based 16S rRNA and genome database

2017년 8월

서울대학교 대학원

생명과학부

윤 석 환

Development of prokaryotic taxonomy-based 16S rRNA and genome database

Advisor: Professor Jongsik Chun, Ph. D.

By Seok-Hwan Yoon

Submitted in Partial Fullfillment

Of the Requirements for the

Degree of Doctor of Philosophy

August 2017

School of Biological Sciences

Seoul National University

원핵미생물의 분류체계에 기반한 16S rRNA 유전자 및 유전체 데이터베이스의 개발

지도 교수 천 종 식

이 논문을 이학박사 학위논문으로 제출함

2017 년 5 월

서울대학교 대학원

생명과학부

윤 석 환

윤석환의 박사 학위논문을 인준함

2017 년 6 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

ABSTRACT

In prokaryotic taxonomy, the 16S ribosomal RNA (rRNA) gene sequence-based approach has served as an alternative standard method to DNA-DNA hybridization (DDH), for which the 97% 16S rRNA gene sequence similarity was considered to be equivalent to the 70% DDH value for species demarcation. While the 16S rRNA-based method is unable to perfectly classify and identify bacterial and archaeal species using 16S rRNA gene, it is currently the most general tool to evaluate the taxonomic position of a prokaryotic strain at the same genus or species levels. Therefore, the 16S rRNA-based approach is still important in the classification of prokaryotes and the use of a database with taxonomically well-curated sequences such as EzTaxon-e is essential for accurate species identification.

There has been a recent evolution of DNA sequencing technologies, called next-generation sequencing (NGS), which has been facilitating Culture-independent microbial community analysis using 16S ribosomal RNA gene as well as the use of genome sequencing data for more informative and precise classification and identification of Bacteria and Archaea. Because the current species definition is based on the comparison of genome sequences between type and other strains in a given species,

building a genome database with accurate taxonomic information is a premium need to enhance our efforts in exploring prokaryotic diversity and discovering new species as well as for routine identifications.

In this study, an integrated database, called EzBioCloud, was constructed to hold the taxonomic hierarchy of Bacteria and Archaea that are represented by quality-controlled 16S rRNA gene and genome sequences. The various bioinformatics pipelines, tools, and algorithms which were applied during the construction of the database were also developed to optimally utilize the database contents. For a more efficient 16S rRNA-based analysis, the pairwise sequence alignment algorithm was improved and a high-performance microbial community analysis pipeline was newly developed in order to better facilitate the analysis of massive NGS data and to produce better results than conventional methods. For whole genome based analyses, quality assessment methods for genome assembly and a genome annotation pipeline were developed and evaluated. The full-length 16S rRNA extraction method and efficient average nucleotide identity (ANI) calculation algorithm were utilized in the identification of public prokaryotic genomes.

In order to construct the integrated genome database, whole genome assemblies in the NCBI Assembly Database were first screened to determine low-quality genomes and then subsequently subjected to a composite

identification bioinformatics pipeline that employed gene-based searches followed by the calculation of average nucleotide identity. The resulting database consisted of 61,700 species/phylotypes including 13,132 with validly published names, and 62,362 whole genome assemblies that were taxonomically identified at the genus, species and subspecies level. Genomic properties, such as genome size and GC content, and the occurrence in human microbiome data were calculated for each genus or higher taxa. This comprehensive database of taxonomy, 16S rRNA gene, and genome sequences, with its accompaniment of bioinformatics tools, should accelerate genome-based classification and identification of Bacteria and Archaea. The database and related search tools are available at http://www.ezbiocloud.net/.

Keywords: Bioinformatics, Pipeline, Database, 16S rRNA, Genome, Taxonomy, Microbiome, Next-generation sequencing, Prokaryote

TABLE OF CONTENTS

ABSTRACT ..... **i**

TABLE OF CONTENTS ..... **iv**

LIST OF TABLES ..... **vii**

LIST OF FIGURES ..... **viii**

ABBREVIATIONS ..... **x**

CHAPTER 1 General introduction **1**

1.1. Taxonomy of prokaryotes 2

1.1.1. Principle of prokaryotic taxonomy 2
1.1.2. Prokaryotic species concept 4

1.2. Next generation sequencing (NGS) 8

1.2.1. 454 Pyrosequencing 8
1.2.2. Illumina-Solexa sequencing 10
1.2.3. Pacific Bioscience SMRT sequencing 11

1.3. Use of 16S rRNA gene in microbiology 13

1.4. Prokaryotic genomics 17

1.5. Objectives of this study 21

CHAPTER 2 Development of bioinformatics pipelines and tools for EzBioCloud database 23

2.1. Introduction 24

2.1.1. 16S rRNA based prokaryote identification algorithm 25
2.1.2. Microbial community analysis 27
2.1.3. 16S rRNA sequence in genome with short-read sequencing data 31
2.1.4. Public genome data of prokaryotes 31
2.1.5. Quality of genome assembly 32
2.1.6. Average nucleotide identity 33

2.2. Materials and method 36

2.2.1. Improvement of 16S rRNA sequence based identification algorithm 36
2.2.2. Development of microbial taxonomic profiling (MTP) pipeline 38
2.2.3. Method for extracting full-length 16S rRNA genes from short-read sequencing data 42
2.2.4. Pipeline for prokaryotic whole genome analysis 44
2.2.5. Methods for the quality assessment of genome 48
2.2.6. Efficient calculation method for average nucleotide identity 52

2.3. Results 54

2.3.1. Advanced microbial taxonomic profiling (MTP) pipeline 54
2.3.2. Comparison of full length 16S rRNA extraction methods 62
2.3.3. Annotation of public genomes 66
2.3.4. Quality of bacterial genomes 68
2.3.5. Evaluation of algorithms for average nucleotide identity 75

2.4. Discussion 81

CHAPTER 3 Development of EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies 84

3.1. Introduction 85

3.2. Methods 87

3.2.1. Data collection 87
3.2.2. Identification of genome sequences 90
3.2.3. Calculation of genomics features for each taxon 93
3.2.4. Bacterial community analysis of human microbiome ... 93
3.2.5. Operating system and software development 95

3.3. Results 96

3.3.1. Comparison of databases 96
3.3.2. Hierarchical taxonomic backbone 99
3.3.3. Identification of genome projects 103
3.3.4. Genome-derived information 107

3.4. Discussion 108

CHAPTER 4 General conclusions 111

REFERENCES 115

국문초록 **..... 130**

LIST OF TABLES

Table 1. Popular primer names and sequence for variable regions and manly used NGS platforms information. 16
Table 2. Parameters of search engine software in identification algorithm. 37
Table 3. Bioinformatics tools and parameters for the MTP pipeline. 41
Table 4. Bioinformatics tools, parameters and databases of pipeline for whole genome analysis. 47
Table 5. 54 Bacterial core gene (BCG) list and product. 51
Table 6. Version and run parameters used in comparison of four average nucleotide identity methods. 53
Table 7. The result of 16S rRNA extraction using the previous method. 64
Table 8. The result of 16S rRNA extraction using developed method. 65
Table 9. The number of predicted CDSs for 4 genomes of species Rhodopirellula baltica. 67
Table 10. Top ten genus with the lowest number of BCGs. 70
Table 11. The most frequent contaminants from 674 contaminated genomes predicted by ContEst16S. 72
Table 12. 18 body parts of healthy human subjects and analyzed sample count of each bacterial community. 94
Table 13. Comparison of 16S rRNA sequence databases. 98
Table 14. The count of 16S rRNA sequences by source and corrected information. 102
Table 15. The count of genome identification results. 105

LIST OF FIGURES

Figure 1. A bacterial species concept using “Type strain”. 6
Figure 2. 16S rRNA hypervariable regions and popular primer sets for each NGS platform. 15
Figure 3. Growth of the number of prokaryotic genome sequencing data. . 18
Figure 4. Overview of pipelines used by QIIME and MOTHUR. 30
Figure 5. Overview of the MTP pipeline. 40
Figure 6. Comparison of 16S rRNA sequence extraction method..... 43
Figure 7. Overview of prokaryotic whole genome analysis (WGAS) pipeline. 46
Figure 8. Comparison of MTP pipelines. 56
Figure 9. Performance of alignment algorithm for various read length. 57
Figure 10. Comparison of taxonomic profiling for a *Facalibacterium* case. 60
Figure 11. Comparison of taxonomic profiling for a *Bacteroides* case. 61
Figure 12. Correlation between genome size and number of BCGs. 71
Figure 13. CheckM scores of low-quality or contaminated genomes. 74
Figure 14. Correlation between the ANIb and other algorithms in the whole range of ANI. 78
Figure 15. Correlation between the ANIb and other algorithms in the range of >90% ANI. 79
Figure 16. Running times of four ANI algorithms. 80
Figure 17. Scheme of data collection for EzBioCloud database. 89
Figure 18. Outline of an algorithm for identifying genome sequence. 92
Figure 19. OrthoANIu-based dendrogram of the genus *Acinetobacter* including 13 tentatively named species. 101
Figure 20. Pie on pie chart of 16S rRNA sequence source and corrected 16S rRNA sequences on EzBioCloud database. 102

Figure 21. Pie on pie chart of genome identification of NCBI genomes. .. 105
Figure 22. Example of misidentified and unidentified WGAs. 106

ABBREVIATIONS

DNA	Deoxyribonucleic acid
NTP	Nucleoside triphosphate
RNA	Ribonucleic acid
rRNA	ribosomal RNA
tRNA	transfer RNA
tmRNA	transfer-messenger RNA
NGS	Next-generation sequencing
DDH	DNA-DNA Hybridization
ANI	Average nucleotide identity
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
ICSB	International Committee on Systematic Bacteriology
ICSP	International Committee on Systematics of Prokaryotes
OGRI	Overall genome related index
SMRT	Single molecule real time
ZMV	Zero-mode waveguides

SSU Small subunit

RDP Ribosomal Database Project

OLC Overlap layout consensus

DBG De Bruijn Graph

CRISPR Clustered regularly interspaced short palindromic repeats

CDS Coding DNA sequence

BLAST Basic Local Alignment Search Tool

QIIME Quantitative Insights Into Microbial Ecology

OTU Operational taxonomic unit

MTP Microbial taxonomic profiling

SRA Sequence Read Archive

CCS Circular consensus

KEGG Kyoto Encyclopedia of Genes and Genomes

BCG Bacterial core genes

HMM Hidden Markov Model

ContEst16S Contamination Estimator by 16S

WGA Whole genome assembly

WGAS Whole genome analysis

RefGD	Reference Genome Database
PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
AWS	Amazon Web Services

CHAPTER 1

General introduction

1.1. Taxonomy of prokaryotes

1.1.1. Principle of prokaryotic taxonomy

Taxonomy is the science of classification of organisms. Prokaryotic taxonomy consists of three separate, but interrelated areas: classification, nomenclature, and identification. Classification refers to arranging organisms into separated groups (taxa) based on similarity or relationship. Nomenclature is the assignment of names to the taxonomic groups according to international rules (Lapage*, et al.*, 1992). Identification is the practical use of a classification method to determine the identity of an isolate as a member of a pre-defined taxon or as a member of a previously unidentified species.

1.1.1.1. Classification and identification

Classification is the process of arranging organisms into groups (taxa) based on their similarities or relationships (not confined to relationships by ancestry). The result of classification is an orderly arrangement or system designed to express interrelationships of organisms and to reveal a variety of different functions which may cause groupings of organisms. Early bacterial

classifications relied on phenotypic typing schemes, which generally employed morphological, anatomical, gram staining, biochemical characteristics and so on. During this time, bacteria were often classified based on a single characteristic or a series of single characteristics, using a method termed monothetic classification (Sneath and Sneath, 1962). This artificial monothetic classifying method was known to carry the serious risk of misidentification when an encountered organism was aberrant in one of the key characteristics selected (Steel, 1965). In contrast, numerical classification was a polythetic procedure, incorporating high information content introduced to bacteriology (Sneath, 1957), which could generally accommodate strain variation and was objective in the sense that it wasn't overtly sensitive to the addition of more strains or characteristics (Chun, 1995). After decades, today, ribosomal RNA sequence-based numerical classification is commonly used in microbiology because phenotypic characteristics are not informative enough, or too unstable to be used as phylogenetic markers (Ludwig and Schleifer, 1994).

1.1.1.2. Nomenclature

There is no official classification of bacteria, but there is a valid nomenclature (EUZéBY, 1997). This bacteriological nomenclature started in 1980, when the Approved Lists of Bacterial Names were published in the International

Journal of Systematic Bacteriology (SKERMAN*, et al.*, 1980). Initially, the International Committee on Systematics of Bacteriology played a role in setting the cornerstone of bacterial nomenclature when it published the *International Code of Nomenclature of Bacteria*. In 1999, in order to update the Code and adjust it to fit modern requirements (De Vos and Trüper, 2000), the name of the *International Code of Nomenclature of Bacteria* was changed to the *International Code of Nomenclature of Prokaryotes*, and the name of the International Committee on Systematic Bacteriology (ICSB) was changed to the International Committee on Systematics of Prokaryotes (ICSP). To this day, the ICSP provides recommendation reports for nomenclature and type strains from time to time, but it cannot formally set up the definition of bacterial species. It is rather decided by community efforts or consensus among scientists, and this collective decision is widely accepted by microbiologists.

1.1.2. Prokaryotic species concept

The number of species of bacteria and archaea with valid names is surprisingly small despite their early evolution and their genetic, as well as ecological diversity. This is due to the difficulty of growing samples in pure culture, and the occurrence of extensive horizontal gene transfer that blurs the distinction of species (Staley, 2006). At present, the most commonly

accepted species concept is the polyphasic species definition, which takes into account both phenotypic and genetic differences. It is called the *"Phylo-phenetic species concept",* coined by Rosselló-Mora and Amann (Rosselló-Mora and Amann, 2001). A phylo-phenetic species is "*a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity with respect to many independent characteristics, and is diagnosable by a discriminative phenotypic property*". The definition means that a species is a group of bacterial strains that share a most recent common ancestor and have similar genome sequences. In addition, two different species should be differentiated by phenotypic characteristics like biochemical, morphological and physiological properties.

However, in practice, "Phylo-phenetic species concept" becomes unreliable when clear divisions of clusters are not apparent among properties of strains in different species. To solve this problem, bacterial taxonomists have introduced the concept of a "type strain". A type strain is a living culture that serves as a fixed reference point for the assignment of bacterial and archaeal names. When multiple strains are discovered for a single species, one of them, a likely representative strain, can be chosen as the type strain. As most of bacterial species can be described with only one or a few strains, the type strain of a species is often designated to the strain which was first discovered. These type strains may not be very typical for a given species, which can lead to strains with very different properties being assigned to the same species, as illustrated in **Figure 1**.

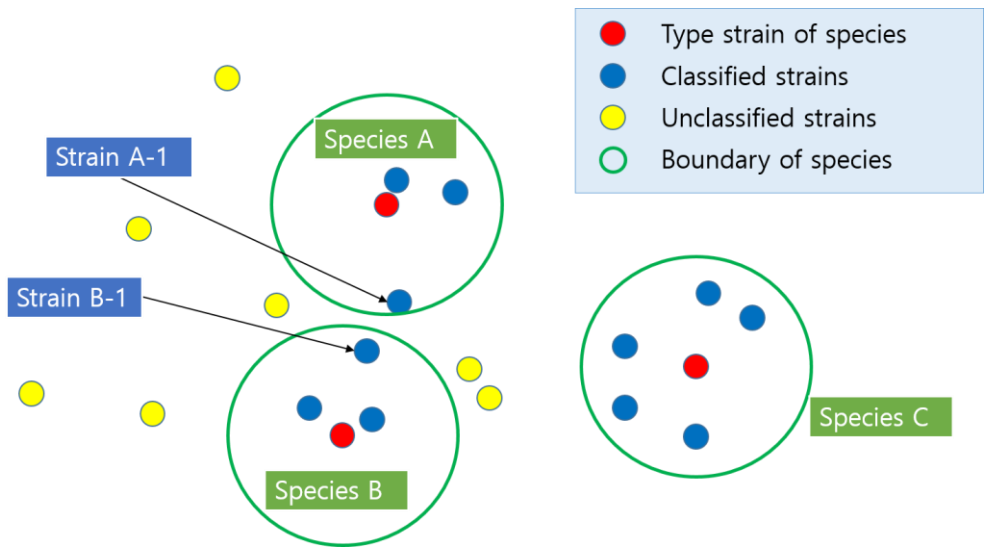

Figure 1. A bacterial species concept using "Type strain". The green circles in Figure 1 represent species boundaries, which have inclusion cutoffs that depend on the criteria for determining species or marker genes. In the figure, we observe that although the distance between the designated type strain of species A (red) and strain A-1 (blue) is greater than the distance between strains A-1 and B-1, strain A-1 belongs to species A and B-1 belongs to species B according to this particular type strain and predefined species boundary designation.

Without considering phenotypic properties, a quicker diagnostic *ad hoc* threshold to separate species is less than 70% DNA-DNA hybridization (DDH) (Wayne*, et al.*, 1987). DDH provides an overall, albeit indirect, measure of genomic similarity between two strains, and serves well as a surrogate for genome sequence comparison. DDH threshold corresponds to less than 97% 16S DNA sequence identity (Stackebrandt and Goebel, 1994), and the cutoff value is still widely used for identification of prokaryotic species using 16S DNA sequence.

Recently, the development of next generation sequencing technology has enabled many individual researchers to perform prokaryotic genome sequencing. Genome comparison is more accurate than other marker gene-based methods like the 16S rRNA-based method because genomes contain much greater genetic information than single genes. “Overall Genome Related Index” (OGRI) is a term for any computational method to calculate similarity between two genome sequences, first coined in 2014 (Chun and Rainey, 2014). Although there are many different algorithms of OGRI that can be used for comparing two strains, Average Nucleotide Identity (ANI) (Goris*, et al.*, 2007) has been the most widely accepted. The generally accepted cutoff value for the species boundary is about 95% ANI. Recently, an improved version of ANI called OrthoANI was introduced (Lee*, et al.*, 2016) and it demonstrated better performance than the original ANI.

1.2. Next generation sequencing (NGS)

Sanger sequencing is a chain termination method developed by F. Sanger and A. R. Coulson in 1977 (Sanger*, et al.*, 1977). It is a method of inserting DNA polymerase and dideoxy NTP, which does not have an oxygen atom at the 3' position, to terminate DNA synthesis and to decode the base sequence. This Sanger sequencing technology has been around for decades as a gold standard in biology and medical research, but it has been difficult to analyze large numbers of nucleotides, to the extent that it cost $3 billion over 15 years to decode the human genome with 3 billion nucleotides. However, since the advent of the first commercially available NGS technology in 2005, many NGS technologies have enabled sequencing DNA at an unprecedented speed and surprisingly low cost. In addition, these technologies are suitable for metagenomics analysis by DNA amplification using universal adapters without the cloning process that was previously required for Sanger sequencing.

1.2.1. 454 Pyrosequencing

The 454 Pyrosequencing technology was the first commercially available NGS technology. The principle of pyrophosphate detection was first

described in 1985 and the first system based on DNA sequencing technology using this principle was reported in 1988. In 2005, the GS device series was released as the first commercially available next-generation DNA sequencer commercialized by 454 Life Sciences (later acquired by Roche Diagnostics), that offered a parallelized version of pyrosequencing. This method amplifies DNA inside water droplets in an oil solution, a process called emulsion PCR, which amplifies a single DNA template attached using a single primer-coated bead in oil droplets. The sequencing machine contains PicoTiterPlate which includes many picoliter-volume wells, each containing a single bead and sequencing enzymes. In pyrosequencing, luciferase is used to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs (Egholm*, et al.*, 2005). In the sequencing process, light signal intensity is proportional to the number of pyrophosphates released and hence the number of bases incorporated. However, this approach is prone to errors that result from incorrectly estimating the length of homo-polymeric sequence stretch. The error type of insertion or deletion from homo-polymer in pyrosequencing was well known to scientists. The overall error rate is known to be about 0.5% and the errors caused by homo-polymer is known to account for 39% of the total errors (Huse*, et al.*, 2007). Despite the error rates, this technology has been used primarily in microbial community analysis because of its ability to sequence the longest nucleotide sequences of NGS technologies for years

after its launch. However, as of 2016, Roche Diagnostics no longer supports pyrosequencing devices, and researchers are not using it anymore.

1.2.2. Illumina-Solexa sequencing

In 2006, another next-generation sequencing platform called Genome Analyzer was commercialized by Solexa (later acquired by Illumina). The principle of the sequencing method is based on reversible dye-terminators technology which was developed by Shankar Balasubramanian and David Klenerman in 1998 (Bentley*, et al.*, 2008). In this method, primers and DNA molecules are initially attached to a slide or flow cell and amplified with polymerase so that local clonal DNA clusters are formed. The vast number of DNA clusters that can be formed in a flow cell leads to an extremely large number of reads, enabling the Illumina sequencing platform to be incredibly high-throughput. To determine the sequence, four types of reversible chain terminator nucleotides, each labeled with a different fluorescent dye, are supplied and unused nucleotides are washed away. Because of this process, unlike pyrosequencing, the DNA chains are extended one nucleotide at a time and homo-polymer errors are not generated. Although Illumina platform exhibits a lower error rate and higher throughput when compared to pyrosequencing, older devices such as HiSeq had primarily been used for genome and transcriptome analyses, as device-generated read lengths were

too short to be used in microbial community analysis studies. However, the advent of 250bp paired-end sequencing through MiSeq enabled Illumina to produce results on par with pyrosequencing, making Illumina the most widely used platform for microbial community analysis.

1.2.3. Pacific Bioscience SMRT sequencing

Single molecule real time (SMRT) sequencing technology(Eid*, et al.*, 2009) was introduced by Pacific Bioscience in 2010 (McCarthy, 2010) and it is sometimes referred to as “third-generation” sequencing (Hayden, 2009). The sequencing is based on the sequencing by synthesis approach and it is performed using unmodified polymerase, which attaches to zero-mode waveguides (ZMW; small well-like containers with the capturing tools located at the bottom of the well) and fluorescently labeled nucleotides flowing freely in the solution. During DNA synthesis, fluorescence from the labeled nucleotides at the bottom of the well is detected by the sequencing machine. This approach allows reads of 20,000 nucleotides or more, with average read lengths of 5 kilobases. Recently, Pacific Biosciences announced the launch of a new sequencing instrument called the Sequel System, with much more throughput than its previous PacBio RS II system. PacBio sequencing technology can be used to analyze microbial communities that require precise analysis (Mosher*, et al.*, 2014) and to complete microbial genomes

(Liao*, et al.*, 2015) by obtaining the longest possible reads among all of the commercialized NGS platforms.

