이학박사학위논문

차세대 염기서열 분석 장비로 생성한 메타지놈 데이터 분석을 위한 최적의 생물정보학 시스템 개발

Development of Optimized Bioinformatic Analysis Systems for Next Generation Sequencing Data from Metagenome

2014년 2월

서울대학교 대학원

생물정보학 협동과정

전윤성

Development of
Optimized Bioinformatic Analysis Systems
for
Next Generation Sequencing Data from
Metagenome

by Yoon-Seong Jeon

Advisor: Professor Jongsik Chun, Ph. D.

A Thesis Submitted for the Partial Fulfillment of

the Degree of Doctor of Philosophy

February 2014

Interdisciplinary Program in Bioinformatics

Seoul National University

ABSTRACT

Metagenome is total DNA directly extracted from environment, and the purpose of metagenomics is to reveal the function of the metagenome as well as the taxonomic structure in the metagenome. There are two analysis approaches for metagenomics, namely amplicon based approach and random shotgun based approach. Both approaches require large scale sequencing reads which could not be satisfied through Sanger sequencing. However, high throughput sequencing of reads at relatively low cost by Next Generation Sequencing (NGS) technologies meets the requirement of metagenomics. In addition, the advent of NGS technologies gave rise to the development of bioinformatic algorithms necessary for processing this large and complex sequencing data. Consequently, the large amount of sequencing data obtained from NGS and corresponding proper bioinformatic algorithms facilitated the metagenomics to become essential tool for microbiology. However, limitations incurred by NGS sequencing errors, short read length, and lack of analysis system still hinder accurate metagenome analysis. Therefore, evaluation of currently used NGS error handling algorithms and development of systematic pipeline with more efficient algorithms are required to improve the accuracy of analysis.

In this study, bioinformatic pipelines were constructed for both metagenome analysis approaches. The pipelines were dedicated to improve the accuracy of the final end result by minimizing the effect of errors and short read length. For the amplicon based metagenomics, two different

analysis pipelines were developed for both 454 pyrosequencing and Illumina MiSeq. During the construction of 454 pyrosequencing pipeline, new error handling algorithm was developed to treat homo-polymer and PCR errors. Upon completion of the pipeline construction, household microbial community was analyzed using 454 pyrosequencing data as a case study. As for Illumina MiSeq data, the most appropriate sequencing conditions and sequencing target region were settled. Paired end merging programs were evaluated and correlation of the sequencing errors and quality was studied to correct the errors within 3' overlap regions. Novel iterative consensus clustering method was developed to correct the errors occurring ubiquitously in a single read.

For shotgun metagenomics approach, bioinformatic analysis system for Illumina MiSeq paired end data was constructed. Unlike the targeted amplicon sequencing reads, most of the shotgun sequencing reads are not merged; thus short reads are used for both functional and taxonomical profiling. However, a short read has less information than longer contigs, so the use of short reads is likely to cause biased characterization of the metagenome. Therefore, the development of analysis system did focus on creating longer contigs by means of mapping and de novo assembly. For raw read mapping, a dynamic mapping genome set construction method was developed. A list of mapping genomes was selected from the taxonomic profile inferred from the ribosomal RNA profiles. The genome sequence of the selected genomes were downloaded from Ezbiocloud. By mapping raw reads to the genome sequences, the longer contigs can be obtained in case of the relatively simple metagenome such as fecal matter.

However in case of the complex metagenomes such as soil sample, both mapping and de novo assembly did not perform properly due to a lack of sequencing coverage and numerousity of uncultured microorganisms in the metagenome. In addition to the pipeline construction, visualization tools were also developed to display resultant taxonomic and functional profile at the same time.

Newly developed JAVA-based standalone sequence alignment editing application was named as EzEditor. As both, conserved functional coding sequences and 16S rRNA gene have been used copiously in bacterial molecular phylogenetics, the codon-based sequence alignment editing functions are required for the coding genes. EzEditor provides simultaneous DNA and protein sequence alignment editing interface which enables us with the robust sequence alignment for both protein and rRNA sequences. EzEditor can be applied to various molecular sequence involved analysis not only as a basic sequence editor but also for phylogenetic application.

Keywords: Bioinformatics, Next generation sequencing, Metagenome, amplicon, shotgun, analysis pipeline, EzEditor, phylogenetics

TABLE OF CONTENTS

ABSTRACT ..... **I**

TABLE OF CONTENTS ..... **IV**

ABBREVIATIONS ..... **VI**

FIGURE LIST ..... **VII**

TABLE LIST ..... **XII**

Chapter 1 General Introduction **1**

1.1 Bioinformatics **2**

1.2 Next Generation Sequencing **5**

1.3 Metagenomics **11**

1.4 Objectives of This Study **21**

Chapter 2 Amplicon-based Metagenome Analysis Systems **23**

2.1 Introduction **24**

2.2 Analysis System for 454 Pyrosequencing **35**

2.2.1 Methods **36**

2.2.2 Results **39**

2.3 Analysis System for Illumina MiSeq **60**

2.3.1 Methods **62**

2.3.2 Results **68**

2.4 Summary and Discussion **93**

Chapter 3 Shotgun-based Metagenome Analysis System 99

3.1 Introduction 100

3.1.1 Tools for Metagenomics 101

3.2 Methods 118

3.3 Results 125

3.4 Summary and Discussion 165

Chapter 4 EzEditor: A versatile Molecular Sequence Editor for Both Ribosomal RNA and Protein Coding Genes 169

4.1 Overview 170

4.2 Features of EzEditor 172

4.2.1 Algorithms and Models Implemented in EzEditor 177

4.2.2 Miscellaneous Functions 178

4.3 Summary and Discussion 181

Conclusions 183

References 187

APPENDIX I. Estimated Diversity Index of Household Microbiome 217

국문 초록 (Abstract in Korean) 221

감사의 글 (Acknowledgement) 227

ABBREVIATIONS

Avg.: Average

BLAST: Basic local alignment search tool

CDS: Coding sequence

DNA: Deoxyribonucleic acid

EM: Expectation Maximization

EmPCR: Emulsion PCR

HMM: Hidden markov model

NCBI: National center for biotechnology information

NGS: Next Generation Sequencing

OTU: Operational taxonomic unit

PE: Paired end

PcoA: Principal coordinate analysis

PCR: Polymerase chain reaction

RDP: Ribosomal Database Project

rRNA: Ribosomal ribonucleic acid

TBC: Taxonomy-based clustering

UPGMA: Unweighted pair-group method with arithmetic means

FIGURE LIST

Figure 1. Typical analysis steps for bacterial community study. 15

Figure 2. Nine hypervariable regions in 16S rRNA gene..... 27

Figure 3. Flowgram of 454 platform. 29

Figure 4. Illustration showing formation of a chimera during PCR reaction. 34

Figure 5. Scheme of clustering-based homopolymeric error handling algorithm. 37

Figure 6. Bar graph showing ratio of qualified reads obtained after two mock community data sets were analyzed by different denoising programs..... 40

Figure 7. Time and memory usage of denoise programs. 43

Figure 8. Sensitivity of the chimera detection programs and different db. 45

Figure 9. Overall microbial community analysis pipeline for analyzing pyrosequencing data. 47

Figure 10. The average compositions of bacterial communities obtained from the vegetable compartments of refrigerators and from toilets by using culture-independent method 52

Figure 11. A pie chart diagram showing the proportion of species shared between human skin and gut microbiomes with bacterial community from refrigerator and toilet..... 56

Figure 12. Similarities between bacterial communities that originated from refrigerator, toilet, human skin, and gut samples as visualized by a PCoA plot. 58
Figure 13. Average quality of Illumina MiSeq PE. Quality score is getting lower as the length of the sequence is getting longer. 61
Figure 14. Illumina MiSeq paired sequencing scheme. 64
Figure 15. Evaluation scheme for paired end read merging. 65
Figure 16. Number of reads obtained from a sequencing run varies with both DNA library concentration and PhiX ratio. 73
Figure 17. The number of merged reads is reduced when trim 50 nucleotides at the end of the read. 76
Figure 18. The number of substitutions per read increases as the trim length increases. 77
Figure 19. Correlation between errors and quality score of Illumina MiSeq read. 81
Figure 20. Plot showing the number of mismatches within overlap region of merged read and the similarity to the template sequence. 82
Figure 21. Errors are observed ubiquitously and no pattern is shown. Integrative Genomics Viewer. 84
Figure 22. Taxonomic composition bias of three different error correction methods. 88
Figure 23. Comparison of taxonomic composition at genus and phylum

levels assigned to two different amplicon sets with original composition. 89
Figure 24. Taxonomic composition of pig fecal sample. Two different analysis methods were compared with the original data set. 91
Figure 25. Illustration of overall microbial community analysis pipeline for Illumina MiSeq paired reads. 92
Figure 26. Microbial community recovered from 454 junior and Illumina MiSeq. Inner pie represents phylum and outer pie genus level composition. 96
Figure 27. General metagenome analysis pipeline implemented by the open metagenome pipelines including MG-RAST, EBI and IMG/M. 115
Figure 28. The number of contigs obtained from each mapping program. 127
Figure 29. N50 and N90 statistics of obtained from each program..... 128
Figure 30. The number and length of the contigs varies upon the given Kmer. 132
Figure 31. The number and length of contigs of metagenome de novo assemblers. 133
Figure 32. The ratio of missing ORFs of gene calling programs. 137
Figure 33. Positive Predictive Value indicates gene prediction sensitivity of gene calling programs. 138
Figure 34. Metagenome functional annotation DB consistis of Ezgenome and Pfam..... 139

Figure 35. Computational performance of mapping and assembly programs.141

Figure 36. Schematic illustration of overall metagenome shotgun analysis pipeline. 142

Figure 37. Comparison of the number of phylotypes at each taxonomic level between MG-RAST and Chunlab..... 144

Figure 38. COG annotation profile of MG-RAST and Chunlab pipeline. ... 148

Figure 39. SEED annotation profile of MG-RAST and Chunlab pipeline... 149

Figure 40. Visualization - Summary table of metagenome analysis. 150

Figure 41. Visualization - Taxonomic hierarchy inferred from functional profile. 153

Figure 42. Visualization - Taxonomic composition information with the relative abundance of Seed category, COG category, and Gene Ontology. 154

Figure 43. Visualization - Subcategory of selected SEED and COG category in *Proteobacteria*. 156

Figure 44. Visualization - Comparative analysis is available by loading more than two metagenomes at the same time. 157

Figure 45. Visualization - Tables for annotated gene profiles refined from SEED subsystem, NCBI, and COG..... 158

Figure 46. Visualization - Example of visualization of taxonomy composition of functional coding gene. 159

Figure 47. Time and memory usage of mapping and assembly programs for

each sample. 164

Figure 48. Main work space of EzEditor. The work space can be loaded with multiple data files at once. 172

Figure 49. All information except DNA sequences is shown in SelectPanel. Meta data of selected sequence in the left panel is shown in the right panel..... 174

Figure 50. Align Panel of 16S ribosomal RNA. Secondary structure pairing information of selected sequence in DNA alignment panel is shown. 175

Figure 51. Align Panel of functional coding sequence. Protein sequence alignment is shown in the panel below DNA alignment panel. 176

Figure 52. Secondary structure pairing information can be used to assess the robustness of the sequence alignment. 179

Figure 53. Sequence similarity to all the other sequences in the dataset is shown. 180

TABLE LIST

Table 1. Comparison of next-generation sequencing platforms. 10

Table 2. Biased taxonomic composition at genus level..... 41

Table 3. Noise reads are processed applying the quality and length cutoff parameters. 42

Table 4. Comparison of simulated amplicon by different combinations of two variable regions. 70

Table 5. Comparison of Illumina MiSeq paired end merging programs. 75

Table 6. Merging statistics given different threshold of end trimming length cutoff. 80

Table 7. The number of clusters generated after each round of clustering..... 86

Table 8. The number of recovered phylotypes by two NGS platforms. 97

Table 9. Various algorithms and tools used for metagenome analysis. 110

Table 10. Database of functional coding sequence for metagenome annotation. 112

Table 11. Detailed process of metagenome analysis pipeline of the public metagenome pipelines. 116

Table 12. List of 5 known bacterial strains comprising mock metagenome. 121

Table 13. Sequencing results of Illumina MiSeq machine. 125

Table 14. Result of taxonomic profiling using HMM profile of three types

of ribosomal RNA. 126

Table 15. Consensus contig sequences obtained from reference mapping process. 130

Table 16. Metagenome de novo assembly result. 135

Table 17. The number of predicted orfs in both chromosomes and raw reads. 136

Table 18. Content of the annotation database for prokaryotes. 140

Table 19. 10 most abundant genus from Chunlab pipeline and MG-RAST. 145

Table 20 Summary of metagenome shotgun sequencing of a soil and fecal sample. 160

Table 21. Estimation of sequencing coverage. 161

Table 22. Result of reference mapping and de novo assembly of soil and fecal metagenome shotgun reads 163

Table 23. Data field of EZE file and example data. 173

Chapter 1 General Introduction

1.1 Bioinformatics

The core definition of bioinformatics can be defined as the science rooted in life science which helps us to understand life events with the aid computational science along with other contributing disciplines such as statistics and mathematics (Huerta *et al.*, 2000). According to the definition of the bioinformatics by the National Institute of Health, USA, "bioinformatics includes research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, archive, or visualize such data. And, from the perspective of biological science, bioinformatics can be called as computational biology meaning biology involving computational science encompassing the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to study of biological systems". The term 'Bioinformatics' is relatively recent invention and was firstly coined by P. Hogeweg and B. Hesper (1978) to refer to the study of information processes in biotic systems such as biological modeling, building database and developing biological sequence analysis algorithms. There are many research fields related to bioinformatics which are usually the inter-disciplinary such that two or more disciplines converge (Searls, 2010).

Early bioinformatics era was primarily focused on the sequence

analysis which involved saving, retrieving, analyzing or predicting the composition or the structure of the biomolecules (Luscombe *et al.,* 2001). But in the genomics era, bioinformatics did step toward more complicated tasks such as short fragmented sequence assembly (Miller *et al.*, 2010; Pop, 2009) and higher level of data mining and data modeling, and with the advent of post genomic era, bioinformatics shifted its focus to the comparative or functional genomics of the completed genomes (Binnewies *et al.,* 2006; Horner *et al.*, 2010). Many algorithms designed and developed at that time were the conceptual nourishment for the current bioinformatical analysis algorithms.

Bioinformatics and it's related research fields have undergone a huge change with the advent of new sequencing technology which was named as Next Generation Sequencing or NGS (Metzker, 2009; Schuster, 2007; Shendure *et al.*, 2008). Compared to traditional Sanger DNA sequencing technology, NGS has following remarkable characteristics, 1) large volume of output, 2) unparalleled speed of sequencing and 3) lower cost of per base sequencing (Liu *et al.,* 2012b). Because analysis of this large volume of output data is not feasible without higher level of computing resources, bioinformatics has become the indispensable discipline to make NGS more practical (Bateman *et al.,* 2009). Besides, sequencing is no more a bottleneck as the cost of sequencing is getting cheaper and cheaper so that the small scale academic laboratories can use the NGS sequencing technology, but the computation is more of a bottleneck (Desai *et al.*, 2012). The Moore's law (Schaller, 1997) describing the trend whereby the number of transistors that can be loaded on integrated circuit doubles approximately

every 2 years, and this law is applied to the sequencing cost indicating that more and more biological laboratories will be getting into the reach of the NGS technology, moreover the trend of NGS is going toward the longer reads and higher throughput meaning that the analysis of the NGS data will never be done without the aid of bioinformatics (Metzker, 2009).

Management and organizing the processed data is another responsibility of bioinformatics (Luscombe *et al.*, 2001). As the sequencing cost is getting affordable and development of bioinformatics technique allows to process such data, more and more researchers generate biological data and then submit their data to the public domain such as NCBI (http://www.ncbi.nlm.nih.gov) or EBI (http://www.ebi.ac.uk) ending up growth of these public database at an astonishing rate. Of course, the expansion of the database is having a positive effect on the biological research however extracting and refining data from these public databases to build up well organized and curated secondary database is mandatorily essential for the biological research.

Visualization of the biological data is also an important task of bioinformatics (Hamady *et al.*, 2010; Tao *et al.*, 2004). NGS data is one of the major agents promoting evolution of biological research into the big data science. Consequently, there emerges an urgent and growing need for improved methods and tools to be used for gaining insights and understanding from the biological data .

1.2 Next Generation Sequencing

Since the Sanger sequencing method was introduced in 1977 (Sanger *et al.*, 1977), this first generation sequencing technology of enzymatic dideoxy technique has been the gold standard of sequencing methodology in biology and medicine and led to a number of monumental accomplishments including the completion of the human genome project (Collins *et al.*, 2004). But, in the past decades, NGS technologies have been introduced and have gradually overcome the inertia of a field that relied wholly on Sanger-sequencing for more than 30 years. As NGS technologies have delivered on its promise of sequencing DNA at an unprecedented speed and cost, which led to impressive scientific achievements and novel biological applications (Schuster, 2007), and have substantially widened the scope of biological disciplines. Together with enormous advancement of bioinformatic analysis algorithms and data processing technology, NGS is now accelerating and altering a wide variety of biological inquiry. The principal sequencing chemistry and output specification of each NGS platform are different with each other, hence the application of each NGS platform may be different to each other also. Thus, it is of importance to understand the background chemistry and feature of output reads including the length and pattern of errors or error rate (Yang *et al.*, 2012). Currently available NGS technology can be divided into two categories according to template preparation method or randomly broken DNA fragment (Metzker,

2009). There are clonally amplified templates and single-molecule templates. Clonally amplified templates technology includes emulsion PCR (Tawfik *et al.*, 1998) which is used in Roche-454 (Margulies, 2005), Life/APG (Valouev *et al.*, 2008), and Polonator (Shendure *et al.*, 2005), and solid-phase amplification used in Illumina/Solexa (Bentley, 2006). Single-molecule template includes Helicos Biosciences (Harris *et al.*, 2008) and Pacific Biosciences (Eid *et al.*, 2009). It requires relatively small amount of genomic DNA material (<1 nanogram) and does not rely on PCR which could cause the bias . Also quantitative applications, such as RNA-seq (Wang *et al.*, 2009), perform more effectively with non-amplified template sources, which do not alter the representational abundance of mRNA molecules.

454 Pyrosequencing

Principle behind the pyrophosphatic detection, as the basic principle of pyrosequencing, was described firstly in 1985 (Nyrén *et al.*, 1985) and the first system based on the principle for DNA sequencing was reported in 1988 (Hyman, 1988) followed by the further development into a routinely functioning technique leading to a commercialized technique in parallel microtiter plate (Ronaghi *et al.*, 1996). In 2005, 454 Life Sciences (later acquired by Roche; http://www.454.com) introduced the GS device as their first system of the next-generation DNA sequencer in the market. Cloning required for Sanger sequencing was prevented by making use of the Emulsion PCR (emPCR). The principle of the 454 pyrosequencing is

sequencing-by-synthesis that measures the intensity of the light released by chemiluminescence (Ansorge *et al.*, 1986; Ansorge *et al.*, 1987) accompanied by nucleotide incorporation . The sequence of DNA is determined from a pyrogram which corresponds to the order of correct nucleotides that had been incorporated. Since chemiluminescent signal intensity is proportional to the amount of pyrophosphates released and hence the number of bases incorporated, the pyrosequencing approach is prone to errors that result from incorrectly estimating the length of homo-polymeric sequence stretch. This error type of indel (undercall/overcall) is the typical error of 454 pyrosequencing. The error rate of the pyrosequencing was known to be about 0.5% and the errors caused by homo-polymeric nucleotides is known to account for 39% of the total errors (Huse *et al.*, 2007). But recently the error rate was reported to be over 1.0% (Gilles *et al.*, 2011).

Illumina-Solexa Genome Analyzer

In 2006, another next generation sequencing platform was commercialized by the Solexa (later acquired by Illumina; http://illumina.com). The principle of the system is based on sequencing-by-synthesis chemistry which was originally developed by Shankar Balasubramanian and David Klenerman, co-founder of Solexa. This sequencing-by-synthesis method uses novel reversible chain terminator nucleotides for the four bases each labeled with a different fluorescent dye, and a special DNA polymerase enzyme able to incorporate them. Illumina

Solexa is high-throughput resulting in an extremely large number of reads. The read length of the Illumina Solexa is shorter than 454 pyrosequencing. But the paired end read generated by Illumina MiSeq platform makes the length of the merged 250bp paired end read approach the length of the pyrosequencing reads. The error type prevalent with Illumina solexa is substitution errors (miscall) and is different from that of the 454 pyrosequencing. Another known feature of the Illumina Solexa reads is that the quality of the base calling is getting worse as the read length is getting longer, hence the errors are observed more frequently within the 3' region of the read than the 5' end.

ABI SOLiD System

The ABI SOLiD (Sequencing by Oligo Ligation and Detection) next generation sequencing system, a commercial platform using a unique sequencing chemistry so called sequencing-by-ligation is based upon ligation and catalysis by DNA ligase, was introduced in the market in 2007.

Single molecule sequencing

Single molecule sequencing is sometimes referred to as third-generation sequencing (Check, 2009) partly because it eliminated cumbersome sample preparation steps, including complex ligations and polymerase chain reactions for amplification. Pacific Bioscience introduced Single Molecule Real-Time (SMRT) DNA sequencing technology (McCarthy, 2010). The method of real time sequencing involves imaging the continuous

incorporation of dye-labeled nucleotides during DNA synthesis. RS2 instrument of the Pacific Bioscience is capable of 2GB nucleotide per 2-hour run with maximum read length 4K according to the brochure from the company, albeit the error rate is more than 10% and is still is in dire need of improvement. Helicos single-molecule sequencing utilizes sequencing-by-synthesis methodology and novel technology called Virtual Terminator nucleotides (Bowers *et al.*, 2009).

Table 1. Comparison of next-generation sequencing platforms.

Platform	Template	Chemistry	Avg. read length (bp)	Run time (day)	Gb per run
GS Flx Titanium	EmPCR	Pyrosequencing	330	0.35	0.45
Illumina HiSeq 2500	Solid phase	Reversible terminator	2 x 100	2-11	600
Illumina MiSeq	Solid phase	Reversible terminator	2 x 250	0.2~2.5	1
Life SOLiD 5500 xl	EmPCR	sequencing by ligation	50	1	20
Helicos Bioscience Heliscope	Single molecule	Reversible terminator	32	8	37
Pacific bioscience RS2	Single molecule	Real-time	4K	0.1	2

1.3 Metagenomics

Metagenome means all the genetic material existing in an environmental sample. The study of metagenome or metagenomics, firstly coined by Jo Handel man (1998), is a discipline that aims to fully characterize a metagenome by revealing the composition of the microbial inhabitants and their biological functions. The metagenome has been considered as a key to understanding our environment not only because they are ubiquitous but also they are essential to all life as they are the primary source for nutrients. Further, there are ten times more bacterial cell inhabiting our body than our own cells (Berg, 1996). Understanding microbes are vital for completely understanding human . However, as the 'Great Plate-Count Anomaly' (Staley *et al.*, 1985) highlighted, our understanding of microbes has been highly skewed towards a small fraction of readily culturable bacteria. A couple of studies (Schmidt *et al.*, 1991; Stein *et al.*, 1996; Vergin *et al.*, 1998) addressed this issue by making use of directly extracted environmental DNA to unravel the phylogenetics and functional diversity, which setup the beginning of the metagenomics. There have been several landmark studies which demonstrated the power of metagenomics including the study of Sargasso Sea (Venter *et al.*, 2004) identifying more than hundred novel phylotypes. Tyson and colleges reconstructed the genome sequences of unculturable bacteria from a low diversity metagenome to reveal complete metabolic pathway and hence to

elucidate their functional and nutritional properties (Tyson *et al.*, 2004). These two landmark studies evidently stated the power of metagenomics and shows that this relatively newcomer to science has become one of the important and indispensable tools to expand our understanding of the microbial world as well as ourselves.

Metagenomics study could be divided into two research areas driven by technical applications (Gilbert *et al.*, 2011; Scholz *et al.*, 2012); Environmental single-gene surveys and random shotgun studies of all environmental genes. The former also could be called targeted or focused metagenomics where single targets are amplified using polymerase chain reaction and then the products/amplicons are sequenced. Random shotgun metagenomics is a study in which total DNA is isolated from a sample and then sequenced resulting in a profile of all genes within the community. As mentioned previously, characterization of metagenome involves the community structure analysis which is usually performed while making use of the amplicon based metagenomics targeting a phylogenetic marker such as 16S rRNA gene (Fan *et al.*, 2012; Huang *et al.*, 2009). On the other hand, to elucidate the functional aspect of the metagenome, random shotgun metagenomics is applied. There have been several approaches (Arthur Brady *et al.*, 2009; McHardy *et al.*, 2007; Sharma *et al.*, 2012) to retrieve the taxonomic hierarchy or microbial diversity information from random shotgun metagenome reads, targeted metagenomics is considered to study bacterial diversity albeit the skewed result caused by PCR amplification. Since the beginning of the metagenomics which was the Fosmid, BAC-derived method (Gilbert *et al.*, 2011), metagenomics has experienced a

huge changes and this change was continued and accelerated by the advent of the new sequencing technology, Next Generation Sequencing. The first NGS platform introduced to the market was Roche 454 GS (Margulies, 2005) and thereafter many NGS platforms have been introduced in the market and enormous development has been achieved making large scale metagenomic studies practical and cost-effective. Further improvement in NGS is expected toward the longer read length, higher throughput and declining the cost of the sequencing, which brings the costly and time consuming metagenomics within small scale laboratories.

Amplicon-based Metagenomics

Understanding the metagenome begins with illustration of the composition, organization and spatial distribution of the microbes in the metagenomic community (Temperton *et al.*, 2012). PCR based single gene metagenomics has been exploited to explore the microbial diversity and taxonomy targeting the 16S ribosomal RNA. Although NGS has overcome the drawback of insufficient sample size and thus low coverage driven by the Sanger sequencing, short read length is the limitation by which the full length of the target sequence could not be covered (Wommack *et al.*, 2008). It is considered to be ideal to use full length of the target gene for the community analysis; however, no NGS platform so far provides such a long read length. Thus, the only one out of nine hypervariable regions (Kumar *et al.*, 2011a) or the combination of two to three regions are amplified for the community analysis. This short sequence

tags surrogating the full length 16S rRNA gene is reported to provide a stable estimate of the abundance of the each phylotype in the microbial community, although the selection of the region affects the diversity estimation (Liu *et al.*, 2008; Liu *et al.*, 2007). It is the 454 pyrosequencing that dominated NGS adopting diversity studies and numerous bacterial community studies (Caporaso *et al.*, 2011; Jeon *et al.*, 2013; Sogin *et al.*, 2006). Recently, Illumina MiSeq, generating 250bps read whose paired end merged length approximately approaches the length of the 454 pyrosequencing, is subjected to the test for the feasibility of bacterial diversity estimation (Degnan *et al.*, 2012; Kozich *et al.*, 2013).