1.3. Use of 16S rRNA gene in microbiology

16S rRNA stands for 16S ribosomal ribonucleic acid (rRNA), where S (Svedberg) is a unit of measurement (sedimentation rate). This rRNA is an important constituent of the 30S small subunit (SSU) of prokaryotic ribosomes as well as mitochondria and chloroplasts. 16S rRNA genes are used in reconstructing phylogenies, due to the slow rates of evolution of this region of the gene (Woese and Fox, 1977). To be used as a DNA barcode, a gene should be ubiquitous and should contain sufficient phylogenetic information. All members of Bacteria and Archaea are known to have the 16S rRNA gene, which is about 1,500bp long, and the genetic variation within this gene found among prokaryotes is adequate to be used in phylogenetic analysis for broad taxonomic applications. It is successfully used to infer phylogenetic relationships among phyla, andalso used for the comparison of species in the same genus. The gene acts as a DNA barcode, and can also be easily amplified by PCR. 16S rRNA gene has multiple conserved regions that can be used as priming sites. This becomes a significant advantage for NGS-based short read sequencing. After many years of international collaboration, several 16S rRNA sequence databases like SILVA (Quast*, et al.*, 2013), Greengenes (DeSantis*, et al.*, 2006), RDP (Cole*, et al.*, 2013) and EzTaxon (Kim*, et al.*, 2012) contain almost all known species of Bacteria and Archaea. By searching the 16S sequence against these databases, anyone,

even without knowledge of serious taxonomy, can identify newly isolated strains. Sequence variation in bacterial 16S rRNA gene is known to be not uniformly distributed. The gene sequence contains nine hypervariable regions (V1-V9) ranging from about 30-100 base pairs long that can provide species-specific signature sequences useful for identification of bacteria (Pereira, et al., 2010). While the entire 16S sequence allows for comparison of all hypervariable regions, at approximately 1500 base pairs long it can be more expensive for studies seeking to identify or characterize diverse bacterial communities (Yang, et al., 2016). NGS is suited for elucidating bacterial community structure, as it eliminates the requirement of tedious E. coli cloning and allows high throughput DNA sequencing. Because different lengths of DNA are sequenced by various NGS platforms, a suitable pair of PCR primers should be used **(Figure 2, Table 1)**. Illumina platforms and 454 pyrosequencing do not cover all hypervariable regions of 16S rRNA, but these NGS platforms are cheaper and allow for deeper community coverage. For this reason, these platforms have been widely used in many previous studies for microbial community analyses. Recently, full-length 16S rRNA sequences have been obtained at a reasonable price using the Pacbio platform, and studies using it have been increasing. However, the Illumina MiSeq platform is still very popular because of its lower price.

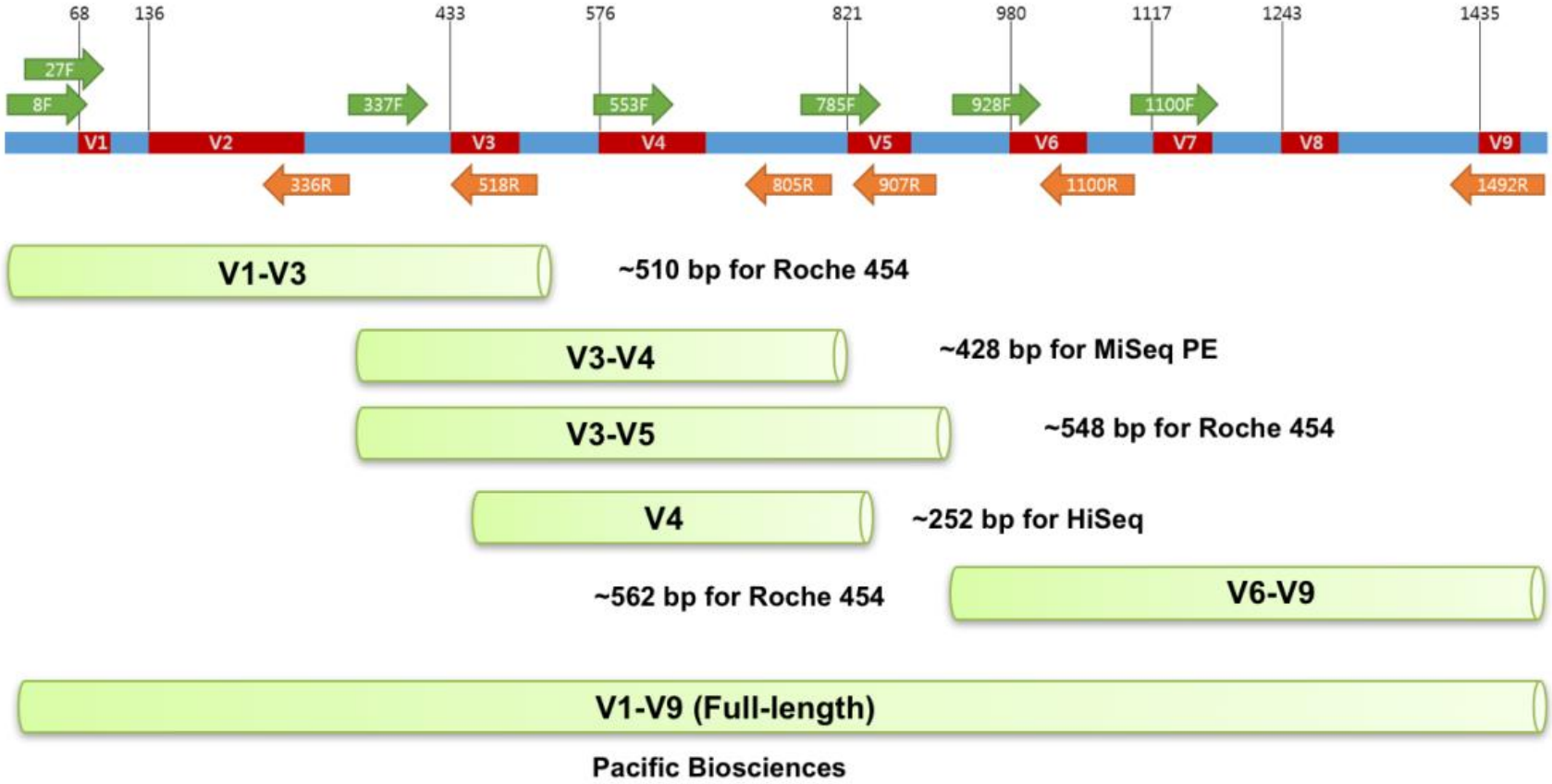

Figure 2. 16S rRNA hypervariable regions and popular primer sets for each NGS platform.

Table 1. Popular primer names and sequence for variable regions and manly used NGS platforms information. The full-length 16S gene is usually amplified by the pair of primers 27F and 1492R, followed by Sanger DNA sequencing. To obtain accurate sequence, both DNA strands should be sequenced using multiple primers given in the above table.

Name of Primer F=forward, R=reverse	Variable region	NGS platform (mainly used)	Sequence
8F	V1 ~	-	AGAGTTTGATCCTGGCTCAG
27F	V1 ~	454, Pacbio	AGAGTTTGATCMTGGCTCAG
336R	~ V2	-	ACTGCTGCSYCCCGTAGGAGTCT
337F	V3 ~	454	GACTCCTACGGGAGGCWGCAG
518R	~ V3	454	GTATTACCGCGGCTGCTGG
533F	V4 ~	Illumina	GTGCCAGCMGCCGCGGTAA
785F	V5 ~	-	GGATTAGATACCCTGGTA
805R	~ V4	Illumina	GACTACCAGGGTATCTAATC
907R	~ V5	454	CCGTCAATTCCTTTRAGTTT
928F	V6 ~	454	TAAAACTYAAAKGAATTGACGGG
1100F	V7 ~	-	YAACGAGCGCAACCC
1100R	~ V6	-	GGGTTGCGCTCGTTG
1492R	~ V9	454, Pacbio	CGGTTACCTTGTTACGACTT

1.4. Prokaryotic genomics

Since the first two bacterial genome sequencing were completed in 1995 (Fleischmann*, et al.*, 1995; Fraser*, et al.*, 1995) and the first archaeal genome in 1996 (Bult*, et al.*, 1996), technical improvements including next-generation and third-generation sequencing technology have led to dramatic reduction in the price of prokaryotic genome sequencing. Along with the cost reduction, next-generation sequencing technology remarkably reduced the read length. In contrast, third-generation sequencing technology allowed for longer read lengths. The astonishing reduction of sequencing cost has made bacterial genome sequencing affordable to a large number of laboratories, leading to a democratization of sequencing (Shendure and Ji, 2008). As a result, the number of publicly released bacterial genome sequencing data has grown explosively **(Figure 3)** and the large number of genome data caused a cost shift from sequencing to assembly, analysis, and data management (Land*, et al.*, 2015).

Single genome analysis using sequencing data largely consists of three steps: assembly, gene prediction, and function annotation. Software used in the genome assembly step may differ by sequencing platform, but overlap layout consensus (OLC) and de Bruijn graph (DBG) are the two most widely used algorithms (Li*, et al.*, 2012).

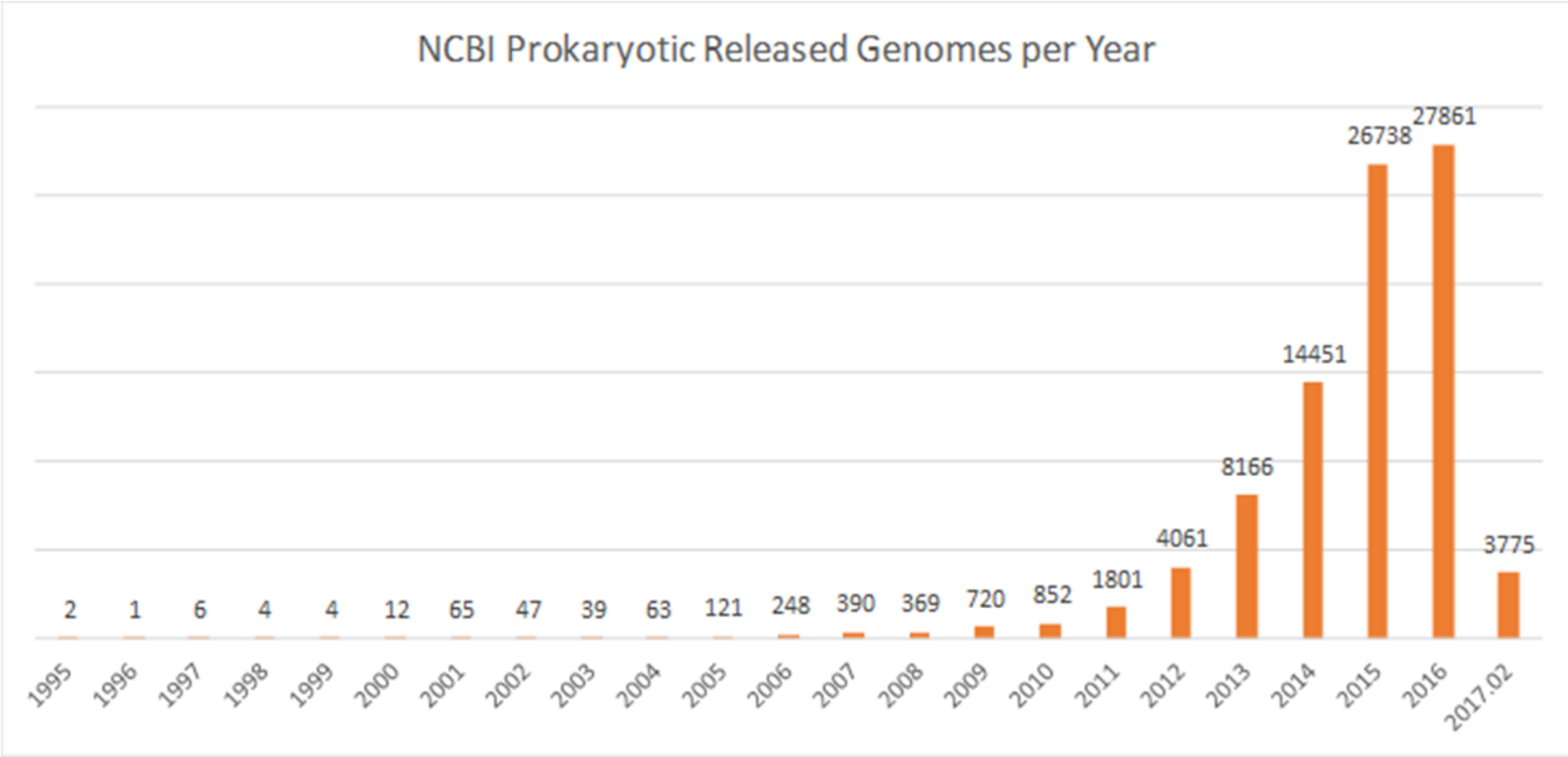

Figure 3. Growth of the number of prokaryotic genome sequencing data. Each vertical bar represents the accumulated number of genome sequencing data.

OLC is an algorithm that is suitable for contiguous genome assembly from long sequences and generally works in three steps. First, overlaps among all reads are found, then all reads and overlaps are placed on a graph, and finally, the consensus sequence is inferred. The DBG method is mainly used for creating an assembly from short sequences and involves finding the optimal path in a de Bruijn graph after splitting the sequences into k-mers. The DBG method is sensitive to sequencing errors and lacks robustness when assembling from repetitive sequences, but it is much faster and less computationally demanding than the OLC method. Newbler and HGAP assemblers implement the OLC algorithm and are used for 454 pyrosequencing data and Pacific Bioscience sequencing data, respectively. Many public assemblers that implement the DBG algorithm, such as SPAdes, Velvet, SOAPdenovo, IDBA, etc. are used for Illumina sequencing data. Illumina Miseq platform with DBG-based assembly has been widely used for the creation of Draft Genome data for prokaryotic genome analysis, and recently, Pacific Bioscience sequencing technology with OLC-based assembly is being used to generate Complete Genome data.

For prokaryotic genome data, the gene prediction step involves finding tRNA, rRNA, non-coding RNA, CRISPR and CDS regions, for which a variety of programs are used. Finally, the Functional Annotation step queries predicted CDSs to a database of existing protein sequences using a similarity search program such as BLAST. Depending on the search program and database used, the results from this step may vary. Because this is a

bioinformatic analysis, further experimentation using methods such as RNA-seq can be employed to verify whether or not the actual gene functions in the microorganism.

An explosive increase in microbial genome data has resulted in a corresponding increase in comparative genomics research. One of the earliest and crucial generalizations of prokaryotic comparative genomics is the readily recognizable evolutionary conservation of protein sequences encoded in the majority of the genes in each sequenced genome (Koonin and Galperin, 1997). Comparative genomics has also shown that widespread horizontal gene transfer occurs between prokaryotes (Jain*, et al.*, 1999). Through comparative genomic studies of these microorganisms, we can determine what genes are common among different genomes, and which genes are unique. Using this information, we can derive phylogenetic relationships, identify useful genes, and discover methods to combat pathogenic genes.

1.5. Objectives of this study

In the next-generation sequencing era, two important challenges of bioinformatics are providing accurate analysis results in a short time using massive data and helping to interpret these results using applications. The purpose of this study is the development of a united, comprehensive prokaryotic taxonomy, 16S rRNA, and genome database as well as the development of supporting software, applicable pipelines, and tools for microbiome analysis. An additional purpose was to provide a web-based application that would allow many researchers to easily browse and use database contents.

Chapter 2 focuses on the development of several advanced tools and pipelines using improved algorithms. In order to process the genomic data of many microorganisms, methods that have the same or higher level of accuracy as well as faster analysis speed than the existing algorithms are needed. First, I improved the speed of the in-house pairwise alignment algorithm used in various tools to improve the performance of pipelines and tools. The improved algorithm was applied to an identification engine using 16S rRNA, a taxonomic profiling pipeline using amplicon-based NGS data, and many steps requiring pairwise similarity calculation. To construct a microbial genome database, I developed a full 16S rRNA extraction algorithm and a genome annotation pipeline for identification, and applied a quality

evaluation algorithm and a contamination detecting algorithm. I further developed and optimized the taxonomic profiling pipeline and ANI calculation method to utilize the constructed database.

Chapter 3 focuses on the constructed database using several tools and algorithms developed in chapter 2. To build the database, tens of thousands of 16S rRNA and genome data were collected, and these data were applied to the database through refinement and reanalysis processes through improved analysis methods. In order to guide the utilization of the constructed database, the thousands of human microbiome data were analyzed based on the 16S rRNA database. The genomic properties of various microbial species and genus were analyzed based on the genome database, and many prokaryotic microbial species were more accurately classified. All of the analyzed information was visualized through the website.

CHAPTER 2

Development of bioinformatics pipelines and tools for EzBioCloud database

2.1. Introduction

The primary goal of bioinformatics is to increase the understanding of biological processes. After the introduction of the first NGS technology in 2005 (Egholm *et al.*, 2005), research on bioinformatics has been actively carried out with the development of sequencing technology. Major bioinformatics researches in the field include gene finding, sequence assembly, pairwise and multiple sequence alignment, protein structure prediction, protein structure alignment, evaluation of gene expression and protein–protein interactions, the modeling of evolution, genome-wide association studies, phylogenetic analysis and so on. The results of these bioinformatics research are in the form of algorithms, databases, web-based visualization, and standalone tools. Algorithm development is a very core area, and speed and accuracy are the main goals of the study. However, algorithm development further depends on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, graph theory, machine learning, and statistics. Databases are essential for bioinformatics research and many databases exist for various information types such as DNA or protein sequences, molecular structures, taxonomic information and other linked meta information. The most famous database of bioinformatics is NCBI (https://www.ncbi.nlm.nih.gov/) (Sayers*, et al.*, 2011) and the database contains information for almost all fields of bioinformatics,

but also contains some non-refined information. A pipeline is not only a bioinformatics term but also a computer science term. It can be a well-defined model for bioinformatics analysis processes with a specific structure defined by the topology of data-flow interdependencies, and a particular functionality arising from the data transformations applied at each step (Gil*, et al.*, 2007). There are also a number of web-based and software applications for performing analysis as part of a pipeline, or as a standalone tool.

2.1.1. 16S rRNA based prokaryote identification algorithm

Many studies have been published regarding algorithms to identify prokaryotes using their 16S rRNA sequences. Among these algorithms, the identification algorithm published in the EzTaxon database paper is a reliable one that has been cited more than 4800 times (Kim *et al.*, 2012). This identification method queries a phylogenetically classified 16S rRNA database using a search program such as BLAST and subsequently performs pairwise alignment to calculate similarity values. Although a similarity cutoff of 97% is still used as the identification criterion at the species level, this is not officially defined, and a paper published in 2015 proposed a species level cutoff of 98.7% for 16S rRNA similarity through comparison with ANI values (Kim*, et al.*, 2014). Because the identification algorithm has

search and alignment processes, slight differences in similarity values can be observed when using different search engines or pairwise alignment algorithms, and performance improvements can be made by optimizing either the search engine or the pairwise alignment algorithm, or both.

A complete 16S rRNA gene sequence is defined as the DNA sequence region between PCR primers 27F and 1492R for Bacteria (Lane, 1991), and between PCR primers A25F and U1492R for Archaea (Dojka*, et al.*, 1998). This definition is based on the most widely used primers in 16S rRNA gene sequencing. The complete 16S rRNA gene sequence serves as a reference against which partial 16S rRNA gene sequences (obtained from high throughput sequencing) can be compared. Complete 16S rRNA gene lengths vary depending on species, and a complete or nearly complete sequence is generally required for taxonomic analyses. To determine whether a 16S rRNA gene segment that was sequenced from a sample is complete or nearly complete, the measure is used called completeness. The definition of completeness is an objective measure of the degree of coverage of a query 16S rRNA gene sequence with respect to the full-length, complete 16S rRNA gene sequence. Mathematically, completeness is defined by below equation as (Kim *et al.*, 2012).

$$\text{Completeness}(\%) = \frac{L \times 100}{C}$$

where L is the length of a query sequence and C is the length of the most similar sequence that is regarded as complete. The most similar sequence in the database of complete sequences is identified by using a software called USEARCH (Edgar, 2010). The suggested minimum threshold for using a 16S rRNA gene sequence for taxonomic purposes is 95% completeness, as incomplete or partial sequences with low completeness scores will have insufficient resolving power, resulting in erroneous identification results.

2.1.2. Microbial community analysis

Microbial community analysis is a general method for understanding of the role of the microbiome in environmental places or human body. The structure of microbial communities is commonly investigated using next-generation sequencing (NGS) data of 16S rRNA amplicons. There are various tools already available to analyze microbial community using 16S rRNA sequencing data including QIIME (Quantitative Insights Into Microbial Ecology) (Caporaso*, et al.*, 2010), MOTHUR (Schloss*, et al.*, 2009), MG-RAST (Metagenomics - Rapid Annotation using Subsystems Technology) (Meyer*, et al.*, 2008), MEGAN (Mitra*, et al.*, 2011), the RDPipeline (Ribosomal Database Project Pipeline) (Cole *et al.*, 2013), Vegan (Dixon and Palmer, 2003), and MICCA (Albanese*, et al.*, 2015). Some of these tools

consist of the entire process of microbial community analysis whereas the others can be used only for specific steps in the analysis. The pipelines with entire processes contain various algorithms for quality-control, chimera filtering, taxonomic assignment, Operational Taxonomic Unit (OTU) clustering, diversity calculation and results visualization. As a common process of any pipeline, quality control includes trimming sequences by quality score, paired-end merging, primer trimming both in the 5' and 3' ends of reads from NGS platforms, and output average length specific minimum length filtering (Jeon*, et al.*, 2013). The amplified 16S rRNA sequencing data contains chimeric sequences by PCR amplification (Edgar*, et al.*, 2011). Thus, these chimeric sequences are selected and excepted from taxonomic profiling results by various algorithms in chimera filtering. The taxonomic assignment is a massive parallel identification process for finding consensus or representative sequences of OTUs or non-redundant sequence clusters. The OTU clustering is de novo sequence clustering or reference based clustering or a combination of these two methods, where OTU usually means species with 97% identity (Janda and Abbott, 2007). The calculated diversity measures include alpha diversity indices and rarefaction curve (Schloss *et al.*, 2009). Among the tools mentioned above, QIIME and MOTHUR were reported as two outstanding pipelines due to their comprehensive features and support documentation. These two pipelines are also the most frequently used pipelines and the overview of the workflow used by the pipelines are shown in **Figure 4** (Plummer*, et al.*, 2015). An important common trait of

these two pipelines is that the taxonomic assignment step is performed after OTU clustering. This is done in order to decrease computational cost and bias of massive parallel identification, but it is not suitable for EzBioCloud, as it is a database for identification at thje species level, built upon the type strain concept. Thus, in this study, I developed a new pipeline for species level identification named microbial taxonomic profiling (MTP) pipeline.

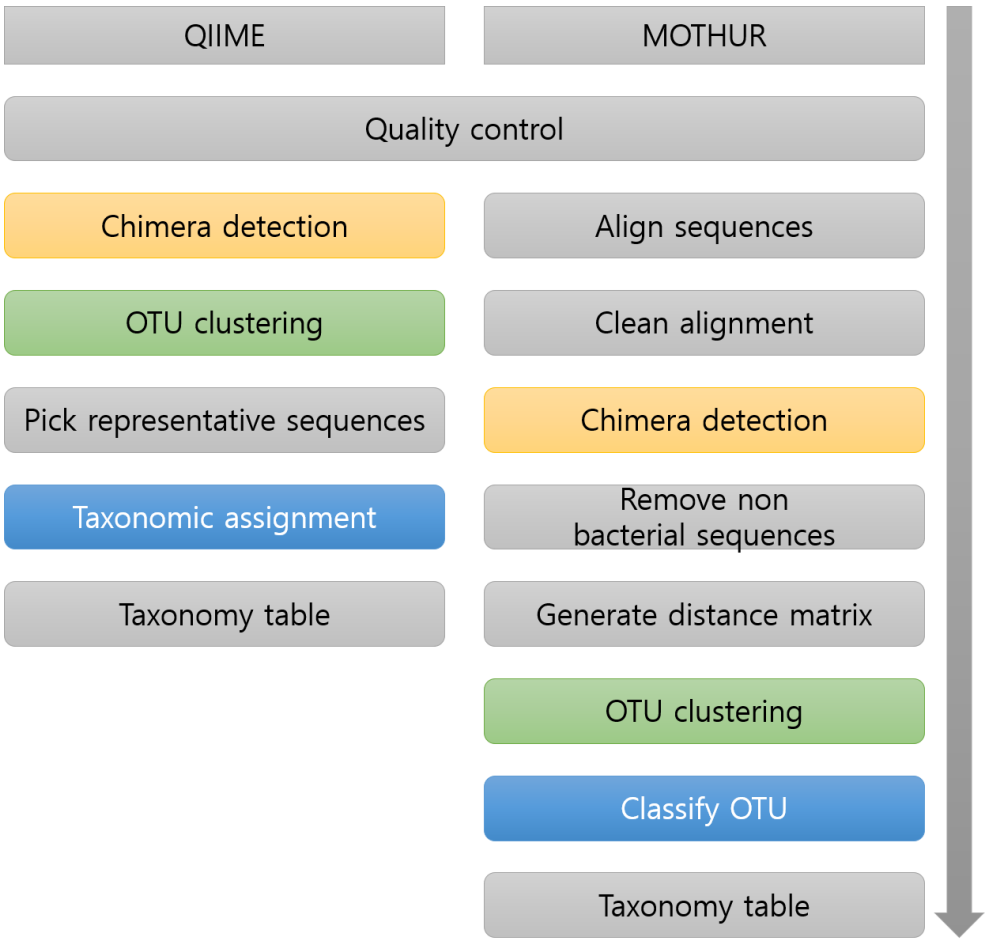

Figure 4. Overview of pipelines used by QIIME and MOTHUR. Several steps in microbial community analysis are shared between two pipelines (e.g. quality control, chimera detection, classification or assignment taxonomy).

2.1.3. 16S rRNA sequence in genome with short-read sequencing data

Among NGS technologies, Illumina's HiSeq and MiSeq platforms are widely used in whole genome sequencing because they provide cost-effective, high-throughput results (Goodwin*, et al.*, 2016). The most commonly used method for 16S rRNA isolation from the whole genome is to perform de novo assembly of sequencing data followed by extraction of 16S rRNA sequence. However, we often discover only partial 16S rRNA sequences in contig data due to some of the inherent limitations of a short-read assembly from NGS data, such as contamination, new insertions, collapsed repeats, etc. (Alkan*, et al.*, 2011). Using a partial 16S rRNA sequence for identification may yield low-similarity, inaccurate results. To solve this problem, I have developed an algorithm that extracts full-length 16S rRNA sequences from short-read sequencing data better than conventional methods.

2.1.4. Public genome data of prokaryotes

The NCBI (http://www.ncbi.nlm.nih.gov) genome database contains a significant number of prokaryotic genomes. Advances in NGS technology enabled prokaryotic genome sequencing and analysis by individual researchers or small-scale laboratories, leading to an explosive growth in the

number of public genomes, as shown in **Figure 3**. Much of this data is the result of many researchers analyzing data from various NGS sequencing platforms using different assembly and annotation methods. Due to this variety, while NCBI abounds with prokaryotic genome data, data quality is not maintained at the same level. In addition, because of the somewhat stringent genome submission requirements regarding fields such as annotation format and description, some researchers have only uploaded raw sequencing data to the Sequence Read Archive (SRA) Database (Leinonen*, et al.*, 2010) or have opted to publish their data to other public websites. In this study, I developed an annotation pipeline that can be applied to all prokaryotic genomes, regardless of the problems mentioned above.

2.1.5. Quality of genome assembly

Genome sequencing data can be assembled in a variety of ways, and the quality of the assembly depends on the quality of the raw data, the type of assembly software, and the provided parameters. In order to make robust inferences from the increasing availability of draft genomes, it is critical to distinguish between genomes of varying quality (Mardis*, et al.*, 2002). The quality of isolate genomes has traditionally been evaluated using assembly statistics such as N50 (Salzberg*, et al.*, 2012). However, the accuracy and completeness of genome sequencing cannot be determined with N50, as it

only considers the length of the assembled contigs. In addition, contamination may occur during isolation, library preparation, and during sequencing (Laurence*, et al.*, 2014). To measure the quality of these genome assemblies, it is useful to use ubiquitous and single-copy genes as marker genes. Single-copy marker genes present multiple times within a recovered genome have also been used to estimate potential contamination (Albertsen*, et al.*, 2013). In many studies using genomic data, comparative genomics is an important tool for finding shared and unique genes between different genomes. However, depending on the quality of the genome, comparing the same two genomes may yield different results each time, and genome quality has widespread impacts on genome assembly, gene prediction, and functional annotation. Therefore, quality control is a very important and necessary step to reduce the bias of comparative genomics.

2.1.6. Average nucleotide identity

Due to the recent advancement in DNA sequencing technologies, using genome sequence data in prokaryotic taxonomy has gained a great momentum. One of the major achievements of genomics is in providing a gold standard in demarcating species of Bacteria and Archaea in lieu of DNA-DNA hybridization (DDH) (Oren and Garrity, 2014). For this purpose, several comparative measurements between two genome sequences, called overall

genome relatedness indices (OGRI), were developed and proposed to provide a cut-off or define boundaries between species (Chun and Rainey, 2014). Among them, average nucleotide identity (ANI) is most widely used with a proposed species boundary cutoff of 95~96% (Goris *et al.*, 2007; Richter and Rosselló-Móra, 2009; Kim *et al.*, 2014).

The currently accepted concept of ANI was proposed by Goris et al. (Goris *et al.*, 2007) in order to replace DDH by mimicking the experimental procedure of DDH. ANI is defined as a pairwise measure of overall similarity between two genome sequences. In the original method (Goris *et al.*, 2007), two genomes are differently treated as query and subject, respectively. The query genome sequence is fragmented in silico into 1,020 bp long sequences, and these fragments are then searched against the intact subject genome to find homologous regions. Identity values are calculated between query fragments and homologous regions of the subject genome. The final ANI value is the mean of identity values of all fragments of the query genome. The algorithm is composed of two tasks: searching query genome fragments against the subject genome, and calculating similarity between query genome fragments and their homologous counterparts in the subject genome.

The method of Goris et al. (Goris *et al.*, 2007) used BLAST program (BLASTN to be precise) (Altschul*, et al.*, 1990), which is recognized as a standard in prokaryotic taxonomy. Richter and Rosselló-Mora (Richter and Rosselló-Móra, 2009) suggested that MUMmer program (Kurtz*, et al.*, 2004),

which performs an ultrafast genome alignment, can also be used instead of BLAST. These algorithms using either BLAST or MUMmer calculate ANI values with a directional specificity, meaning that when a pair of genomes are compared, the calculated ANI value can be different depending on which genome was selected as a query, even though the differences are minor for most cases. An improved algorithm, named OrthoANI, was proposed to overcome this problem (Lee *et al.*, 2016). This new method also reduced the computational time as it does not require reciprocal calculations.

As the number of genomes in public databases are exponentially growing, there is an urgent need to critically evaluate the currently available software tools for ANI calculations in the light of accuracy and computational costs.

2.2. Materials and method

2.2.1. Improvement of 16S rRNA sequence based identification algorithm

To improve this algorithm, I used the database search program UBLAST (Edgar, 2010) instead of BLAST (Altschul *et al.*, 1990) **(Table 2)** and improved the pairwise alignment algorithm for similarity calculation. I had been using the pairwise alignment algorithm of ClustalW2 (Larkin*, et al.*, 2007) in existing algorithms. This algorithm uses pairwise alignment (Myers and Miller, 1988) to generate a consensus template, and subsequently performs another pairwise alignment with the generated template. If there are 3 or more sequences submitted for multiple sequence alignment, variable weights are used to calculate the alignment scores. Because only two sequences are used in pairwise alignment for the identification algorithm developed in this study, a static value was applied to the alignment score to improve alignment speed.

Table 2. Parameters of search engine software in identification algorithm.

Program	Version	Parameters
USEARCH	8.1.1861_i86 linux32	-evalue 1e-07 –threads 8 -strand both -accel 1
BLAST+	ncbi-blast-2.2.30+	blastn -evalue 1.0E-7 –num_threads 8 -num_alignments 5

2.2.2. Development of microbial taxonomic profiling (MTP) pipeline

In this study, microbial taxonomic profiling (MTP) pipeline was developed as a new pipeline for species level identification. The low-quality reads with under 25 average quality score were abandoned. For the Illumina MiSeq data, Illumina adapter sequences were trimmed and the low-quality reads with under 25 average quality score were abandoned by trimmomatic 0.32 (Bolger*, et al.*, 2014), and paired-end sequences were merged by PANDAseq 2.8.1 (Masella*, et al.*, 2012). For the Pacbio RS II data, Circular consensus (CCS) reads (Koren*, et al.*, 2012) were generated by SMRT Portal software (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/) with data specific minimum full pass and predicted accuracy cutoff parameters. The primer sequences were discarded and all reads were directed as 5` to 3` by primer location using in-house JAVA program. This process is applied differently depending on the NGS platform or PCR primer used, and is called pre-processing. After pre-processing, Non-specific amplicons that do not encode 16S rRNA and 16S rRNA sequencing variable regions for overall specific amplicons are detected by HMMER's hmmsearch program (Eddy, 2011) with 16S rRNA profiles. All sequences were processed by denoising, which is a correction of sequencing errors with adequate modeling using DUDE-Seq (Lee*, et al.*, 2015) and non-redundant reads were extracted by UCLUST-clustering (Edgar, 2010). Taxonomic identification was

assigned against the database using USEARCH (8.1.1861_i86linux32) (Edgar, 2010) followed by more precise pairwise alignment (Myers and Miller, 1988). The chimera sequences were detected by UCHIME (Edgar *et al.*, 2011). Only sequencing reads with lower than 97% similarity to the database were considered for chimera detection. Operational taxonomic units (OTUs) in the sample were investigated using open-reference method (Rideout*, et al.*, 2014) which used in QIIME pipeline with CD-HIT (Fu*, et al.*, 2012) and UCLUST (Edgar, 2010). The alpha diversity indices and rarefaction curves were estimated by in-house code. The overview of the MTP pipeline is provided in **Figure 5** and bioinformatics tools and parameters are shown in **Table 3**. In order to verify the newly developed pipeline, I performed a comparison analysis against the QIIME pipeline using human microbiome data, which includes genus *Faecalibacterium* and *Bacteroides*, known to live mainly in the intestinal environment.

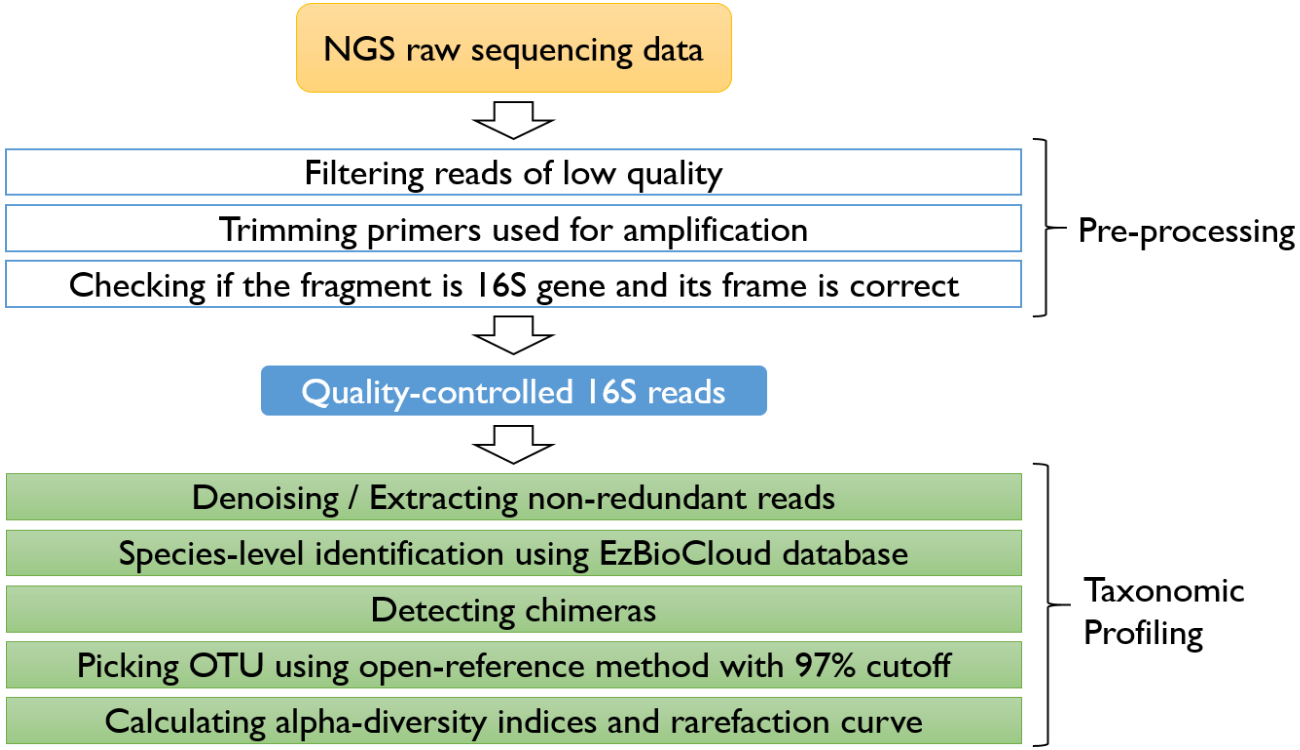

Figure 5. Overview of the MTP pipeline. In the MTP pipeline, taxonomic assignment is performed before OTU clustering, Because the purpose of MTP pipeline using EzBioCloud is species-level identification on type strain based species concept.

Table 3. Bioinformatics tools and parameters for the MTP pipeline.

Pipeline steps	Software and version	Platform	Parameters
Filtering reads by quality	Trimmomatic 0.32	Miseq	PE –threads 8 AVGQUAL:25 MINLEN:150
Merging paired-end	PANDAseq 2.8.1	Miseq	Default parameter
CCS processing	SMRT portal 2.3.0	Pacbio RS II	Minimum full pass:0~6 , Minimum predicted accuracy:90~99
Primer trimming	In-house JAVA code	All	Extention length:10, Cutoff identity:0.8
Check 16S amplicon and variable region	HMMSEARCH 3.1b2	All	--cpu 8 –noali –E 10
Denosing	Dude-seq	All	-k 5 (Only for Miseq : -p Illumina.pi)
Extracting unique reads	USEARCH 8.1.1861_i86linux32	All	-cluster_fast -sort length -id 1.0 -threads 8
Taxonomic assignment	USEARCH 8.1.1861_i86linux32 and In-house JAVA code	All	-usearch_global -threads 8 -evalue 1.0E-7 -strand plus -id 0.5 -maxaccepts 0 -maxrejects 0 -maxhits 10 -blast6out
Detecting chimera	USEARCH 8.1.1861_i86linux32	All	-uchime_ref -mindiv 1.5 -strand plus -threads 8
Taxonomic profiling	In-house JAVA code	All	species (x ≥ 97%), genus (97> x ≥94.5%), family (94.5> x ≥ 86.5%), order (86.5> x ≥82%), class (82> x ≥78.5%), and phylum (78.5> x ≥75%)
Picking OTUs	USEARCH 8.1.1861_i86linux32	All	-cluster_fast -sort length -id 0.97 -threads 8
	CD-HIT 4.6.1	All	cd-hit-est -T 8 –c 0.97
Estimating alpha diversity	In-house JAVA code	All	No specific parameters

2.2.3. Method for extracting full-length 16S rRNA genes from short-read sequencing data

In the first step, full-length sequencing reads containing 16S rRNA were extracted from unassembled short-read sequencing data by using a search (Nawrocki and Eddy, 2013) mechanism implementing the small subunit ribosomal RNA profile and covariance model from the Rfam database 12.0 (Nawrocki*, et al.*, 2015). In the second step, rockhopper2 (Tjaden, 2015) was used for assembly. Using the rockhopper2 algorithm, candidate 16S rRNA contigs are assembled from k-mers found in the sequencing reads. Because of Bruijn graph algorithm using fragmented k-mer, some candidate contigs may not be supported by full-length reads. Thus, sequencing reads are mapped to candidate contigs in order to filter candidate contigs into a set of high-quality final contigs that are well supported by full-length sequencing reads. Because any full-length sequencing read containing partial 16S rRNA were extracted in the first step, k-mer assembly in the second step may yield sequences longer than full 16S rRNA sequence. Thus, the small subunit ribosomal RNA profile of the Rfam database 12.0 (Nawrocki *et al.*, 2015) was used to re-extract only the 16S rRNA sequence from the assembled contigs. **Figure 6** shows the overall process of this method and previous method.

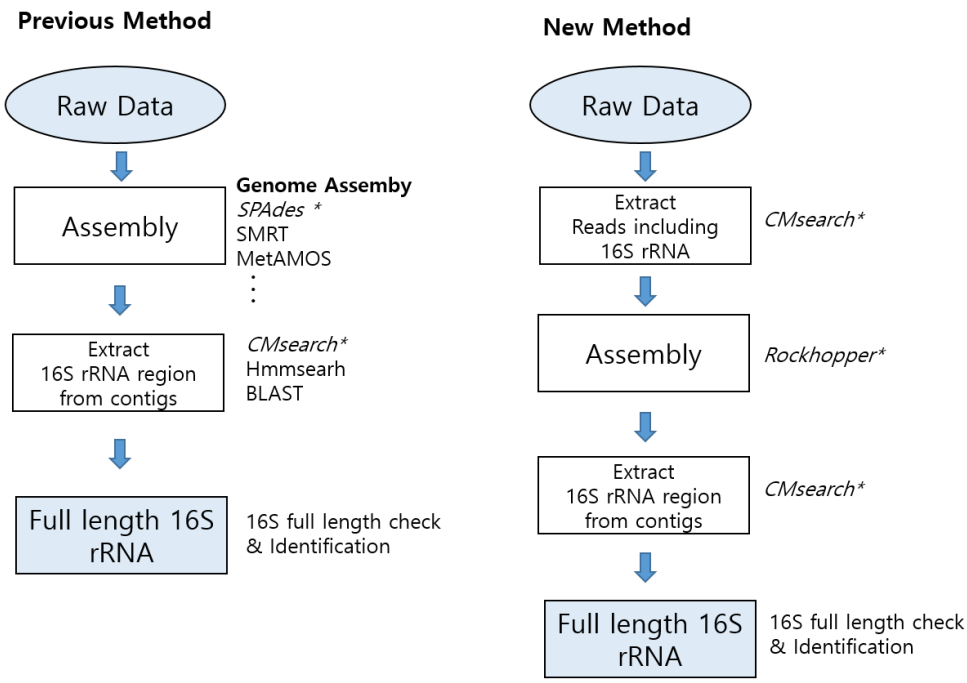

Figure 6. Comparison of 16S rRNA sequence extraction method. A comparison of the steps and software used for 16S rRNA extraction between the previous method and the new method can be seen in Figure 6. For a performance comparison of the two methods, 215 raw genome sequencing data files from NCBI SRA database were downloaded and analyzed.

2.2.4. Pipeline for prokaryotic whole genome analysis

The *de novo* assembly is the first step of prokaryotic whole genome analysis using NGS sequencing data. In the case of short-read sequencing data such as the output of Illumina sequencing platform, the software named SPAdes (Bankevich*, et al.*, 2012) was used for *de novo* assembly. SPAdes is a stable assembly software, which has been continually developed since launch to its most recent update on March 1st, 2017 to version 3.10.1. Many researchers use this software for assembly, and the original paper detailing it has been cited more than 1800 times to date. Version 3.7.1 was used in whole genome analysis (WGAS) pipeline for this research. SMRT Portal was used for the assembly of Pacific Bioscience genome sequencing data. For extracting 16S rRNA from short-read sequencing data, full-length 16S rRNA extraction method was used. This method doesn't require de novo assembly prior to 16S rRNA extraction and can be run in tandem with assembly. It then compiles 16S rRNA extracted from raw data together with the 16S rRNA extracted from de novo assembly results to calculate 16S rRNA completeness. Finally, the sequence with the higher completeness value becomes the representative 16S rRNA sequence for a particular genome.

The second step is finding the various patterns of gene's start and end location using assembled genome sequence called "Gene-finding". The gene-finding process for prokaryotic genome includes finding transfer RNA

(tRNA), ribosomal RNAs (rRNA), non-coding RNA, clustered regularly interspaced short palindromic repeat (CRISPR), protein-coding sequences (CDSs). CDSs were predicted by Prodigal 2.6.2 (Hyatt*, et al.*, 2010). Genes coding for tRNA were searched using tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997). The rRNA and other non-coding RNAs were searched by a covariance model search (Nawrocki and Eddy, 2013) with Rfam 12.0 database (Nawrocki *et al.*, 2015). CRISPRs were detected by PilerCR 1.06 (Edgar, 2007) and CRT 1.2 (Bland*, et al.*, 2007). The CDSs were classified into groups based on their roles, with reference to orthologous groups (EggNOG 4.5; http://eggnogdb.embl.de) (Huerta-Cepas*, et al.*, 2015). For more functional annotation, the predicted CDSs were compared with Swissprot (Consortium, 2014), KEGG (Kanehisa*, et al.*, 2015) and SEED (Overbeek*, et al.*, 2005) databases using UBLAST program (Edgar, 2010). The overview of whole genome analysis (WGAS) pipeline is shown in **Figure 7** and bioinformatics tools, parameters and databases for the pipeline is provided in **Table 4**.