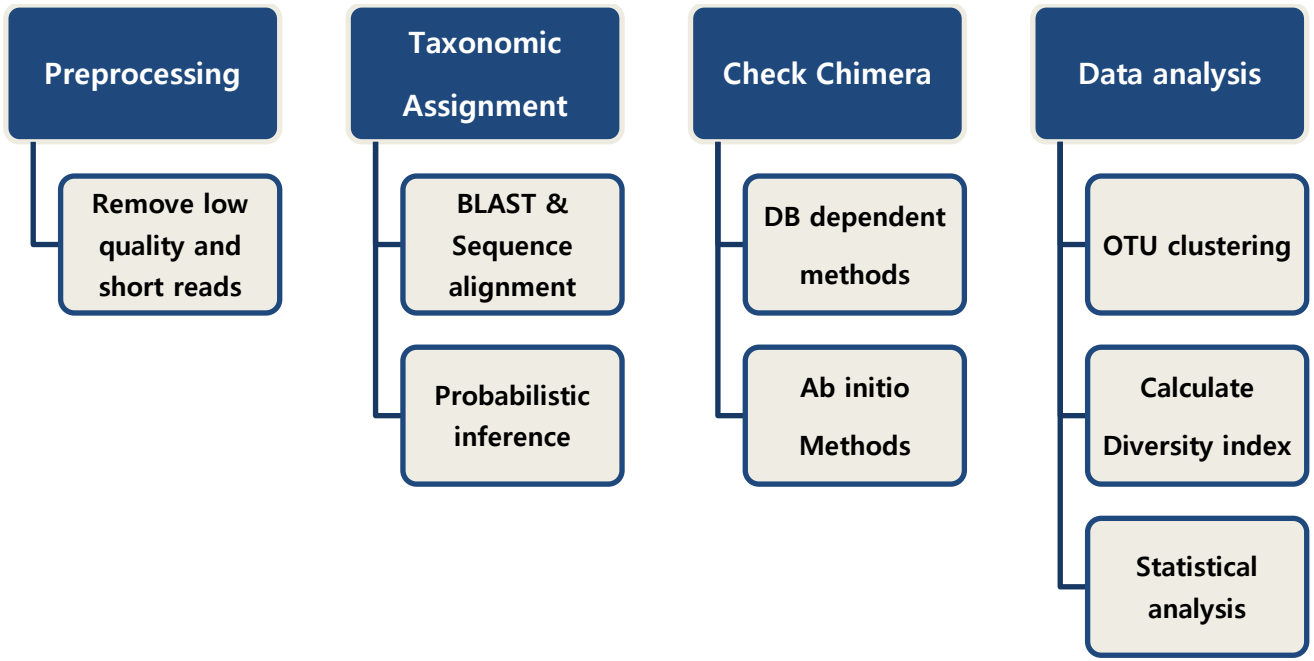


Figure 1. Typical analysis steps for bacterial community study.

In addition to the short read length, another factor that can influence the bacterial community analysis is the sequencing errors. Unlike genome analysis where the sequencing does not depend on the PCR amplification, here errors in sequencing reads could be corrected by sequencing coverage. Errors within a read in the community analysis could be identified as a novel sequence and in effect can inflate the bacterial community (Kunin *et al.*, 2010). Therefore, errors should be removed or corrected in a proper manner. The NGS platform specific errors also could plaque analysis results. Pyrosequencing specific error type is insertion/deletion caused by a homo-polymeric sequence (Margulies, 2005) while the substitution error (Dohm *et al.*, 2008) which occur more frequently in the Illumina reads are substitutions observed more at the 3' end. Several algorithms and programs were introduced to handle the pyrosequencing errors (Bragg *et al.*, 2012; Quince *et al.*, 2011; Reeder *et al.*, 2010) whereas little algorithm is released for the Illumina error correction because the applications of the Illumina platform have been limited to the re-sequencing or genome project where the sequencing coverage is more important condition than error frequency. Therefore, another error handling approach is required for the construction of Illumina platform based analysis pipeline.

The artificial chimeric read originating during the PCR amplification process is known to account for the 10% of all the data and many algorithms have been designed (Haas *et al.*, 2011; Huber *et al.*, 2004; Rosen *et al.*, 2012; Wright *et al.*, 2012) to detect the chimeric sequence. In addition, the chimeric sequence detection depends on the quality of the

database but the public databases are reported to have chimeras in one out of 20 sequences (Hugenholtz *et al.*, 2003), thus the selection of the database for the chimera detection is another important issue.

Typical bioinformatic analysis steps for bacterial community studies is composed of related steps as shown in Figure 1. Preprocessing step includes de-multiplexing of pooled samples (Hamady *et al.*, 2008) and noise filtering process. For the taxonomic assignment, BLAST search against database is widely used method and RDP uses probabilistic approach implementing Naïve Bayesian algorithm (Wang *et al.*, 2007). There are well curated public databases such as Greengenes (DeSantis *et al.*, 2006), RDP (Maidak *et al.*, 2001), Silva (Pruesse *et al.*, 2007) and Eztaxon-e (Kim *et al.*, 2012). Blast search is often followed by the pairwise alignment (Eztaxon-e and Greengenes) or multiple alignment (Huse *et al.*, 2008). QIIME (Caporaso *et al.*, 2010) and mothur (Schloss, 2009b) are open, widely used analysis pipelines providing versatile functions required in each analysis step. Development of various type of pipelines with different combination of software are still in need. Besides, as the NGS techniques continue to improve toward higher throughput and longer read length, development of corresponding suitable analysis pipelines will become necessarily.

Random Shotgun Metagenomics

The ultimate goal of metagenomics is to explain the functions of metagenome but the amplicon-based approach does not provide sufficient

and accurate information about their functions because primer bias may skew our knowledge towards only the amplified target strains even though there are non-amplified target sequences of rare species. Further amplification efficiency may also alter the true abundance of the strains. To circumvent these limitations of the PCR-based targeted metagenomics, random shotgun metagenome approach could be an attractive alternative strategy for the functional characterization of the metagenome. Although this random shotgun sequencing approach offers promise of more comprehensive insight concerning the metagenome, bioinformatics analysis is far more complicated and demanding here than the amplicon based approach largely due to the huge number of short and partial sequencing reads and the inherent feature of the metagenome. With the advent of the next generation sequencing, it is possible to say that sequencing is becoming no more a bottleneck at least for the metagenome analysis. For example, for the soil sample which is considered to have the most complex microbial community, it was estimated that the minimum 6 billion nucleotides is required to recover the genome sequence of the dominant organisms.

Bottleneck of the metagenomics analysis at the moment is the limitation of computational resources and the incomplete reference protein sequence databases. The intrinsic characteristics of NGS output data, which comprises of a higher throughput and short read length, have led to the development of numerous bioinformatic algorithms and softwares. However, because the initial development of NGS was motivated and driven for the isolated single genome analysis, most of the bioinformatics

algorithms have been focusing on the treating the sequencing reads originated from the single homogenous genome sequence. The metagenome data coming from mixed heterogeneous microbial community containing, sometimes more than 10,000 genomes is totally different data in terms of bioinformatic analysis. Although, these softwares were applied to the metagenome analysis (Kunin *et al.*, 2008) at the beginning of the metagenomics, most of these were not feasible for metagenome analysis and hence a whole new, different algorithm was required for its analysis (Desai *et al.*, 2012). During the past few years, metagenomics has seen an explosion in computational methods applicable to the studies. But it is fair to say that most of these programs were developed for shotgun metagenomics and still are at the premature stage. In addition to the lack of appropriate reads processing programs, insufficient reference database is another hardship in metagenome analysis. One of the notable advantages of metagenomic study is that the study can reveal those previously unculturable genomes and their component genes and hence, the more robust microbial community diversity and functional properties can be attained. But at the same time, the availability of unculturable bacteria is another obstacle dwelled in the metagenome analysis. The patterns and protein sequences deposited to the public databases to date were based on the previously reported data that was extracted from the culturable genomes. Thus the component genes identified from the unculturable genomes are likely to be left as hypothetical and this phenomenon continued only to widen the gap between characterized proteins and the hypothetical proteins (Tyson, 2008) as more metagenomic researches are being undertaken.

Several metagenome PaaS (platform as a service) including MG-RAST (Glass *et al.*, 2010), IMG/M (Markowitz *et al.*, 2008), CAMERA (Seshadri *et al.*, 2007) and EBI (https://www.ebi.ac.uk/metagenomics) are available with unique pipelines and databases. Also, many metagenome consortium such as Global Ocean Sampling (Rusch *et al.*, 2007), Human Microbiome Project (Turnbaugh *et al.*, 2007), Earth Microbiome Project (Gilbert *et al.*, 2010) and MetaHit (Ehrlich, 2011) have their unique in-house pipelines and reference database. So, metagenome analysis pipeline can vary depending on various factors including target environment and sequencing platform. However, as most of the softwares used in the metagenome analysis are still immature, development of various metagenome pipelines and analysis software corresponding to the development of NGS is required.

1.4 Objectives of This Study

The purpose of this study was to develop bioinformatic analysis systems required to analyze metagenome sequencing data obtained from NGS machines. As two metagenome analysis approaches, targeted amplicon sequencing approach and random shotgun approach, share little common data processing steps, two different types of pipelines corresponding to each metagenome analysis approach needs to be developed. In addition, for the targeted amplicon analysis, two major NGS machines have their own specific unique features in output data including read length and error types . So, two distinct analysis systems corresponding to both NGS platforms were developed for a targeted metagenome approach. All the pipelines developed in this study focused on the sequencing error handling to improve the accuracy of the analysis.

In order to develop 454 amplicon analysis pipelines, a novel homo-polymeric error handling algorithm was implemented and compared to other error handling programs to evaluate its specificity and sensitivity. In Illumina MiSeq paired end analysis pipeline, paired end overlap region error and non-overlap region error needed to be handled separately. Errors in overlapping region between a pair of reads were studied to assess the effect of the merging process by means of a paired end merging evaluation program which was developed in this study. For a non-overlap region error, novel consensus clustering method was implemented to avoid bias caused

by sequencing error.

Analysis system for random shotgun metagenomics was also developed. Since longer contigs are helpful in reducing the bias caused by short reads, a raw read mapping process was incorporated into the analysis pipeline. By means of rRNA profile, the pipeline dynamically configured a mapping database to which the raw reads were aligned. During the construction of the pipelines, all components programs in each step such as chimera detection, de novo assembly and mapping programs were evaluated to organize an optimized analysis pipeline.

Sequence alignment editor, EzEditor was developed to provide codon based alignment editing functions capable of phylogenetic analysis based on conserved coding sequences obtained from metagenome analysis.

Chapter 2 Amplicon-based Metagenome Analysis Systems

2.1 Introduction

Microbial community analysis has experienced a huge change since the next generation sequencing technology was introduced. Before the NGS era, culture-independent PCR-based microbial community analysis using Sanger sequencing method had been the main stream of the analysis. However, the higher per base sequencing cost and relatively small number of reads per single sequencing run enforced the Sanger sequencing method to be replaced with NGS.

PCR based amplicon based approach, sequences only a small part of target gene which can surrogate the entire target gene sequence (Liu *et al.*, 2008; Liu *et al.*, 2007). The target sequencing region must be highly conserved, so that the nucleotide variations such as substitutions or indels within the conserved target region could be the evidence of the evolutionary events (Woese, 1987). This infers that the bacterial diversity study is so vulnerable to the errors that even a single nucleotide substitution or indel error may lead to skewed diversity estimation and this susceptibility to errors is gets worse with the shorter reads. Meanwhile, as the NGS produces large quantity of reads, the basic question arises concerning the possible detrimental effects of this shift in quantity over quality of the obtained data (Kunin *et al.*, 2010). In practice, metagenome community study of 16S rRNA gene using NGS could be plagued by technical error, thus it is important to separate the noise from the actual data

not just for the correct assessment of a microbial community, but also to separate the novel organism from sequencing noise.

Several bacterial diversity studies using NGS reported that the extent of rare microbial populations in several environment, “rare biosphere”, was many orders of magnitude lager and more diverse than those previously appreciated (Huber *et al.*, 2007; Roesch *et al.*, 2007; Sogin *et al.*, 2006), but Kunin and colleges (2010) addressed that the intrinsic NGS sequencing errors may resulted in skewed community structure. Some other studies also reported that the majority of the rare biospheres were highly composed of non-authentic novel sequences, and the inflated diversity caused by the errors was reduced to the factor of 10 (Huse *et al.*, 2008; Quince *et al.*, 2009). In addition, the PCR related errors such as artificial chimeric sequences also were known to affect the species richness and evenness (Engelbrektson *et al.*, 2010). Ideally, the sequencing accuracy is the fundamental prerequisite for the bacterial diversity estimation however sequencing errors are common and are difficult to avoid. Therefore, treating the errors either by detection or by correcting them is one of the most critical step for the amplicon based microbial community analysis.

In this chapter, bacterial community analysis pipeline for both 454 pyrosequencing and Illumina MiSeq platforms were constructed while focusing on reducing both the NGS sequencing errors and PCR related errors. 454 pyrosequencing platform has been used predominantly for amplicon analysis due to the longest read length among NGS platforms. So, during the construction of the 454 pipeline, not only novel denoising step was developed but those previously developed programs were evaluated

and compared to assess an optimal pipeline. As for the pipeline of Illumina MiSeq platform, the paired end merging process was analyzed in the perspective of the error correction and quality of the merged reads. To reduce bias caused by the substitution errors occurring outside the overlapping regions, novel iterative consensus clustering approach was implemented.

Target gene and hypervariable regions

16S ribosomal RNA gene is the major target genetic marker for bacterial and archaeal community analysis whose length is about 1500bps approximately. Because the read length of NGS, even the longer 454 read, is not enough to cover the whole region of the target gene, only the small hypervariable regions of the 16S rRNA gene have been targeted for the NGS sequencing. As shown in Figure 2, nine hyper variable regions in 16S ribosomal RNA gene are known to be flanked by the conserved regions. The entropy of each region is different from each other, thus certain regions show better discriminative capability for a specific bacterial lineage (Chakravorty *et al.*, 2007; Kumar *et al.*, 2011a). However, no region is known to be the best sequence tag and this is the reason why the many different studies used different regions for bacterial and archaeal community analysis (Huse *et al.*, 2008; Liu *et al.*, 2008; Sundquist *et al.*, 2007). Further, because species richness and community evenness is affected much more by regional variation than the other variables such as DNA extraction and PCR related bias (Acinas *et al.*, 2005; Engelbrektson *et*

al., 2010; Kumar *et al.*, 2011a), target region should be chosen carefully according to the purpose of the study. In this study, the combined V1~V3 region and V4~V5 region were used for the construction of the bacterial community analysis pipeline for 454 pyrosequencing and Illumina MiSeq, respectively.

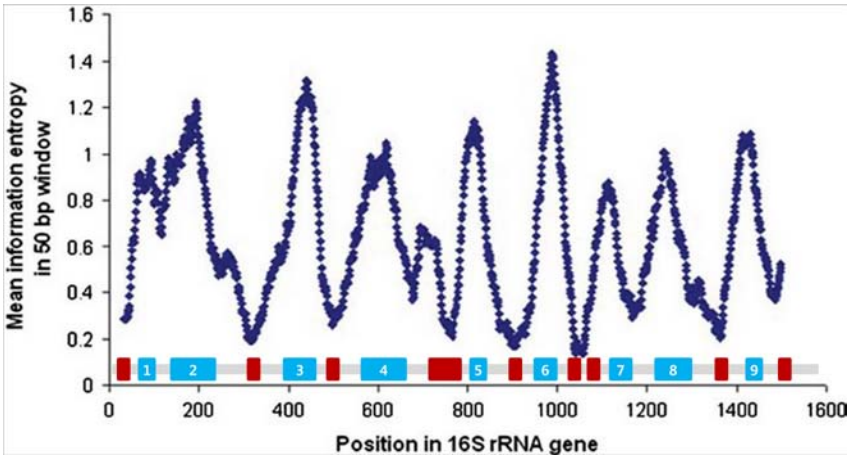
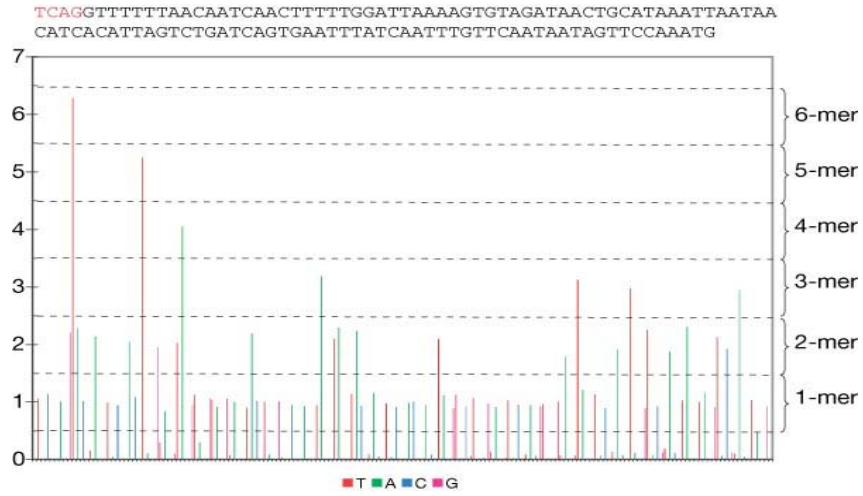


Figure 2. Nine hypervariable regions in 16S rRNA gene.

454 Pyrosequencing errors

Unlike genome sequencing projects in which sequencing errors can be corrected by assembly and sequencing depth (Goldberg *et al.*, 2006; Moore *et al.*, 2006), each read in a pyrotag analysis is interpreted as a unique identifier of a community member and therefore errors will potentially inflate diversity estimates. Kunin and colleges (2006), reported that the number of rare biosphere revealed by the pyrosequencing is overestimated due to the intrinsic error rate of the pyrosequencing. 454 platform presents a flowgram containing the processed data (Fig. 3) which is a series of intensity values for each flow. During each flow, the incorporation of zero, one, or more instances of a single base, and the repeat in a predetermined order is called a flowcycle. The signal intensity (the height of peaks) for a flow is rounded to an integer to give the number of monomers of the corresponding base that were incorporated. For example, the flowgram (T:0.9, A:0,4, C:1.7, T:0.1, T:0.2, A:2.1, C:0.8, G:0.3) would correspond after rounding to a sequence as TCCAAC. Because 454 system reads the flowgram instead of individual bases directly, the major source of error is that the light intensities do not faithfully reflect the homo-polymer length. For example, interpreting 1.7 as 2 when there was only one base (overcall) or 0.4 as 0 even when there was a base at that position (undercall). The average accuracy rate of the pyrosequencing ranges from 99.5% (Huse *et al.*, 2007) to 98.93% (Gilles *et al.*, 2011). Similar to quick and dirty noise handling approach in genome project, OTU generation method is

commonly used to minimize the effect of the errors. This method comprises of two steps, multiple alignment followed by a complete linkage clustering.



(Margulies, 2005)

Figure 3. Flowgram of 454 platform.

This may reduce the effect of the errors but stringent quality based trimming and clustering threshold no greater than 97% should be used to avoid the overestimation of the rare biosphere (Kunin *et al.*, 2010). Thus, bacterial diversity studies with 454 pyrosequencing should begin with filtering out the 'noise' reads. Several denoising applications including PyroNoise (Quince *et al.*, 2011), Denoiser (Reeder *et al.*, 2010), Acacia (Bragg *et al.*, 2012), DADA (Rosen *et al.*, 2012), Pyrocleaner (Mariette *et al.*, 2011), DRISEE (Keegan *et al.*, 2012) have been released so far. Among these denoising programs, PyroNose, Denoiser and Acacia are widely used and compared in this study to evaluate their specificity and sensitivity.

PyroNoise is a part of Amplicon Noise pipeline which simultaneously accounts for both PCR and pyrosequencing error. PyroNoise implements a flowgram clustering method to accounts for pyrosequencing error. PyroNoise algorithms are divided into two steps. In the first step, the algorithm removes read that do not pass the strict conditions. Any sequence that has a signal intensity less than 0.5 are truncated. In case of 454 Titanium, those reads are removed which have their noisy flow occurring before 360. Approximately, 15% of the reads are removed at the first step. In the second step, the algorithm implements the distance generated using the flowgram for each signal, that reflects the probability a sequence is generated from true sequence given pyrosequencing error. Then, a true sequence is inferred using maximum likelihood on the basis of expectation-maximization (EM) algorithm. AmpliconNoise takes post process to remove PCR noise including chimeras. PyroNoise is computationally intensive algorithm which is impractical for analyzing larger data sets. Denoiser (Reeder *et al.*, 2010) is a faster algorithm that uses frequency based heuristics. Basic idea of the Denoiser is that the empirical rank-abundance curves of actual microbial communities tend to be dominated by a relatively small number of abundant taxa. To avoid inefficient all-on-all comparisons for clustering, a subset of reads representing the clusters is subjected to a comparison of the clusters. Algorithm of the Denoiser. 1) a read which is the prefix of other read is removed, 2) initial sequence distribution is computed, 3) sorting the prefix clusters in descending order of abundance and used the distribution to cluster similar reads comparing additional unclustered read to the most abundant clusters. Although

Denoiser has the relatively short Denoiser's running time is less compared to the PyroNoise, it is still a time consuming and computation intensive step which is hardly feasible on a personal computer. Acacia (Bragg *et al.*, 2012) uses frequency based heuristics, thus is one of the fastest denoising program. Unlike previous two denoising programs which do not modify the raw reads and select an error-free read to represent a cluster, Acacia creates a consensus sequence representing clusters. To reduce the number and complexity of alignment, each read in a cluster is aligned to the dynamically updated cluster consensus sequence. Acacia reduces the number and complexity of alignment by avoiding all-against-all pairwise alignment within a cluster. Rather, Acacia aligns each read in a cluster to the dynamically updated cluster consensus sequence. The alignment algorithm is modified to only consider the informative region (homopolymeric region) to reduce the running time.

Illumina paired end read merging

The point of paired end reads is to take advantage of longer reads without actually being able to sequence read that long. The paired end reads can be used for other purposes like contig assembly or scaffolding in genome assembly to construct scaffolds using the paired end information. To get the benefit of the longer read of paired ends, two counterpart reads pair should have an overlapping region in order to successfully merge them. There are several paired end merging programs released online. COPE (Connecting Overlapping Paired End Reads) (Liu *et al.*, 2012a), FLASH

(Fast Length Adjustment of Short Reads to improve genome assembly) (Magoč *et al.*, 2011), PANDASEQ (Magoč *et al.*, 2011), and PEAR (Zhang *et al.*, 2013). In the merging process, unmerged reads increase due to the sequencing errors, which in turn increase as the sequencing read length gets longer ending up in decreasing the number of informative merged reads. PANDASEQ uses the probabilistic model approach to get as many merged reads as possible. PANDASEQ builds a probabilistic model using the length, quality, and the nucleotide frequency in the overall Illumina read assuming that the entire paired end reads can be merged. PANDASEQ scores the alignments of overlap region using the probabilistic model to make a decision for the true nucleotide between the mismatched base pair. PANDASEQ works well for the short fragment library, however it exhibits a higher false positive ratio because it tries to merge all the unmerged reads whose fragment size is over the sum of the paired end read length. FLASH merges the paired reads that maximizes the overlap length-to-matches ratio. FLASH requires the mean DNA size and standard deviation of the fragment size as input parameters indicating that it can only merge paired end reads into fragments of nearly identical size. It is known that FLASH performs poorly when the overlaps between reads are short (Zhang *et al.*, 2013). COPE is designed to work for the deep genome sequencing datasets, thus it deploys the Kmer approach to filter out the infrequent Kmer considered as sequencing errors. COPE's approach is similar in that it finds the best overlap besides taking the quality scores into the consideration. COPE consumes lots of computing memory and execution time is relatively longer. PEAR merges reads by maximizing the assembly score of the read overlap

via a scoring matrix that penalizes mismatches with a negative value and rewards matches with a positive value. PEAR scores all possible overlaps for each pair of corresponding paired-end reads to determine the overlap with the highest assembly score so that it can conduct a statistical test to assess the statistical significance of the merged reads to ignore the merged reads which could not pass this test. It also ignores the merged reads which are shorter than user defined length threshold. In this study, paired end merging program using pairwise sequence alignment was developed for the legitimate assessment of the correlation between substitution errors and corresponding sequencing quality score.

Chimeric read detection

Chimera is the error generated during PCR amplification when an incomplete extension occurred in one round of PCR and then the resulting sequence fragment acts as a primer for different sequence in the next round (Fig. 4). Individual samples contain chimeras ranging from few up to 45% (Huber *et al.*, 2004). Bimera, a chimera formed from two parent sequences, is the most common type of chimera accounting for 89% of all chimeras and trimera is for 11% and quadramera is for 0.3% respectively (Quince *et al.*, 2011). Because chimera is one of the caveats for overestimating microbial diversity, detection and removal of chimera is of much importance. Among many chimera detection programs (Ashelford *et al.*, 2006; DeSantis *et al.*, 2006; Edgar, 2011; Haas *et al.*, 2011; Huber *et al.*, 2004; Pruesse *et al.*, 2007; Quince *et al.*, 2011; Wright *et al.*, 2012) released

so far, Chimera Slayer, UCHIME, Decipher and Perseus were evaluated because they are applicable to NGS short reads. In addition, a couple of studies reported that a significant number of chimeric 16S rRNA sequences of diverse origin were identified (Ashelford *et al.*, 2005; Hugenholtz *et al.*, 2003) in public databases, so the chimera detection capability of each public database should also be evaluated.

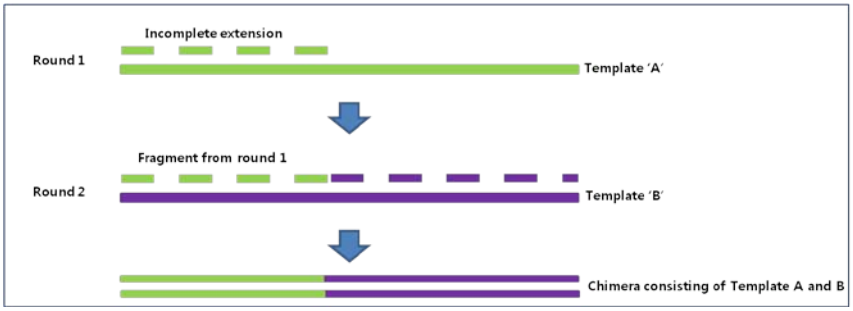


Figure 4. Illustration showing formation of a chimera during PCR reaction.

2.2 Analysis System for 454 Pyrosequencing

454 pyrosequencing platform is the first NGS platform introduced to the market and has the longest read length than any other NGS platform. This advantage makes the 454 pyrosequencing platform to be considered as the appropriate platform for the metagenome community studies. Practically, 454 pyrosequencing has dominated the field and numerous data processing programs were developed for it. Here, an evaluation and comparison of the known pyrosequencing error handling programs was carried out and a novel homopolymeric error handling algorithm was developed. In addition, chimera detection algorithms and sensitivity of reference databases was also evaluated to construct the bacterial community analysis pipeline.

Pyrosequencing

16S rRNA gene fragments corresponding to the V1~V3 regions were amplified from the genomic DNA of mock community samples by using a previously described method (Hur *et al.*, 2011). PCR amplifications were performed in a final volume of 50 μL containing 10X *Taq* buffer, dNTP mixture (Takara, Japan), 10 μM of each barcorded fusion primer (http://oklbb.ezbiocloud.net/content/1001), and 2 U of *Taq* polymerase (ExTaq, Takara) by a C1000 Touch thermal cycler (Bio-Rad, Hercules, CA, USA). After initial denaturation at 94°C for 5 min, the product was

amplified by 30 cycles of denaturation (30 s, 94°C), primer annealing (30 s, 55°C), and extension (30 s, 72°C), with a final extension step of 7 min at 72°C. The PCR product was confirmed by 2% agarose gel electrophoresis and visualized under a Gel Doc system (Bio-Rad). Amplified products were purified with a QIAquick PCR purification kit (Qiagen, Valencia, CA, USA) and quantified using a PicoGreen dsDNA Assay kit (Invitrogen, Carlsbad, CA, USA). Equimolar concentrations of each amplicon from different samples were pooled and purified using an AMPure bead kit (Agencourt Bioscience, Beverly, MA, USA) and then amplified on sequencing beads by emulsion PCR. Recovered beads from emulsion PCR were deposited on a 454 Picotiter Plate and sequenced with a Roche/454 GS Junior system following manufacturer's instructions.