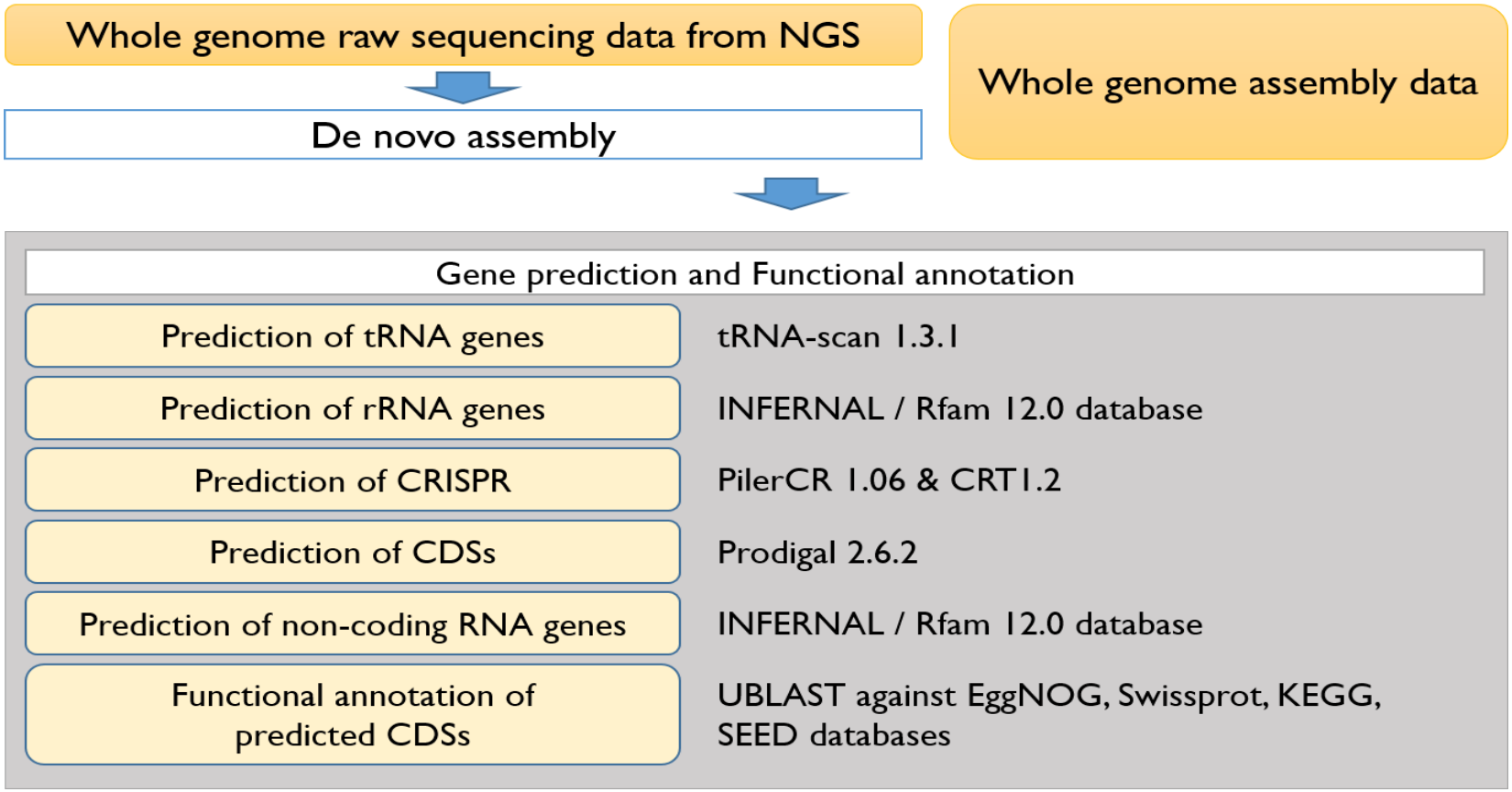

Figure 7. Overview of prokaryotic whole genome analysis (WGAS) pipeline.

Table 4. Bioinformatics tools, parameters and databases of pipeline for whole genome analysis.

Pipeline steps	Software and version	Platform	Parameters
Assembly	SPAdes 3.7.1	Miseq	--careful -k 21,33,55,77,99,127
	SMRT portal 2.3.0	Pacbio RS II	-noSplitSubreads -minReadLength 200 -maxScore -1000 -maxLCPLength 16
Finding tRNA	tRNA-scan 1.3.1	All	tRNA-scan-SE -bact
Finding rRNA	CMsearch / Rfam 12.0	All	-E 1.0E-5 –Z 700 –noali
Finding CRISPR	PiperCR 1.06	All	Default
	CRT 1.2	All	Default
Finding ncRNA	CMsearch / Rfam 12.0	All	-E 1.0E-5 -Z 700 –noali
Finding CDS	Prodigal 2.6.2	All	-f gff -m -c -g 11
Functional Annotation	USEARCH 8.0.1517_i86linux64	All	-ublast -maxaccepts 1 -evalue 1.0E-5 -accel 1.0 -ka_dbsize 700000000
Databases	**Realesd data or version**	**DB size**	**Annotated information**
SWISSPROT	2015.12.10	462655	UniProt id, functional note, product, description
EggNOG	4.5	7509316	EggNOG id, category, functional description
KEGG	2015.12.10	5756709	KEGG id, pathway, functional description
SEED subsystems	2015.12.10	2265008	Subsystems hierarchy, functional description

2.2.5. Methods for the quality assessment of genome

The complete genome sequence of *Escherichia coli* K-12 MG1655 strain is one of the most studied prokaryotic genome sequences (Blattner*, et al.*, 1997). It is often used as a reference genome for Bacteria in general because the locations and functions of genes in this genome have been studied extensively and proven experimentally. In this study, I carried out reference-based annotation on about 64,280 publicly released bacterial genomes for quality assessment by performing BLAST searches (E-value cutoff: 1e-05) for genes that matched the genes in the *Escherichia coli* K-12 MG1655 genome. The genes of genomes assessed as above were assigned annotations according to the gene abbreviations used for *E. coli* genes, and this data was used to calculate gene frequencies for all genomes in the complete genome list. A Hidden Markov Model (HMM) based profile was created by taking the high-frequency genes, extracting the corresponding sequences for those genes in each genome, sorting those sequences by gene type, and performing an alignment. Using these HMM profiles (Eddy, 2011), highly specific genes within the complete genome list set were extracted and as a result, 54 ubiquitous and single copy genes were listed (Seong-In Na, personal communication). These 54 genes were designated as Bacterial Core Genes (BCGs) **Table 5**, and I built a database containing information on the presence/absence of BCGs and their cardinality in each genome of the genome list. If the annotated genome had a complete genome

at the species level or genus level, the complete genome was used as a reference for BCG counts, and any genome with less than 80% BCGs of its complete counterpart was categorized as a low-quality genome.

In order to find out the contaminated genome, the method of ContEst16S (Lee*, et al.*, 2017), a fast and simple method using 16S rRNA gene, was also applied. The sequence similarity was obtained for all pairwise combinations of two or more 16S rRNA sequences over 500bp in a single genome. If a pair of fragments is not aligned at all or by at least 400bp, the calculation is ignored. Among considered pairs, if the two fragments differed by >5% in sequence similarity, each fragment was identified against EzBioCloud 16S rRNA database containing type strains and representatives of phylotypes. If two identified results showed >97% similarities to different sequences in genus level, the genome containing these 16S rRNA sequences was determined as 'Contaminated'. It does not mean that all other cases are not contaminated. This algorithm can only detect the surely contaminated genome sequences, but cannot determine if it is absolutely free of contamination.

CheckM, published in 2015, is another method of assessing the quality of a genome using a set of pre-calculated marker genes by lineage from trusted reference genomes (Parks*, et al.*, 2015). This method does not judge whether the genome is of low-quality, but provides only the score values for completeness, contamination and strain heterogeneity. In this study, the

score values analyzed using the CheckM were compared with those using the ContEst16S and the BCGs.

Table 5. 54 Bacterial core gene (BCG) list and product.

Gene Name	Product	Gene Name	Product
rpsT	30S ribosomal protein S20	**rplR**	50S ribosomal protein L18
rpsB	30S ribosomal protein S2	**rplF**	50S ribosomal protein L6
tsf	Elongation factor Ts	**rpsH**	30S ribosomal protein S8, chloroplastic
frr	Ribosome-recycling factor	**rplE**	50S ribosomal protein L5
recR	Recombination protein RecR	**rplX**	50S ribosomal protein L24
pth	Peptidyl-tRNA hydrolase	**rplN**	50S ribosomal protein L14
pheS	Phenylalanine--tRNA ligase alpha subunit	**rpsQ**	30S ribosomal protein S17
rplT	50S ribosomal protein L20, chloroplastic	**rplP**	50S ribosomal protein L16
infC	Translation initiation factor IF-3	**rpsC**	30S ribosomal protein S3
rnc	Ribonuclease 3	**rplV**	50S ribosomal protein L22
rplS	50S ribosomal protein L19	**rpsS**	30S ribosomal protein S19, chloroplastic
trmD	tRNA (guanine-N(1)-)-methyltransferase	**rplB**	50S ribosomal protein L2
rpsP	30S ribosomal protein S16	**rplW**	50S ribosomal protein L23
smpB	SsrA-binding protein	**rplD**	50S ribosomal protein L4
pgk	Phosphoglycerate kinase	**rplC**	50S ribosomal protein L3
rpsO	30S ribosomal protein S15	**rpsJ**	30S ribosomal protein S10
truB	tRNA pseudouridine synthase B	**rpsG**	30S ribosomal protein S7
rpmA	50S ribosomal protein L27	**rpsL**	30S ribosomal protein S12, chloroplastic
rplU	50S ribosomal protein L21	**rplK**	50S ribosomal protein L11
rpsI	30S ribosomal protein S9	**rplA**	50S ribosomal protein L1
rplM	50S ribosomal protein L13	**rplJ**	50S ribosomal protein L10
rplQ	50S ribosomal protein L17	**rplL**	50S ribosomal protein L7/L12
rpsK	30S ribosomal protein S11	**rpoB**	DNA-directed RNA polymerase subunit beta
rpsM	Cyanelle 30S ribosomal protein S13	**rpoC**	DNA-directed RNA polymerase subunit beta
secY	Protein translocase subunit SecY	**rpsF**	30S ribosomal protein S6
rplO	50S ribosomal protein L15	**rpsR**	30S ribosomal protein S18
rpsE	30S ribosomal protein S5	**rplI**	50S ribosomal protein L9

2.2.6. Efficient calculation method for average nucleotide identity

Four ANI algorithms were compared: (i) ANIb, the original algorithm using BLASTN (Goris et al., 2007), (ii) ANIm, the algorithm using MUMmer (Kurtz et al., 2004; Richter and Rosselló-Móra, 2009), (iii) OrthoANIb, the orthologous ANI algorithm using BLASTN (Lee et al., 2016) and (iv) OrthoANIu, the orthologous ANI algorithm using USEARCH program (Edgar, 2010). Mean values were obtained from the reciprocal calculations for ANIb and ANIm methods.

A total of 49,734 quality-controlled genome sequences belonging to 132 genera were chosen from EzBioCloud database (http://www.ezbiocloud.net/) (Yoon*, et al.*, 2016) for ANI calculation. ANI values were computed only for pairs of genomes that belonged to the same genus in which at least 20 genomes were available. The computing run-times were measured using computers with 8 Core i7-6700 3.4GHz CPU (Intel).

All four algorithms were implemented using JAVA programming language (http://www.java.com/) and run on the Linux operating system. The R package was used for all statistical analyses (https://www.r-project.org/). The versions and parameters for each program are given in **Table 6.**

Table 6. Version and run parameters used in comparison of four average nucleotide identity methods.

Program	Version	Parameters
USEARCH	8.1.1861_i86 linux32	-usearch_local -id 0.5 -strand both -evalue 1.0E-15 -maxaccepts 1 -xdrop_g 150 -mismatch -1 -match 1 -dbaccelpct 100 -qmask none -dbmask none
BLAST+	ncbi-blast-2.2.30+	blastn -evalue 1.0E-15 -dust no -xdrop_gap 150 -penalty -1 -reward 1
MUMmer	3.23	nucmer --mum -l 20 -b 200 -c 65 -g 90 --optimize -p

2.3. Results

2.3.1. Advanced microbial taxonomic profiling (MTP) pipeline

2.3.1.1. Computational efficiency

Improving the pairwise alignment algorithm resulted in approximately 1.5 times faster alignment speeds with no difference in the calculated similarity values. I ran the alignment algorithm on 1000 unique 16S rRNA V3-V4 region sequences (primarily used in Illumina Miseq NGS platform) and on 1000 unique full 16S rRNA V1-V9 sequences (primarily used by Pacbio RS II NGS platform), and found that for V3-V4 regions, the current algorithm took 0.0175 seconds per alignment on average, versus 0.0114 seconds using the improved version; for full-length sequences, the current algorithm took 0.1946 seconds per alignment on average, whereas the improved algorithm took 0.1256 seconds. **Figure 9** shows the performance of alignment algorithms for various length reads obtained from NGS.

The improved pairwise alignment algorithm was also used in the development of the microbial taxonomic profiling (MTP) pipeline, and in comparison with the pipeline (Jeon et al., 2013) that was developed to

analyze 454 pyrosequencing 16S rRNA amplicon data using EzTaxon database (Kim et al., 2012), the MTP pipeline showed improved processing speeds for analyzing high-throughput sequencing data from NGS platforms such as Illumina Miseq, as shown in **Figure 8**.

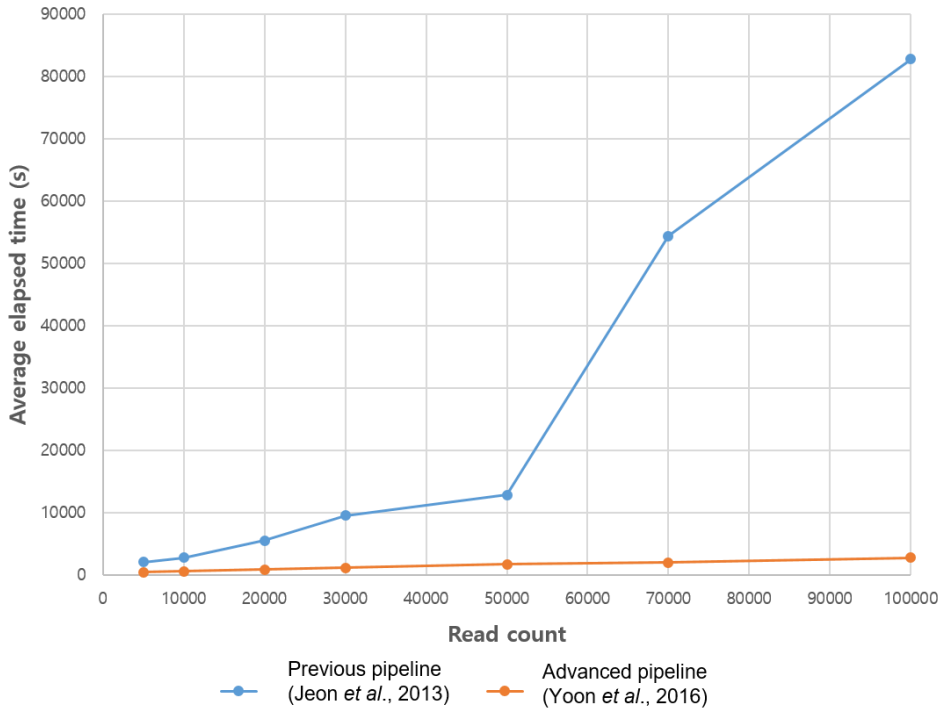

Figure 8. Comparison of MTP pipelines. The computing environment used for comparisons consisted of a i7-6700 3.40GHz 8-core CPU and 32Gb of RAM. Each average elapsed time was calculated after the rigorous test, where the test was performed over 30 times. The pipelines showed the most marked performance difference when analyzing high-throughput data with more than 50,000 reads, and the advanced MTP pipeline developed in this study demonstrated a controlled linear increase in computing time with an increase in read count.

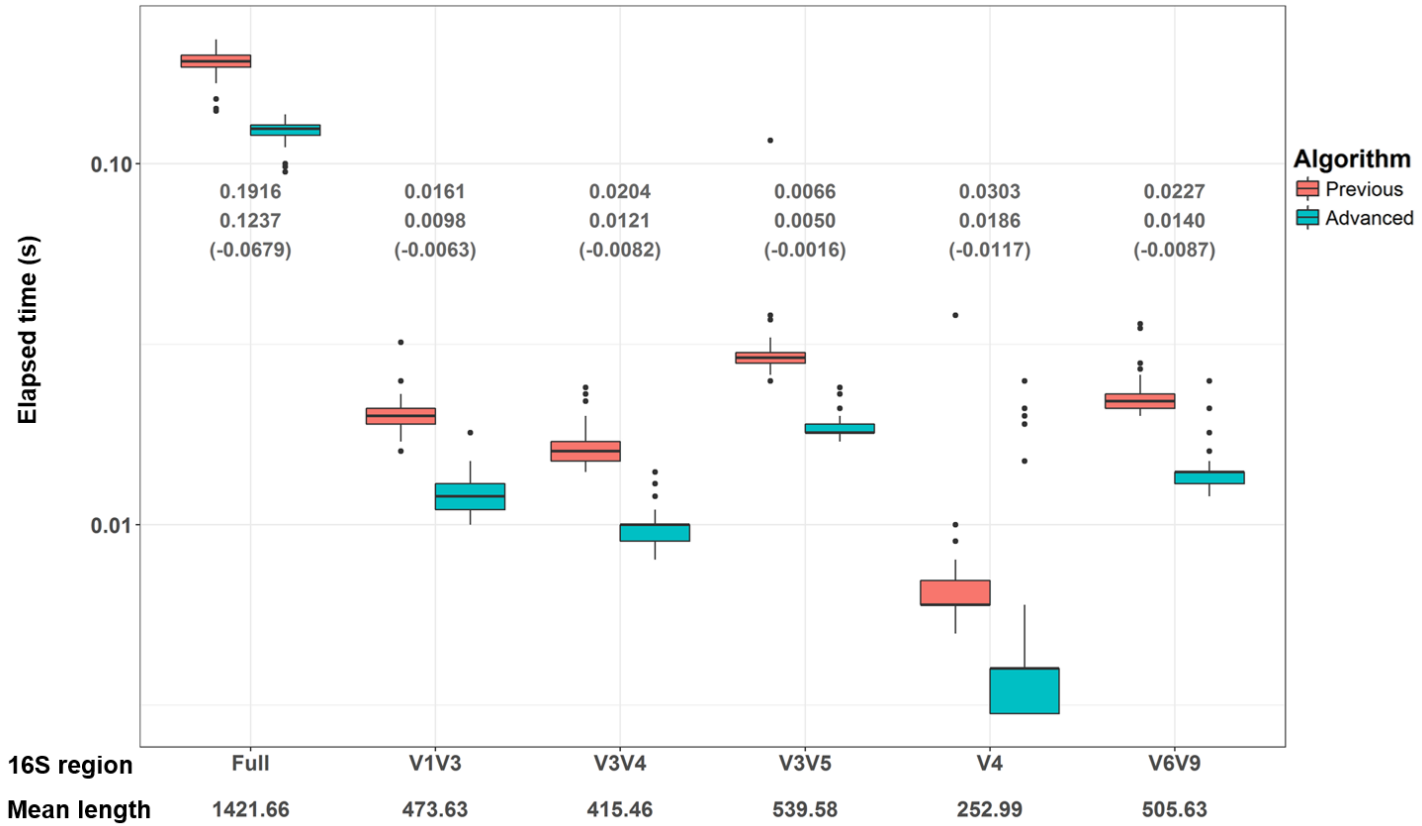

Figure 9. Performance of alignment algorithm for various read length.

2.3.1.2. Taxonomic profiling comparison between MTP and QIIME pipeline

EzBioCloud 16S database and MTP pipeline are designed for optimal performance in species-level identification, even when there is a clear limitation due to the lack of differences. The combination of EzBioCloud and bioinformatics pipelines allows us a species-level exploration of human microbiome data as well as environmental samples. The genus *Facalibacterium* is known as a major player in the human gut environment (Louis and Flint, 2009). Unfortunately, there is only one species name that is officially recognized, that is, *Faecalibacterium prausnitzii*. This genus is further classified into 48 species on the basis of 16S and genome sequences in EzBioCloud database (Yoon *et al.*, 2016). In the comparison shown in **Figure 10**, the genus level compositions are very similar for *Faecalibacterium* **(Figure 10(A))**. However, species-level compositional differences are dramatic **(Figure 10(B))**. Using EzBioCloud and MTP, I can identify 5 different *Faecalibacterium* species. Another genus that is abundant in human fecal samples is the genus *Bacteroides*, also a major inhabitant of the human gut (Arumugam*, et al.*, 2011). The compositions at the genus level are similar to the *Faecalibacterium* case **(Figure 11)**. However, species-level compositions are significantly different from results between two databases. *B. coprocola*, *B. massiliensis*, *B. nordii*, *B. thetaiotaomicron*, *B.vulgatus*, and DQ798855_s, which is an uncultured phylotype, were recognized only by

EzBioCloud and MTP **(Figure 11(B))**. In the QIIME (Caporaso *et al.*, 2010) pipeline with Greengenes (DeSantis *et al.*, 2006), all of these species were recognized as unclassified species. At the time of writing, EzBioCloud 16S database contains 272 species of the genus *Bacteroides* and 557 species of the genus *Prevotella*. It is far larger than any other database at the species level classification and it can support more accurate microbial community analysis.

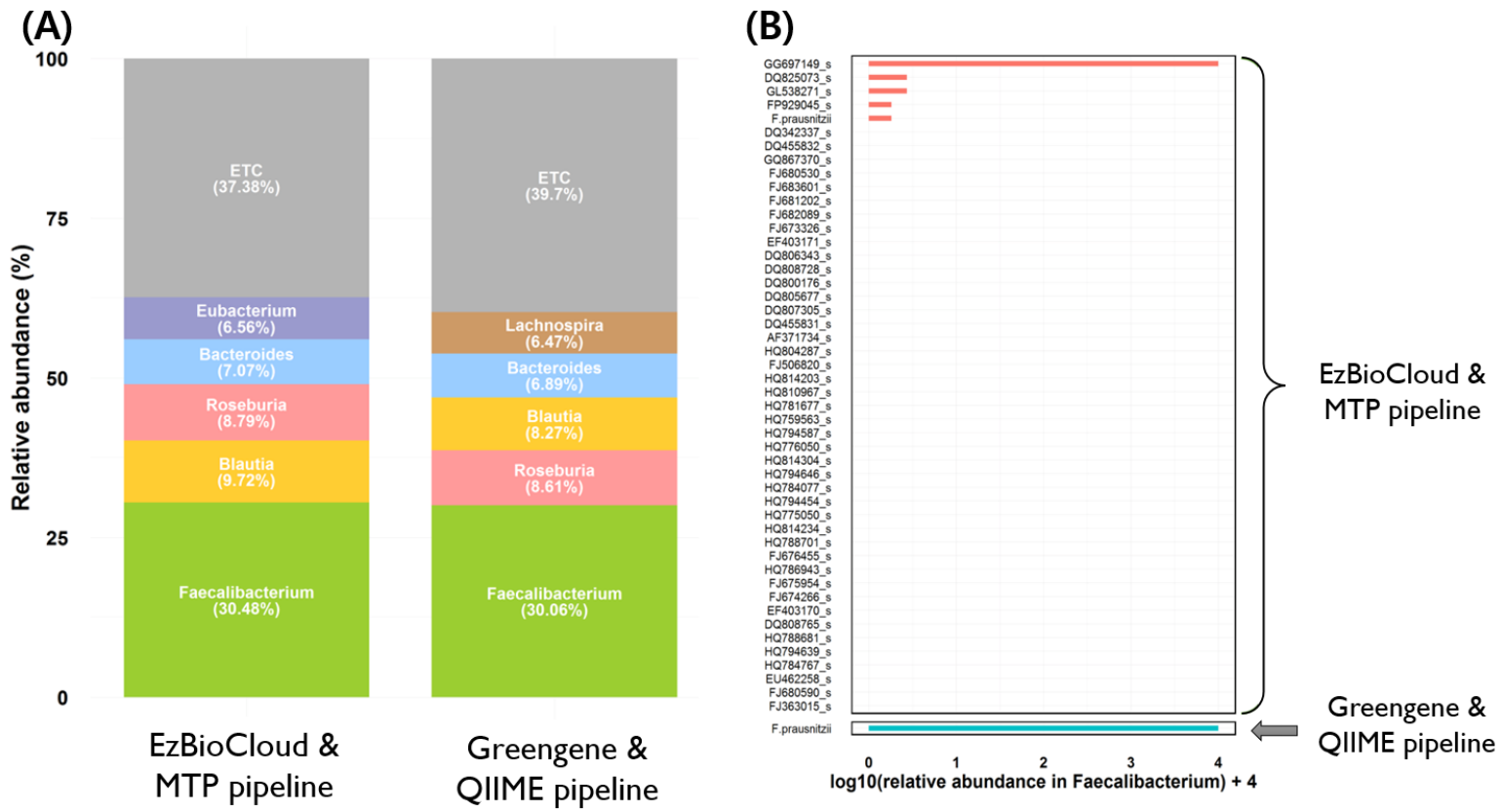

Figure 10. Comparison of taxonomic profiling for a *Facalibacterium* case.

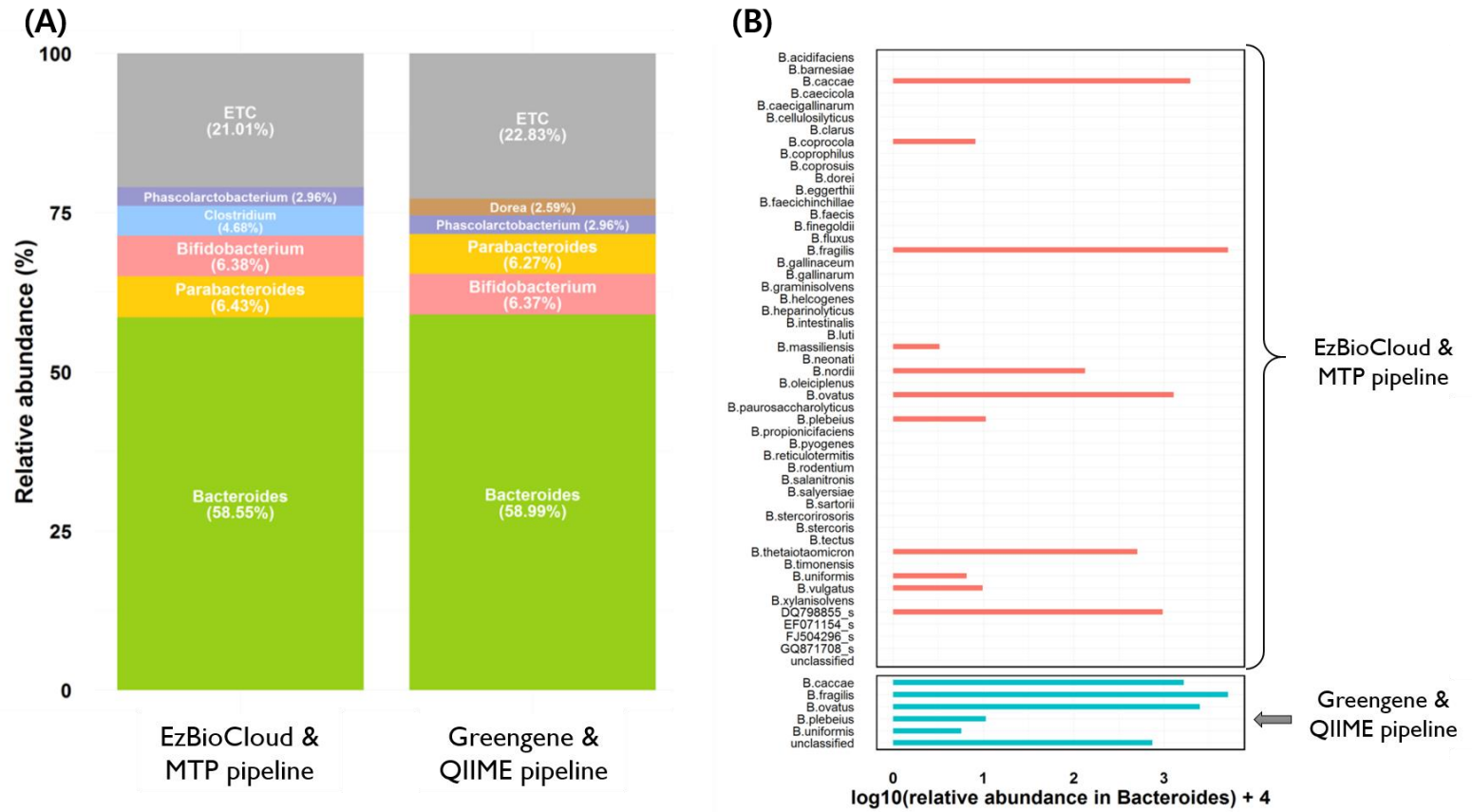

Figure 11. Comparison of taxonomic profiling for a *Bacteroides* case.

2.3.2. Comparison of full length 16S rRNA extraction methods

Depending on the quality of the data, full-length 16S rRNA sequence fragments may not be fully included in draft genome sequencing data. Thus, a cutoff of 0.9 completeness was used as the criteria for a full-length 16S rRNA sequence in this study. Using a total of 205 genome sequencing data samples, the newly developed method extracted 204 full-length 16S rRNA sequences, compared to only 192 extracted using the current existing method. Both methods failed to extract 16S rRNA sequence from 1 sample data, and the 12 data samples with differing results were evaluated by identification.

Using the current method, 10 of the 12 data samples had yielded partial 16S rRNA, resulting in low completeness scores, but their identification results still matched the organism name listed in NCBI. However, the existing method failed to produce a high fidelity 16S rRNA sequence for SRR1200912, resulting in a low 58.8% similarity hit, failing to meet the 97% similarity cutoff for identification. For SRR3176161, 16S rRNA identification yielded a top hit that was a different species, as shown in **Table 7**. In stark contrast, using the newly developed method resulted in full-length 16S rRNA sequences with completeness values of 1, and identification results showed similarities of at

least 99% in all cases. Furthermore, all identification results matched the data samples' original names at the species level, as shown in **Table 8**.

The existing method extracted partial 16S rRNA sequences from 12 data samples out of 205, only 0.06% of the time. However, the accurate identification of genomes plays a critical role in comparative genomics and industrial applications. Therefore, an accurate full-length 16S rRNA sequence extraction method likes the one developed in this study, is essential.

Table 7. The result of 16S rRNA extraction using the previous method. In previous method, genome sequencing raw data was de novo assembled by SPAdes 2.7.1 and 16S rRNA sequence was extracted by CMsearch from contigs. SRR1200912 and SRR3176161 shows critical errors in identification.

SRA accession	Seq length	Completeness	Similarity	Taxon name of top hit	NCBI organism name
SRR264183	968	0.667	1	Bacteroides ovatus	Bacteroides ovatus ATCC 8483
SRR446823	1083	0.727	0.996	Salmonella enterica subsp. enterica	Salmonella enterica subsp. enterica serovar Montevideo str. CT_02035321
SRR2143479	1001	0.671	0.998	Salmonella enterica subsp. enterica	Salmonella enterica subsp. enterica str. ADRDL-LA-38-2014
DRR014735	1130	0.755	1	Bacillus anthracis	Bacillus anthracis (isolated in Zambia)
SRR1200912	1137	**0**	**0.588**	Orientia tsutsugamushi	Orientia tsutsugamushi str. TA716
SRR3217427	1067	0.704	0.998	Citrobacter freundii	Citrobacter sp. AATXR
SRR3176161	781	0.526	1	**Enterococcus hirae**	**Enterococcus faecium**
SRR1055838	608	0.411	1	Clostridium difficile	Clostridium difficile 5.3
SRR1185964	893	0.615	0.997	Clostridium difficile	Clostridium difficile 19.3
DRR014739	1041	0.693	1	Bacillus anthracis	Bacillus anthracis (isolated in Zambia)

Table 8. The result of 16S rRNA extraction using developed method. In developed method, SRR1200912 and SRR3176161 which showed errors were normally identified.

SRA accession	Seq length	Completeness	Similarity	Taxon name of top hit	NCBI organism name
SRR264183	1529	1	1	Bacteroides ovatus	Bacteroides ovatus ATCC 8483
SRR446823	1540	1	0.994	Salmonella enterica subsp. enterica	Salmonella enterica subsp. enterica serovar Montevideo str. CT_02035321
SRR2143479	1542	1	0.993	Salmonella enterica subsp. enterica	Salmonella enterica subsp. enterica str. ADRDL-LA-38-2014
DRR014735	1552	1	1	Bacillus anthracis	Bacillus anthracis (isolated in Zambia)
SRR1200912	1504	1	0.997	Orientia tsutsugamushi	Orientia tsutsugamushi str. TA716
SRR3217427	1542	1	0.998	Citrobacter freundii	Citrobacter sp. AATXR
SRR3176161	1560	1	0.999	Enterococcus faecium	Enterococcus faecium : Colony4
SRR1055838	1503	1	0.997	Clostridium difficile	Clostridium difficile 5.3
SRR1185964	1503	1	0.997	Clostridium difficile	Clostridium difficile 19.3
DRR014739	1552	1	1	Bacillus anthracis	Bacillus anthracis (isolated in Zambia)

2.3.3. Annotation of public genomes

A total of 75,386 public whole genome assemblies (WGAs) were re-annotated by the WGAS pipeline **(Figure 7).** As a result of the WGAS pipeline, the number of predicted protein-coding sequences (CDSs) in 24,922 genomes was different from the annotation of NCBI. In the case of the *Rhodopirellula baltica* SH1 (GCF_000196115) genome, the NCBI annotation resulted in 7,325 CDSs, but the WGAS pipeline predicted 5,475 CDSs **(Table 9)**. In the genome of the other 3 strains of *Rhodopirellula baltica* species (WH47, SH28, SWK14), the number of CDSs did not show any significant difference between NCBI annotation and WGAS pipeline. The annotation pipeline used for GCF_000196115.1 on the NCBI GenBank file was PEDANT (Frishman*, et al.*, 2001), and the pipeline used for the other three strains was used for the different analysis pipeline for each of MicHanThi (Quast, 2006). In the case of the above, annotation of the public genome with different pipelines will provide inconsistent analysis results, and differences in these annotation results may lead to a critical bias in comparative genomics studies.

Table 9. The number of predicted CDSs for 4 genomes of species *Rhodopirellula baltica*.

NCBI assembly accession	Strain name	Genome size (bp)	Pipeline for NCBI CDSs	Number of NCBI CDSs	Number of WGAS CDSs
GCF_000196115	SH1	7,145,576	PEDENT	**7,325**	**5,475**
GCF_000195185	WH47	7,033,319	MicHanThi	5,675	5,589
GCF_000304635	SH28	7,145,707	MicHanThi	5,774	5,576
GCF_000330745	SWK14	7,488,930	MicHanThi	6,065	5,880

2.3.4. Quality of bacterial genomes

A Total of 64,280 public genomes were checked by bacterial core genes (BCGs) and 306 species that belong to 137 different genera did not have all 54 genes from the BCG list. **Table 10** shows the list of top ten genera with the lowest number of BCGs. Among them, the *Nasuia* genus had the least amount of BCGs by each complete genome having 32 BCGs. These genomes are also the smallest and are only 110kb long. However, the genome size did not indicate any meaningful correlation with the BCG numbers by only showing a 0.06 R2 value **(Figure 12)**. By the assessment algorithm, 1394 genomes (2.17% of the total analyzed genomes) were selected as 'LOW BCG'. Among the low BCG genomes, 59 were genomes of the same species with erroneous assembly sizes, and the other 1335 were genomes that had low BCG counts despite their normal genome sizes. Through these results, I can confirm that a genome's assembly statistics alone isn't a sufficient indicator of genome data quality and that this proposed method can add insight to, or aid the quality check process.

There were 674 (1.05%) genomes identified as contaminated by the ContEst16S algorithm and out of those genomes 169 of them carried a eukaryotic rRNA which results in 25.07% of the contaminated genomes. By analyzing 674 contaminations, all top five contaminants were found to be

originated from the eukaryotic small subunit while the *Enterococcus* genus was the main contaminant for Bacteria **(Table 11)**.

Table 10. Top ten genus with the lowest number of BCGs. The Nasuia genus Nasuia was determined to have the smallest genome size and the fewest BCGs while the complete genomes of the Tremblaya genus had an average of 37.5 BCGs. Despite their large genome size, the genus Chloracidobacterium and the Desulfobulbus genus have relatively few BCGs. Collectively, these observations prove that the existences of the BCGs have no significant correlation with the genome size.

Genus Name	Number of BCGs	Genome size (bp)
Nasuia	32	112091
Carsonella	34	159662
Hodgkinia	36	140570
Tremblaya	37.5	138927
Chloracidobacterium	40	**3695372**
Sulcia	41	245530
Desulfobulbus	43	**3851869**
Zinderia	44	208564
Uzinura	46	263431
Thermoanaerobacter	46	**2306092**

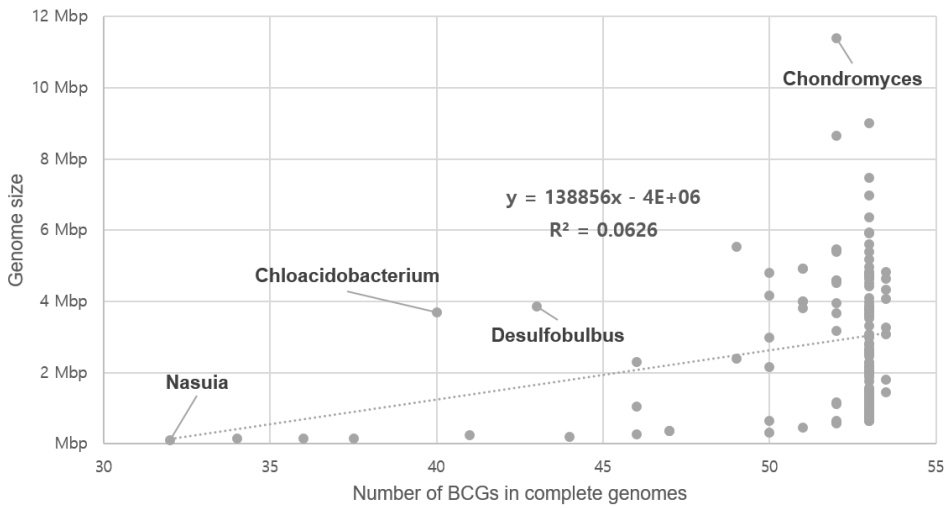

Figure 12. Correlation between genome size and number of BCGs. Although there is a trend of bigger genome sizes resulting in higher BCG numbers, some genera with big enough genome sizes showed small BCG numbers such as the *Chloacidobacterium* and the *Desulfobulbus*. Additionally, numerous small genomes that are no bigger than 2Mbp had 50 or more BCGs which all contributes to the insignificant correlation between the genome size and the BCG numbers with a R^2 value of 0.06.

Table 11. The most frequent contaminants from 674 contaminated genomes predicted by ContEst16S.

Organism name	Domain	Count	Ratio
Saccharomyces cerevisiae	Eukarya	45	4.85%
Homo sapiens	Eukarya	28	3.02%
Ophiorrhiza pumila	Eukarya	19	2.05%
Arabidopsis thaliana	Eukarya	13	1.40%
Glycine max	Eukarya	11	1.19%
Agathobacter rectalis	Bacteria	7	0.76%
Enterococcus faecalis	Bacteria	7	0.76%
Staphylococcus epidermidis	Bacteria	7	0.76%
Bacillus anthracis	Bacteria	6	0.65%
Enterococcus faecium	Bacteria	6	0.65%
Others	-	778	83.93%
Total		**927**	**100.00%**

Out of the 64,280 genomes, 1394 genomes were determined to be low-quality genomes by using the BCG and a comparative analysis was conducted between this and the CheckM completeness scores. Also, a comparative analysis of the 674 genomes that were identified as contaminated by the ContEst16S and the CheckM contamination scores was carried out **(Figure 13)**. The distribution of the completeness scores varied heavily with the low-quality genomes such that in the case of *Shigella sonnei* GCA_001413795.1 genome which has a 99.84 CheckM complete score only had 33 BCGs. According to these results, there is not a single perfect way of determining low-quality genomes and therefore the method that uses BCG can also be an appropriate way of doing this. The median contamination score of all the genomes that were deemed contaminated by the ContEst16S algorithm was 1.05 which means half or more of these couldn't have been discovered by the CheckM contamination score. Thus ContEst16S can be a useful method in finding contaminated genomes.

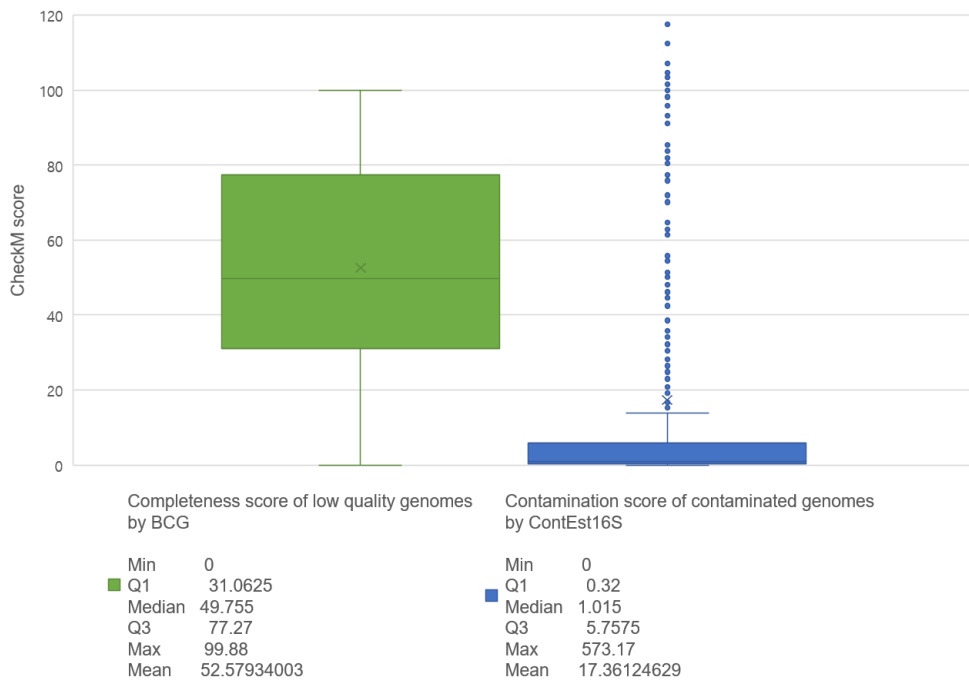

Figure 13. CheckM scores of low-quality or contaminated genomes. This figure shows the distribution of CheckM completeness score of the low-quality genomes that were determined with BCG and the distribution of CheckM contamination scores of the contaminated genomes that were determined with the ContEst16S.

2.3.5. Evaluation of algorithms for average nucleotide identity

In this study, three external programs were used for ANI calculation, namely BLASTN, MUMmer and USEARCH. USEARCH is a software tool that is known to perform orders of magnitudes faster than BLASTN with comparable search results. I reckon that mean ANIb values from reciprocal calculations are the standard for taxonomic purposes. Therefore, other three ANI algorithms were evaluated to see if they produce similar values to that of the ANIb.

A total of 107,442 pairs of genomes were subjected to ANI calculations by four algorithms. I was not able to obtain ANIm values for fifty-two pairs due to the unrecoverable errors caused by MUMmer program. When three algorithms were compared to ANIb in the whole range of ANI values **(Figure 14)**, OrthoANIb and OrthoANIu showed good correlations to ANIb; Both R2 and r values were over 0.999. An exception, ANIm, exhibited relatively low correlation values and showed many false positives for ANI species cutoff(>95%) even though ANIb values were lower than 70%. However, in the range of both compared ANI values over 90%, all three methods correlated well with standard ANIb values **(Figure 15)** with high R2 and r coefficients, even though ANIm still showed the least correlation relative to other methods. Because the major purposes of ANI in prokaryotic taxonomy

is the demarcation of species where the cutoff of around 95-96% is often applied (Richter and Rosselló-Móra, 2009; Chun and Rainey, 2014), the comparative study summarized in **Figure 14** and **Figure 15** clearly indicates that all three algorithms are suitable to replace ANIb in the range that matters most for the taxonomic purposes. However, ANIm should not be used for distantly related genomes (ANI of <90%).

Computational cost or run-time is a critical issue when a large number of genomes is considered. First, I compared the run-times of four algorithms using the whole dataset (107,442 pairs of genomes). The average size of genomes used for calculations is 3,735,585 ± 1,486,776 bp. Statistical comparison of run-times among four algorithms are given in **Figure 16(A)**. In terms of mean run-times, ANIm (11.08 s, 4.7X faster than ANIb) was the fastest, followed by OrthoANIu (18.57 s, 2.8X), OrthoANIb (21.65 s, 2.4X) and ANIb (52.61 s). The latter two are slower than the former two because BLASTN program runs substantially slower than MUMmer and USEARCH.

The run-times to compute ANI may depend on the sizes of genomes. Indeed, when only genomes with ≥ 7Mbp were considered, the differences in run-times among four methods were considerably enlarged **(Figure 16(B))**. Again, ANIm was the fastest at 40.51 s (52.9X faster than ANIb), followed by OrthoANIu (98.44 s, 21.8X), OrthoANIb (1371.99 s, 1.6X) and ANIb (2142.86 s). It is clear that ANIm and OrthoANIu are more suitable than BLAST-based algorithms when ANI values between large genomes are computed.

A web-service that can be used to calculate OrthoANIu between a pair of genome sequences is available at http://www.ezbiocloud.net/tools/ani. For large scale calculation and integration in bioinformatics pipelines, a standalone JAVA program is available for download at http://www.ezbiocloud.net/tools/orthoaniu.