2.2.1 Methods

Denoising Algorithms and Chimera Detection Database

A novel clustering-based error correction algorithm named CDenoiser is introduced here. This new algorithm intends to retain the advantage of NGS i.e., the capability of detecting rare biosphere. As illustrated in Figure 5, the error handling algorithm is as following. 1) condensing homopolymers to a mono/single nucleotide. 2) clustering exactly the same reads or substring of other longer strings. 3) creating consensus sequence of the clusters. 4) sorting the clusters in the order of the descending cluster size while allowing 2 mismatches. 5) trimming ends of the consensus

sequence showing a low coverage (depth < 2).

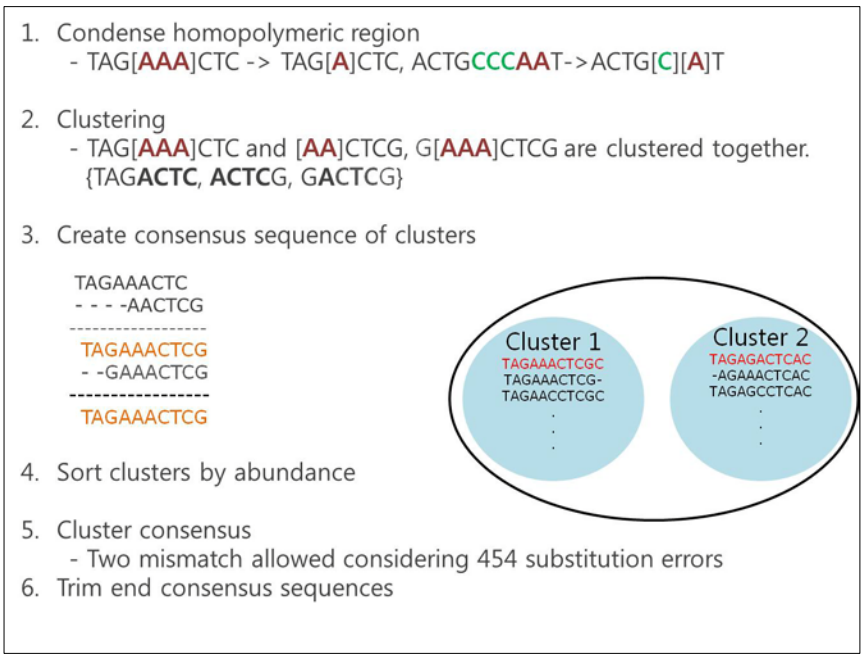


Figure 5. Scheme of clustering-based homopolymeric error handling algorithm.

Comparison of Denoiser

Two mock communities were fabricated with known strain sequences. One was composed of 19 sequences and the other consisted of 47 sequences. Tested programs were PyroNoise, Denoiser, Acacia and CDenoiser. After denoising, filtered noise reads were processed to identify the non-noise sequencing reads. Roche instrument generates SFF file as a result of a sequencing run. SFF files were converted into flowgram file for

PyroNoise and Denoiser, and then split into fasta and quality files for Acacia and CDenoiser. In detail, Perl script contained in QIIME package was used to run PyroNoise and Denoiser. SFFINFO script distributed by Roche was used to convert the SFF file into fasta and quality files for Acacia and CDenoiser. Eztaxon-e (Kim *et al.*, 2012) database was used to assign the taxonomic position to both the quality reads and noise reads.

Comparison of Chimera Detection Programs and Reference Database

Chimera Slayer, UCHIME, DECIPHER and Perseus were tested with artificial bacterial community sequences. 10 artificial data sets containing as much as 10~20% known chimeric reads were fabricated. The proportion of each type of chimera in the artificial communities mimicked the chimera composition reported by Quince et al (2011). The validation procedure was carried out as follows: 1) fabrication of a mock community with known sequences, 2) retrieving V1~V3 regions of each sequence, 3) creating 10 artificial chimeric data sets using the retrieved sequences mixed with known non-chimeric reads. RDP gold database was used for Chimera Slayer and UCHIME, which are database dependent chimera detection programs. To test the database dependency of these 2 programs combination of UCHIME and widely used public rRNA database including Ribosomal Database Project (RDP), Greengenes (GG) and Eztaxon-e were tested using the same artificial mock community data sets.

2.2.2 Results

Comparison of accuracy of noise filtering programs

The number of reads obtained after demultiplexing from the 454 pyrosequencing of mock19 and mock47 samples were 10,347 and 21,430 respectively. One of the advantages of NGS sequencing is the high throughput, which enables us to explore the low abundant species which were previously unknown for the environmental samples. Thus it is required for the denoising programs not to remove the non-noise reads. The number of reads after filtering the noise reads from the two mock communities sequencing reads by each programs was compared (Fig. 6). The sequence cluster based denoising programs saved more reads than the flowgram clustering methods in both the mock data sets indicating that the flowgram clustering methods might be significantly removing the non-noise reads in addition to noise reads. Among the sequence clustering based programs, Acacia removed substantially more reads than the CDenoiser in the mock19 data set while the filtered reads were almost the same in the mock47 data set for both programs.

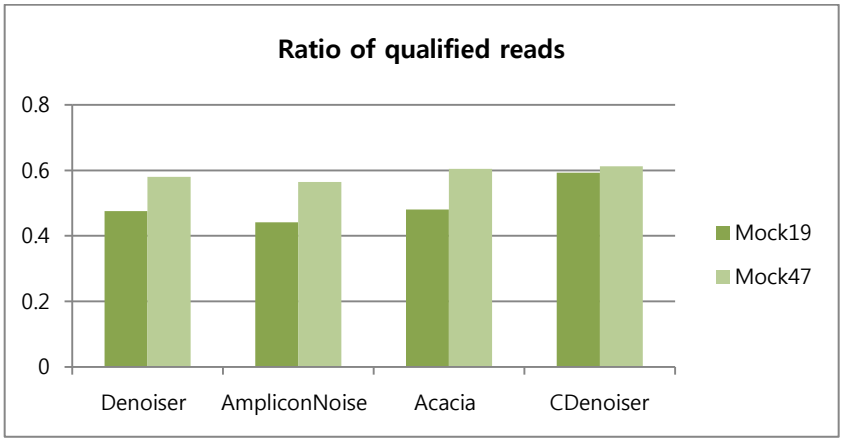


Figure 6. Bar graph showing ratio of qualified reads obtained after two mock community data sets were analyzed by different denoising programs.

All the remaining qualified reads after each noise filtering programs were assigned taxonomic positions using Eztaxon-e database. Taxonomic assignment method followed the way described in the previous study (Jeon *et al.*, 2013). From the comparison of the taxonomic composition of each data set to the original mock community (Table 2), the detected number of false positive genus was largest by CDenoiser than any other programs in mock19 data set while the ratio of false positive genus by CDenoiser accounted for 0.9% which is the smallest among all the programs. In mock47 data sets, both the detected false positive genus and it's ratio increased in all the cases indicating it is possible that the noise filtering programs may not filter true positive noise as the complexity of the bacterial community increases.

Table 2. Biased taxonomic composition at genus level.

	Mock19		Mock47	
	Count[(a)]	**Ratio**	**Count[(a)]**	**Ratio**
Denoiser	40	2.1%	59	6.8%
AmpliconNoise	37	2.2%	49	4.0%
Acacia	41	2.0%	57	4.1%
CDenoiser	45	0.9%	45	2.1%

[a] The number of detected genus (phylogenetically different) which are not in mock community.

The reads removed from each program were analyzed to confirm whether the reads are true noise (Table 3). After processing the reads using the same criteria, shorter that 300bp and average quality lower than 25, it turned out that a larger number of non noisy reads were removed by the flowgram clustering for the mock47 data set. Out of all the programs tested here, Acacia removed the least number of noisy reads. As the CDenoiser actually corrects the homopolymeric reads instead of deleting them, all the reads could be processed for further downstream analysis allowing us to minimize the loss of information and hence enabling us to explore the low abundance taxa.

Table 3. Noise reads are processed applying the quality and length cutoff parameters.

	AmpliconNoise		Denoiser		Acacia	
	Mock19	**Mock47**	**Mock19**	**Mock47**	**Mock19**	**Mock47**
Noise reads	2,161	9,030	1,525	8,405	3,753	2,234
Low quality[(a)]	1,973	2,387	1,434	2,079	3,663	2,211
Chimera	44	2,042	28	1,934	25	6
Qualified reads[(b)]	144	4,601	63	4,392	65	17
Non noise reads[(c)]	72	3,277	35	3,135	33	13

[a] The number of short (300bps) and low quality (Q<30) reads in noise reads.

[b] The number of normal reads in 'Noise' reads.

[c] The number of non noise reads which should not be discarded.

Time and memory usage of tested programs were measured and results are shown in Figure 7. Denoiser and PyroNoise used 16 CPUs however CDenoiser and Acacia used single CPU because they don't provide parallel computing algorithms.

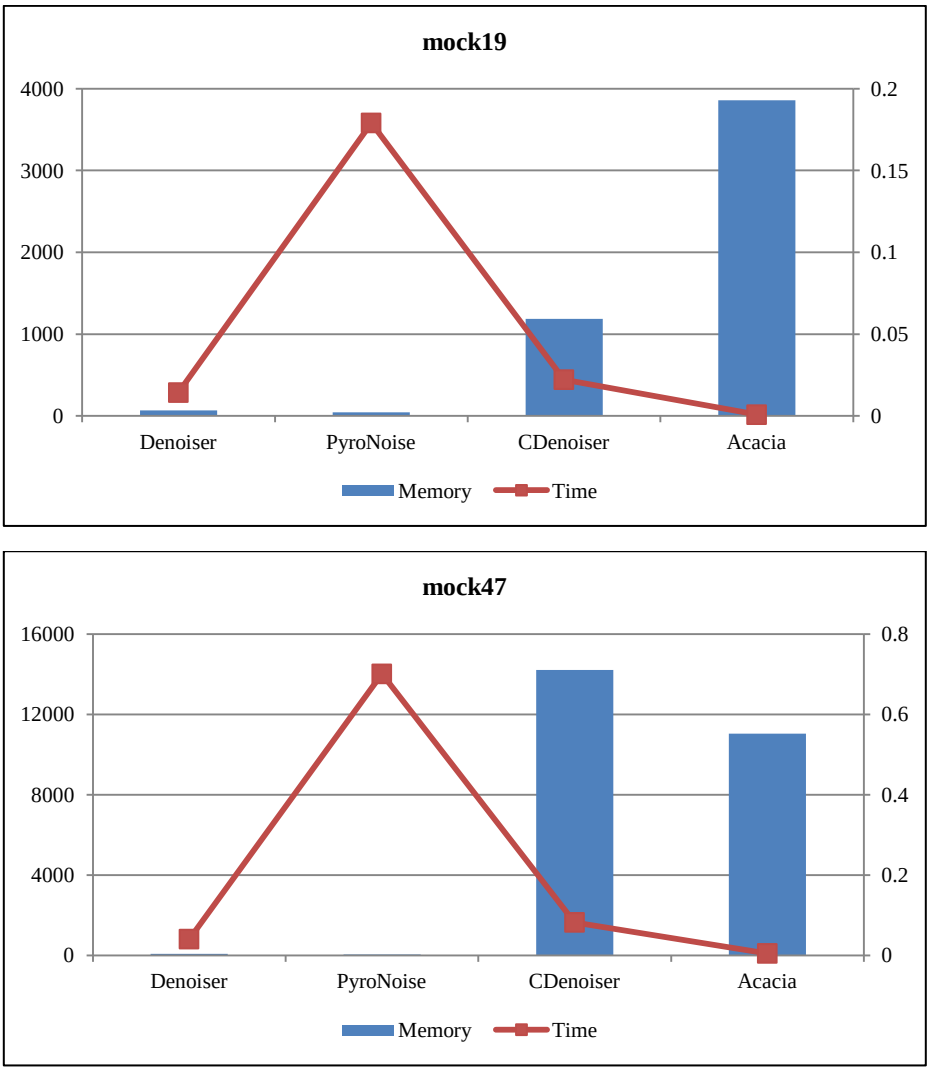


Figure 7. Time and memory usage of denoise programs.

Sensitivity of chimera detection programs and reference DB

ChimeraSlayer (CS) and UCHIME are the database dependent programs and DECIPHER and Perseus are ab initio chimera detection programs. Perseus is a part of Amplicon Noise pipeline which is used for chimera detection and can process sequences both ways i.e., by deploying reference database and also in ab initio mode. When checked in ab initio mode , it reported more than 99% of reads as chimeric sequence in all of the 10 artificial sets, so Perseus was not taken into the comparison. Detection sensitivity and the ratio of detecting true chimera reads was compared. Figure 8 indicates that DECIPHER showed the best chimeric reads detection sensitivity while ChimeraSlayer did the worst performance. Since, all DB dependent detection programs are likely to be the dependent on DB, therefore it was worth comparing the detection sensitivity for the DB-dependent programs given different reference databases. Also, UCHIME showed/s higher sensitivity than ChimeraSlayer, UCHIME ran repeatedly 10 times each for Greengenes (GG) and Eztaxon-e. The size of the RDP, Greengenes and Eztaxon-e DBs are 5,181 , 4,938 and 23,046 sequences, respectively.

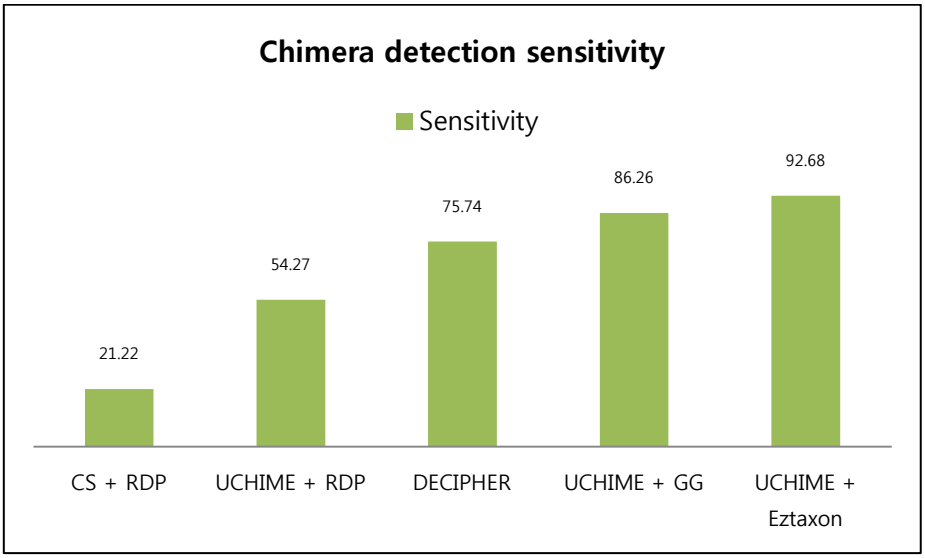


Figure 8. Sensitivity of the chimera detection programs and different db.

In the case of the Greengenes database, UCHIME reported an average of 10% false positive chimeric sequences, whereas other cases reported no false positives. This may because of the lack of non-chimeric sequence in the database as the size of the greengenes is the smallest among all three databases. Combination of UCHIME and EzTaxon-e showed the best sensitivity and accurate chimera detection result (Fig. 8). Even with this combination, we were unable to detect around 7.5% (average) chimeric reads including all data sets. Thus, those true negative reads were identified using Eztaxon-e database. None of these reads were identified as showing a similarity $\geq$ 97%. This suggests that only the reads identified showing $<$ 97% could be considered as the potential chimeric reads.

Overall pipeline for microbial community analysis of 454 pyrosequencing results

The overall 454 pyrosequencing analysis pipeline is illustrated in Figure 9. Because multiple samples are pooled using DNA barcode tagging and then sequenced at once, demultiplexing should be done before analysis begins, and thereafter, quality filtering is carried out with each of these demultiplexed single file. The thresholds for the quality filtering are both average quality score and read length. Preprocessing raw reads is followed by the assembly process as a way of denoising. Thereafter, taxonomic assignment using EzTaxon is performed. In this process, top 5 blast hits are identified and pairwise alignment of query and each of 5 subject hits are carried out to identify the most similar one among those 5 hits.

Chimera detection is performed after taxonomic assignment. Only the reads whose identification similarity is < 97% are subjected to the chimera detection process. For the statistical analysis, CD-Hit (Li *et al.*, 2006) was used for OTU clustering whose output is passed on to the mothur (Schloss, 2009b) platform for the diversity indices calculation.

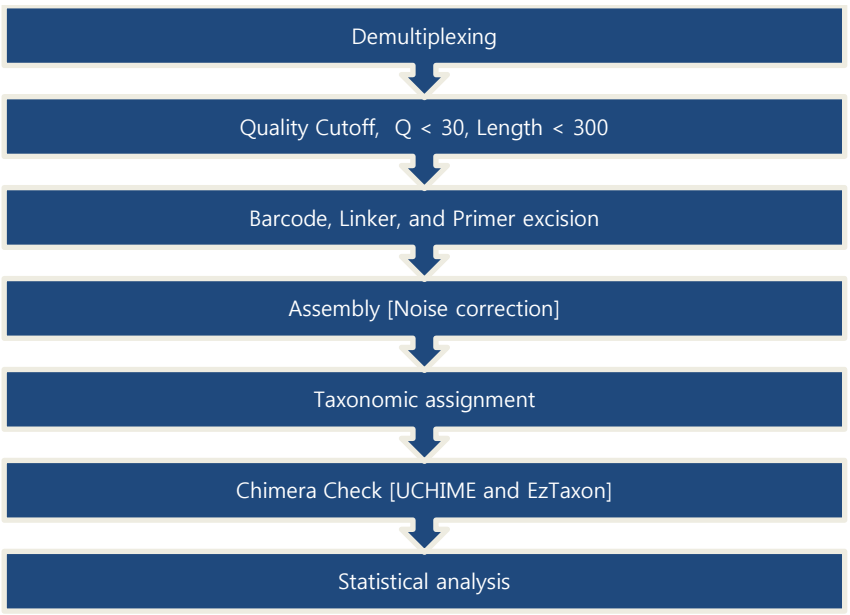


Figure 9. Overall microbial community analysis pipeline for analyzing pyrosequencing data.

Application of the Pipeline

As a reference study of the 454 amplicon analysis pipeline, a household microbial community analysis was performed. Indoor microbes have been studied in the context of human health by using culture-dependent and independent techniques. Most of these studies focused on the bacterial contamination of surfaces in kitchens and restrooms, which are easily colonized by microbes (Flores *et al.*, 2011; Flores *et al.*, 2013; Kembel *et al.*, 2012; Ojima *et al.*, 2002b; Rintala *et al.*, 2008). Some pathogenic bacteria can survive on the surfaces in these environments for some time, and contamination of food by these pathogenic bacteria can cause illness.

Microbial contaminants of refrigerators have also been previously studied (Barker *et al.*, 2000; Carpentier *et al.*, 2012; Evans *et al.*, 2004; Jackson *et al.*, 2007). Moisture and nutrients (food particles) in refrigerators provide favorable growth conditions for bacterial contamination from unwashed raw foods, leaking packages, and hands. In particular, higher bacterial counts and temperatures in vegetable compartments could cause critical problems (Carpentier *et al.*, 2012). Recently, a German outbreak caused by Shiga-toxin producing *Escherichia coli* O104:H4 illustrated that unwashed vegetables could be a risk element (Buchholz *et al.*, 2011). Therefore, the study of bacterial contamination in the vegetable compartments of refrigerators is important for public health.

Most of the previously reported culture-dependent studies of kitchen and refrigerator microbes focused on pathogen detection (Evans *et al.*, 2004; Jackson *et al.*, 2007; Ojima *et al.*, 2002a; Ojima *et al.*, 2002b; Sinclair *et al.*, 2011). The recent advent of next generation sequencing techniques provides unprecedented data on the microbial composition, and the ecology of various environments, including indoor spaces (Flores *et al.*, 2011; Flores *et al.*, 2013; Hewitt *et al.*, 2012; Kembel *et al.*, 2012). Analyses of microbes in various environments by high-throughput sequencing can benefit various fields, including source-tracking. Identification of sources of bacterial contamination in indoor environments is important for managing food safety. Human skin is a primary source of bacteria in indoor environments, and individuals can transmit bacterial pathogens by touching indoor spaces (Flores *et al.*, 2011; Flores *et al.*, 2013). Comparing various parts of the human microbiome with microbial communities in indoor environments

can identify bacterial species commonly found in both environments and thereby can track the source of contamination or transmission.

In this study, I characterized bacterial communities within vegetable compartments of refrigerators and on toilet seats by using pyrosequencing based on 16S rRNA genes. The comparison of bacterial communities analyzed in this study with already published human microbiome data, provides further insight into shared species and sources of bacteria on the surfaces of refrigerators and toilets. Opportunistic pathogens were shared between the human skin microbiome and microbial populations in refrigerators and toilets.

The bacterial communities in swab samples were analyzed using high-throughput 16S rRNA gene pyrosequencing. Diversity indices calculated by three different methods are presented in Appendix I. In refrigerator and toilet samples, the richness and diversity of the communities obtained from metagenomic DNAs were higher than those obtained from plate washing DNAs. Although the values calculated by the TBC method were higher than those calculated by the CD-HIT and TDC-TBC methods, the diversity trends in each sample were similar among the three methods. Four phyla, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, and *Actinobacteria*, were dominant (over 98% of total reads from each sample) in mean bacterial communities, which was obtained by pooling the culture-independent results from the refrigerator and toilet surfaces of 10 households (Fig. 10a). These major phyla were also identified in previous indoor studies (Aydogdu *et al.*, 2010; Flores *et al.*, 2013; Kembel *et al.*, 2012). Although the compositions of dominant phyla were similar on surfaces of refrigerators

and toilers, the phyla proportions were different. Proteobacteria was the most dominant phylum in refrigerator (63.6% of total reads) and toilet samples (42.2%). The relative abundance of Firmicutes in toilet samples (36.2% of total reads) was higher than refrigerator samples (15.7%). A total of 30 phyla were detected in refrigerator samples, while 16 phyla were obtained from toilet samples. This could be due to differences in survivability that depend on the moisture or temperature of surfaces and the frequency of transmission. The compositions of the top 10 most prevalent genera in each sample showed clear differences between bacterial communities of refrigerators and toilets (Fig. 10b). *Pseudomonas* and *Pantoea* from taxa Gammaproteobacteria were identified as the dominant genera in refrigerator samples. Although the genus *Pseudomonas* was also dominant in toilet samples, the proportion of *Pantoea* was relative low and *Bacillus*, *Staphylococcus*, and *Streptococcus* from phylum Firmicutes were dominant genera. The bacterial communities present in the individual samples obtained from each house are presented in Figure 11. The number of toilet samples were smaller than that of refrigerator samples as sufficient DNA couldn't be isolated from many of the collected swab samples of toilet seat surfaces. This is probably because toilet surfaces are cleaned more frequently than the vegetable compartments of refrigerators in general households. The compositions of bacterial communities in refrigerators of most identical houses obtained by plate washing method were similar to the compositions obtained by culture-independent methods, except for #6 house. However, only 5 of 30 phyla were detected in the plate washing results, and the proportions of each member in bacterial communities were

different between two different methods. The differences between culture-based and culture-independent results were significant in toilet samples obtained from identical houses (#1 and 3): *Firmicutes* and *Actinobacteria* were more abundant in culture-based plate washing results. This difference could be due to the selectivity of Plate Count Agar (PCA) or Nutrient Agar (NA) media. The genus *Staphylococcus* was the most dominant bacteria obtained by culture-based plate washing method in toilet samples (average 45.9% of total reads). The phylum and genus compositions in the refrigerator and toilet samples were unique because of the people and their behaviors (e.g., frequency of cleaning, cleaning products used, kinds of refrigerators and toilets, and usage patterns) varied in each household.

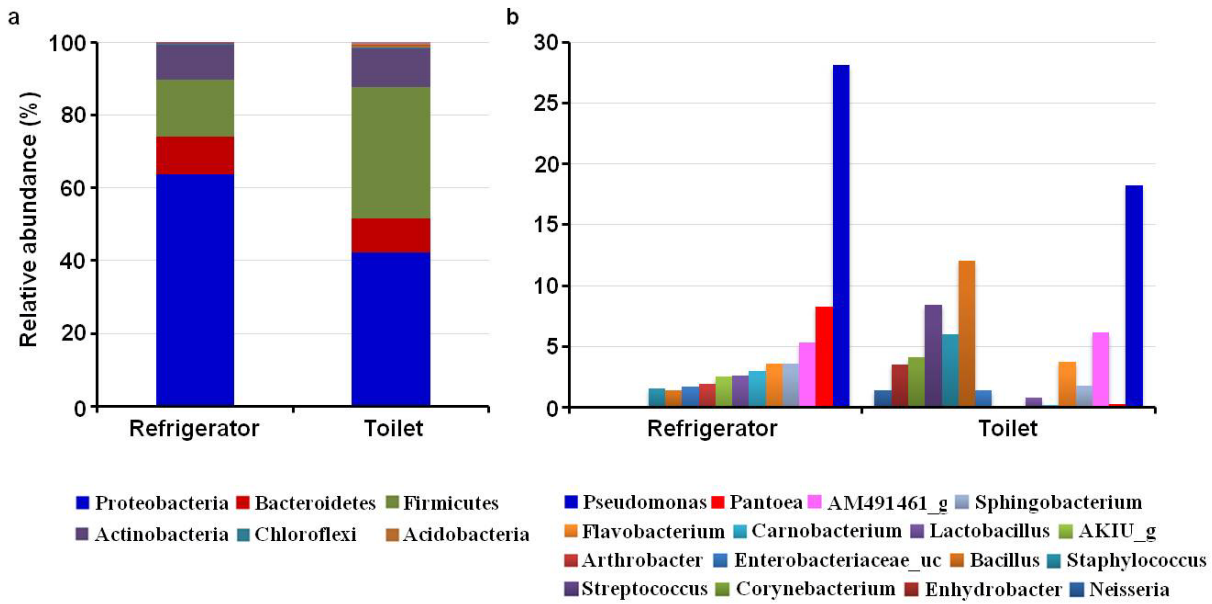


Figure 10. The average compositions of bacterial communities obtained from the vegetable compartments of refrigerators and from toilets by using culture-independent method

Several studies have reported that most indoor bacteria could be of human origin, particularly from human skin such as hands (Flores *et al.*, 2011; Flores *et al.*, 2013; Rintala *et al.*, 2008). To identify bacterial species present simultaneously on human skin and in the two indoor environments, bacterial communities obtained in this study were compared with microbiota from human skin and fecal samples. Human microbiome data was downloaded from the Human Microbiome Project (Methe *et al.*, 2012) webpage. Skin and gut microbiome data were selected because of the possibility of direct contact with the surfaces of refrigerators and toilets. On an average, 15.6% of the bacterial species obtained from human skin and 4.9% of the species obtained from human gut samples were shared with refrigerator's community (Fig. 12). The proportion of bacterial species shared between toilets and human skin samples (51.6%) was higher than the proportion shared by toilets and the human gut microbiome (15.4%). These results indicate that the human skin microbiome could be a significant source of bacterial transmission by touch or exposure . This is similar to the results of public restrooms, where human skin was identified as the principal source of bacteria (Flores *et al.*, 2011). The proportion of bacteria shared by human skin and the surface of the toilet was higher than that shared by human skin and the refrigerator because of the higher frequency of human contact with toilets. The species shared between human skin and refrigerators were similar to those shared between human skin and toilet surfaces. These result supports the previous findings that most indoor bacteria possibly originated from human skin and indicates that particular bacteria can attach to and survive for long periods on indoor

surfaces (Flores *et al.*, 2011; Flores *et al.*, 2013). Out of the shared species found in the gut microbiome, *Bacteroides vulgates* was the most abundant on the surfaces of refrigerators and toilets, but the composition of shared species was different on the two surfaces (Fig. 11). This could be due to direct or indirect exposure of fecal bacteria to the surfaces of refrigerators or toilets. *Propionibacterium acnes* was the most abundant species shared between human skin and the surfaces of refrigerators and toilets. This species is a member of the normal flora of the skin, oral cavity, large intestine, and other human body sites. It mainly plays a role in acne, and it can cause postoperative and device-related infections as an opportunistic pathogen (Marinelli *et al.*, 2012; Perry *et al.*, 2011). *Staphylococcus epidermidis* and *Staphylococcus hominis* (other isolates) are commensal bacteria on human skin; they inhibit virulent bacteria such as *Staphylococcus aureus*. However, they are also opportunistic pathogens that cause nosocomial infections by dwelling inside medical devices (Fey *et al.*, 2010; Rogers *et al.*, 2009). *Bacteroides vulgatus* was the most abundant species shared between the human gut microbiome and the surfaces of refrigerators or toilets. Although this bacterium is one of the predominant bacteria in the gut of a healthy person, it was isolated from a patient with Crohn's disease and identified as an antibiotic-resistant pathogen (Kumar *et al.*, 2011b; Ruselervanembden *et al.*, 1989). The distribution patterns of these opportunistic pathogens pose considerable issues for explaining potential contamination of foods or residential environments. Bacterial communities on the surfaces of refrigerator vegetable compartments could be transferred to the vegetables and cause food borne illness, such as the

recent German outbreak of *E. coli* in 2011 (Buchholz *et al.*, 2011).

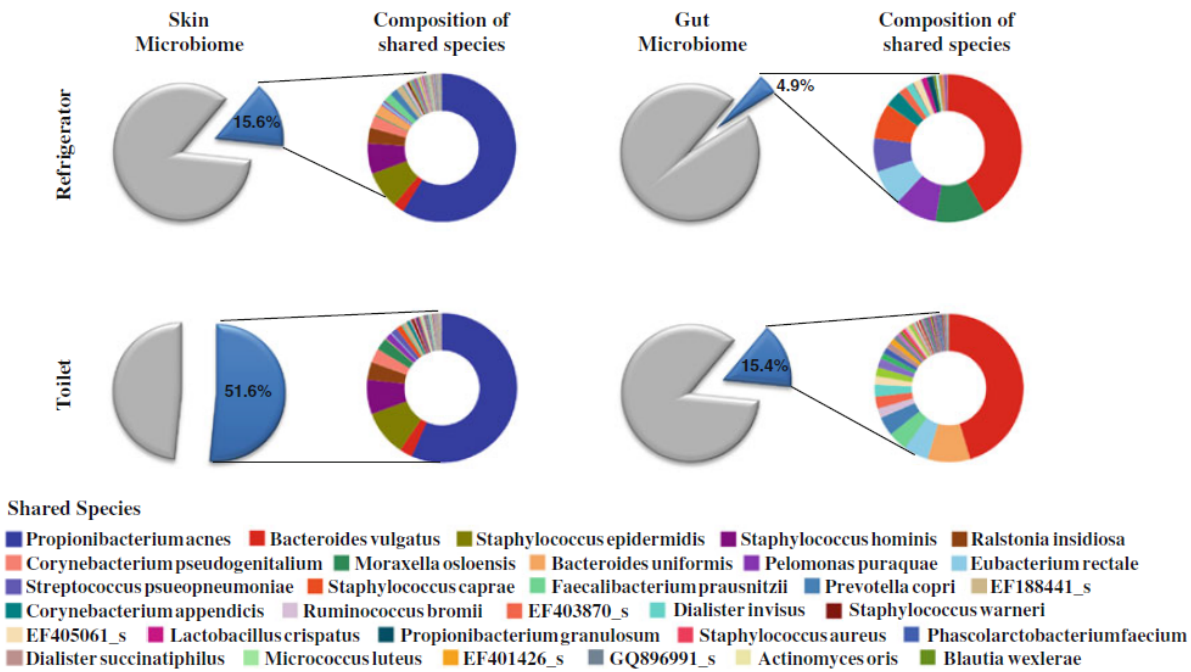


Figure 11. A pie chart diagram showing the proportion of species shared between human skin and gut microbiomes with bacterial community from refrigerator and toilet.

PCoA (Principal Coordinates Analysis) plots based on four different statistical calculations of distance were compared to analyze the relationships among the samples (Fig. 12). Although there were variations in the bacterial communities obtained from refrigerator, toilet, and skin, these communities were more related to each other than to communities from fecal samples in PCoA plots based on UniFrac distance (Fig. 12a). Bacterial communities obtained from refrigerators and toilets were similar in PCoA plots using the Bray-Curtis and Sorenson abundance coefficients (Fig. 12b and d). This might be simply because of samples obtained from refrigerators and toilets in the same house were exposed to the same people. Bacteria from skin samples were more similar to bacteria from refrigerator or toilet samples than fecal samples in UniFrac, Bray-Curtis, and Sorenson analyses. These three statistical analyses of community similarity were consistent with species shared between the samples (Fig. 11). However, bacterial communities from fecal samples were similar to bacterial communities from toilet and refrigerator samples in the Jaccard abundance analysis (Fig. 12c). The bacterial communities obtained from toilets were more similar to those of fecal samples than to the bacterial communities of other samples. These analyses again revealed that microbes on and within the human body could be a source of bacteria in indoor environments. The significances of differences among bacterial communities were analyzed by Libshuff comparison ($p < 0.05$).

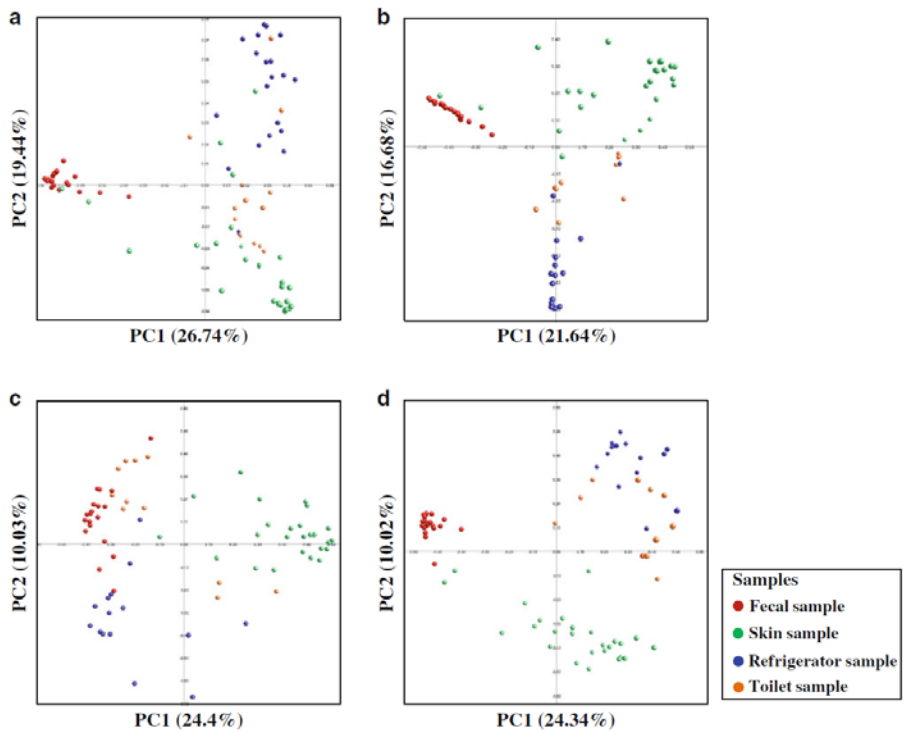


Figure 12. Similarities between bacterial communities that originated from refrigerator, toilet, human skin, and gut samples as visualized by a PCoA plot.

The initiation of food-borne illness has been reported to occur more frequently in private homes than in commercial operations (Scott, 1996; Scuderi *et al.*, 1996). Refrigerators in kitchens could be colonized by bacteria, and these bacteria might contaminate other stored foods or attach to and survive on the internal surfaces of the refrigerator, thereby posing risks of indirect, long term contamination during subsequent food preparation activities (Michaels *et al.*, 2001; Ojima *et al.*, 2002a; Ojima *et al.*, 2002b; Sinclair *et al.*, 2011). In this study, most bacteria detected were probably not pathogens or opportunistic pathogens, and genera belonging to common pathogens were detected in only a very small fraction of communities. However, their presence could influence other microorganisms, since they survive on and are transmitted to the surfaces of indoor environments. This potential risk can be prevented by wrapping stored foods and regularly cleaning indoor environments, including refrigerators. The expansion of studies on indoor microbial communities by using high-throughput molecular methods will advance our understanding of microorganisms in indoor environments and improve preventive measures for public health.

2.3 Analysis System for Illumina MiSeq

Illumina platform has different background sequencing chemistry from that of 454 pyrosequencing platform. The difference in sequencing chemistry gives rise to different specification of output reads. The number of sequencing reads per single run is larger but the read length is shorter than the 454 pyrosequencing. Particularly, in case of the MiSeq platform, the number of output reads is about 10 times larger than 454 Titanium. The read length of Illumina GA2X was as short as 37~125 bps which has been the main reason why the Illumina reads has been applied to rather the resequencing or genome projects than the population study. Recent development of sequencing instrument makes Illumina MiSeq platform generate paired end data whose single end read length is about 250bps and merging the paired end reads approaches the length of the 454 read. However, Illumina reads are frequented by low quality reads in 3' region (Fig. 13) which is an obstacle in community studies. There are a couple of published microbial population studies online using merged Illumina paired end reads (Bartram *et al.*, 2011; Kozich *et al.*, 2013). However, no studies validated the effect of the merging process by which the overall accuracy of the read could be improved through the correction of the erroneous base call around low quality region of the paired reads. In this chapter, evaluation for the selection of a suitable region of 16S rRNA gene for community study with reads of MiSeq system (currently 300bp paired

sequence developed) was conducted, and the proper sequencing conditions of PhiX concentration and sequencing library was also determined. In addition to selecting optimized sequencing conditions, analysis method of correcting erroneous sequences in overall sequence regions was also improved. Recommended conditions for sequencing run and improved analysis method will be helpful using MiSeq platform for amplicon based bacterial community analysis.

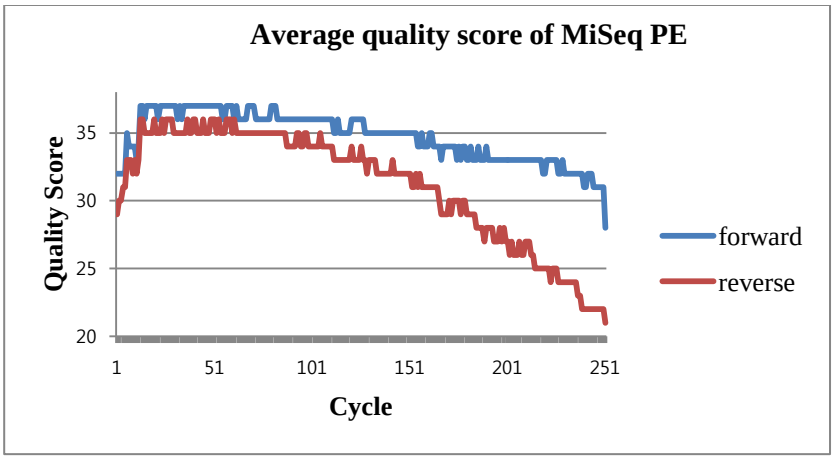


Figure 13. Average quality of Illumina MiSeq PE. Quality score is getting lower as the length of the sequence is getting longer.

2.3.1 Methods

Target Region Selection

To determine the proper primer sites for amplicon sequencing on MiSeq instrument, *in silico* test was performed by combination of two variable regions. Average length, average dissimilarity, and coverage (the ratio of successfully amplified sequence by primer sets in the database) of regions were calculated by means of sequences in EzTaxon-e database.

Amplification was conducted with ExTaq polymerase (Takara, Shiga, Japan) by a C1000 Touch thermal cycler (Bio-Rad, Hercules, CA, USA). The condition of amplification was followed as described earlier (Jeon *et al.*, 2013). Amplified products were purified with a QIAquick PCR purification kit (Qiagen, Valencia, CA, USA) and quantified using a PicoGreen dsDNA Assay kit (Invitrogen, Carlsbad, CA, USA). 1ug of each purified amplicon was used to construct library using by Truseq library kit (Illumina) according to the manufacturer's instruction. The concentrations of libraries were calculated with primers target to Illumina adapters and SYBR Green (KAPA SYBR FAST Universal qPCR kit; KAPA biosystems, Woburn, MA, USA) using the CFX96 Real-Time PCR Detection system (Bio-Rad). Different concentrations (4, 6, and 8 pmol) of amplicon libraries were compared to determine cluster densities. Libraries were mixed with PhiX control libraries (Illumina) and denatured using NaOH following manufacturer's instructions. Different proportions (50%, 10%, and 5%) of PhiX in final sequencing libraries were compared to determine proper

mixed concentration of PhiX control in sequencing run.

After the excision of the primers sequence chosen from the *in silico* evaluation, the length of both reads was about 230bps. According to the estimated average length of the amplicon, the length of overlap region should be around 80bps. In the next section, this overlapped length will be used as a factor for setting up a parameter for evaluation of the trimming length (Fig. 14).

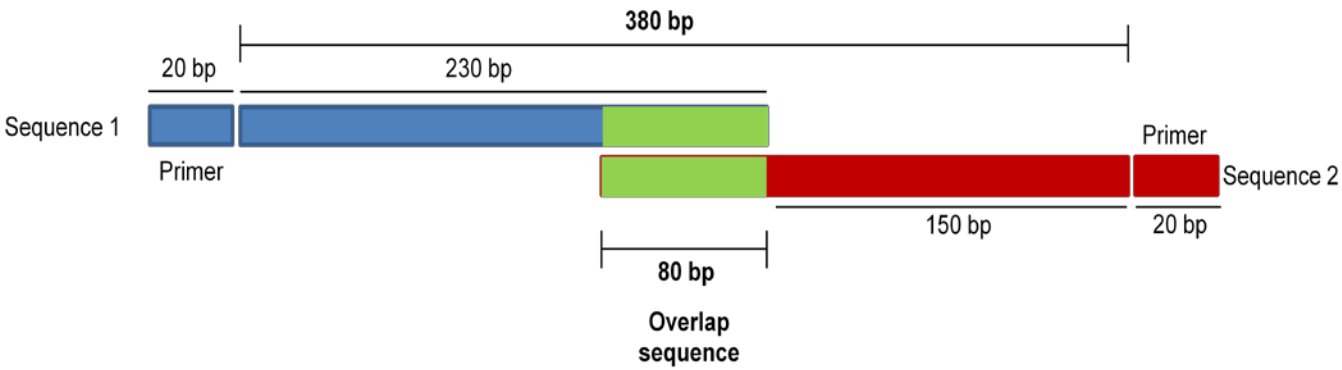


Figure 14. Illumina MiSeq paired sequencing scheme.

Error Correction by Paired End Merging

3’ end of the Illumina reads has higher error frequency, so, the merging process may require trimming out the end of the each read so only the nucleotides having relatively higher quality take part in the merging process. To make it clear whether the truncation of the low quality end improves the merging quality, the merging process was repeated with the trim length parameter varying from 0 to 50 as shown in Figure 15.

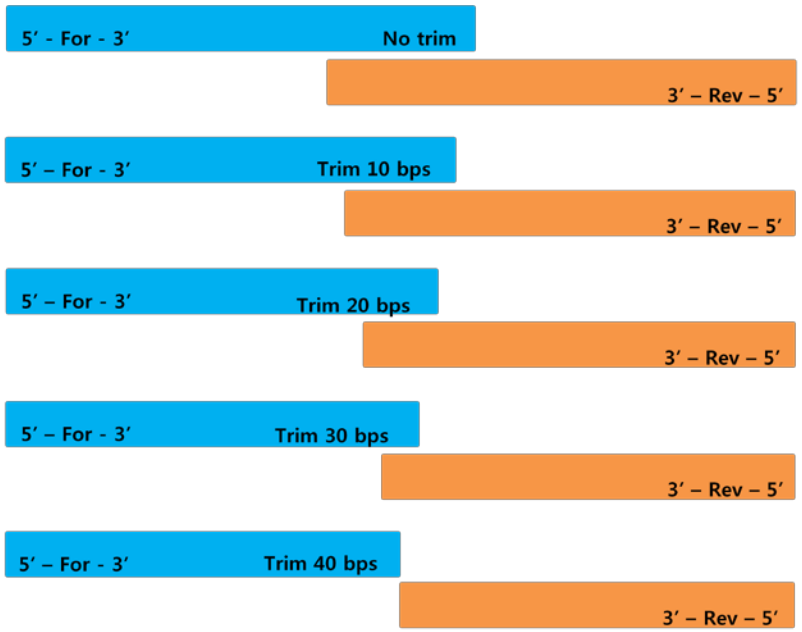


Figure 15. Evaluation scheme for paired end read merging.

Local pairwise sequence alignment was used in merging process and similarity was measured for the aligned overlapped region. Unusual higher

gap open penalty was given to the algorithm as not to allow any gaps to be introduced to the alignment. Since the overlapped region of the two reads is originated from the same fragment, any gap created during the pairwise alignment is likely to be an error. A consensus sequence of overlapped region is constructed according to the following rules. 1. Read pair showing less than 0.5 % similarity of overlap region was eliminated; 2. When a mismatch occurred, nucleotide having higher quality score between the two bases was chosen as the consensus nucleotide; 3. When a gap was introduced within the alignment, the pair was also discarded.

To minimize the effect of the error which occurred randomly and frequently in 5' non overlap region of both read pair, the errors should be corrected or the read containing them should be discarded before the downstream analysis begins. Since the nucleotide sequencing quality has been reported to be related with the quality score, the error rate and corresponding quality score was measured by following steps. 1) BLAST search of merged reads against database which consists of 47 known sequences; 2) Pairwise alignment of the merged read to the most similar sequences obtained by BLAST search; 3) Divide overall region of merged read into 10 sub regions, since the lengths of merged reads vary; 4) Calculate error rates of each sub region per read by,

$$\text{Error rates in each region per read} = \text{Ec}/\text{Mr}$$

where Ec is a sum of mismatched nucleotides to reference sequence within each region and Mr is the total number of merged reads.

Error Correction by Iterative Consensus Clustering

The effect of erroneous sequences on determination of microbial community was analyzed by the scatter plot using R software (ver. 2.15.2). Similarities of merged reads to mock community sequences was used to evaluate the effect of erroneous sequences. Decreased quality score toward 3' end sequence was related to the mismatched sequences in overlap regions of paired sequence. The correlation between the number of mismatched nucleotide in overlap regions and the similarity of merged reads to reference sequences was analyzed in a scatter plot. The distribution of mismatched sequences with mock community sequence was analyzed using Burrows-Wheeler Aligner (BWA) (Li *et al.*, 2009) and displayed by Integerative Genome Viewer (Robinson *et al.*, 2011). To correct erroneous nucleotides in reads, similar sequences (97% sequence similarity cutoff) were clustered by USEARCH program (Edgar, 2010), and a consensus was made by selecting majority nucleotide of mismatched column within cluster. This step was repeated until the number of cluster did not change. Corrected sequences were assigned their taxonomic position using the EzTaxon-e database, and the chimeras were detected by UCHIME program (Edgar *et al.*, 2011).

2.3.2 Results

Hyper variable selection through In silico Estimation

To decide suitable regions of 16S rRNA gene for amplicon sequencing by MiSeq 250bps paired reads, combination of two variable regions were analyzed for their average amplified length, average dissimilarity values of the target region among amplified sequences, and the proportion of detectable sequences by primers in the database (Table 4). The size of amplicon varied with each new combination. The shortest amplicon was obtained using V6/V7 combination (193.7 ± 14.0 bp), whereas the longest one was V3/V4 regions (416.8 ± 11.2 bp). However, the V3/V4 region was reported to generate significant amplification bias in a previous study (Claesson *et al.*, 2010). Therefore, this V3/V4 region was not considered as a target region. Combinations shorter than 250bp (V5/V6 and V6/V7) were also removed from candidate regions for sequencing. The short length of amplicon could be generated from complete overlap between paired reads, and this diminishes the effect of 250bp paired end sequences. Dissimilarity value of each combined region indicates the diversity of sequences in target region. Although dissimilarity values of all combined regions was over 20%, the dissimilarity of regions of V1/V2 and V4/V5 was relatively higher than that of other regions. The dissimilarity values of V7/V8 and V8/V9 were lower than that of other regions, thus these were regions removed from target region candidates. V4/V5 region showed the highest detection ratio (coverage of sequences, 86.05% of sequences in

database detectable), while relatively lower proportions was detected in V1/V2 (43.87%). Low proportion of detected sequences by V1/V2 region is due to the lack of prior sequences (primer sequence) to V1 region in database. The limited information of V1 region sequences was due to the lack of forward primer sequences. Most sequences in public database do not contain the sequence information of V1 forward primer region. Finally, two target regions, V2/V3 (388.8 ± 19.2 bp) and V4/V5 (372.0 ± 7.4 bp), were selected for the evaluation due to the longer amplified sizes than V1/V2 (314.5 ± 29.6 bp) and relatively high coverage of sequences in the database. V4/V5 region has already been reported to show best performance in accuracies, classification and consistency across RDP-Classifier and MEGAN based on BLAST searches (Claesson *et al.*, 2010).

Table 4. Comparison of simulated amplicon by different combinations of two variable regions.

Combined region	Length of Amplicon (bp)	Dissimilarity (%)[a]	Detectable sequences in database (%)[b]
V1_V2	314.5 ± 29.6	27.5 ± 4.6	43.87
V2_V3	388.8 ± 19.2	25.6 ± 4.5	68.47
V3_V4	416.8 ± 11.2	24.0 ± 4.6	75.55
V4_V5	372.0 ± 7.4	26.8 ± 5.4	86.05
V5_V6	249.6 ± 7.4	24.6 ± 4.8	84.64
V6_V7	193.7 ± 14.0	23.6 ± 4.6	57.90
V7_V8	277.3 ± 6.9	20.8 ± 4.1	73.92
V8_V9	301.2 ± 41.6	22.0 ± 4.2	50.39

[a] Dissimilarity indicates the diversity of sequence in amplified region among different phylotypes.

[b] Detectable sequences indicates the coverage of primer sets to sequences in database of EzTaxon-e

Comparison of different concentrations of mixed PhiX and amplicon library

50% of PhiX was mixed with amplicon libraries to increase genetic diversity in previous MiSeq machine. Currently, Illumina improved the Real Time Analysis software (RTA) in MiSeq Control software (MCS). This improvement reduces the mixed ratio of PhiX in sequencing library and increases data quality in low genetic diversity samples. Three different proportions of PhiX in sequencing libraries were compared to check the performances of each condition (Fig. 16). Cluster density was increased by reducing PhiX ratio in library (Fig. 16A; 445 at 50% ratio and 899 at 5% ratio), while the quality score over 30 was decreased in 5% ratio (Fig.16B). The decreasing percentage of over Q30 was about 10% of total reads, whereas the increasing number of obtained target sequence read was over 10 million (Fig. 16C). The number of undetermined reads which were not sorted by index were decreased by lowering PhiX ratio (Fig. 16D). This indicates that low proportion of PhiX in library generates more qualified sequence reads and it shows possibility of application low proportion of PhiX in amplicon sequencing. This result is consistent with previous reports which compared different PhiX ratio (Kozich *et al.*, 2013). They obtained 9.0×10^6 pair reads with 80.1% of $\geq$ Q30 at 8.0% of PhiX mixed and 10.5×10^6 pair reads with 74.6% of $\geq$ Q30 at 6.2% of PhiX mixed conditions. Different concentrations of amplicon libraries were compared in each run. Average of obtained amplicon read from each sample is shown in

Figure 16E. The difference of library concentration was not observed in PhiX 50% mixed sequencing run, whereas the number of paired reads increased by increasing concentration of library in PhiX 10% and 5% mixed runs. Over 6×10^5 reads per sample were generated from 4pM of library, and over 12×10^5 reads were obtained from 8pM of library. The increasing number of reads from 6pM to 8pM (2.7×10^5) was fewer than those from 4pM to 6pM (3.8×10^5). This indicates that no more increasing read number can be obtained after a range of library concentration. 10pM of library concentration in previous report generated similar read numbers to as 8pM in the present study. In previous study, 10pM of library generated 2×10^5 reads more than 5pM of library, however the error rate increased and the proportion of over Q30 decreased in 10pM library condition. This shows that increasing read number also make us more susceptible to high error rates. Proper concentration of library is necessary to maximize the number of qualified reads. Therefore, raw read data generated by 8pM of library concentration and 10% of PhiX mixed ratio was used to improve sequencing analysis method and evaluations in this study.

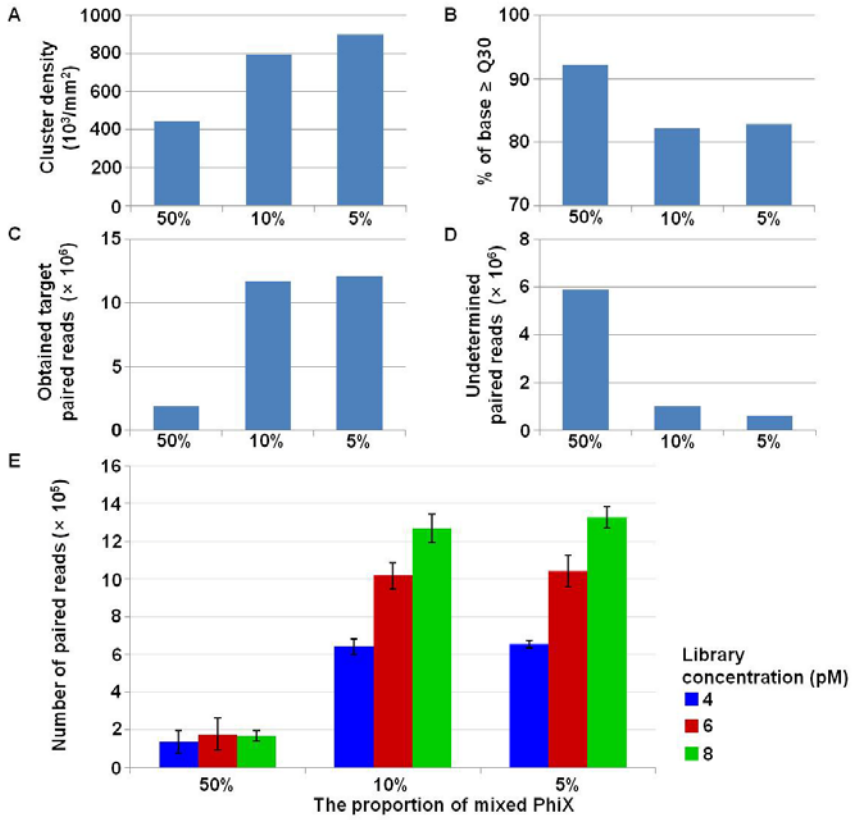


Figure 16. Number of reads obtained from a sequencing run varies with both DNA library concentration and PhiX ratio.

Evaluation of paired end merging programs

Many paired end merging programs were developed and used as introduced in the previous chapter. PANDASEQ, COPE, and FLASH were compared in terms of their running time, memory usage and merging accuracy. In paired end merging process, it is important to know that not all the read pairs can be merged because the sequencing library can't be prepared to have same length and hence to have overlap region. In addition, the read pairs can't be merged when the width of the target region is longer than the sum of the read pairs thus no overlap region exists between the pair. The evaluation was performed with V4/V5 region having 1,084,877 sequencing pairs (Table 5). Simple paired end merging program using JAVA was developed to assess the merging process and 798,989 reads were merged with a similarity of overlap region over 50%. Since the read pair originated from a single template fragment, it is fair enough to decide that the pairs with less than 50% similarity in an overlapping region are not merged. From the comparisons, COPE showed better performance in terms of running time, memory usage, and accuracy. In this study, however, merging program that I developed was used for the evaluation of the correlation between errors and quality of the sequencing reads. This program took 37 min 39 sec and used an average 740Mb of memory to merge the same paired end sequencing data.