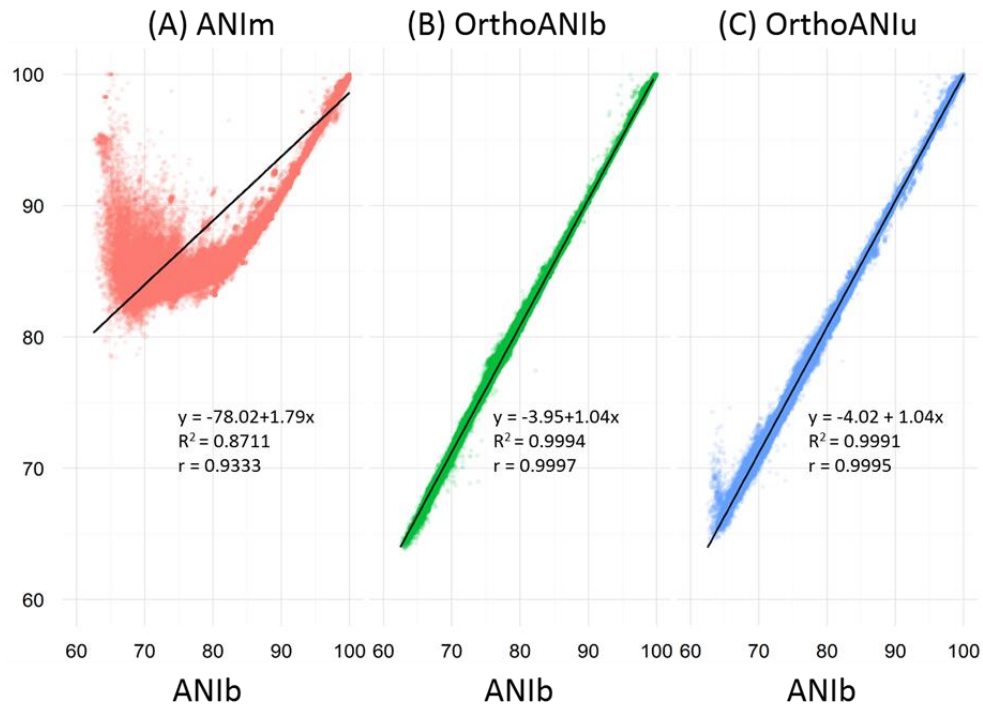

Figure 14. Correlation between the ANIb and other algorithms in the whole range of ANI. The plots were generated from 107,442 pairs of genomes.

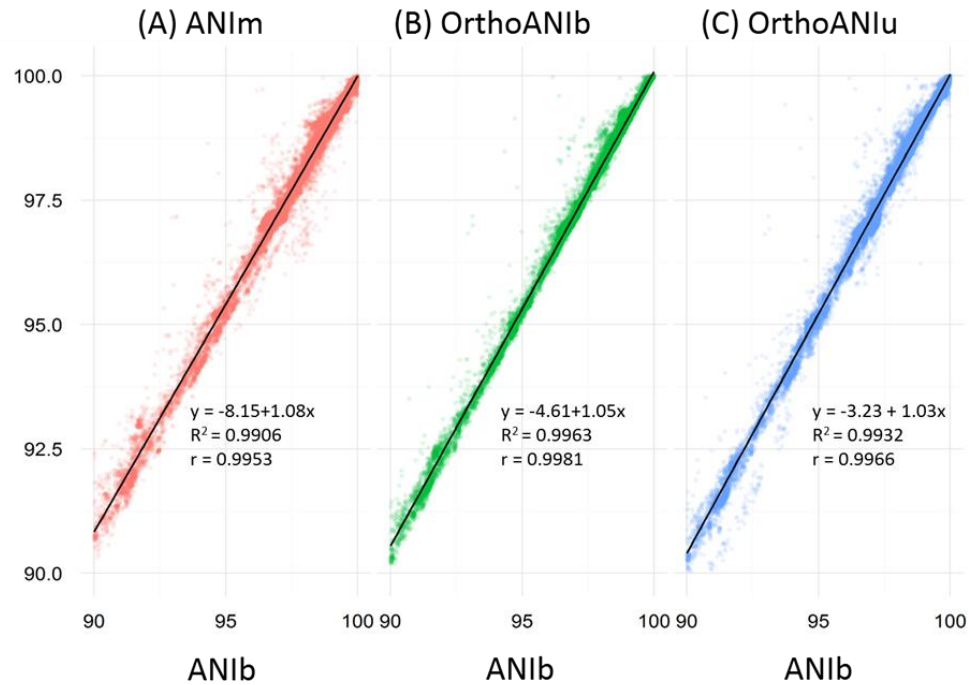

Figure 15. Correlation between the ANIb and other algorithms in the range of >90% ANI. The plots were generated from 54,236 pairs of genomes.

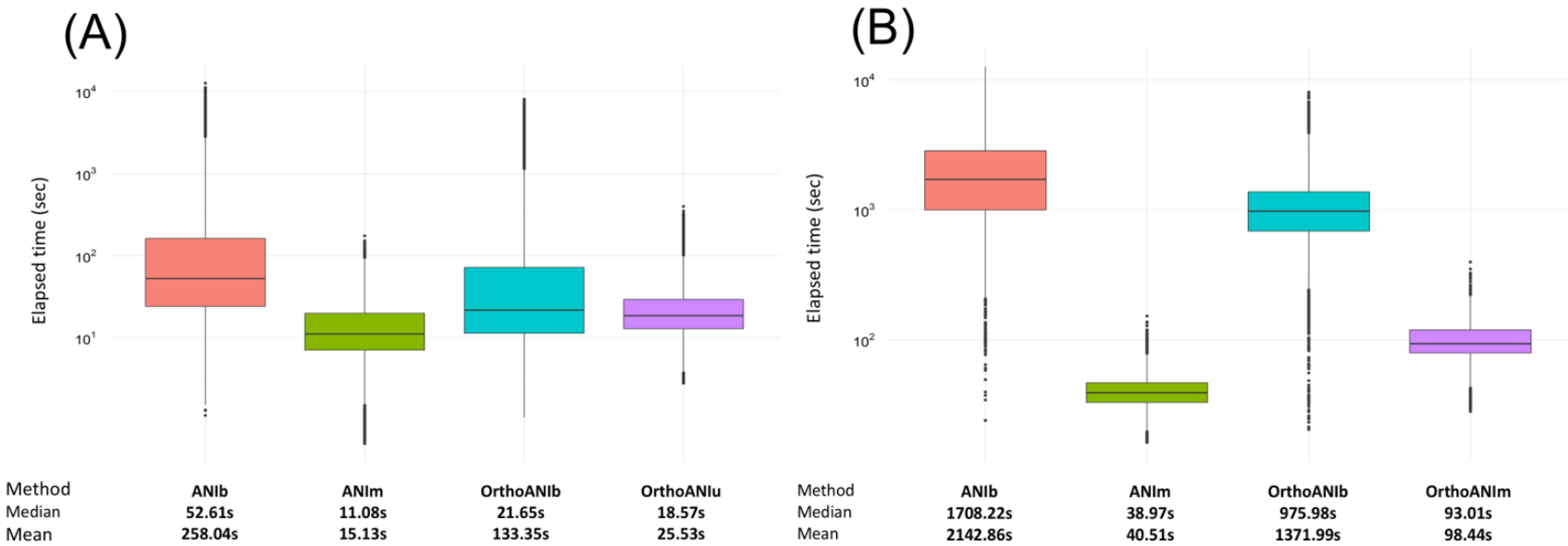

Figure 16. Running times of four ANI algorithms. Boxplots showing run-times of four algorithms based on (A) 107,442 pairs of genomes with all range of genome sizes and (B) 4,509 pairs of genomes whose sizes are $\geq$ 7 Mbp.

2.4. Discussion

Various methods for prokaryotic NGS data analysis were introduced and evaluated in this chapter. By using the improved MTP pipeline, every sequencing platforms' 16S rRNA amplicon sequencing data can be analyzed to the species level in a short amount of time with the help of EzBioCloud 16S rRNA database. Especially for MiSeq, which recently has been the most popular high-throughput platform for microbial community analysis, a quick and precise analysis can be carried out. One important issue with 16S rRNA amplicon analysis using MiSeq is alpha diversity. Despite its low sequencing error rate, due to its ultra-high-throughput data, it can contain a high number of errors. These occurrences of random errors cause an overestimated alpha diversity. With these types of errors, simple error correction methods such as noise reduction and quality checking are often not enough and omitting singletons from the OTU calculation is more appropriate (Dickie, 2010). However, noise can still be detected from 16S rRNA analyses and a fraction of species that have an extremely similar sequence cannot be distinguished from each other. Nevertheless, 16S rRNA analysis is still the most effective way of understanding the community structure of a sample up to date and therefore continuous effort on improving related algorithms is necessary to increase the credibility of the 16S rRNA analysis.

Recently with the surge of NGS technology and bioinformatics, 16S rRNA analysis is being replaced by analysis using the genome. Yet 16S rRNA is still an important marker gene and with a well-defined database, it can be used for identification and to determine the contamination of a genome. Therefore, extracting a complete 16S rRNA as a phylogenetic marker is essential compared to other genes. However, the majority of MiSeq draft genome sequencing data created by de novo assembly does not include a complete 16S rRNA and the short sequence length causes identification errors. To resolve this issue, a new method for extracting a full-length 16S rRNA was introduced. Certainly, there are ways such as using the long-read sequencing of the Pacbio sequencing platform which is expected to overcome the shortcomings mentioned. Nonetheless, long-read sequencing is not free from contamination and a drop in genome quality can occur. Both BCG-based algorithm, which is used for finding low-quality genomes, and ContEst16S, which is used for finding contaminated genomes, were used for quality assessment and they focused on higher specificity rather than sensitivity to determine the genomes which were definitely contaminated or low-quality. Among all tested data by BCG-based algorithm, 25,212 genomes were found to have at least 1 repeated BCG. This rather common phenomenon may be attributed to the inclusion of partial BCG sequences elsewhere in the genome, or to the presence of pseudogenes that have similar sequences to BCGs. Another possibility is that a genome may have ended up with multiple copies of a BCG through lateral gene transfer.

However, 71 genomes were found to have more than twice the number of genes annotated as all of the available BCG annotations. In a further study, I hope to harness this repeated BCG statistic to assess genome contamination and to find oversized genomes. There is no perfect way of detecting all contamination events from whole genome assembly unless the genome sequencing is completely finished. For example, a contamination event by a taxonomically closely related organism cannot be differentiated, with high confidence, from micro-sequence-heterogeneity of ribosomal RNA operons or sequencing errors.

In the evaluation of ANI methods for comparative genomics, all four methods for computing ANI perform well in the range of 90~100% where the suggested species boundary, i.e. 95~96%, is encompassed. However, ANIm did not correlate with others in the lower range of values (below 90%). ANIm and OrthoANIu run significantly faster than algorithms based on BLASTN program (ANIb and OrthoANIb), particularly for large genomes. Overall, OrthoANIu is suitable for all range of ANI calculations while being computationally efficient enough for large-scale comparative genomics. Therefore, I recommend OrthoANIu in lieu of ANIb for large-scale comparative studies.

CHAPTER 3 Development of EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies

3.1. Introduction

One of the goals of the modern taxonomy of Bacteria and Archaea is the objective definition of species, insofar as it applies to classification and identification. The process of determining taxonomy has continually improved over time, with the advent of new technologies. PCR followed by sequencing of 16S rRNA genes (16S) has revolutionized our understanding of phylogeny of Bacteria and Archaea. With the introduction of comprehensive 16S databases that cover almost all known species (DeSantis *et al.*, 2006; Kim *et al.*, 2012; Cole *et al.*, 2013; Quast *et al.*, 2013), the rate of discovering new species was significantly improved. However, even though a bioinformatic comparison of 16S provides an objective and reliable way of identifying a given strain, it has a critical limitation in its use at the species level; even almost identical 16S may not guarantee that two strains belong to the same species (Fox*, et al.*, 1992; Kim *et al.*, 2014). To overcome this problem, an experimental approach called DNA-DNA hybridization (DDH) has been used to complement 16S-based classification (Wayne *et al.*, 1987). More recently, the use of genome data was recommended to replace error-prone, laborious DDH. Several overall genome relatedness indices (OGRIs) were proposed to define species boundaries (Chun and Rainey, 2014). For example, average nucleotide identity (ANI) (Richter and Rosselló-Móra, 2009) and OrthoANI (Lee *et al.*,

2016) suggested a species boundary of 95-96%. Because genome sequences can be used for assessing suprageneric phylogeny, recognizing species (Chun and Rainey, 2014), and differentiating clinical clones with few single nucleotide polymorphisms (Snitkin*, et al.*, 2012), it is evident that their use in the taxonomy of Bacteria and Archaea will greatly improve not just taxonomy, but also other microbiological disciplines. As in the case of 16S, the construction of a quality-controlled genome database of all type strains is a prerequisite for the wider application of genomics-based taxonomy (Kyrpides*, et al.*, 2014). At present, almost 70,000 genome sequences are available in the primary public databases, such as NCBI Assembly Database (https://www.ncbi.nlm.nih.gov/assembly). Even through these genomes have the great potential as a resource for basic, applied and clinical microbiology, their metadata such as taxonomic names require substantial curation. Here, I introduce an integrated database with a complete taxonomic hierarchy of Bacteria and Archaea that is represented by 16S and genome sequences. All genomes were downloaded from www.microbiologyresearch.org by taxonomically identified at the genus, species, or subspecies levels using the combination of gene-based search and OrthoANIu (Yoon*, et al.*, 2017) calculations. Integration of over 2,000 quality-filtered genomes allows us to generate comprehensive reports of GC content, genome sizes, and other significant genomic features of each taxon. The database and related search tools are available at http://www.ezbiocloud.net/.

3.2. Methods

3.2.1. Data collection

The up-to-date reference 16S rRNA sequences were maintained as described earlier (Kim *et al.*, 2012). The 16S rRNA sequences of valid species were collected from the published list of International Journal of Systematic and Evolutionary Microbiology (IJSEM) and the sequences of uncultured species were collected directly from the NCBI nucleotide database. The collected sequences were selected with the best quality for each species using the following strategy. For cases in which multiple sequences were available for a type strain, the sequence extracted from its whole genome assembly (WGA) was selected. As for PCR-derived sequences, the quality of sequencing was manually checked by secondary structure-aware alignment using the EzEditor program (Jeon*, et al.*, 2014). Maximum-likelihood phylogenetic trees of each taxonomic groups, such as phyla, classes, orders or families, were generated from manually aligned 16S sequences using RAxML software (Stamatakis, 2014). All 16S sequences were taxonomically assigned to the species level as a part of the complete taxonomic hierarchy which consisted of phylum, class, order, family, genus, and species (subspecies if applicable).

To construct a whole genome database, I constructed a system that periodically collects genome list from the NCBI Genome database and adds updated genomes. NCBI BioSample IDs were used to identify genomes determined to be of low-quality by metagenome or single cell genome and NCBI quality check information was also used. The GenBank accession list was collected through the assembly report of the FTP link address for the updated genome and the GenBank files were downloaded using NCBI E-utilities (Sayers, 2010). The taxonomic information and assembly information of the genome were inserted into the database by parsing each GenBank files, and strain information were used to determine the type strain by using straininfo.net information. For each genome, the contig sequence data was sorted by length and then MD5 checksum string for the sequence was used to verify that the identical data was not duplicated. A large number of standard strain genomic data not in the NCBI genome database were collected from the JGI genome database. In addition, unassembled sequencing raw data was collected from the NCBI SRA database and the analysis results were inserted into the database using the whole genome assembly (WGAS) pipeline **(Figure 7)**. The overall scheme of 16S rRNA and genome data collection is shown in **Figure 17**.

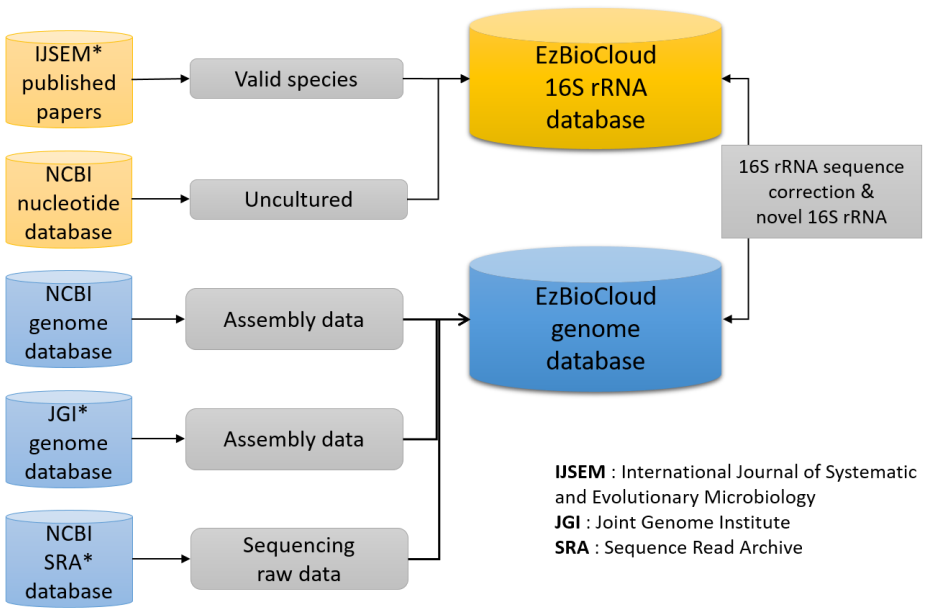

Figure 17. Scheme of data collection for EzBioCloud database.

3.2.2. Identification of genome sequences

Taxonomic identification of each WGA was carried out using the algorithm outlined in **Figure 188**. Prior to this, all WGAs were processed by a genome annotation pipeline using a combination of software tools and databases **(Table 4)**. Two types of databases were used, namely (i) 16S database that is also used in the 'Identify' engine described in the previous section, and (ii) the Reference Genome Database (RefGD). The latter was compiled to hold tetra-nucleotide compositions (Teeling*, et al.*, 2004), and gyrB and recA sequences from all available genome sequences of type or representative strains. Tetra-nucleotide compositions were calculated from each WGAs using an in-house JAVA program. 16S, gyrB and recA genes in WGAs were predicted while processed in WGAS pipeline **(Figure 7)**. RefGD entries then served as the targets of USEARCH-based searches.

A list of phylogenetically related taxa to each WGA in the NCBI Assembly Database was generated using a combination of different approaches. The 16S, gyrB and recA sequences of a query WGA, wherever possible, were searched against the respective databases, and the best hits were added to the list. The correlation values (=z score) based on tetra-nucleotide composition were calculated against all WGAs in the RefGD (Teeling *et al.*, 2004) and the best hits were also added to the list. The final identification was carried out by comparing average nucleotide identity (ANI) values

between the query WGA and those in the list for which I used 95% as the cutoff for species. For ANI calculation, I adopted OrthoANI algorithm (Lee *et al.*, 2016) with USEARCH instead of BLASTN to reduce the computation time (Yoon *et al.*, 2017). I attempted to taxonomically identify all WGAs at least to the genus level. If this was not possible due to the lack of 16S, gyrB and recA genes, it was assigned as 'Unidentified'.

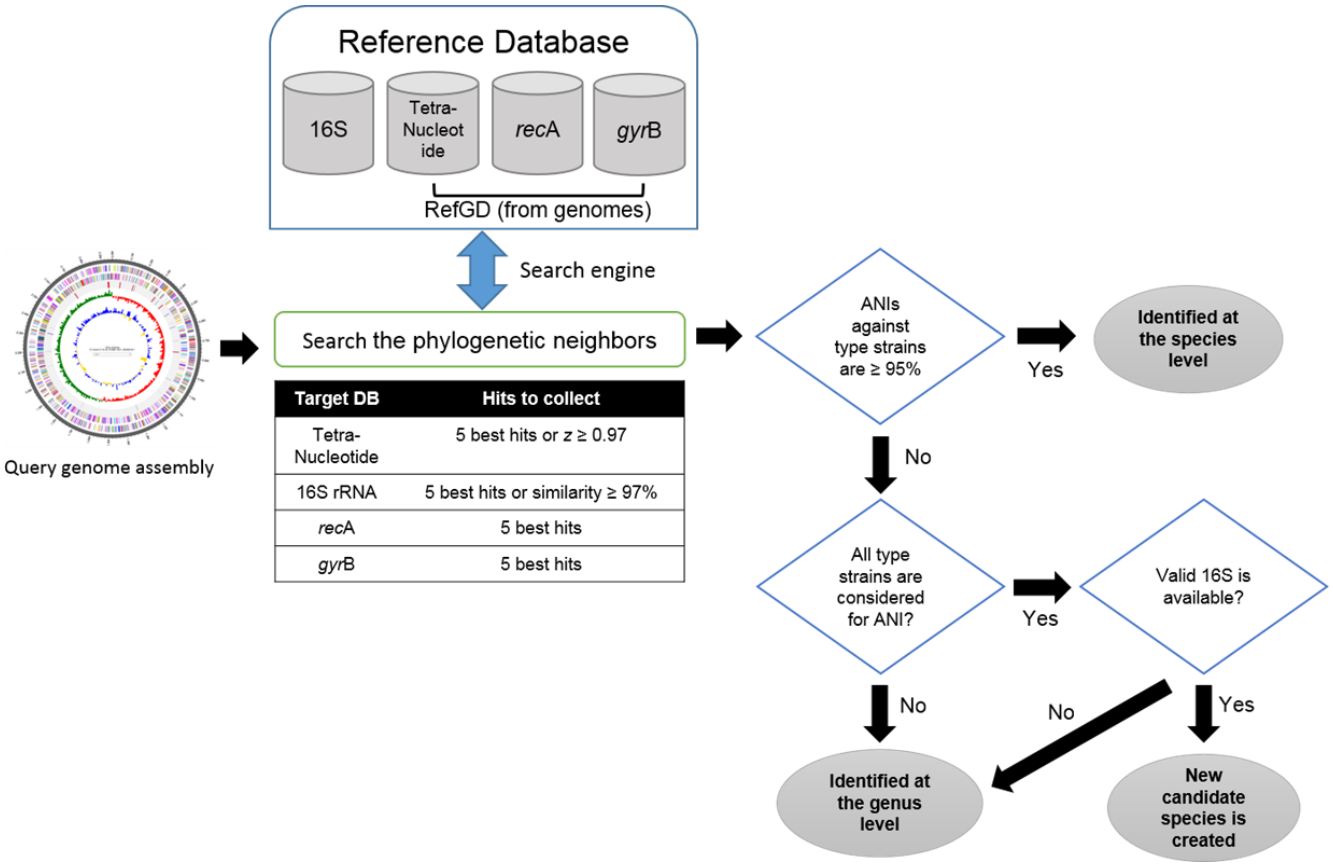

Figure 18. Outline of an algorithm for identifying genome sequence.

3.2.3. Calculation of genomics features for each taxon

Several genomic features of taxonomic importance, including genome sizes, DNA G+C content, the number of genes and lengths of CDS (coding sequences)/intergenic regions, were calculated and statistically compiled for each taxon. The R package was used for all statistical analysis. Information on the number of 16S genes in genomes was obtained from complete genome sequences. If a species didn't have any complete genomes, PICRUSt (Langille*, et al.*, 2013) was used to predict the values.

3.2.4. Bacterial community analysis of human microbiome

The bacterial community dataset of the US NIH human microbiome project was obtained from http://hmpdacc.org/ and processed by a bioinformatics pipeline given in **Figure 5**. Frequencies of each taxon (from phyla to genera) at 18 body parts of healthy subjects were compiled and visualized as box-plots at the web-page for each taxon. The body parts and the number of analyzed samples are shown in **Table 12**.

Table 12. 18 body parts of healthy human subjects and analyzed sample count of each bacterial community.

Body sites	Number of Samples
Anterior nares	433
Attached Keratinized gingiva	545
Buccal mucosa	536
Hard palate	502
Left Antecubital fossa	280
Left Retroauricular crease	458
Mid vagina	216
Palatine Tonsils	563
Posterior fornix	216
Right Antecubital fossa	284
Right Retroauricular crease	487
Saliva	487
Stool	574
Subgingival plaque	558
Supragingival plaque	575
Throat	520
Tongue dorsum	570
Vaginal introitus	244
Total	**8,048**

3.2.5. Operating system and software development

The entire system was built on the standard Linux operating system and uploaded to Amazon Web Services (AWS) servers. JAVA, JavaScript and R programming languages were used, and MySQL was used as the database management system.