Paired end merging

The average lengths of the selected target variable regions were 389bp

(V2/V3) and 372bp (V4/V5), respectively. Primer sequence (about 20bp) at 5' end of each read was excised because the primer sequences are highly conserved and thus uninformative. In the Figure 14, the span of overlapped region was 80bps according to the sequencing scheme.

Table 5. Comparison of Illumina MiSeq paired end merging programs.

Program	System Running Time(sec)[(a)]	Merged reads	Memory Usage (Mb)	Accuracy[(b)]
PANDASEQ	37.133	1,044,057	19.5	74%
FLASH	5.37	1,066,823	0.2	71%
COPE	15.185	889,910	14.2	74%

[a] The system time was measured by Linux 'time' command.

[b] Accuracy was measured as a ratio of the reads whose pairwise similarity to their template sequence is over 97%.

Thus, the majority of the paired reads were expected to fail in getting merged when the trim length is longer than 40 bps. This was confirmed in both the V2/V3 and V4/V5 sequencing regions as shown in Figure 17. The test was no longer needed to proceed in case the trim length is over 50bps.

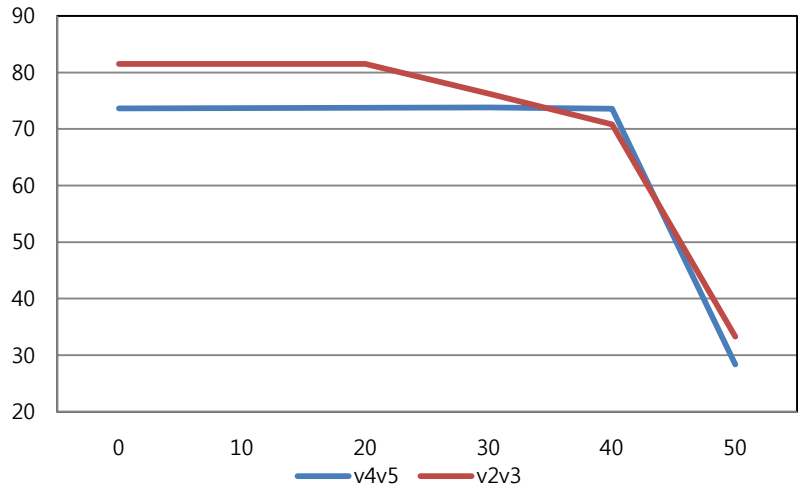


Figure 17. The number of merged reads is reduced when trim 50 nucleotides at the end of the read.

Several previous studies (Claesson *et al.*, 2010; Kozich *et al.*, 2013; Nakamura *et al.*, 2011) reported that the Quality score decreases toward 3' end with increasing sequence length, and more frequent errors are observed in second read than first read. Paired end merging process was repeated with different trim length cutoff to assess the effect of the 3' trim length on the number of merged reads (Table 6). The number of read pairs whose overlapped sequences were identical increased as the trim length increased. This indicates that more erroneous nucleotides are present toward 3' end regions. The number of reads with identical overlap region increased highly while trimming 30bp and 40bp in V2/V3 and V4/V5 regions respectively. This shows that more nucleotide substitutions occurred before 30bp or 40bp terminal region of reads than 3' end region. The number of corrected

sequence reads within overlap region increased in no-trim data set. This indicates that trimming away the 3' end sequences deprives the substitution errors of the opportunity of being corrected via merging process. Even though the total number of merged reads remains almost unchanged, the correction of erroneous sequences can improve the overall accuracy of merged sequences. The average number of substitutions for each trim length was measured using pairwise alignment. Figure 18 shows that the substitutions increased as the trim length increased, confirming the most of the substitutions can be corrected by long overlap between paired sequences.

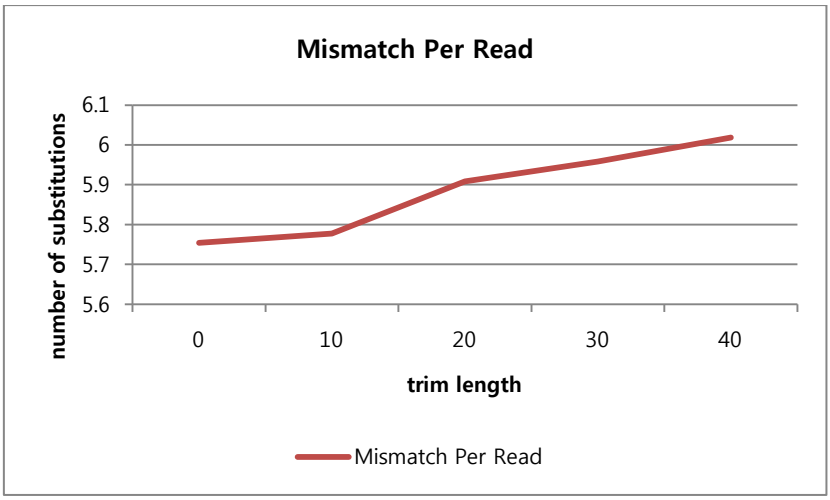


Figure 18. The number of substitutions per read increases as the trim length increases.

Increasing sequence quality by correcting mismatched nucleotide by paired reads merging was evaluated by comparing the corrected merged

sequences to the corresponding reference sequences in the mock community. Merged reads were compared with reference sequences by pairwise alignment (Fig. 19). Since the length of the merged reads varied, all of the reads were divided into 10 regions to identify any potential correlation between the error rate and quality by regions. The error rate was calculated as per the proportion of mismatches in each region. In the overlapping region, no trimming could reduce error rate more by correcting sequences than trimming a certain length of 3' end sequences. These results show that the correction of sequence in the overlapping region is necessary to increase the quality of sequences. We compared the error rates within each region of merged reads with average quality score of corresponding region (Fig. 19A). The highest error rate was observed in second reads after 30bps from primer sequences (region 9). In particular, the average quality score of this region was higher than a quality score of 35. This result shows that there are no correlation between erroneous reads and quality score on the contrary to the previous report that errors generated by MiSeq were related with low-quality score (Kozich *et al.*, 2013). This difference could be produced by different sequencing conditions such as customized primers and library composition for sequencing run. However, Kozich and colleagues (year) also showed an overestimation of OTUs after filtering chimeras and sequencing errors in all of tested regions. 21 isolates were mixed for mock community and 20 OTUs were expected in final analysis result, however they obtained over 98 OTUs in V4/V5 region results. This indicates the presence of erroneous sequences with high quality score. This overestimated information can make biased community composition by

amplicon sequencing.

Table 6. Merging statistics given different threshold of end trimming length cutoff.

Region	Trim length (bp)	Reads with identical overlap region	Reads with mismatches in overlap region	Total merged reads	Average length of merged read (bp)
V2/V3	0	338,288	831,636	1,169,924	380.0
	10	492,620	677,388	1,170,008	380.0
	20	660,071	509,335	1,169,406	379.9
	30	712,577	381,920	1,094,497	376.5
	40	670,371	346,122	1,016,493	370.9
V4/V5	0	213,206	585,783	798,989	372.0
	10	270,621	528,818	799,439	372.0
	20	335,317	464,747	800,064	372.0
	30	410,010	390,546	800,556	372.0
	40	539,858	258,653	798,511	371.9

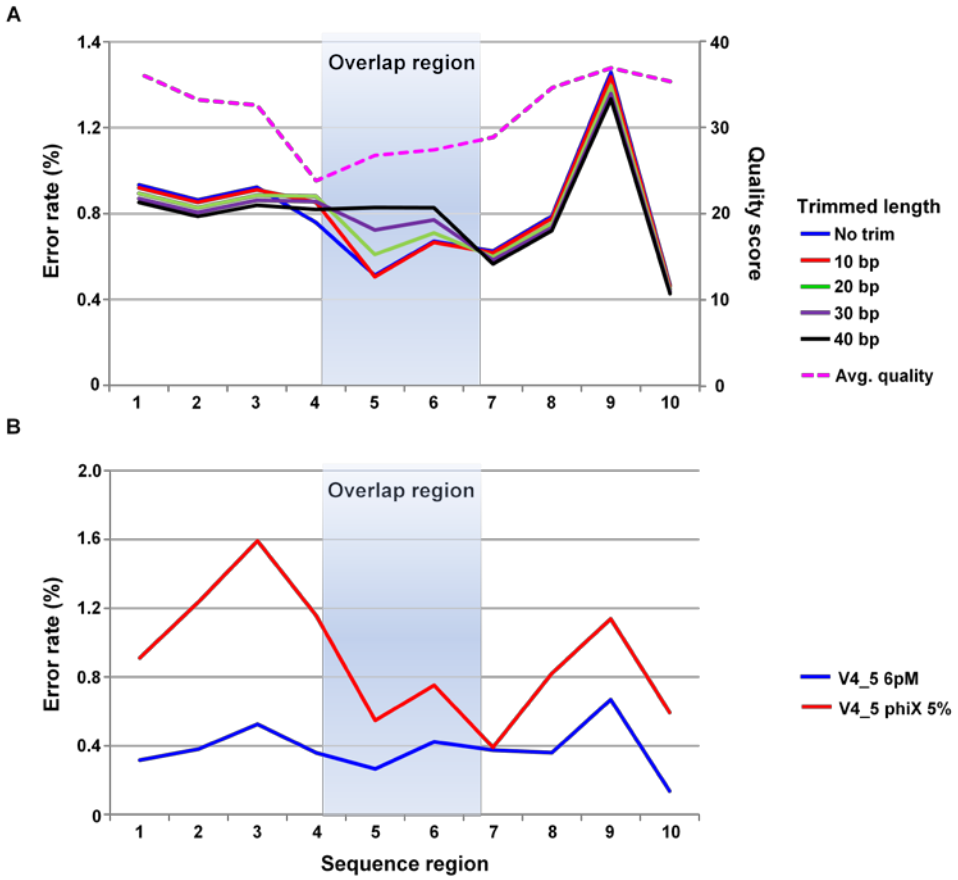


Figure 19. Correlation between errors and quality score of Illumina MiSeq read.

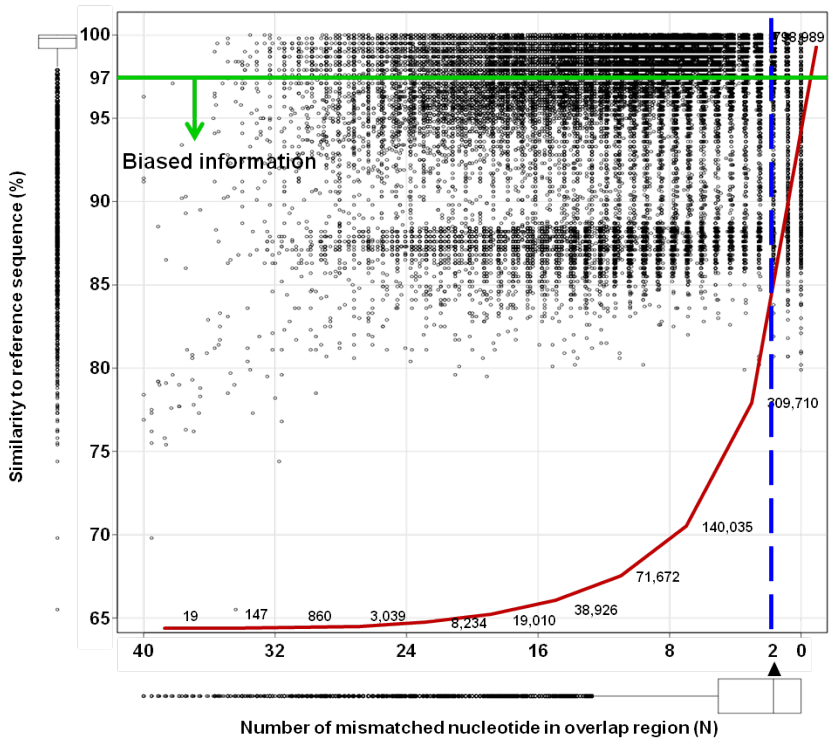


Figure 20. Plot showing the number of mismatches within overlap region of merged read and the similarity to the template sequence.

Correcting erroneous sequence in overall read

Clustering step could reduce the sequencing error and the effect of clustering in analysis of NGS reads was reported previously (Kozich *et al.*, 2013; Schloss *et al.*, 2011). The correction of randomly occurring nucleotide error is available by making a consensus sequence from the multiple alignment of member reads of the cluster. Correction of errors in overlap region was achieved by merging of paired reads, while correction of sequences in non-overlap region was possible by choosing majority sequences in heterogeneous columns of multiple alignment in the cluster (Fig. 21). Raw reads obtained from MiSeq machine after quality check were clustered at 97% cutoff value and consensus sequences of each cluster was made. Then, consensus sequences obtained from the first clustering and correcting step were clustered and new consensus sequences were created again for each cluster. Clustering and correcting sequences were repeated until the number of clusters does not change. The reduced number of clusters was compared to that of different samples (Table 7). In general, the number of clusters did not change after the third clustering step in all the samples (different target region, library concentration, and sequencing run). Over 1,000 clusters were created after first clustering step in all of the samples, while the numbers of clusters were reduced after every iterated clustering step. The numbers of clusters from V4/V5 region amplicon (75 clusters in 8pM library and 72 clusters in 6pM library) were lower than those of V2/V3 region amplicon (148 clusters) and V4/V5 region amplicon in different sequencing run (5% PhiX mixed run; 333 clusters).

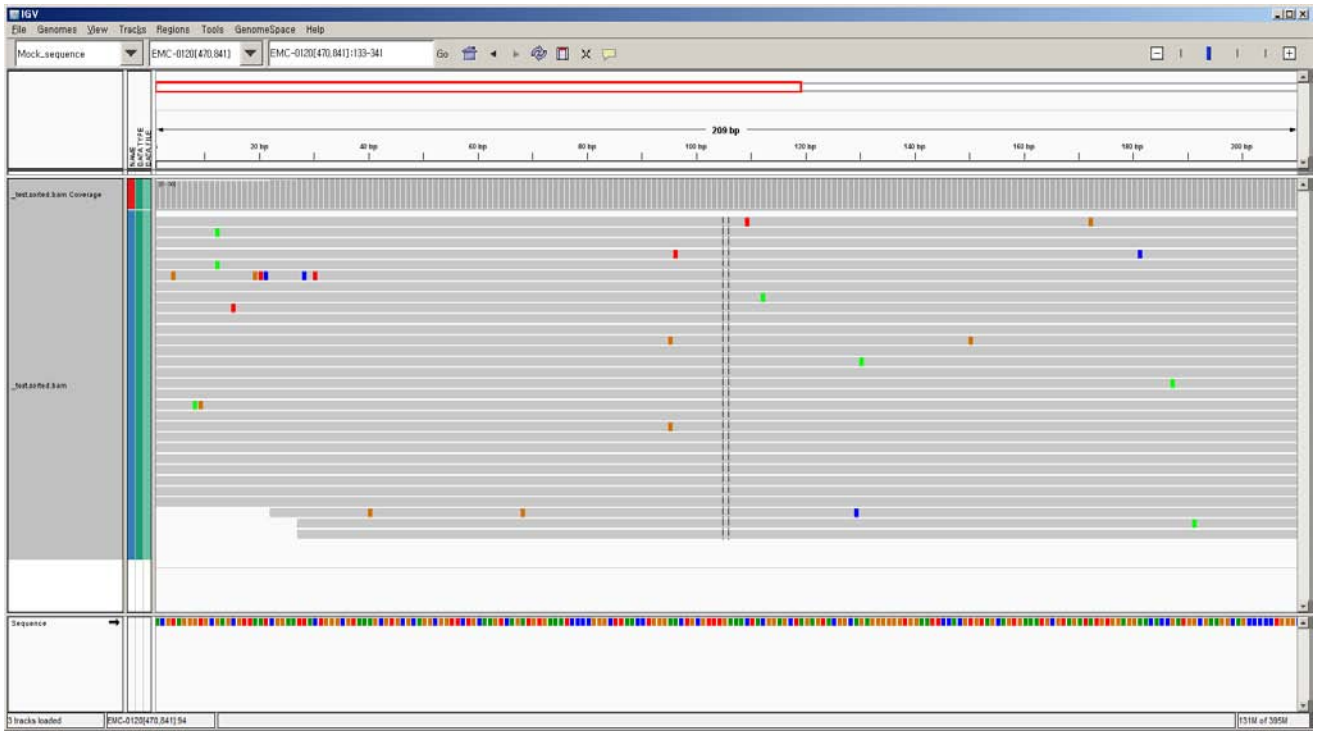


Figure 21. Errors are observed ubiquitously and no pattern is shown. Integrative Genomics Viewer.

The number of merged reads obtained from 5% PhiX mixed run V4/V5 sample (1,014,630 reads) was higher than that of same library from 10% PhiX mixed run (798,989 reads). This shows that high number of merged reads could make high number of biased information, and the determination of proper concentration of mixed PhiX is necessary. Chimera check was conducted after the first round of clustering and final round of clustering. Generated consensus sequences after the first round of clustering contains lots of chimeric sequences of raw reads obtained from MiSeq machine. Therefore, removing chimeric sequences after first round of clustering could reduce the sequences and thus computation for further analyses. Detecting chimeras and removing them after the final clustering step are necessary for the final consensus reads to be assigned taxonomic information. The effect of clustering and correcting sequences can be evaluated by comparing the ratio of bias in taxonomic assignment between the total merged reads and reads with less than 2 mismatches in overlap regions (Fig. 20). Biased community compositions of the total merged read set were found to be over 3% at phylum level and over 6% at genus level. This biased information was generated by reads under 97% similarities to mock community sequences (Fig. 20). Over 2% genus level composition bias was observed in result of reads with fewer than 2 mismatched nucleotides within the overlap region. This reduced bias was due to the reduced number of analyzed reads (489,279 reads) by discarding the merged reads with more than 2 mismatches. In contrast to these two results, reads after clustering and correcting step made around 0.5% of biased composition at the genus level, and 608,419 analyzed reads still remained

even after repeated clustering and correcting steps (Fig. 22). Clustering and correcting steps can significantly reduce biased community composition using MiSeq system.

Table 7. The number of clusters generated after each round of clustering.

The number of Clusters					
Mock community sample library (n=47)	**1st Round**	**Chimera removed**	**2nd Round**	**3rd Round**	**4th Round**
8pM V4/V5 region with 10% PhiX	2,803	2,054	94	75	75
6pM V4/V5 region with 10% PhiX	1,710	1,249	86	73	72
8pM V2/V3 region with 10% PhiX	3,502	1,387	194	154	148
8pM V4/V5 region with 5% PhiX	9,282	3,801	416	337	333

Comparison of community compositions obtained from different

primer sets and different analysis methods

Sequences of the amplified products by two selected primer sets chosen from *in silico* test after assigning their taxonomic composition and structures were compared with original mock community at phylum and genus levels (Fig. 23). The amplified product of V4/V5 was more similar to mock community composition than V2/V3 in UPGMA tree based on UniFrac distance . The proportion of each genus was different to original template concentration as the efficiency of amplification was different for each genus. However, the primer set of V4/V5 was more similar to the original community, and this region has already been reported as having highest classification accuracy (Claesson *et al.*, 2010). The amplified region of V2/V3 was different from original template even at the phylum level. So, target region of V4/V5 was considered more suitable for amplicon sequencing on MiSeq platform. The composition of bacterial community obtained from pig fecal sample was compared to that of different analysis methods (Fig. 24). Phylum composition of totally merged reads analysis was similar to the composition of clustered and corrected reads analysis but the phylum composition obtained from reads with less than 2 mismatches in overlap region was different from them.

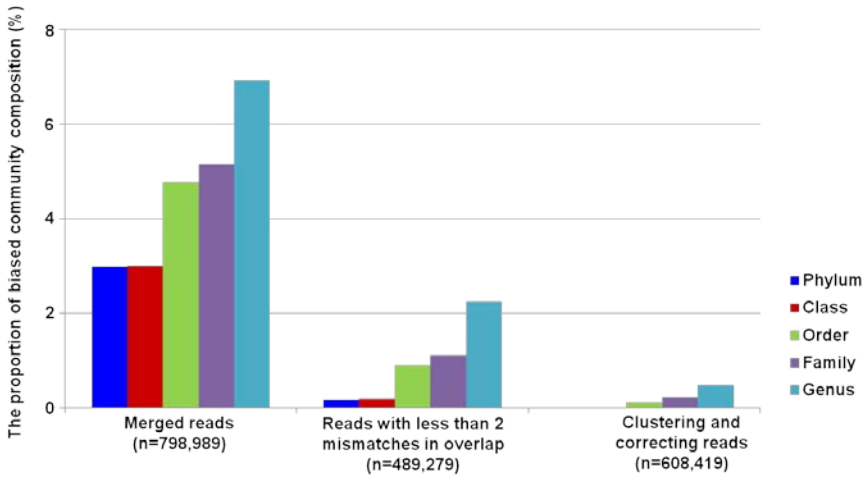


Figure 22. Taxonomic composition bias of three different error correction methods.

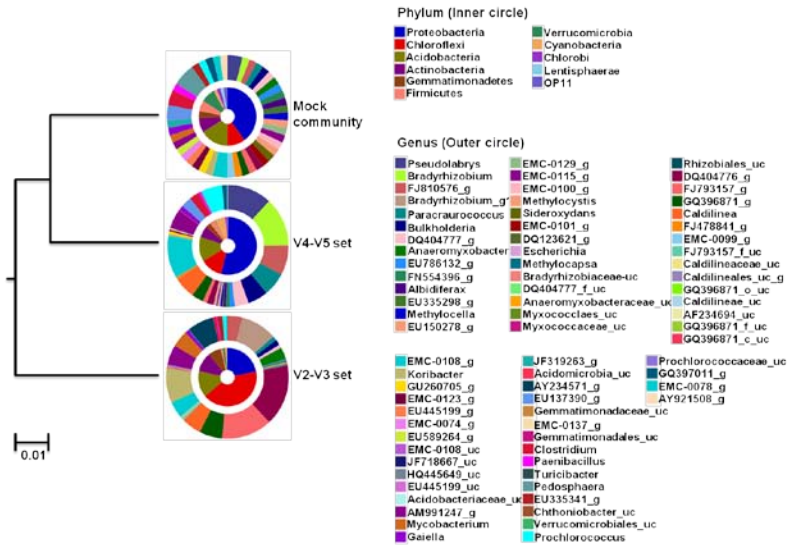


Figure 23. Comparison of taxonomic composition at genus and phylum levels assigned to two different amplicon sets with original composition.

This difference could be caused due to discarding most of the reads while following the 2 mismatches rule within overlap region (Fig. 20). The total number of analyzed reads in 2 mismatches allowed analysis (69,343) was smaller than total merged reads analysis (702,875) and corrected reads (615,236). Two mismatches allowance in overlap region can also give biased information about community composition. Analysis by iterative clustering had most of merged reads (87.5% of merged reads), and this can correct biased information as presented in Figure 22. The number of observed phylum in analysis of merged reads was 19, while 11 phyla were detected in clustered and corrected reads analysis. The number of genus in analysis of merged reads was 537, while 173 genera were detected in

clustered and corrected reads analysis. This shows the overestimated diversity can be corrected by this improved method.

Overall pipeline for Illumina MiSeq paired end amplicon analysis

The final pipeline consists of paired end merging, initial raw reads clustering, iterative consensus clustering, taxonomic assignment and chimera detection as shown schematically in Figure 25. Preprocessing for quality filtering was carried out in the paired end merging process. Errors can be corrected twice, once while paired end merging step and then again during the consensus clustering step. Database used for the taxonomic assignment and chimera detection step is the same as used for 454 pyrosequencing pipeline.

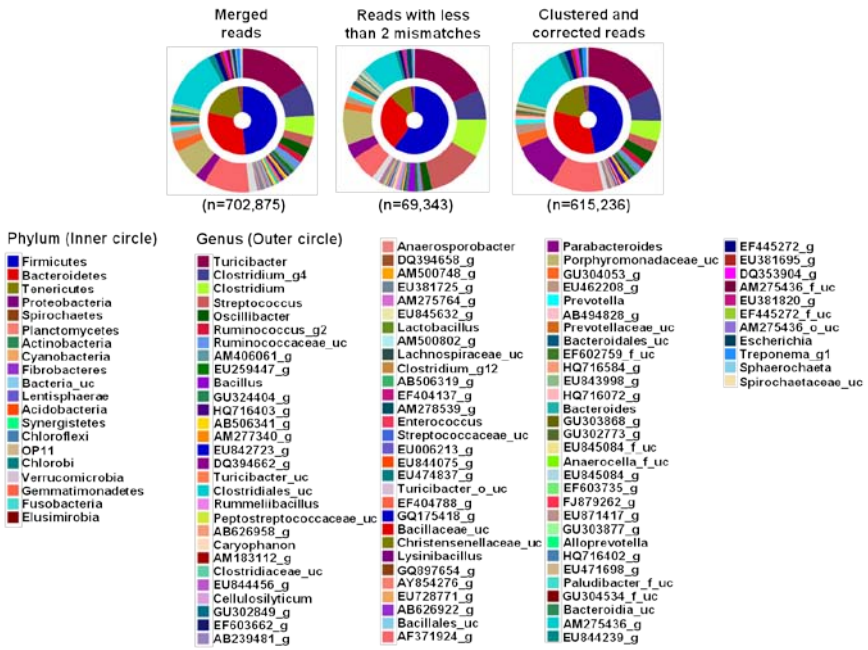


Figure 24. Taxonomic composition of pig fecal sample. Two different analysis methods were compared with the original data set.

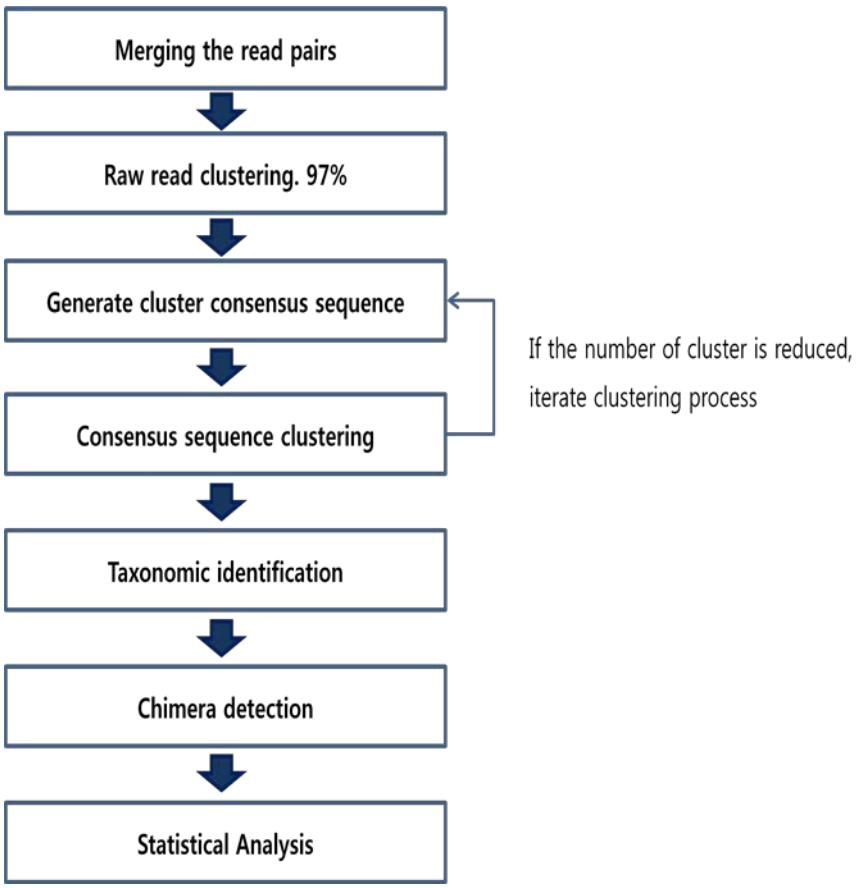


Figure 25. Illustration of overall microbial community analysis pipeline for Illumina MiSeq paired reads.

2.4 Summary and Discussion

Since the NGS technologies application to the metagenomic community studies, 454 pyrosequencing has dominated this research field. However, as a result of the rapid evolution of the sequencing technology, Illumina MiSeq can also be applied to the bacterial community study as the read length is getting longer and approaching the length of the 454 sequencing read. In this study, amplicon based microbial community analysis pipelines were constructed focusing on reducing bias by detecting and then removing or correcting errors. Since the types of errors are unique to each other platform, the pipelines were organized with different constituent programs. For 454 pyrosequencing, several homopolymeric error handling programs are present in addition to Chunlab's software. These programs were evaluated and compared to the Chunlab's newly developed clustering-based noise correction algorithm which showed better performance than the other algorithms. Chimera detection programs were also evaluated. DB dependent detection programs showed better performance than the ab initio programs. Among DB detection program, UCHIME showed higher sensitivity than the other ones. Size and quality of the database used by the DB dependent detection programs were shown to affect the detection capability and UCHIME detected the chimera most accurately with Eztaxon database.

During the development of pipeline for Illumina MiSeq system, we

evaluated the suitable amplicons sequencing conditions for the sequencing machine, and improved the curation pipeline of erroneous reads for an accurate analysis. Large number of sequences (at least 100,000 paired reads) per sample generated from MiSeq can provide high sequence depths as previously reported (Kozich *et al.*, 2013). High sequencing depth allows us to obtain more accurate community composition in sample, and reduces sequencing cost per read. However, a large number of sequences of MiSeq platform also makes biased community compositions (Fig. 20 and 22) so, the erroneous nucleotides correction step is necessary to reduce this anomaly. Variable region of V4/V5 can provide more accurate information about the microbial community for 250bps paired MiSeq reads than other regions in *in silico* and sequencing analyses. In our study, 8 8pM of sequencing library and 10% of PhiX mixing proportion turned out to be most proper for obtaining as many qualified reads as possible on MiSeq platform. Improved analysis pipeline deploying clustering and correcting step can reduce overly estimated information of community composition without filtering out too much normal reads. This makes only 0.5% of genus level bias in our mock community data (Fig. 22; Table 7). Proposed pipeline (Fig. 25) for correcting erroneous sequences and decided sequencing conditions will be useful for constantly developing newer schemes as more numbers and longer sequence reads of Illumina platform.

Swine fecal samples were analyzed using both NGS platforms to compare the recovered taxonomic structure by two platforms. Figure 26 shows a recovered bacterial community by two platforms. The microbial community composition at genus level and phylum level in the chart

showed that these two platforms recover a similar taxonomic composition. However the number of phylotypes recovered by Illumina MiSeq was much larger than 454 pyrosequencing which could be due to difference in the number of the reads generated between the two platforms (Table 8). Therefore, Illumina MiSeq can be said to detect more number of bacterial taxonomies than 454 platform largely due to the larger number of output reads.

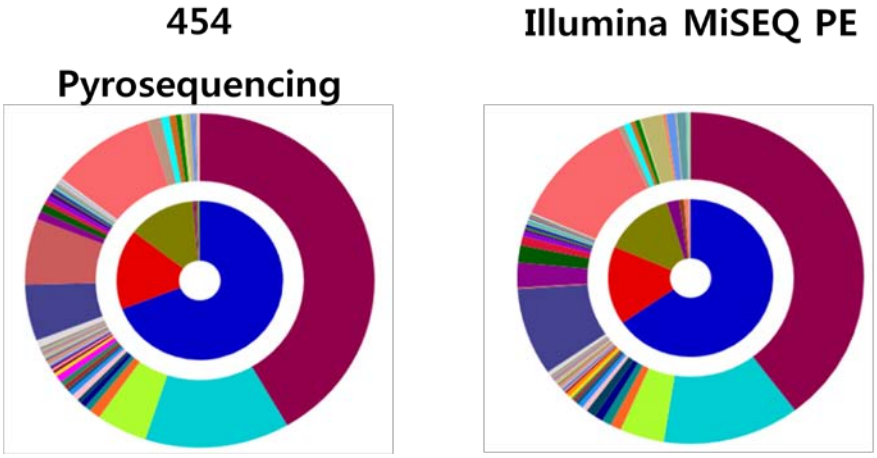


Figure 26. Microbial community recovered from 454 junior and Illumina MiSeq. Inner pie represents phylum and outer pie genus level composition.

Table 8. The number of recovered phylotypes by two NGS platforms.

	454 pyrosequencing	Illumina MiSeq
Total reads	23,057	586,328
After filtering	8,986	250,676
Phylum	7	15
Class	11	26
Order	15	50
Family	44	98
Genus	157	313
Species	440	1,372

Chapter 3 Shotgun-based Metagenome Analysis System

3.1 Introduction

Shotgun metagenome analysis approach is different from the targeted amplicon metagenome analysis in that the shotgun metagenome analysis does not target a specific gene family but it breaks down the entire genome sequences into large amount of tiny small fragments. From these randomly fragmented genomic DNA, both taxonomical and functional attributes of the metagenome can be achieved, providing us with more comprehensive understanding of the dynamics of the environmental microbial world. This random shotgun metagenomics is further divided into two categories, read based analysis and assembly based analysis (Scholz *et al.*, 2012). The read based analysis has a big drawback in terms of CPU time usage and too short read length which may cause a skewed result (Feldmeyer *et al.*, 2011; Wommack *et al.*, 2008). In the assembly based analysis, assemblers can be confused by two distinct characteristics of metagenomic data, 1) uneven representation of the organisms within a sample and 2) polymorphism between closely related members of an environment (Pop, 2009). Thus, most of the assemblers could construct unrelated contigs for each section of the genome that is consistent across multiple individuals in the sample resulting in a fragmented reconstruction and obscuring the population structure within the environment.

In this chapter, bioinformatics pipeline for metagenome shotgun

analysis was constructed using both the two shotgun metagenome analysis approaches. Illumina MiSeq paired end reads were merged for the taxonomic profiling and then from this taxonomy profile, genome database was configured dynamically for the raw read mapping, and the remaining unmapped reads were de novo assembled. The gene prediction was performed on both the assembled contigs and the remaining raw reads. Annotation database was compiled using Pfam (Punta *et al.*, 2012) EzGenome (http://ezgenome.ezbiocloud.net/) database. The visualization application was developed to provide the graphical presentation of the taxonomical and functional profiles of the metagenome.

3.1.1 Tools for Metagenomics

Recovering individual genomic sequence by assembly

Metagenome assembly algorithms developed so far could be categorized into two groups, traditional OLC algorithms and Eulerian path traverse algorithm together with de-bruijn graph algorithm (Pevzner *et al.*, 2001) where the size of k-mer is a trade-off between specificity and sensitivity. De bruijn algorithms outperforms the OLC algorithms with respect to the time parameter, however they need more memory space than OLC algorithms. Meta-IDBA (Peng *et al.*, 2011), Meta-Velvet (Namiki *et al.*, 2012) and SOAPDenovo (Luo *et al.*, 2012) implement de Bruijn graph. ABySS (Simpson *et al.*, 2009) is also a Kmer based de bruijn graph algorithm which can be executed on a distributed parallel computing

environment. MAP (Lai *et al.*, 2012) exploits the OLC algorithm. Genovo (Laserson *et al.*, 2010) takes a input sequence no more than a few millions of reads limiting the usage of this assembler to relatively smaller number of reads set. Assembly of metagenomics reads with overlap-based algorithms has a critical drawback of heavy time complexity $O(N^2)$ indicating that the time twice as much as the number of input reads, is required. On the other hand, it is important to note that the K-mer based assembly method reduces the time of assembly, but at the cost of requiring significant RAM which is proportional to the size of the genomes being assembled or the amount of the data. Further, reads are split into smaller reads length K, reads themselves are no longer the target of assembly leading to the potential introduction of assembly errors. Although these algorithms are designed to assemble multiple genome data and successfully assemble metagenome data to a certain extent, they are still far from the satisfactory when applied to high level of complex metagenome.

Taxonomic inferences

16s rRNA gene has been used to profile taxonomic structure in the metagenome. As PCR based diversity analysis of metagenome may obscure the real level of diversity and microbial composition of given sample (Wintzing *et al.*, 2006) induced by the amplification bias, metagenomics applying shotgun sequencing directly to DNA obtained from environment could avoid the PCR and primer pitfalls though it still has a limitation incurred by the length of the read which is too short to cover the full length

of the target sequence. Meta-RNA (Huang *et al.*, 2009), a part of CAMERA (Seshadri *et al.*, 2007) toolkit, or i-RDNA (Mohammed *et al.*, 2011b) could be the choice to identify the 16s rRNA sequence from metagenomic sequences. However, due to the high level of similarity of 16S ribosomal RNA, this short read cannot provide enough information to infer phylogenetic composition precisely. In addition, assembly of these short reads of highly conserved marker may result in co-assembled interspecies chimeric reads that do not exist in the given sample, may exacerbate the problem. Thus reconstruction of phylogenetic marker gene sequence (Fan *et al.*, 2012; Miller *et al.*, 2011) from metagenome data to overcome the difficulties has been addressed.

Detection or recovery of 16s rRNA from the massive data is followed by the identification of those detected sequences. This process is similar to the PCR-based metagenomic reads identification. In short, BLAST search is carried out against well curated 16S rRNA database such as Eztaxon-e, RDP, Greengenes, and SILVA. Either probabilistic identification method or sequence alignment between the query and blast hit sequences is performed for more accurate classification.

Measuring species richness and diversity index by OTU (Operational Taxonomic Unit) analysis could be calculated through various software packages. Mothur (Schloss, 2009b) and QIIME (Caporaso *et al.*, 2010) are the most prevailed analysis packages. Mothur is feasible both on Window-based operation system and Unix-based system without tricky installation process. The QIIME, though based on Unix-like operation system, provides Virtual Machine based VirtualBox in which no more inter-dependent

programs or third-party programs needs to install so that users can install and run the process more easily. However, installation and execution QIIME through command-line still need computer experties. EstimateS (Colwell, 1997) is also a widely used software, however preprocess is needed to convert sequence data to an input format acceptable to EstimateS.

Binning

After finding out how many OTUs are there, we want to know what each of those OTUs is doing and how much they make a contribution to the function of the metagenome as a community. The assembled contigs or reads should be associated with OTUs by so called binning process. There are typically two type of binning strategies, composition based binning and homolog based binning. Composition based strategy uses basic sequence descriptor such as GC content, codon usage or oligonucleotide composition. This approach is superior to homology-based one, in that this reduces the running time however it needs suitable pre-trained data set to decide the model parameter of composition. Through this approach, OTUs in the sample are erroneously classified if they are closer to each other (Wooley *et al.*, 2010). TETRA (Teeling *et al.*, 2004) adopts a composition based approach using k-mer frequencies and Markov Model. PhyloPythia (McHardy *et al.*, 2007) uses composition information together with sample-derived population models. CompostBin (Chatterji *et al.*, 2008) bins raw reads without training reference data set. Phymm uses the larger database as a training set and combine Interpolated Markov Model. Homology-based

binning exploits the reference database, mostly NR, and a various specific algorithm. This approach shows a relatively higher accuracy than the composition-based one however it needs a lot more time and computer resources than composition-based one. MEGAN (Hudson *et al.*, 2007), one of the representative homology-based binning program implements the Lowest Common Ancestor (LCA) algorithm. SOrT-ITEMS (Mohammed *et al.*, 2009) has similar approach to MEGAN but it makes use of orthology information. Meta-Bin utilizing a ORF information and BLAT (Kent, 2002). SPHINX (Mohammed *et al.*, 2011a) adapts hybrid approach to take an advantageous aspect from both approaches. AbundanceBin (Wu *et al.*, 2008) takes a different approach than above two approaches utilizing the different abundances of species in the given environmental sample. This implements a Expectation-Maximization algorithm (EM) (Do *et al.*, 2008).

Gene Prediction and Functional Annotation

Predicting genes from assembled contigs or raw reads is the key step for functional profiling. Identification of the functional coding sequences could reduce the volume of data by ignoring the out-of-frame translation. Gene calling is tackled by fragmentation of ORF sequences and short read length and lack of reference sequences, so gene calling methods such as Glimmer for a single genome are not applicable for these mixed genome data set (Desai *et al.*, 2012). Also, gene prediction with the metagenomic shotgun reads is hampered by the sequencing error disrupting coding fragment. There are two strategies for gene prediction method, Evidence-

based and Ab-initio (without reference, model-based) gene calling (Kunin *et al.*, 2008). Evidence-based (reference-dependent) methods simply find hits against reference protein database thus this approach is not suitable for metagenome data where unknown level of unknown genes exist. Though the model-based ab initio method has another challenging aspect that the model built through the training sequence couldn't be adapted for every sequence in the metagenome; this machine learning gene prediction methods are widely used because of their capacity of detecting unknown genes. Ab initio gene prediction fidelity approach employed machine learning techniques such as Neural-nets (Lippmann, 1987) or Hidden Markov Model (Eddy, 2011) or the variant of HMM, Interpolate Markov Model. MetageneAnnotator (Noguchi *et al.*, 2008), Glimmer-MG (Kelley *et al.*, 2012), GeneMark (Lukashin *et al.*, 1998), fgenesb (http://www.softberry.com), Orphelia (Hoff *et al.*, 2009), FragGeneScan (Rho *et al.*, 2010), and Prodigal (Hyatt *et al.*, 2010) are the machine learning based gene prediction programs available for metagenome sequence data. Sequencing errors causing frame shift may alter the entire downstream sequence ending up being assigned different/wrong functional annotation. In addition, prediction accuracy can be compromised by other factors, such as genomic islands of differing GC ratio, pseudo genes and genes with programmed frameshift (Pati *et al.*, 2010). Thus the key factor for the accurate prediction of the ORFs is the prediction of true translation frame.

Functional annotation database

Gene prediction, is followed by functional annotation and this process is not much different from single genome annotation process. In the typical annotation process, predicted genes are subjected to a homology search against an existing database consisting of known annotated proteins such as Pfam (Bateman *et al.*, 2002; Bateman *et al.*, 2004; Punta *et al.*, 2012), TIGRfam (Haft *et al.*, 2003), SEED (Aziz *et al.*, 2008) or NCBI COG (Tatusov *et al.*, 1997) database. SEED annotation environment is the first annotation system deploying subsystem based approach (Overbeek *et al.*, 2005). SEED database is composed of 30 categories and 106 sub categories. Gene Ontology (Ashburner *et al.*, 2000) consortium was founded to provide common language for a shared biological elements via integrating the various ways of description and conceptualizing the elements. Three top level mutually exclusive categories, biological process, cellular component, and molecular function are in GO database. All the sub elements should belong to either one or all of the three categories. Prokaryotic COG (Cluster of Orthologous Groups of Proteins) database is a phylogenetic classification of proteins from complete genomes (Tatusov *et al.*, 2001) extended later to include Eukaryotic Cluster of Orthologous Group (KOG) (Tatusov *et al.*, 2003). Each group contains proteins which are thought to be orthologous and coding sequences from a newly sequenced genome are added to the database using designated tool called COGNITOR. Pfam is a large collection of protein families and domains grouped by hidden markov model profile (HMM profile) constructed from the multiple sequence

alignment. Proteins in a family descended from a common ancestor and typically have similar functions and significant sequence similarity and the similarity is the strictest indicator of homology and therefore the clearest indicator of common ancestry, so the multiple alignment of the proteins in a protein family should be achieved to have a certain level of homology. Typical annotation process carries out homology search using blast against the mentioned DBs. However, this homology search method often fails to annotate all the predicted genes due to the novel protein which was predicted. So, alternatively, context annotation methods such as genomic neighborhood (Bohnebeck *et al.*, 2008), gene fusion, phylogenetic profiles and co expression (Kunin *et al.,* 2008) could be the alternatives. One notable caution is as the delineate of metagenome in interest by gene prediction and annotation methods mentioned above is still far from satisfactory, there is significant room for improvement in the methods (Mavromatis *et al.*, 2007).

Visualization

One of the important constituents that metagenome analysis pipeline must provide is to visualize the end results of the metagenome analysis. Since the purpose of the shotgun metagenome analysis is to reveal both the microbial community structure and the functional profile, the visualization application is required to provide intuitive graphical interface representing the community structure and functional profile. Along with the taxonomic and functional profiling independently, both profiles can be related to the

each other, which is the key advantage of the shotgun metagenome analysis enabling us to understand the main agent of the functions fulfilled by the metagenome.

Table 9. Various algorithms and tools used for metagenome analysis.

Category	Program Name	Homepage	Algorithm
Assembly	Meta-IDBA	http://www.cs.hku.hk/～alse/metaidba	Eulerian path with de Bruijn Graph
	Genovo	http://cs.stanford/edu.genovo	Probabilistic model based on Iterated Conditional Model (ICM)
	Meta-Velvet	http://metavelvet.dna.bio.keio.ac.jp	Eulerian path with de Bruijn Graph
	MAP	http://bioinfo.ctb.pku.edu.cn/MAP	Overlap/Layout/Consensus
	SOAPDenovo	http://soap.genomics.org.cn/soapdenovo.html	Eulerian path with de Bruijn Graph
	Abyss	http://www.bcgsc.ca/platform/bioinfo/software/abyss	De Bruijn Graph/Distributed
Binning	TETRA	http://www.megx.net/tetra	Composition Based
	Phylopythia	http://cbcsrv.watson.ibm.com/phylopythia.html	
	CompostBin	http://phylogenomics.wordpress.com/software/	
	Phymm	http://www.cbcb.umd.edu/software/phymm/	

	MEGAN	http://ab.inf.uni-tuebingen.de/software/megan/	Homology Based
	SoRT-ITEMS	http://metagenomics.atc.tcs.com/binning/	
	Meta-Bin	http://metabin.riken.jp/	
	MetaPhlan	http://huttenhower.sph.harvard.edu/metaphlan	
	SPHINX	http://metagenomics.atc.tcs.com/SPHINX/	Hybrid
Gene calling	MetaGeneMark	http://exon.gatech.edu/GeneMark/metagenome/	Oligo Nucleotide Frequencies
	MetaGene-Annotator	http://metagene.cb.k.u-tokyo.ac.jp/	Ribosomal Binding Site
	Orphelia	http://orphelia.gobics.de/	Fragment length-specific model
	FragGeneScan	http://omics.informatics.indiana.edu/FragGeneScan/	Codon Usage and Error Model
	Glimmer-MG	http://www.cbcb.umd.edu/software/glimmer-mg/	Interpolate Markov Model

Table 10. Database of functional coding sequence for metagenome annotation.

Name	Homepage	Key Feature
TIGRFAM	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	Curated multiple sequence alignments HMMs
PFAM	http://pfam.sanger.ac.uk/	Multiple alignments and HMMs.
COG	http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups
KOG	http://genome.jgi.doe.gov/Tutorial/tutorial/kog.html	EuKaryotic Orthologous Groups
KEGG	http://www.genome.jp/kegg/pathway.html	Collection of manually drawn pathway maps
KO	http://www.genome.jp/kegg/ko.html	KEGG Orthologous group
GO	http://www.geneontology.org/	Gene ontology
SEED	http://www.theseed.org/wiki/Main_Page	Fellowship for Interpretation of Genomes based on subsystems.
IMG	http://img.jgi.doe.gov/cgi-bin/w/main.cgi	The Integrated Microbial Genomes system of all publicly available genomes from three domains of life
eggNOG	http://eggnog.embl.de/version_3.0/	Evolutionary genealogy of genes:Non supervised orthologous

Public Metagenome analysis pipelines

As sequencing cost is getting cheaper and cheaper, more researchers are beginning to participate in the metagenome projects but the computing resources required for the raw read processing and analysis listed in the previous section is beyond a small academic laboratory scale. Different metagenome analysis pipelines are provided by several consortiums or web-portal in a form of “Platform as a Service (PaaS)”, behind which is high performance computing resources for big data processing. In addition, as metagenomics workflow became a routine and common process, more and more cloud-computing based analysis systems with fixed analysis protocol were released. European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/metagenomics/), CAMERA (Seshadri *et al.*, 2007), IMG-M (Markowitz *et al.*, 2009), and MG-RAST(Glass *et al.*, 2010) are a few examples of web-based metagenome analysis tools. METAREP (Goll *et al.*, 2010), which is a part of HMP, also a tool for comparative metagenomics either at the level of whole metagenome or at the level of the specific protein coding sequence providing relationship between environments or the time point of the same environment. Smash community (Arumugam *et al.*, 2010), MetAMOS (Treangen *et al.*, 2011) and CloVR (Angiuoli *et al.*, 2011) are standalone metagenome analysis pipelines. One of the strong point of CloVR is that CloVR takes advantage of DIAG cloud (Keahey, 2010) or Amazon cloud system (http://aws.amazon.com). Using the cloud system to process big data gives

good results as well as give high performance in respect to time complexity and cost (Wilkening *et al.*, 2009). Though the process, software and database used in each step are different by the pipelines (Table 9 and 10), there is little difference in overall process of pipelines (Fig. 27; Table 11).

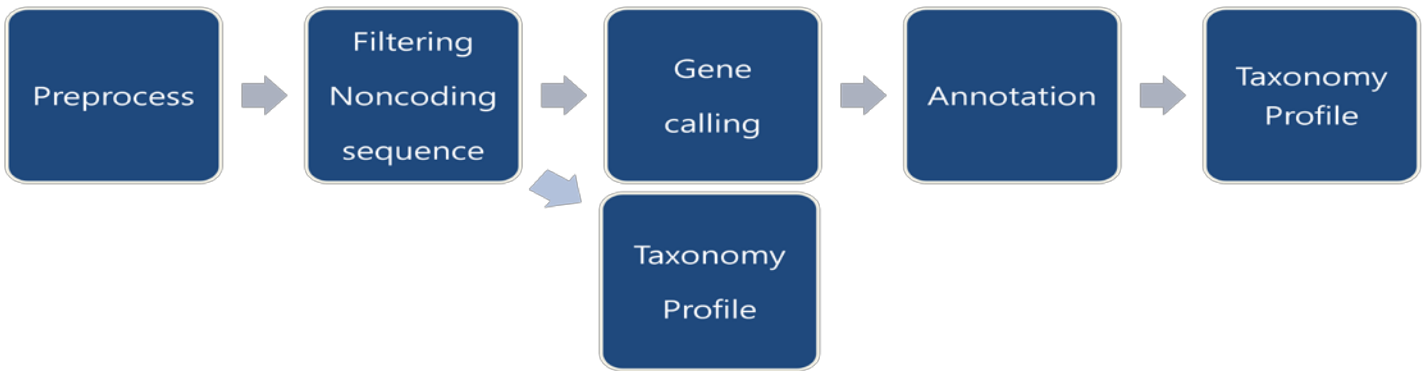


Figure 27. General metagenome analysis pipeline implemented by the open metagenome pipelines including MG-RAST, EBI and IMG/M.

Table 11. Detailed process of metagenome analysis pipeline of the public metagenome pipelines.

	MG-RAST	IMG/M	EBI	CAMERA
Preprocess	Artificial duplicates. Screening reads from some model organisms.	Quality filtering. Trimming. Noisy duplicates read.	Low Quality End, Ambiguous Read filter Duplicate read filter Repeat sequence filter	QC filtering De-replicate
Filtering RNA & Non-coding sequence	Detect rRNA using BLAT and SILVA90	Repeated sequence rRNA-SPATAN tRNA-tRNAScan	Filtering rRNA reads rRNASelector V1.0.0	HMMER and BLAST
Gene calling	FragGeneScan, Protein sequence clustering (>90%) by UCLUST	GeneMark Prodigal MetaGene FragGeneScan	FragGeneScan	RAMMCAP FragGeneScan MetaGene

Annotation	M5NR database. Similarity-based approach	Public Database	InterProScan 5.0 Public Database	Public Database
16S Taxonomy profiling	Clustering (>97%) Identification by blasting against M5RNA DB (SILVA, GREEN-GENEs, RDP)	No taxonomy profile using rRNA	16S rRNA using QIIME	No taxonomy profile using rRNA

3.2 Methods

Mock metagenome

Artificial mixture of genome was fabricated with 5 known strains (Table 12). The pipeline development was based on the result of the evaluation of each step with this artificial metagenome.

Paired end merging

250bp paired end reads obtained from Illumina MiSeq machine were merged using pairwise sequence alignment program implementing Miller and Myers optimal linear space sequence alignment algorithms (Myers *et al.*, 1988). Two paired reads were merged by aligning the overlap region between the pair. Similarity of overlap region was calculated whereas the read pairs showing an alignment similarity lower than 50% remained unmerged. No mismatch number threshold was applied.

Taxonomy profiling

The pre compiled profile using Hidden Markov Model of ribosomal RNA was used for taxonomy structure inference. 16S ribosomal profile of EzTaxon-e and other 5S and 23S profiles from rRNASelector (Lee *et al.*, 2011) were used.

Gene calling

Ab initio metagenome gene calling methods including MetaGeneMark, Prodigal, and FragGeneScan were evaluated because among many ab initio gene prediction methods, as only these 3 programs outputted translated amino acid sequences. Evidence-based gene annotator is not appropriate for the metagenomics because there are unknown coding sequences within the unculturable organisms in the metagenome. Because genes in both assembled contig sequences and unassembled raw reads should be searched, the programs were evaluated using these two different input data type. The genome sequences of 5 mock strains and simulated raw reads were used as two different input data type. ART (Huang *et al.*, 2011) was used to create artificial simulated data set. The resultant predicted orf sequences were searched against blast database containing only CDSs from the 5 mock strains and the positive predictive value (PPV (Hoff *et al.*, 2008)) was measured to compare the prediction sensitivity of the programs.

Annotation database

The final end result of the metagenome analysis should be to provide both functional and taxonomical information. So the coding sequences searched in the gene calling stage should have not only functional information but also taxonomy of the organism from which the coding sequence originated. With these functional and taxonomical information annotated to each coding sequence, taxonomic binning task could be achieved. Annotation process is homology based search, so the accuracy of

the database is also importance. There are numerous publicly released databases for CSD annotation as listed in Table 8. In the pipeline here, CSD protein sequences from Ezgenome for prokaryotes were configured. To avoid any redundancy in prokaryote database, only the protein coding sequences of one selected representative strain of one species were taken into the database. The representative strains were selected by the following rules. 1. integrity, 2. type strain and 3. date of release. Subsystem (Overbeek *et al.*, 2005) information was integrated into the Ezgenome CDS database. Since metagenome is composed of both eukaryotes and prokaryotes, so it is ideal to analyze eukaryotes in the metagenome; as not analyzing them will compromise our ability to assess a microbial community in its entirety (Ni *et al.*, 2013), Pfam database for eukaryote are integrated into the annotation database.

Table 12. List of 5 known bacterial strains comprising mock metagenome.

Name	NCBI BioProject ID	Size (bp)	Molecules
Enterobacter aerogenes **KCTC 2190**	PRJNA68103	5,280,350	1
Escherichia coli **str._K-12_substr._DH10B**	PRJNA58979	4,686,137	1
Pediococcus pentosaceus **ATCC 25745**	PRJNA57981	1,832,387	1
Staphylococcus aureus **subsp. aureus JH9**	PRJNA58455	2,967,558	2
Vibrio vulnificus **YJ016**	PRJNA58007	5,260,086	3

Dynamic Configuration of Mapping Genome Set

A set of genome sequences for raw reads mapping was configured using the 16S ribosomal RNA profile inferred from both merged and unmerged raw reads in the previous step. Pairwise similarity was calculated between the inferred 16S ribosomal RNA sequence and corresponding raw read sequence. Uncultured strains and strains showing less than 99% pairwise similarity were filtered out. If the selected genome sequence existed in the ezbiocloud, the sequence was chosen as mapping genome. The sequence of the mapping genome was downloaded from the ezbiocloud (http://www.ezbiocloud.net/ezgenome). Representative single strain of a genus was selected if all the selected strains of one genus do not exist in the ezbiocloud.