3.3. Results

3.3.1. Comparison of databases

The EzBioCloud database was constructed on March 13, 2017, including 62,685 purified 16S rRNA sequences of different species and 62,362 genome data. At the time of writing, this database is the most updated database and always up-to-date through an every two-month update using an automated update process. The EzBioCloud database is compared with SILVA, Greengene, RDP, and EzTaxon, which are widely known as 16S rRNA sequence databases. **Table 13** shows that the number of sequences contained in the database is the smallest, but the largest number of taxonomic nodes based on the valid nomenclature. Compared with the Greengene database, which can be identified up to the Species level, The EzBioCloud database contains about 6 times the taxonomic node despite the sequence data size of about 60. This high taxonomic coverage is very advantageous for accurately identifying new species. Moreover, 14,921 chimeric sequences were detected in the Greengene database as a result of reference based chimera search using 16S rRNA sequence of valid species of the EzBioCloud. Despite the increase in new species every year, the 16S rRNA sequence number in the EzBioCloud database is smaller than the EzTaxon-e database (Kim *et al.*, 2012) in which 16S rRNA sequences were

collected in the same way. This is because the 16S rRNA sequence of the EzBioCloud database has been screened for a short or low-quality 16S rRNA sequence. In particular, the novel 16S rRNA sequence from the 454 NGS sequencing technology, which was included in the EzTaxon-e database, was about 400-500 bp in length and was excluded in the EzBioCloud. Among the comparative databases, the only species-level identification available are Greengene, EzTaxon-e, and EzBioCloud, and the EzBioCloud database is the only database containing genome Information.

Table 13. Comparison of 16S rRNA sequence databases.

Database	SILVA	Greengenes	RDP	EzTaxon-e	EzBioCloud
Last updated	2016.09.28	2013.05	2016.09.30	2015.08.01	2017.03.13
No. of qualified 16S rRNA	645,151	**99,322**	3,356,809	64,329	**62,685**
No. of nodes	12,117	3,093	6,128	**16,016**	**17,820**
Lowest rank	Genus	**Species**	Genus	**Species**	**Species**
Genome information	None	None	None	None	**62,362**

3.3.2. Hierarchical taxonomic backbone

EzBioCloud database consists of a hierarchical taxonomic system containing 207 phyla, 433 classes, 1,019 orders, 2,805 families, 11,446 genera, 61,700 species and 387 subspecies. This classification was primarily based on the maximum likelihood phylogeny for 16S data, where 97% similarity cutoff was used for the recognition of phylotypes. Taxa without their type or representative 16S sequences were not included in the database. I extended the database by adding new candidate species that were identified by the identification scheme in **Figure 18** based on the combination of sequence-based search and OrthoANIu calculations. As a result, 1,400 tentatively named species of 16S rRNA sequences (2% of total 16S rRNA sequences) from genome were included in the database. **Figure 19** shows the OrthoANIu-based dendrogram of the genus *Acinetobacter* in which 13 such new candidate species are shown. Of the total 62,685 16S rRNA sequences, the number of 16S rRNA sequences of the valid species is 14,441 (23% of total) and 5,716 (40% of the valid species) species of this have whole genome data of type strain. Using this type strain whole genome data, the 16S rRNA sequence of 4,620 species was corrected to the extracted full-length 16S sequences from contigs of a genome, and some sequences of the genome were not applied to the 16S rRNA database because of the low-quality and completeness of the 16S rRNA sequence **(Table 14**, **Figure 20)**.

Taxonomic hierarchical system of EzBioCloud has the following principles: (i) all terminal taxa (species or subspecies) are represented by at least one 16S sequence, (ii) all terminal taxa are assigned under the complete suprageneric ranks (phylum, class, order, family), and (iii) taxonomic assignment is based on the phylogenetic relationship (maximum likelihood treeing and OrthoANIu), not necessarily following the current formal standing in taxonomy. For example, *Shigella spp.* is placed under the genus *Escherichia* but not *Shigella* in EzBioCloud database, as it is phylogenetically a member of the former.

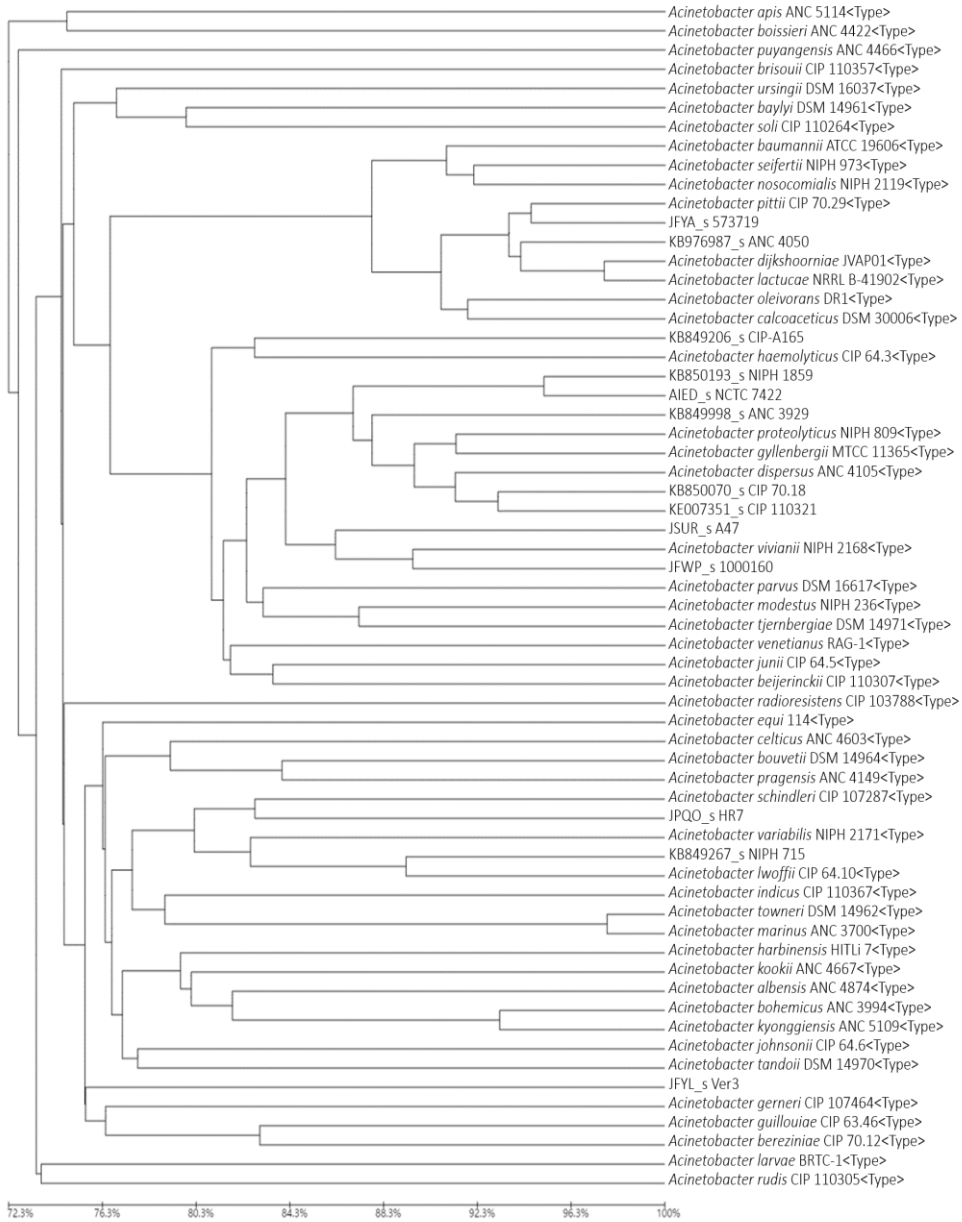

Figure 19. OrthoANIu-based dendrogram of the genus *Acinetobacter* including 13 tentatively named species. The dendrogram is constructed using UPGMA algorithm. The scale bar represents OrthoANIu values. <Type>, type strain.

Table 14. The count of 16S rRNA sequences by source and corrected information.

16S rRNA sequence source	Count
Phylotype of invalid species from PCR	46,844
Phylotype or invalid species from genome	1,400
Valid species name without type strain genome	8,725
Not corrected 16S rRNA by genome	1,096
Corrected 16S rRNA by genome	4,620
Total	**62,685**

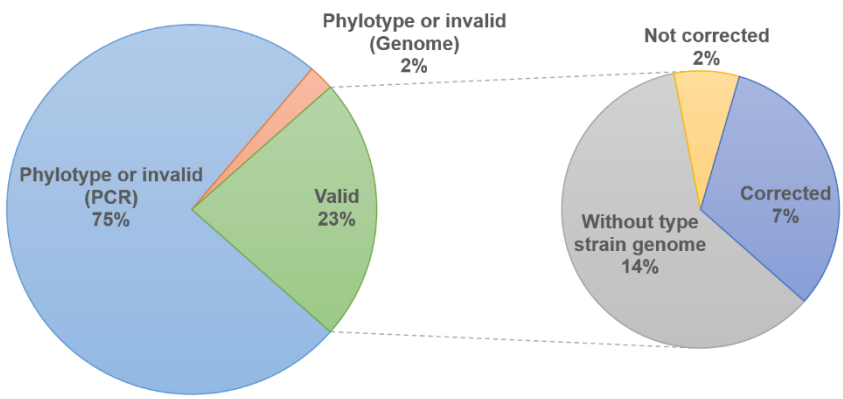

Figure 20. Pie on pie chart of 16S rRNA sequence source and corrected 16S rRNA sequences on EzBioCloud database.

3.3.3. Identification of genome projects

Taxonomic search engine for WGA **(Figure 18)** was designed to ensure that all possible phylogenetically neighboring taxa are chosen for the final ANI calculations. The tetra-nucleotide composition of WGAs has been successfully applied to the rapid comparison between genomic and metagenomic assemblies (Teeling *et al.*, 2004; Richter*, et al.*, 2015). However, this is not a phylogenetic approach and is prone to be biased by large-scale lateral gene transfer. 16S has been widely used for bacterial identification and is ideal for finding phylogenetically related WGAs. However, out of 62,362 qualified WGAs, 4,285 contain no 16S rRNA genes that can be used for such a purpose. Therefore, two of the most widely used protein-coding phylogenetic markers, namely gyrB and recA, are implemented in the search engine in addition to 16S and tetra-nucleotide composition. The genes coding for GyrB and RecA are also known to have higher resolution than 16S in phylogenetic analyses (Thompson*, et al.*, 2004; Kirby*, et al.*, 2010). This composite approach allows the detection of all possible phylogenetically neighboring taxa which are then subjected to OrthoANIu calculations.

With 95% ANI cutoff as species boundaries, 42,136, 15,794 and 4,432 WGAs were identified at the species, subspecies, and genus levels, respectively. Thirty-six WGAs could not be identified by the current version

of RefGD. Also, the taxonomic names of 16,701 WGAs were found to be incorrect **(Table 15**, **Figure 21)**, which was supported by OrthoANIu values. As a result, the taxonomic names of 16,737 WGAs (27 % of the total qualified WGAs) were changed from the original names in NCBI Assembly Database that had been originally assigned by the primary data depositors. Examples of misidentified and unidentified WGAs are given in **Figure 22**. I expect that the portion of WGAs identified at the species/subspecies level will be increased as more genome sequences become available for type strains.

Table 15. The count of genome identification results.

Genome identification of NCBI genomes	Count
Correctly identified	45,661
Unidentified in subspecies level	9,403
Unidentified in species level	2,238
Misidentified with ANI value (<=95% against type strain)	3,087
Misidentified in species level	826
Misidentified in genus level	1,147
Total	**62,362**

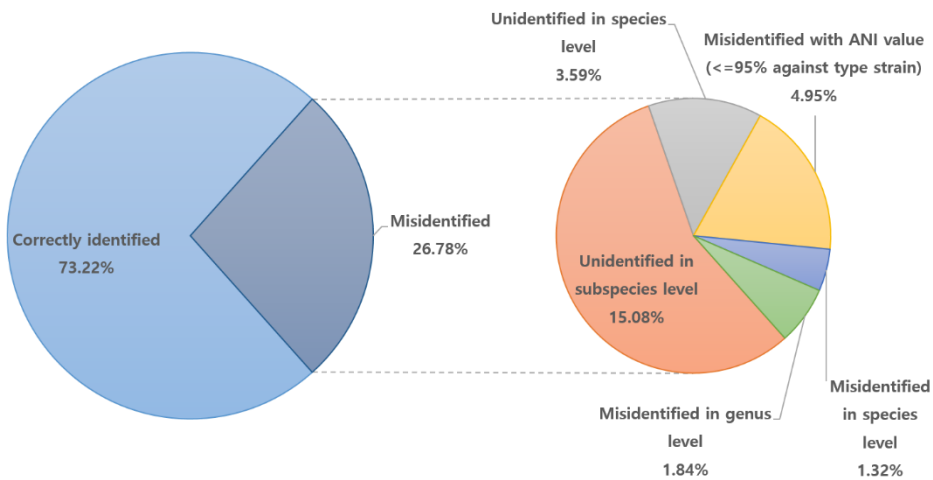

Figure 21. Pie on pie chart of genome identification of NCBI genomes.

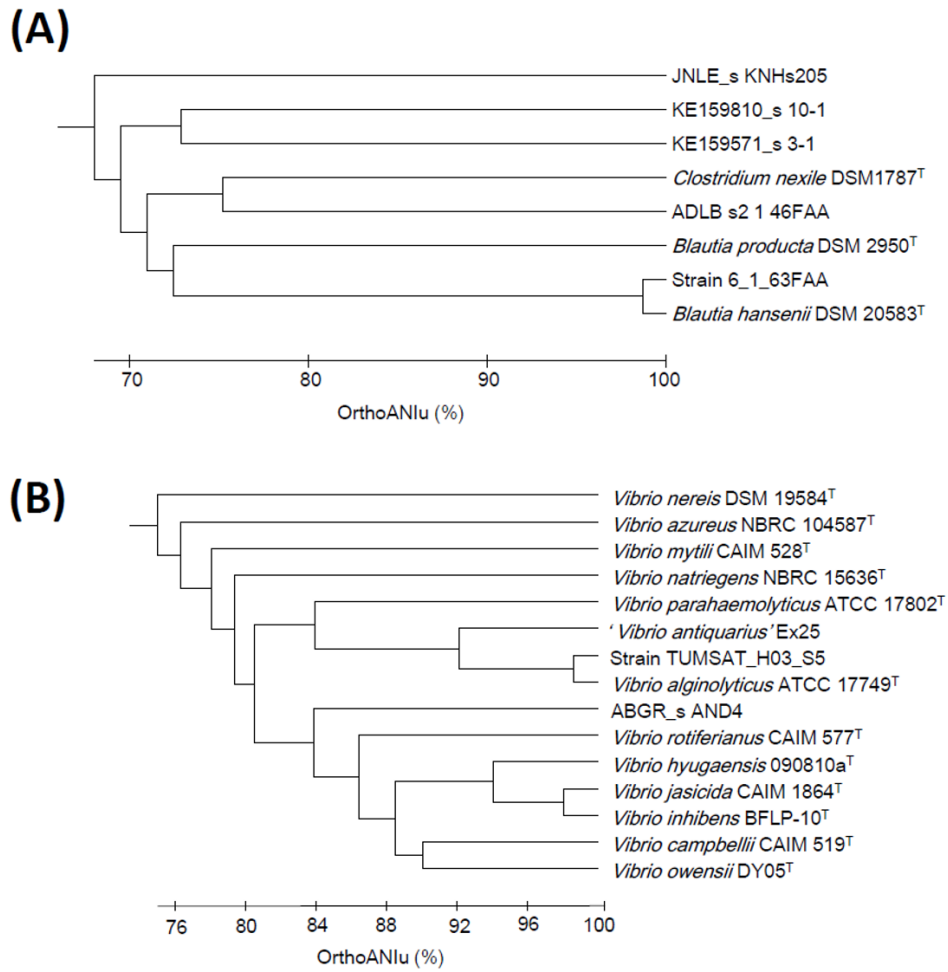

Figure 22. Example of misidentified and unidentified WGAs.

3.3.4. Genome-derived information

The genome is the ultimate source for taxonomy from which a variety of information can be extracted for a better description of the species. For instance, more accurate GC content of DNA can be obtained if calculated from genome sequences instead of experimental methods such as HPLC (Kim*, et al.*, 2015). Because many species are now represented by multiple genomes, taxonomically meaningful information about species can be extracted and statistically compiled. In EzBioCloud, the following information is provided for each taxon, wherever applicable: (i) GC content, (ii) genome size, (iii) the number of CDSs, (iv) the length of CDSs and intergenic regions and (v) the number of 16S rRNA genes. An OrthoANIu-based UPGMA dendrogram of type and reference strains is also provided for each genus if genome data is available. In addition, the occurrence of bacterial taxa, from phyla to genera, in 18 different body parts of the human microbiome is given as box-plot charts.

3.4. Discussion

With EzBioCloud, a database for 16S rRNA and genome data with a well-reflected prokaryotic taxonomy was constructed. Coverage and accuracy of the database are critical factors for classification and identification. Out of all the prokaryotic microorganisms, only less than 1% of them can be cultured and even with all the uncultured microorganisms included in EzBioCloud 16S rRNA database it is nowhere near being able to cover every prokaryotic microorganism. However, a perfect database is impossible and a continuous effort of updating the database is necessary. EzBioCloud 16S rRNA database is the most up to date 16S rRNA database with the largest number of species with their taxonomic information and is scheduled to be updated regularly. Another important factor for a good database other than its size is its quality. In this project, genome data was used for correcting taxonomic errors and increasing the credibility of the relationship between identified microorganisms. The vast amount of genome data, which increased with the development of NGS technology, enables such methods, and at the time of writing there are 96,889 prokaryote public genomes released and this number is growing rapidly. Through this research I discovered that a large percentage of the genomes had errors such as misidentification or contamination. These types of errors can cause critical consequences to industrial and scientific researches and for human microbiome related

researches this can even affect the human health issues directly which is why it must not be taken lightly. Although many public NCBI genome data were corrected by building the EzBioCloud genome database, it is not perfect. Microorganisms that have high importance in the industrial and academic field should be evaluated thoroughly, and microorganisms that are relatively unknown should be supported by further updates. Also since the trend of data piling up is not going to stop any time soon, it is important to employ state of the art statistical and computational algorithms such as deep-learning and big-data mining to improve the database's quality. Nevertheless, EzBioCloud database provides 16S rRNA and Genome data with the best embedded prokaryotic taxonomic structure and widely used species concepts for researchers, businesses, and individuals to aid their research and analyses.

Target search and visualization is another important filed to be challenged in database utilization. Checking every tens of thousands of data in the database is practically impossible and requires a high understanding and skill set in order to make inquiries and handle the data. EzBioCloud implemented an easy searching functionality for tens of thousands of taxonomic names and sequence information through a website. In addition, the website provides genome data analysis results, discovery frequency in human microbiome, and individual genome browser for each taxon by utilizing several visualization techniques. Through these visualized content researchers can intuitively grasp the concepts and attributes of the data and

give insight. The searching and visualization mechanism used in EzBioCloud are both continuously growing fields and therefore updated algorithms that extract meaningful information from more sophisticated and bigger data should constantly be applied on EzBioCloud.

CHAPTER 4

General conclusions

The advancement in sequencing technology suggested a new standard in biological research while facilitating the growth of bioinformatics. Countless bioinformatics algorithms, pipelines, software, and databases were created and improved over time, trying to keep up with the ever-evolving sequencing technology. The growth in the field made lots of new research possible, and microbiome research using metagenomics is one of the fields that have been receiving a great deal of attention. Through this thesis, various bioinformatics pipelines and tools for microbiome analysis were introduced. Also, to assist these tools, a database based on 16S rRNA and genome was built and furthermore was utilized to evaluate algorithms and software performances.

Microbiome study starts by distinguishing microorganisms. To distinguish microorganisms that are not visible to our naked eyes, several different physical, chemical, and molecular techniques are used. Among these, the classification method that incorporates 16S rRNA is the most popular with the development of sequencing technology. 16S rRNA, which is a popular phylogenetic marker gene, was used by many researchers as a mean of storing the microorganisms sequence data on a database and with several reasons was suggested as a taxonomical standard. In this research, a database based on this taxonomical standard was built as well as several pipelines and tools that also makes use of this. The four main bioinformatics algorithms that are used in a typical 16S rRNA amplicon analysis using NGS data are sequence error filtering, pairwise alignment, searching database, and clustering. Each algorithm has excess amounts of papers and tools

published and selecting which tools to be used and combining them as a pipeline has an enormous impact on both the performance and result. Rather than sticking with a general pipeline, it is far better to develop a pipeline that is optimized with the database. Hence, a combination of published software and improved algorithms were used to build the MTP pipeline and with the EzBioCloud database the performance and accuracy were evaluated. As a result, the pipeline was efficient enough to handle the massive data from NGS platforms while maintaining a high accuracy.

Another research area that benefitted from the NGS technology is whole genome sequencing. The price of sequencing has been reducing rapidly which enabled whole genome analysis for numerous microorganisms and resulted in public genome data explosively piling up. The increment in data also increased the amount of data with low-quality and these data can cause an extremely critical bias in comparative genomics approaches. Additionally, individual researchers using different bioinformatics pipelines on the same genome data results in different annotations, hence the same computational annotation pipeline had to be applied to all the data for a comparative analysis. Every public genome that is able to access 16S rRNA database was integrated to verify genomic properties of various microorganisms based on their taxonomical information. And to apply identical computational annotations on every genome a new genome annotation pipeline was developed which was used to reanalyze every genome. Coinciding with the reanalysis, an algorithm for filtering low-quality genomes was implemented

while predicting contaminated genomes by using a published algorithm. Misidentified public genome data were also identified again to ensure all the genomes in the database to have a consistent taxonomical structure and an improved ANI calculation method was developed and evaluated to be used for genome identification and comparative genomics. The public genome data had an unexpectedly large number of errors and the methods applied in order to find qualified genomes did not perfectly rule out all the low-quality genomes. To enhance this procedure, algorithms on each step should be persistently improved.

In order to build the EzBioCloud database, numerous bioinformatics techniques that were previously developed were used; the database contents, all supporting software, and microbiome analysis results are available through the website. The EzBioCloud database is the first ever database based on prokaryotic taxonomy where the 16S rRNA sequence information and the genome information is completely connected and evaluated through various algorithms to maintain highly qualified information. Finally, EzBioCloud database provides taxonomically correct 16S rRNA sequences for microbial community analysis and the connected genome data for functional analysis. Hence, numerous researchers can use the database, as well as its optimized analysis methods to more efficiently conduct their microbiome research.

REFERENCES

Albanese, D., P. Fontana, C. De Filippo, D. Cavalieri and C. Donati (2015). "MICCA: a complete and accurate software for taxonomic profiling of metagenomic data." *Sci Rep* **5**: 9743.

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson and P. H. Nielsen (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." *Nat Biotechnol* **31**: 533-538.