Raw read mapping softwares including SOAPAligner, BWA, and Bowtie are compatible with short paired end read data. These programs were evaluated in terms of the number of generated contigs and N50 statistics (Miller *et al.*, 2010). All these short read mapping programs commonly adopt the seed alignment which is the substring of a read aligned firstly to the reference sequence. Once the seed aligned to a region or multiple regions in the reference sequence, the algorithms try to extend alignment to reach the end of the read. The length of the seed together with the number of allowed mismatches within seed are the important options to which the result of the mapping is very sensitive. In addition, minimum alignment similarity, allowed mismatch within whole alignment, and the

minimum, maximum size of the fragment also are important options in case of the paired end mapping. In the evaluation of the mapping programs, 125 was given as the seed length and 2 mismatches were allowed within the seed alignment. Discordant or unpaired pairs were discarded. The mapping program outputs the result file in SAM (Sequence Alignment Map) file format, thus SAMTools ver1.18 was used to convert the SAM file into fasta sequences.

Evaluation of De novo Assembly

The reads which were not aligned in the mapping stage were subjected to de novo assembly. The assembly of metagenomic shotgun reads is time consuming process and computing intensive work but as the number of reads could be reduced in the previous mapping stage, time complexity and accuracy were expected to be improved. There are a number of methods available to validate the resultant assembly and the error correction methods for an isolated single genome (Engle *et al.*, 1994; Kim *et al.*, 2001; Rouchka *et al.*, 1998). However, no such programs for metagenome assembly evaluation are known so far. Thus, comparison of the assembly result was performed by comparing the number of the contigs and the contig statistics, N50 and N90. Kmer based, combined with De bruijn graph assemblers including Meta-Velvet, Meta-IDBA, Abyss and SOAPDenovo were evaluated. To test the effect of the choice of the Kmer, assembly was repeated given different Kmers, 37, 63, 127, and 185, using Meta-Velvet. After finding the appropriate Kmer, the de novo assemblers

run repeatedly given the selected Kmer.

Visualization of Results of Analysis

Java-based application was developed to summarize and visualize the shotgun metagenome analysis. Standard JAVA runtime environment version 7 and JAVAFX library package was used to draw user interactive charts. The visualization application was designed to examine and compare multiple metagenomes at the same time, enabling comparative metagenome analysis in the perspective of both taxonomical and functional properties of the metagenomes. The visualization of relationship between taxonomy and function was also considered to be visualized.

3.3 Results

Sequencing was carried out on the Illumina MiSeq instrument (500 cycles, 2 x 250 bp on a paired-end protocol) in accordance with the manufacturer's instruction. The number of reads obtained and total base after demultiplexing are shown in Table 13. Reads with average quality value lower than 25 and containing more than one ambiguous nucleotide were discarded. The estimated sequencing depth was about 149X. Sequencing depth, referred to as 'sequencing coverage' in some articles, is of prerequisite importance for metagenome analysis (Tyson, 2008; Wooley *et al.*, 2010).

Table 13. Sequencing results of Illumina MiSeq machine.

Total number of read pairs	7,773,840
Good Quality[(a)]	7,676,083
Dropped Reads	97,757
Forward nucleotides	1,067,941,714
Reverse nucleotides	1,921,795,620
Total number of bases	2,989,737,334
Depth	149X

[a] Contains no ambiguous base 'N' and average quality Q > 25.

Taxonomy profiling and configuration of mapping genome set

The paired end merging resulted in 36,366 merged reads while a greater portion remained unmerged. Ribosomal RNA sequences in both merged and unmerged reads were detected using the Hidden Markov Model profile. As shown in the Table 14, the number of detected genus from the most of rRNA profiles was far more than the number of mock strains indicating that the bias genus appeared as expected. 140 strains were selected from both 16S ribosomal RNA profiles (merged and unmerged) and finally 84 genomes amongst them were selected for mapping genome sequences.

Table 14. Result of taxonomic profiling using HMM profile of three types of ribosomal RNA.

	Unmerged reads			Merged reads		
HMM Profile	5S	16S	23S	5S	16S	23S
Detected Reads	2,761	20,708	41,076	6	136	21
Estimated Genus	21	84	109	3	15	12

Raw read mapping to the reference genome database

Performance of the mapping programs was assessed by the number of contigs; they generated N50, and N90 statistics (Fig. 29; Fig. 30). The

larger the statistics N50 or N90, the better the alignment performed. Thus, the combination of the number of contigs and these contig statistics should be taken into the consideration for the assessment of both aligner and assembler programs. BWA generated the least number of contigs while the SOAP generated largest number. In addition, both N50 and N90 statistics indicated that BWA performed better than the other two aligners generating smaller number of contigs which were longer than other aligners.

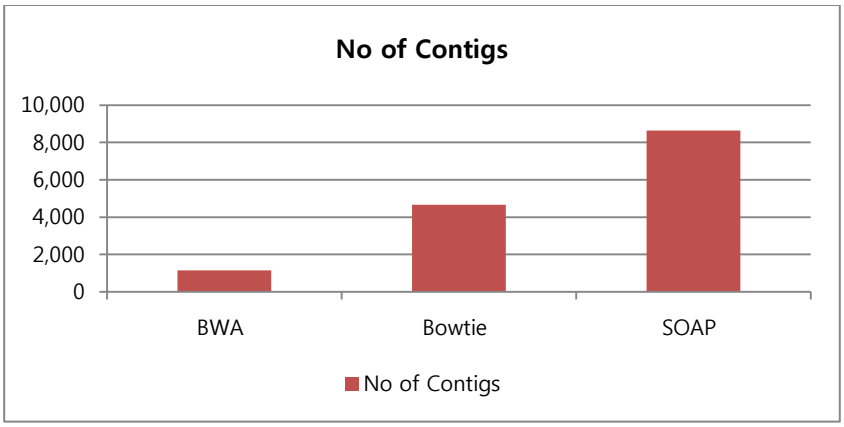


Figure 28. The number of contigs obtained from each mapping program.

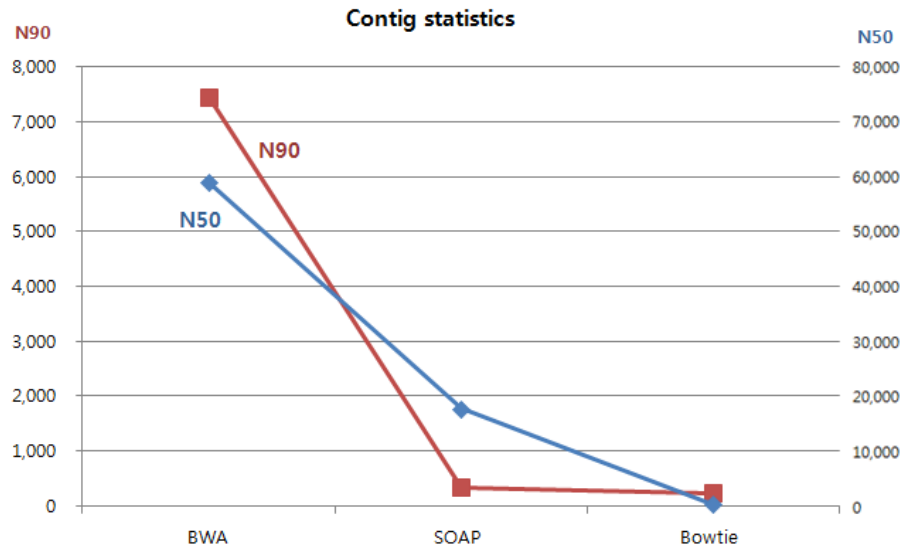


Figure 29. N50 and N90 statistics of obtained from each program.

Using BWA, I tried to map 7,676,083 quality reads to mock strains for confirming the feasibility of the mapping strategy. 74.4% of all the genome sequences were covered by the contigs indicating that the mapping strategy could be used to make contig sequences from raw reads. In addition, taxonomic profiling using short read is likely to be biased because of the short read length and highly homologous ribosomal RNA sequences; hence the mapping database could have false strains which don't really exist in the environment. So, a test was carried out to assess the effect of false strains participation in the mapping database. The test mapping database was comprised of two true strains, *Staphylococcus* aureus and *Pediococcus pentosaceus*, and one false strain, *Klebsiella pneumoniae* which is basonym of *Enterobacter aerogenes* having 98.7% ribosomal RNA sequence

similarity to *Klebsiella pneumoniae* and thus is likely to be a false positive mapping genome. After mapping, the resultant SAM file generated from BWA was converted into fasta sequences. Total 2,168 consensus sequences were obtained and were categorized into their original genome using BLAST to measure the average length of the contigs and to explain how much the individual mapping genome was covered by the resultant contigs (Table 15). The average contig length of *Staphylococcus aureus* and *Pediococcus pentosaceus*, which are true mock genomes, were 2,556 and 1,590 respectively. The mapping coverage of the falsely selected strain, *Klebsiella*, was only 0.4% and the average similarity of these 526 contigs to the *Enterobacter aureus* genome sequence was about 98% indicating that only the reads sequenced from the common region of the *Klebsiella* and *Enterobacter* genome sequences were mapped and hence these contigs made by mapping to the false mapping genome sequence wouldn't alter the accuracy in the downstream analysis.

Table 15. Consensus contig sequences obtained from reference mapping process.

Strains	Chromosome Size (bp)	Covered Span (bp)[(a)]	Contig No.	Avg. Contig Length (bp)	Coverage
Staphylococcus aureus	2,937,219	2,550,239	717	2,556	86.8%
Pediococcus pentosaceus	1,832,387	1,471,556	925	1,590	80.3%
Klebsiella pneumoniae	5,472,672	207,547	526	394	0.4%

[a] Sum of the contigs length

De novo assembly

After read mapping stage, the number of remaining reads was about 70% (5,417,868 pairs) of the total reads, which were to be de novo assembled. To find out the optimal Kmer length for the de novo assembly, de novo assembly ran repeatedly given four different Kmers, K31, K63, K127 and, K185. Figure 30 shows the number of contigs which varies upon the given Kmers indicating that assembler performs better when given the larger Kmers, K127 and K185 than when the smaller Kmers were given. It is natural that the assembler given short Kmer generate larger number of short contigs because a single read is divided into smaller K-length sub sequences. N50 indicates that the K127 generates longer contigs (Fig. 30) satisfying our purpose of making longer contig sequence for the accuracy. In short, when Kmer was configured to the half the read length, assembly performs better than other cases in terms of the number of contigs and the contig length distribution.

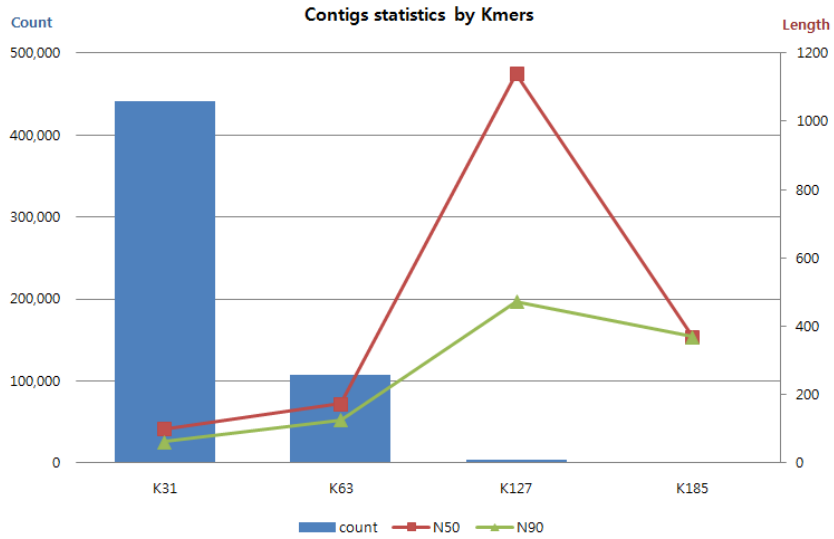


Figure 30. The number and length of the contigs varies upon the given Kmer.

The number of contigs and N50 statistics measured from the contigs obtained through four Kmer-based de novo assembly programs, Meta-velvet, SOAPDenovo, Meta-IDBA, and Abyss were compared. Among the 4 metagenome de novo assembler, Meta-IDBA was considered to be most appropriate de novo assembler for the Illumina MiSeq paired end data with respect to N50 and the number of generated contigs (Fig. 31).

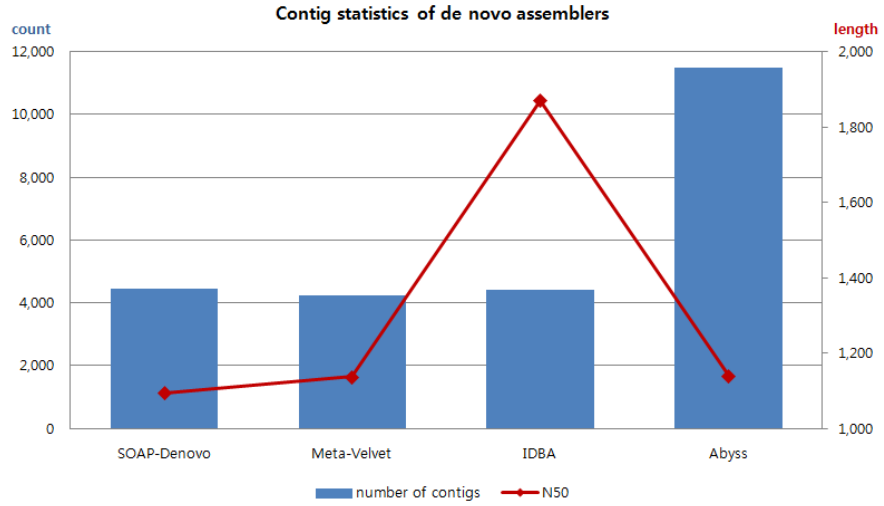


Figure 31. The number and length of contigs of metagenome de novo assemblers.

After de novo assembly, blast search was carried out against the reference genome database to categorize the assembled contigs into corresponding reference genomes. After categorization, the contigs in each category were subjected to blast search against the contigs genome sequence to calculate the average identity of assembled contig to its original genome sequence. The average contig length, the coverage of the assembled contigs of individual genome, and the average identity were calculated (Table 16). The coverage and the average contig length of the Staphylococcus aureus and Pediococcus pentosaceus were relatively low and short because the substantial number of raw reads of these two genomes were already have been mapped in the mapping stage and hence the relatively smaller number of raw reads did participate in the de novo assembly.

Enterobacter aerogenes whose ribosomal RNA reads were highly similar to and thus falsely identified as *Klebsiella* was recovered almost completely and average contig length was longer than average length of other genomes' contigs. Average contig length of *Escherichia coli* and *Vibrio vulnificus* were longer than the average length of the coding sequence of five reference mock genomes indicating that gene prediction bias incurred by the short read length could be reduced.

Theoretically, the identity of the each contig to their reference mock genome sequence is 100%. However, because of many reasons including sequencing errors, sequence homology among other coexisting species causing chimeric assembled reads, the sequence identity could be altered. The average identity of the HSPs (High Scoring segment Pairs) shown in Table 16 indicates the assembled contigs may contain falsely assembled region.

Table 16. Result metagenome de novo assembly.

	Chromosome (bp)	Sum of Contig Length(bp)	Contigs No	Avg. Contig Length (bp)	Coverage (%)	Identity (%)
S. *aureus*	2,937,219	2,863,424	1,655	1,730	97	93.4
P. *pentosaceus*	1,832,387	866,102	1,406	616	47	97.3
E. *aerogenes*	5,280,350	5,303,502	4,446	1,192	100	94.9
E. *coli*	4,686,137	1,491,471	2,736	545	31	95.5
V. *vulnificus*	5,060,086	4,459,405	4,071	1,095	88	91.3
No hit	-	295,895	422	701	-	-
Total	-	15,279,799	14,736	-	-	

Gene calling

The number of sum of CDSs of each strain is 18,544. These CDSs were formatted to create a database for blast search. All the predicted orf protein sequences from both chromosome sequences and raw reads sequences were searched against this database. From the chromosome sequences, MetaGeneMarker (MGM) and Prodigal (PROD) predicted almost the same number of orfs outperforming the FragGeneScan (FGS), but from the raw reads, FGS predicted more number of orfs than the other programs (Table 17) indicating that the FGS performed better with the short reads even though FGS provides the options for assembled contig sequences.

Table 17. The number of predicted orfs in both chromosomes and raw reads.

Input Data\ Program	FGS	MGM	PROD
Chromosome	15,093	18,437	18,491
Raw reads	162,499	156,814	159,605

The prediction coverage and the ratio of the number of missing genes, was calculated as the sum of the total number of CDSs of each strains minus the total number of unique genes in the blast search of predicted orfs. FGS showed the highest ratio of missing genes when chromosome

sequences were input while the ratio of FGS was slightly lower than that of the other programs when the prediction was performed on the raw reads also indicating that the FGS is optimized for the short read (Fig. 32).

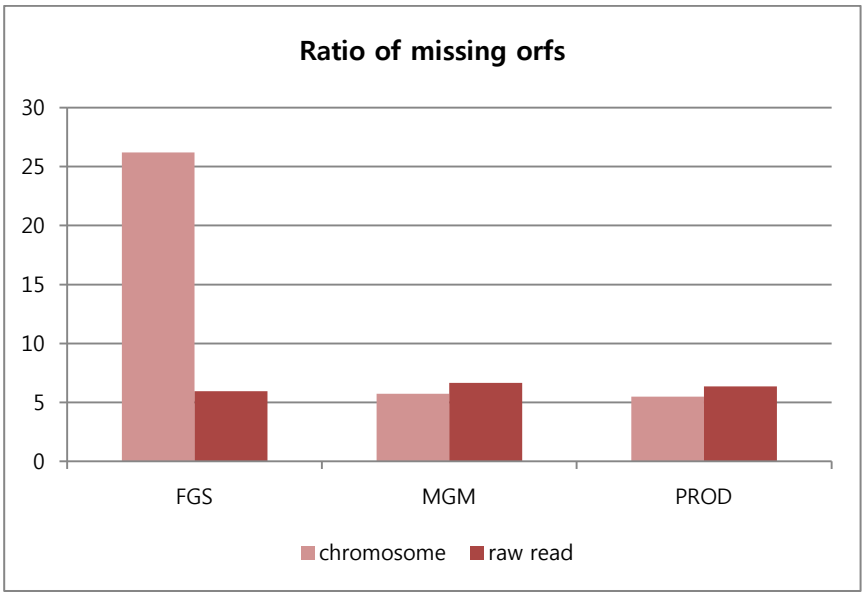


Figure 32. The ratio of missing ORFs of gene calling programs.

The positive predictive value (PPV) is a measure of the prediction sensitivity of each program and was measured as the ratio of the number of true positive orfs in all of the predicted orfs. Figure 33 shows the sensitivity of gene callers. The sensitivity of MGM and PROD were higher than the FGS in both cases and MGM showed higher sensitivity than that of PROD. Meanwhile, as Figure 32 and 33 shows, the use of longer contigs or chromosomes sequences can reduce the bias from the perspective of both sensitivity and ratio of missing orfs, if either MGM or PROD is used for

gene predictions and the pipeline took the MGM as the orf prediction program according to the sensitivity and missing gene ratio.

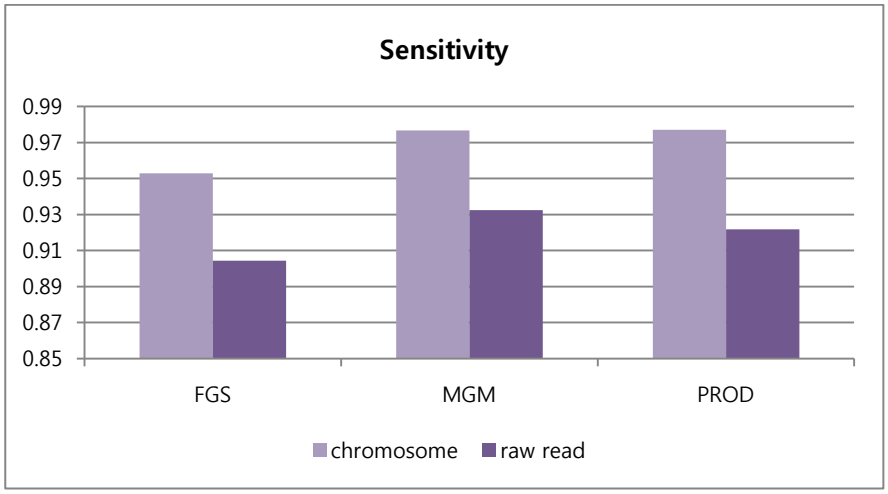


Figure 33. Positive Predictive Value indicates gene prediction sensitivity of gene calling programs.

Annotation database compilation

The protein coding sequence annotation database was compiled from EzGenome and PfamA database (Fig. 34). Total 3,164 representative strains were chosen from the EzGenome database and 10,685,000 prokaryote proteins were retrieved from those selected genome sequences (Table 18). Protein coding sequences retrieved from the EzGenome database carry the taxonomy information from which the coding sequences are originated, as well as other functional information such as COG gene, COG category, product protein, and Gene Ontology accession number. In addition, EzGenome coding sequence database integrates subsystem (Aziz *et al.*,

2008; Overbeek *et al.*, 2005) information via blasting all the retrieved sequences against the Subsystem database. Between PfamA and PfamB, PfamA, which is curated, was incorporated into the annotation database. In the PfamA, the number of eukaryotic CDS sequence in the database is 6,216,335 including 1,318,191 fungal sequences.

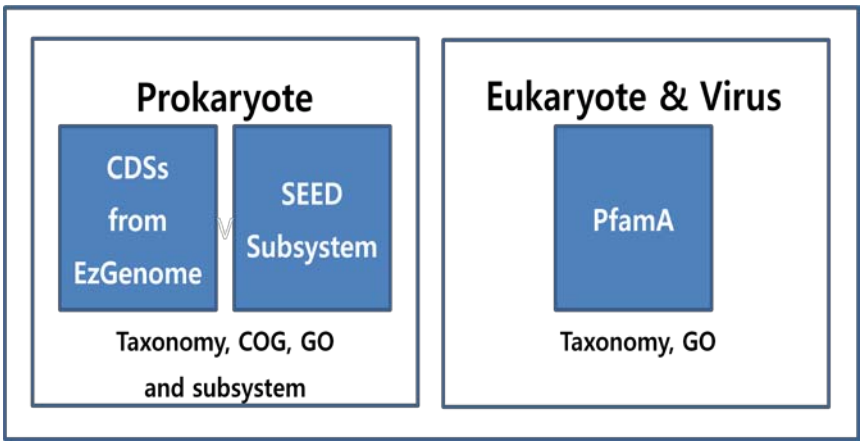


Figure 34. Metagenome functional annotation DB consistis of Ezgenome and Pfam.

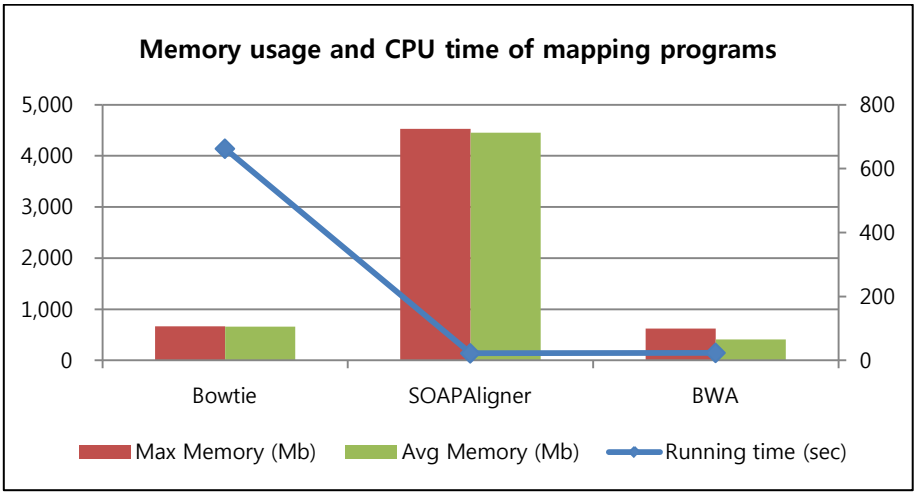
Table 18. Content of the annotation database for prokaryotes.

	Genus with more than 2 strains.		Genus with only one strain.		Total	
Status[(a)]	C	A	C	A	C	A
No. of Projects	567	400	828	1,369	1,395	1,769
Total	967		2,197		3,164	

[a] Assembly status: C for complete and A for assembly.

Overall shotgun metagenomics analysis pipeline

Final overall metagenome shotgun analysis pipeline is illustrated in Figure 36. The steps in red colored box are the steps for the creating longer contiguous sequences. The purpose of this analysis pipeline is to improve the accuracy by making use of longer sequences via the steps of assembly and mapping. Thus, the elapsed time and memory usage of each used programs are of importance. Figure 35 show the statistics of mapping (top) and de novo assembly (bottom) programs respectively.



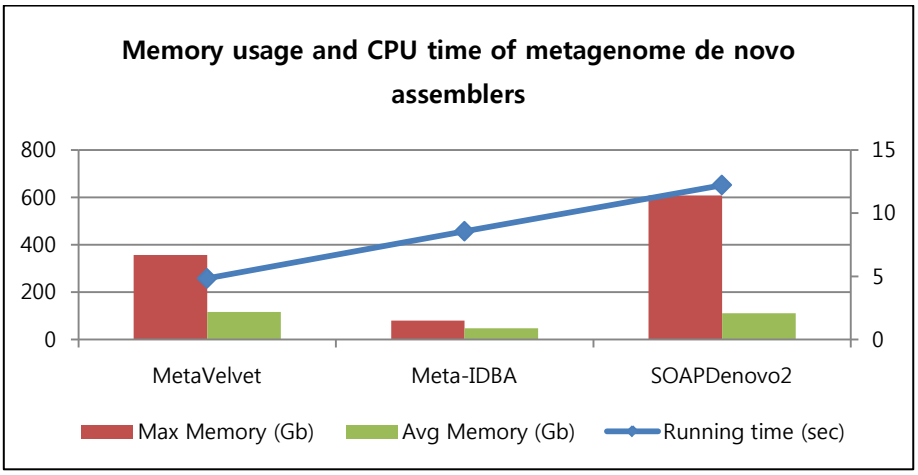


Figure 35. Computational performance of mapping and assembly programs.