Alkan, C., S. Sajjadian and E. E. Eichler (2011). "Limitations of next-generation genome sequence assembly." *Nat Methods* **8**: 61-65.

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

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls and J.-M. Batto (2011). "Enterotypes of the human gut microbiome." *Nature* **473**: 174-180.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham and A. D. Prjibelski (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J Comput Biol* **19**: 455-477.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes and H. R. Bignell (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." *Nature* **456**: 53-59.

Berg, R. D. (1996). "The indigenous gastrointestinal microflora." *Trends Microbiol* **4**: 430-435.

Bianconi, E., A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani and F. Piva (2013). "An estimation of the number of cells in the human body." *Ann Hum Biol* **40**: 463-471.

Bland, C., T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides and P. Hugenholtz (2007). "CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats." *BMC bioinformatics* **8**: 209.

Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode and G. F. Mayhew (1997). "The complete genome sequence of Escherichia coli K-12." *Science* **277**: 1453-1462.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* **30**: 2114-2120.

Bult, C. J., O. White, G. J. Olsen and L. Zhou (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." *Science* **273**: 1058.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich and J. I. Gordon (2010). "QIIME allows analysis of high-throughput community sequencing data." *Nat Methods* **7**: 335-336.

Chun, J. (1995). Computer assisted classification and identification of actinomycetes. (Unpublished doctoral thesis). *University of Newcastle upon Tyne*.

Chun, J. and F. A. Rainey (2014). "Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea." *Int J Syst Evol Micr* **64**: 316-324.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun,

C. T. Brown, A. Porras-Alfaro, C. R. Kuske and J. M. Tiedje (2013). "Ribosomal Database Project: data and tools for high throughput rRNA analysis." *Nucleic Acids Res* **27**: 171-173.

Consortium, H. M. P. (2012). "A framework for human microbiome research." *Nature* **486**: 215-221.

Consortium, H. M. P. (2012). "Structure, function and diversity of the healthy human microbiome." *Nature* **486**: 207-214.

Consortium, U. (2014). "UniProt: a hub for protein information." *Nucleic Acids Res* **43**: D204-D212.

Critchfield, J. W., S. Van Hemert, M. Ash, L. Mulder and P. Ashwood (2011). "The potential role of probiotics in the management of childhood autism spectrum disorders." *Gastroent Res Pract* **2011**: 161358.

De Vos, P. and H. Trüper (2000). "Judicial Commission of the International Committee on Systematic Bacteriology; IXth International (IUMS) Congress of Bacteriology and Applied Microbiology." *Int J Syst Evol Micr* **50**: 2239-2244.

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

Dickie, I. A. (2010). "Insidious effects of sequencing errors on perceived diversity in molecular surveys." *New Phytol* **188**: 916-918.

Dixon, P. and M. Palmer (2003). "VEGAN, a package of R functions for community ecology." *J Veg Sci* **14**: 927-930.

Dojka, M. A., P. Hugenholtz, S. K. Haack and N. R. Pace (1998). "Microbial diversity in a hydrocarbon-and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation." *Appl*

Environ Microb **64**: 3869-3877.

Eddy, S. R. (2011). "Accelerated profile HMM searches." *PLoS Comput Biol* **7**: e1002195.

Edgar, R. C. (2007). "PILER-CR: fast and accurate identification of CRISPR repeats." *BMC bioinformatics* **8**: 18.

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

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince and R. Knight (2011). "UCHIME improves sensitivity and speed of chimera detection." *Bioinformatics* **27**: 2194-2200.

Egholm, M., M. Margulies, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen and Z. Chen (2005). "Genome sequencing in open microfabricated high density picoliter reactors." *Nature* **437**: 376-380.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan and B. Bettman (2009). "Real-time DNA sequencing from single polymerase molecules." *Science* **323**: 133-138.

EUZéBY, J. P. (1997). "List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet." *Int J Syst Evol Micr* **47**: 590-592.

Fleischmann, R. D., M. D. Adams, O. White and R. A. Clayton (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." *Science* **269**: 496.

Fox, G. E., J. D. Wisotzkey and P. Jurtshuk JR (1992). "How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity." *Int J Syst Evol Micr* **42**: 166-170.

Fraser, C. M., J. D. Gocayne, O. White and M. D. Adams (1995).

"The minimal gene complement of Mycoplasma genitalium." *Science* **270**: 197.

Frishman, D., K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner and H.-W. Mewes (2001). "Functional and structural genomics using PEDANT." *Bioinformatics* **17**: 44-57.

Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." *Bioinformatics* **28**: 3150-3152.

Gil, Y., E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau and J. Myers (2007). "Examining the challenges of scientific workflows." *Computer* **40**.

Goodwin, S., J. D. McPherson and W. R. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." *Nat Rev Genet* **17**: 333-351.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme and J. M. Tiedje (2007). "DNA–DNA hybridization values and their relationship to whole-genome sequence similarities." *Int J Syst Evol Micr* **57**: 81-91.

Grice, E. A. and J. A. Segre (2012). "The human microbiome: our second genome." *Annu Rev Genomics Hum Genet* **13**: 151-170.

Hayden, E. C. (2009). "Genome sequencing: the third generation." *Nature* **457**: 768-769.

Hong, S., J. Bunge, C. Leslin, S. Jeon and S. S. Epstein (2009). "Polymerase chain reaction primers miss half of rRNA microbial diversity." *ISME J* **3**: 1365-1373.

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa and M. Kuhn (2015). "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic

and viral sequences." *Nucleic Acids Res* **44**: D286-D293.

Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin and D. M. Welch (2007). "Accuracy and quality of massively parallel DNA pyrosequencing." *Genome Biol* **8**: R143.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* **11**: 119.

Jain, R., M. C. Rivera and J. A. Lake (1999). "Horizontal gene transfer among genomes: the complexity hypothesis." *Proc Natl Acad Sci U S A* **96**: 3801-3806.

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

Jeon, Y.-S., J. Chun and B.-S. Kim (2013). "Identification of household bacterial community and analysis of species shared with human microbiome." *Curr Microbiol* **67**: 557-563.

Jeon, Y.-S., K. Lee, S.-C. Park, B.-S. Kim, Y.-J. Cho, S.-M. Ha and J. Chun (2014). "EzEditor: a versatile sequence alignment editor for both rRNA-and protein-coding genes." *Int J Syst Evol Micr* **64**: 689-691.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2015). "KEGG as a reference resource for gene and protein annotation." *Nucleic Acids Res* **44**: D457-D462.

Kim, M., H.-S. Oh, S.-C. Park and J. Chun (2014). "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes." *Int J Syst Evol Micr* **64**: 346-351.

Kim, M., S.-C. Park, I. Baek and J. Chun (2015). "Large-scale evaluation of experimentally determined DNA G+ C contents

with whole genome sequences of prokaryotes." *Syst Appl Microbiol* **38**: 79-83.

Kim, O. S., Y. J. Cho, K. Lee, S. H. Yoon, M. Kim, H. Na, S. C. Park, Y. S. Jeon, J. H. Lee, H. Yi, et al. (2012). "Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species." *Int J Syst Evol Microbiol* **62**: 716-721.

Kirby, B. M., G. J. Everest and P. R. Meyers (2010). "Phylogenetic analysis of the genus Kribbella based on the gyrB gene: proposal of a gyrB-sequence threshold for species delineation in the genus Kribbella." *Antonie Van Leeuwenhoek* **97**: 131-142.

Koonin, E. V. and M. Y. Galperin (1997). "Prokaryotic genomes: the emerging paradigm of genome-based microbiology." *Curr Opin Genet Dev* **7**: 757-763.

Koren, S., M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie and E. D. Jarvis (2012). "Hybrid error correction and de novo assembly of single-molecule sequencing reads." *Nat Biotechnol* **30**: 693-700.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S. L. Salzberg (2004). "Versatile and open software for comparing large genomes." *Genome Biol* **5**: R12.

Kyrpides, N. C., P. Hugenholtz, J. A. Eisen, T. Woyke, M. Göker, C. T. Parker, R. Amann, B. J. Beck, P. S. Chain and J. Chun (2014). "Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains." *PLoS Biol* **12**: e1001920.

Land, M., L. Hauser, S.-R. Jun, I. Nookaew, M. R. Leuze, T.-H. Ahn, T. Karpinets, O. Lund, G. Kora and T. Wassenaar (2015). "Insights from 20 years of bacterial genome sequencing." *Funct Integr Genomic* **15**: 141-161.

Lane, D. (1991). 16S/23S rRNA sequencing. Nucleic acid techniques

in bacterial systematics. E. Stackebrandt and M. Goodfellow, *New York: John Wiley and Sons*.

Langille, M. G., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber and R. Knight (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." *Nat Biotechnol* **31**: 814-821.

Lapage, S. P., P. H. Sneath, E. F. Lessel, V. Skerman, H. Seeliger and W. Clark (1992). International code of nomenclature of bacteria: bacteriological code, 1990 revision, *ASM Press*.

Larkin, M. A., G. Blackshields, N. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm and R. Lopez (2007). "Clustal W and Clustal X version 2.0." *Bioinformatics* **23**: 2947-2948.

Laurence, M., C. Hatzis and D. E. Brash (2014). "Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes." *PLoS One* **9**: e97876.

Lee, B., T. Moon, S. Yoon and T. Weissman (2015). "DUDE-Seq: Fast, flexible, and robust denoising of nucleotide sequences." *arXiv preprint*, arXiv:1511.04836.

Lee, I., M. Chalita, S.-M. Ha, S.-I. Na, S.-H. Yoon and J. Chun (2017). "ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences." *Int J Syst Evol Micr*: in Press.

Lee, I., Y. O. Kim, S.-C. Park and J. Chun (2016). "OrthoANI: an improved algorithm and software for calculating average nucleotide identity." *Int J Syst Evol Micr* **66**: 1100-1103.

Leinonen, R., H. Sugawara and M. Shumway (2010). "The sequence read archive." *Nucleic Acids Res* **39**: D19-D21.

Li, Z., Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X.

Hu and B. Liu (2012). "Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph." *Brief Funct Genomics* **11**: 25-37.

Liao, Y.-C., S.-H. Lin and H.-H. Lin (2015). "Completing bacterial genome assemblies: strategy and performance comparisons." *Sci Rep* **5**: 8747.

Logares, R., S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. Vargas and G. Lima-Mendez (2014). "Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities." *Environ Microbiol* **16**: 2659-2671.

Louis, P. and H. J. Flint (2009). "Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine." *FEMS Microbiol Lett* **294**: 1-8.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." *Nucleic Acids Res* **25**: 955-964.

Ludwig, W. and K. Schleifer (1994). "Bacterial phylogeny based on 16S and 23S rRNA sequence analysis." *FEMS Microbiol Rev* **15**: 155-173.

Mardis, E., J. McPherson, R. Martienssen, R. K. Wilson and W. R. McCombie (2002). "What is finished, and why does it matter." *Genome Res* **12**: 669-671.

Masella, A. P., A. K. Bartram, J. M. Truszkowski, D. G. Brown and J. D. Neufeld (2012). "PANDAseq: paired-end assembler for illumina sequences." *BMC bioinformatics* **13**: 31.

Mavromatis, K., N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski and M. Land (2007). "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods." *Nat Methods* **4**:

495-500.

McCarthy, A. (2010). "Third generation DNA sequencing: pacific biosciences' single molecule real time technology." *Chem Biol* **17**: 675-676.

Mende, D. R., A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes and P. Bork (2012). "Assessment of metagenomic assembly using simulated next generation sequencing data." *PLoS One* **7**: e31386.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens and A. Wilke (2008). "The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes." *BMC bioinformatics* **9**: 386.

Mitra, S., M. Stärk and D. H. Huson (2011). "Analysis of 16S rRNA environmental sequences using MEGAN." *BMC genomics* **12**: S17.

Mosher, J. J., B. Bowman, E. L. Bernberg, O. Shevchenko, J. Kan, J. Korlach and L. A. Kaplan (2014). "Improved performance of the PacBio SMRT technology for 16S rDNA sequencing." *J Microbiol Meth* **104**: 59-60.

Myers, E. W. and W. Miller (1988). "Optimal alignments in linear space." *Comput Appl Biosci* **4**: 11-17.

Nawrocki, E. P., S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones and J. Tate (2015). "Rfam 12.0: updates to the RNA families database." *Nucleic Acids Res* **43**: D130-D137.

Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches." *Bioinformatics* **29**: 2933-2935.

Oren, A. and G. M. Garrity (2014). "Then and now: a systematic review of the systematics of prokaryotes in the last 80 years."

Antonie Van Leeuwenhoek **106**: 43-56.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz and R. Edwards (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." *Nucleic Acids Res* **33**: 5691-5702.

Pace, N. R. (1997). "A molecular view of microbial diversity and the biosphere." *Science* **276**: 734-740.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." *Genome Res* **25**: 1043-1055.

Pereira, F., J. Carneiro, R. Matthiesen, B. van Asch, N. Pinto, L. Gusmão and A. Amorim (2010). "Identification of species by multiplex analysis of variable-length sequences." *Nucleic Acids Res* **38**: e203-e203.

Plummer, E., J. Twin, D. M. Bulach, S. M. Garland and S. N. Tabrizi (2015). "A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data." *J Proteomics Bioinform* **8**: 283.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez and T. Yamada (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." *Nature* **464**: 59-65.

Quast, C. (2006). MicHanThi-design and implementation of a system for the prediction of gene functions in genome annotation projects. (Unpublished doctoral thesis). *Universität Bremen*.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies and F. O. Glockner (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." *Nucleic Acids Res* **41**: D590-D596.

Richter, M. and R. Rosselló-Móra (2009). "Shifting the genomic gold standard for the prokaryotic species definition." *Proc Natl Acad Sci U S A* **106**: 19126-19131.

Richter, M., R. Rosselló-Móra, F. O. Glöckner and J. Peplies (2015). "JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison." *Bioinformatics* **32**: 929-931.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez and A. Robbins-Pianka (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences." *PeerJ* **2**: e545.

Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil and C. Minor (2000). "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms." *Appl Environ Microb* **66**: 2541-2547.

Rosselló- Mora, R. and R. Amann (2001). "The species concept for prokaryotes." *FEMS Microbiol Rev* **25**: 39-67.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher and M. Roberts (2012). "GAGE: A critical evaluation of genome assemblies and assembly algorithms." *Genome Res* **22**: 557-567.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *Proc Natl Acad Sci U S A* **74**: 5463-5467.

Savage, D. C. (1977). "Microbial ecology of the gastrointestinal tract." *Annu Rev Microbiol* **31**: 107-133.

Sayers, E. (2010). "A general introduction to the E-utilities." *Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information*.

Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio and S. Federhen (2011). "Database resources of the national center for biotechnology information." *Nucleic Acids Res* **39**: D38-D51.

Schloss, P. D. and J. Handelsman (2008). "A statistical toolbox for metagenomics: assessing functional diversity in microbial communities." *BMC bioinformatics* **9**: 34.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks and C. J. Robinson (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." *Appl Environ Microb* **75**: 7537-7541.

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

Sharpton, T. J., S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O'Dwyer, J. L. Green, J. A. Eisen and K. S. Pollard (2011). "PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data." *PLoS Comput Biol* **7**: e1001061.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." *Nat Biotechnol* **26**: 1135-1145.

SKERMAN, V. B. D., V. McGowan and P. H. A. Sneath (1980). "Approved lists of bacterial names." *Int J Syst Evol Micr* **30**: 225-420.

Sneath, P. H. (1957). "Some thoughts on bacterial classification." *Microbiology* **17**: 184-200.

Sneath, P. H. and P. Sneath (1962). The construction of taxonomic groups, *Cambridge University Press*.

Snitkin, E. S., A. M. Zelazny, P. J. Thomas, F. Stock, D. K. Henderson, T. N. Palmore, J. A. Segre and N. C. S. Program (2012). "Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing." *Sci Transl Med* **4**: 148ra116.

Stackebrandt, E. and B. Goebel (1994). "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology." *Int J Syst Evol Micr* **44**: 846-849.

Staley, J. T. (2006). "The bacterial species dilemma and the genomic–phylogenetic species concept." *Philos T Roy Soc B* **361**: 1899-1909.

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* **30**: 1312-1313.

Steel, K. (1965). "Microbial identification." *Microbiology* **40**: 143-148.

Teeling, H., A. Meyerdierks, M. Bauer, R. Amann and F. O. Glöckner (2004). "Application of tetranucleotide frequencies for the assignment of genomic fragments." *Environ Microbiol* **6**: 938-947.

Thompson, C., F. Thompson, K. Vandemeulebroecke, B. Hoste, P. Dawyndt and J. Swings (2004). "Use of recA as an alternative phylogenetic marker in the family Vibrionaceae." *Int J Syst Evol Micr* **54**: 919-924.

Tjaden, B. (2015). "De novo assembly of bacterial transcriptomes from RNA-seq data." *Genome Biol* **16**: 1.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight and J. I. Gordon (2007). "The human microbiome

project: exploring the microbial part of ourselves in a changing world." *Nature* **449**: 804.

Wayne, L., D. Brenner, R. Colwell, P. Grimont, O. Kandler, M. Krichevsky, L. Moore, W. Moore, R. Murray and E. Stackebrandt (1987). "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics." *Int J Syst Evol Micr* **37**: 463-464.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." *Proc Natl Acad Sci U S A* **74**: 5088-5090.

Wylie, K. M., R. M. Truty, T. J. Sharpton, K. A. Mihindukulasuriya, Y. Zhou, H. Gao, E. Sodergren, G. M. Weinstock and K. S. Pollard (2012). "Novel bacterial taxa in the human microbiome." *PLoS One* **7**: e35294.

Yang, B., Y. Wang and P.-Y. Qian (2016). "Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis." *BMC bioinformatics* **17**: 135.

Yoon, S.-H., S.-M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo and J. Chun (2016). "Introducing EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies." *Int J Syst Evol Micr*, DOI:10.1099/ijsem.0.001755.

Yoon, S.-H., S.-m. Ha, J. Lim, S. Kwon and J. Chun (2017). "A large-scale evaluation of algorithms to calculate average nucleotide identity." *Antonie Van Leeuwenhoek*, DOI:10.1007/s10482-017-0844-4.

국문초록 (Abstract in Korean)

원핵미생물의 분류학 연구분야에서 16S rRNA 유전자의 서열을 이용한 방법은 지난 50 년간 널리 사용되어온 DNA-DNA hybridization (DDH) 을 대체하는 표준 방법으로 사용되었다. 원핵미생물 종을 구분하기 위해 사용되어온 DDH 기준 70%의 값은 16S rRNA 유전자 염기서열의 유사도를 기준으로 97%와 동등하게 이용되어왔다. 16S rRNA 유전자를 이용하여 세균과 고세균의 종을 완벽하게 동정하기에는 한계가 있음에도 불구하고 이 방법은 현재까지도 동일한 종 또는 속에 속하는 원핵미생물 균주에 대해 분류학적 위치를 파악하기 위해 가장 많이 사용되고 있는 방법이다. 그러므로 원핵미생물 분류에 있어서 16S rRNA 를 이용한 접근 방법은 여전히 중요하며 여기에 EzTaxon-e 와 같은 분류학적으로 잘 정의된 데이터베이스의 사용은 정확한 종의 동정을 위해 필수적이다. 최근에 차세대 염기서열 분석방법으로 불리는 DNA 염기서열 분석기술의 엄청난 발전에 힘입어 배양을 하지 않고 16S rRNA 를 이용한 미생물의 군집분석과 세균과 고세균의 분류 및 동정에서 좀 더 정밀하고 유용한 정보를 제공하는 게놈 염기서열 데이터의 사용이 가능하게 되었다. 현재 널리 받아들여지고 있는 원핵 미생물종의 정의는 동일한 종의 표준균주와 다른 균주들간에 게놈 염기서열을 비교하는 것에 기반하고 있다. 그러므로 다수의 게놈서열을 이용한 동정을 통해 원핵미생물의

다양성을 확인하고 또한 새로운 종을 쉽고 정확하게 발견하기 위하여 정밀하게 정의된 분류학적 정보를 가지고 있는 게놈 데이터베이스를 구축하는 것은 매우 중요한 일이다.

본 연구에서는 세균과 고세균의 분류학적 계층구조를 바탕으로 서열 정보에 대해 질적으로 관리된 16S rRNA 유전자와 게놈 서열이 통합된 최초의 데이터베이스인 EzBioCloud 데이터베이스를 구축하였다. 또한 데이터베이스 구축 및 활용에 사용되는 다양한 생물정보학적 파이프라인과, 툴, 그리고 알고리즘을 개발하였다. 16S rRNA 를 이용한 분석을 효율성을 높이기 위해 쌍 염기서열 정렬 알고리즘을 개선하였고 대량의 NGS 데이터를 활용한 미생물 군집분석에 적용가능한 빠르고 정확한 파이프라인을 개발하였다. 개발된 알고리즘을 통해 동일한 염기 쌍 정렬 결과를 약 1.5 배 빠르게 도출 가능하였으며 미생물 군집 분석 파이프라인의 경우 수만 리드 이상의 대량 염기서열 데이터에서 동일한 정확도를 가지고 이전보다 매우 빠른 분석속도를 보여주었다. 미생물 전체 게놈 분석을 위해서 어셈블리된 게놈정보의 품질을 평가하는 방법들과 게놈 분석 파이프라인을 개발하고 성능을 평가하였다. 또한 전장 16S rRNA 유전자를 추출하는 방법과 효율적인 Average Nucleotide Identity (ANI) 를 계산하는 알고리즘을 이용하여 공개된 원핵미생물 게놈을 이용한 동정에 사용하였다.

게놈 정보가 통합된 데이터베이스를 구축하기 위해서 NCBI Assembly Database 의 어셈블리된 전체 게놈 정보에 대해 품질이 낮은 게놈을 걸러내고 ANI 계산과 함께 유전자에 기반한 검색을 이용하는 복합적인 생물정보학적 동정 파이프라인을 적용하였다. 이러한 연구 결과결과 13,132 개의 인증된 분류명과 분류학적으로 속, 종 및 아종 수준에서 동정된 62,362 개의 게놈 정보를 포함하는 61,700 개의 종 및 계통형에 대한 데이터베이스를 구축하였다. 또한 데이터 수집속도와 빠른 기술 발전속도를 뒷받침 하기 위해 주기적인 업데이트가 가능하도록 많은 부분에서 자동화된 데이터 수집과정을 적용하였다.

이러한 분류체계 및 16S rRNA 와 게놈 서열 정보가 통합된 데이터베이스와 이를 뒷받침 하는 생물정보학적 도구들은 게놈을 기반으로 하는 세균과 고세균의 동정 및 분류에 대한 연구를 가속화 하며 나아가 기능 유전자에 대한 연구도 뒷받침하게 될 것으로 기대한다. 데이터베이스 컨텐츠 및 관련된 도구들은 http://www.ezbiocloud.net/ 웹사이트를 통해 접근 가능하도록 서버를 구축하여 공개하였다.

주요어: 생물정보학, 파이프라인, 데이터베이스, 16S rRNA, 게놈, 분류학, 마이크로바이옴, 차세대 염기서열 분석방법, 원핵미생물