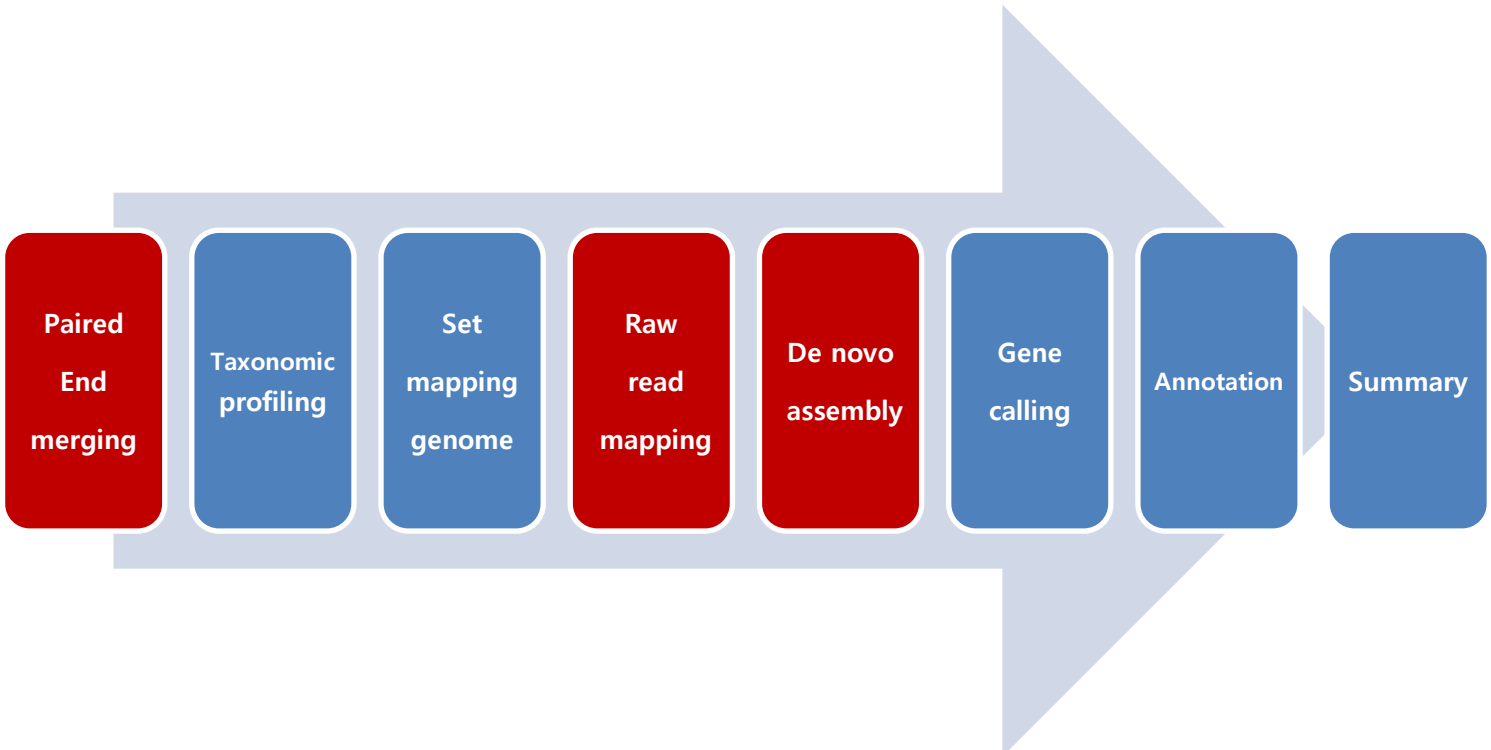


Figure 36. Schematic illustration of overall metagenome shotgun analysis pipeline.

Comparison with MG-RAST

MG-RAST is most widely used metagenome analysis platform (cited 611 times) while CAMERA and IMG/M were cited by 273 and 200 as of 2013 (http://scholar.google.co.kr/). Mock metagenome which is same as what is used in the pipeline development was analyzed through MG-RAST to compare the reported results. MG-RAST provides distribution of each taxonomic level measured with both ribosomal RNA and predicted proteins. The number of phylotypes obtained with MG-RAST was different with our pipeline at all taxonomic levels, with phylum level MG-RAST reporting 56 phylotypes while my pipeline did 46 phylotype; class level 122 versus 52, order level 274 versus 131, family level 528 versus 275 and genus level 1,096 versus 614 phylotypes (Fig. 37). Since only 5 genus comprise the mock metagenome, it can be said that Chunlab's pipeline reduced the biased information more successfully while looking with a perspective of taxonomic distribution estimation. In the meantime, taxonomic profiling only with the contig sequence reduced the bias far more than raw read data set indicating that the longer reads benefitted in reducing bias. Further, comparison of 10 most abundant inferred genus is done (Table 19). Except for *Pediococcus*, reference mock community (5 strains) were detected most correctly using Chunlab's pipeline while the false positive detections were composed of following genus; *Klebsiella*, *Rickettsia* or *Shigella* which were sometimes detected more abundantly than the mixed mock strains. This comparison showed that Chunlab's pipeline inferred the phylogenetic

structure more precisely than the MG-RAST.

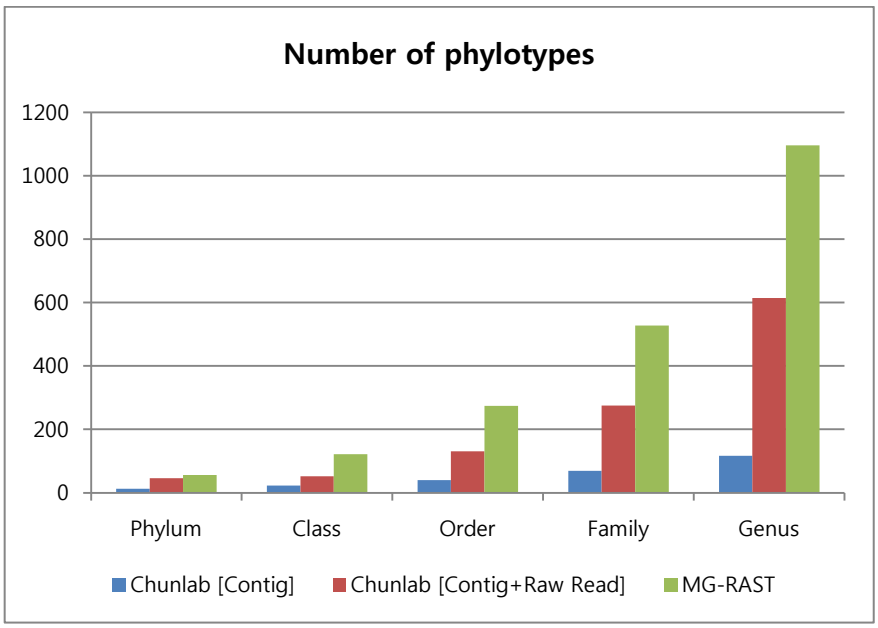


Figure 37. Comparison of the number of phylotypes at each taxonomic level between MG-RAST and Chunlab.

Table 19. 10 most abundant genus from Chunlab pipeline and MG-RAST.

Genus[(a)]	Count
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;	1,200,877
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter;	982,753
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;	396,340
Bacteria;Proteobacteria;Gammaproteobacteria;Vibrionales;Vibrionaceae;Vibrio;	321,287
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Klebsiella;	214,264
Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Pediococcus;	116,866
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Raoultella;	511,36
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Citrobacter;	44,625
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Salmonella;	32,131
Bacteria;Firmicutes;Bacilli;Bacillales;Bacillaceae;Halobacillus;	17,537

Genus[(b)]	Count
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;	1,818,158
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Klebsiella;	1,642,862
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;	877,335
Bacteria;Proteobacteria;Gammaproteobacteria;Vibrionales;Vibrionaceae;Vibrio;	651,321
Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Rickettsiaceae;Rickettsia;	550,492
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Shigella	354,700
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter	290,366
Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Pediococcus;	215,442
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Salmonella;	211,716
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Citrobacter;	177,073

[a] Genus detected by Chunlab pipeline, [b] Genus detected by MG-RAST

Besides the taxonomical structure inference, resultant functional profiling was compared. To compare the functional profile, all CDS from those known mixed strains were assigned COG and SEED categories. Both COG and SEED annotation profiles from Chunlab pipeline and MG-RAST were then plotted with the corresponding annotation profile of mock metagenome. Figure 38 shows the COG annotation profile. Except for the cell motility category which Chunlab pipeline could not detect, abundance of other COG categories were similar to the mock community's COG profile. However, in the SEED annotation profile comparison, except for the protein metabolism category, Chunlab pipeline profile was more similar to the mock community than MG-RAST. Carbohydrate, miscellaneous and clustering-based subsystem were the distinctively deviating categories as shown in Figure 39.

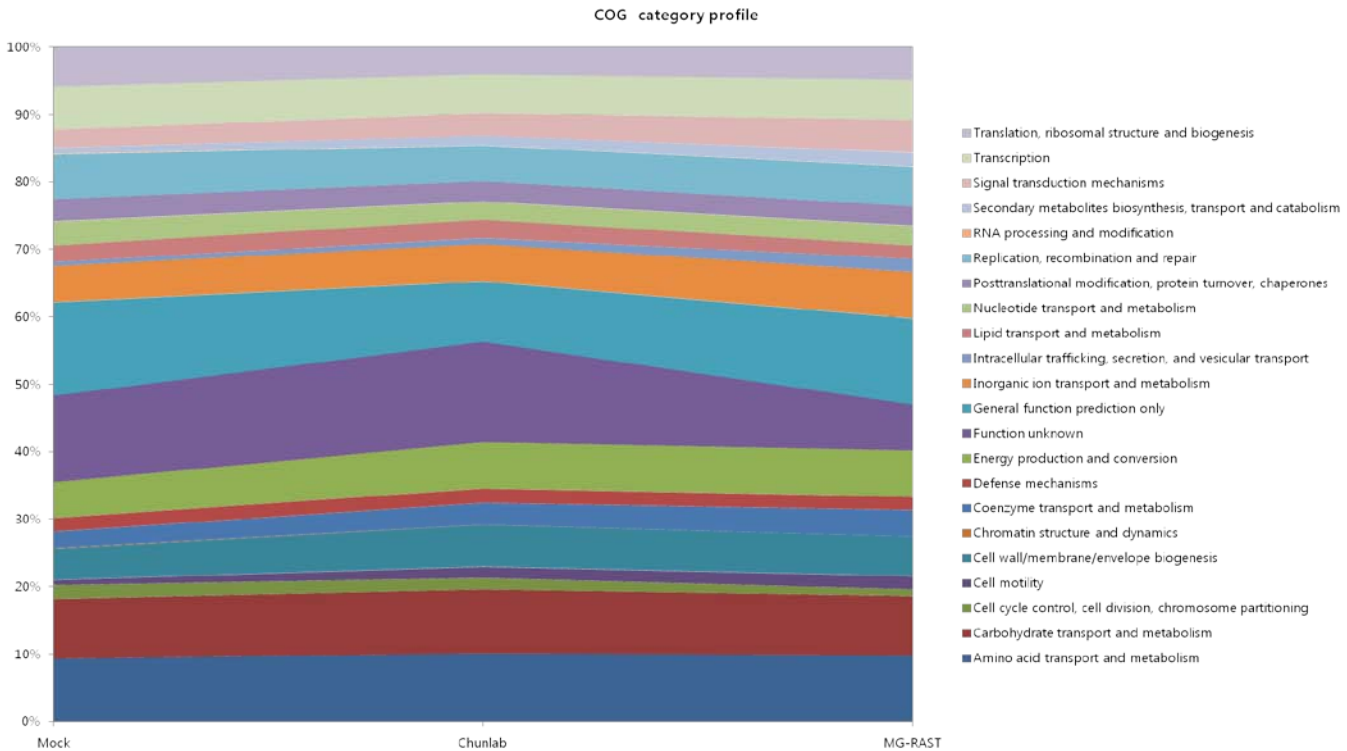


Figure 38. COG annotation profile of MG-RAST and Chunlab pipeline.

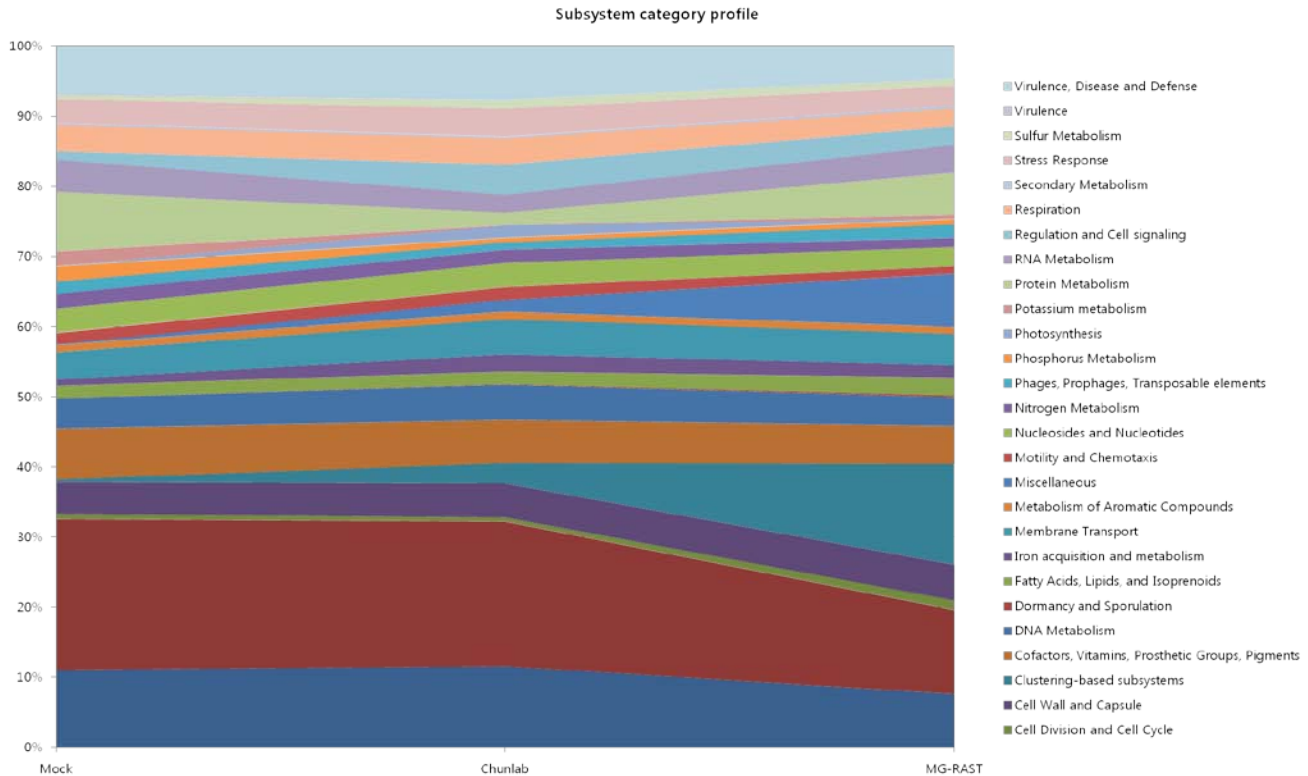


Figure 39. SEED annotation profile of MG-RAST and Chunlab pipeline.

Visualization and summarizing the analysis results

JAVAFX-based application furnishing the intuitive graphical user interface was developed. Taxonomic structure and related functional profiles including COG, SEED, RefSeq, and GO are graphically represented at the same time in relation with the corresponding taxonomies. Multiple genomes can be presented simultaneously enabling the multiple metagenomes comparative analysis. Upon loading the result of metagenome analysis, summary of the metagenomes is introduced through table. Name of each metagenome, the number of total reads, low quality reads, detected ribosomal RNAs, and predicted coding genes are listed in the table (Fig. 40).

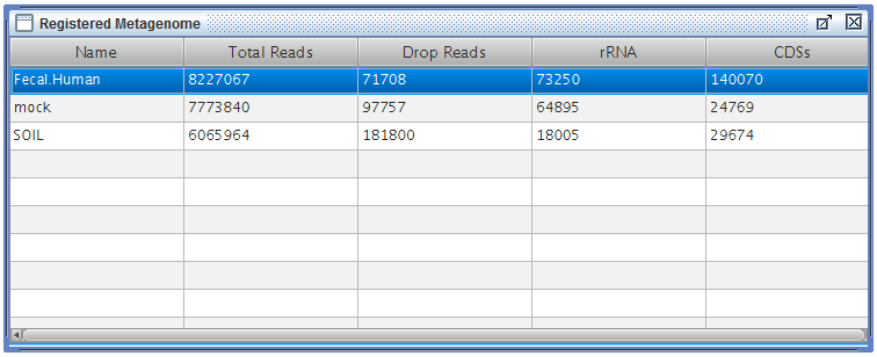


Figure 40. Visualization - Summary table of metagenome analysis.

Taxonomy hierarchy inferred through functional profile can be browsed through pie charts. To provide user interactive browsing, the chart shows the subsidiary composition of each slice in the pie chart by clicking on it.

For example, clicking the pie slice of “Bacteria” in the far left chart, phylum composition of “Bacteria” would be shown in the right of the chart, and in the phylum chart; again clicking Proteobacteria will have the subsidiary class of the Proteobacteria displayed on the right as shown in Figure 41. However, because of the lack of the hierarchical information of the eukaryotes and viruses provided by the Pfam database, this browsing interface is available only for the bacterial domain. For eukaryotes, browsing is available down only to class level.

In the metagenome analysis pipeline, functional annotation of prokaryotic genes was performed against the coding sequence protein database retrieved from EzGenome which incorporates SEED, COG, and GO information. Thus, SEED category and subcategory and COG annotation information indicating the selected phylotype are displayed. By doing so, this browsing application enables to provide users with the insight functional properties in relation to the corresponding taxonomical structure. During annotation process, all predicted functional coding sequences got assigned, not only taxonomic information but also SEED, COG, and GO information which are displayed as an interactive pie chart as shown in Figure 42. Whenever a slice for a phylotype in the taxonomy chart is selected, seed, cog and, go category compositions for that particular phylotype are also shown. Taxonomic hierarchy as: *Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholdeiraceae* is shown in the pipeline’s taxonomy window as is being depicted in Figure 42 as well as seed and cog charts corresponding each taxonomic rank are shown. In the seed and cog charts, subcategory of selected category of

selected phylotype browsing is also available. Figure 43 and Figure 44 are subcategory of seed ‘Carbohydrate’ category and cog ‘Cellular Process and Signaling’ category. In subcategory window, taxonomy composition indicated by the functional category is shown in the table next to the chart.

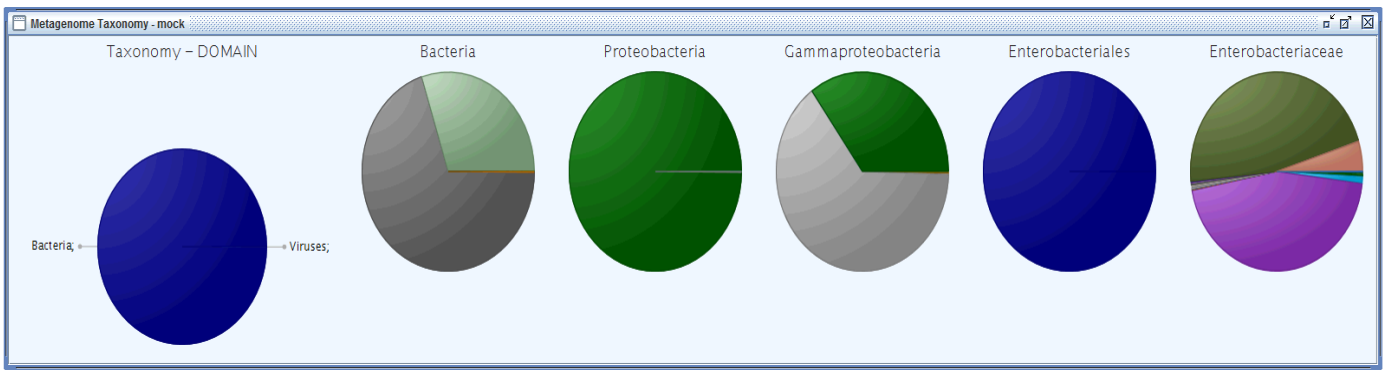


Figure 41. Visualization - Taxonomic hierarchy inferred from functional profile.

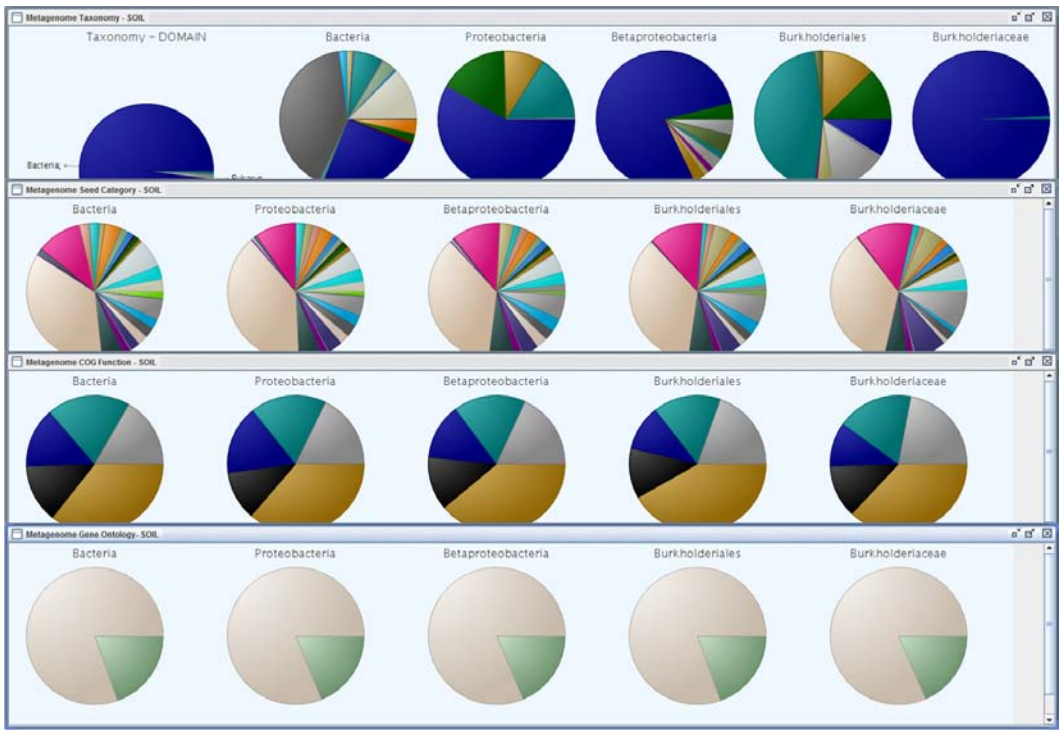


Figure 42. Visualization - Taxonomic composition information with the relative abundance of Seed category, COG category, and Gene Ontology.

For the comparative metagenome analysis, more than two metagenomes taxonomic hierarchies and composition of each hierarchical rank is displayed at the same time (Fig. 44). Selection of any phylotype in either metagenomic sample at any taxonomic rank will change subsidiary chart of both metagenome samples and hence the comparison is easily achieved at a glance. For example, the taxonomical compositions of "*Bacteria; Proteobacteria; Deltaproteobacteria; Desulfobacterales; Desulfobacteraceae*" of both metagenome samples is represented All the taxonomical information including eukaryote and virus inferred from functional annotation could be displayed and exported in TSV (Tab Separated Value) format file. NCBI product, COG's functional information and SEED subsystems information are exported (Fig. 45) in a table. Filtering and sorting is available in each table providing convenient interface to retrieve the frequency information of specific function and product of interest. Taxonomic composition of phylotype indicated by a product or a function can be presented graphically. For example, 'polymerase' was searched in the SEED subsystem table only to filter 'RNA Polymerase Bacterial' and the genus level phylotypes containing 'RNA Polymerase bactrerial' were plotted as represented in Figure 46.

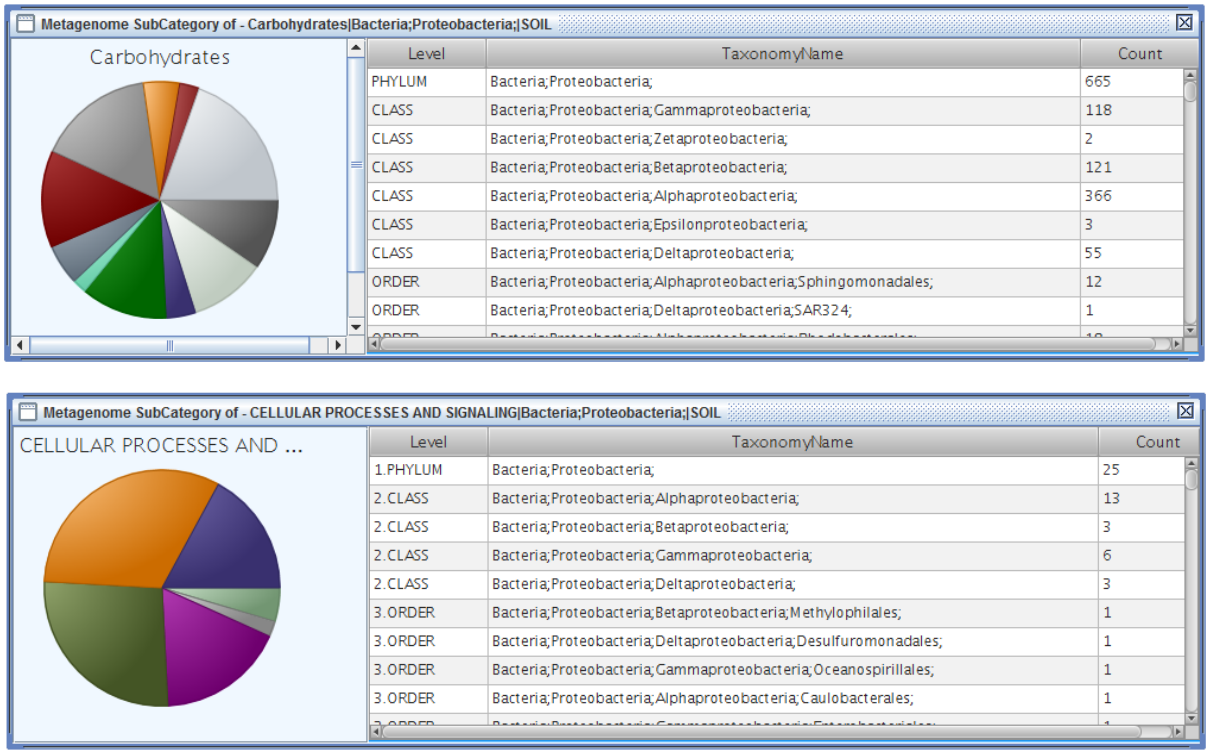


Figure 43. Visualization - Subcategory of selected SEED and COG category in *Proteobacteria*.

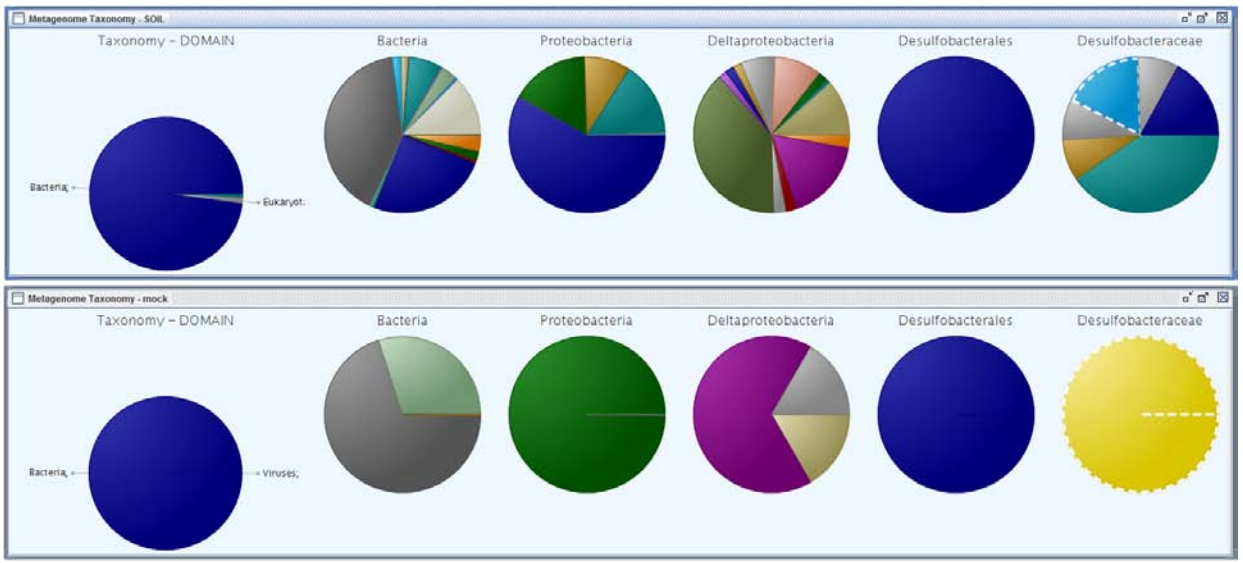


Figure 44. Visualization - Comparative analysis is available by loading more than two metagenomes at the same time.

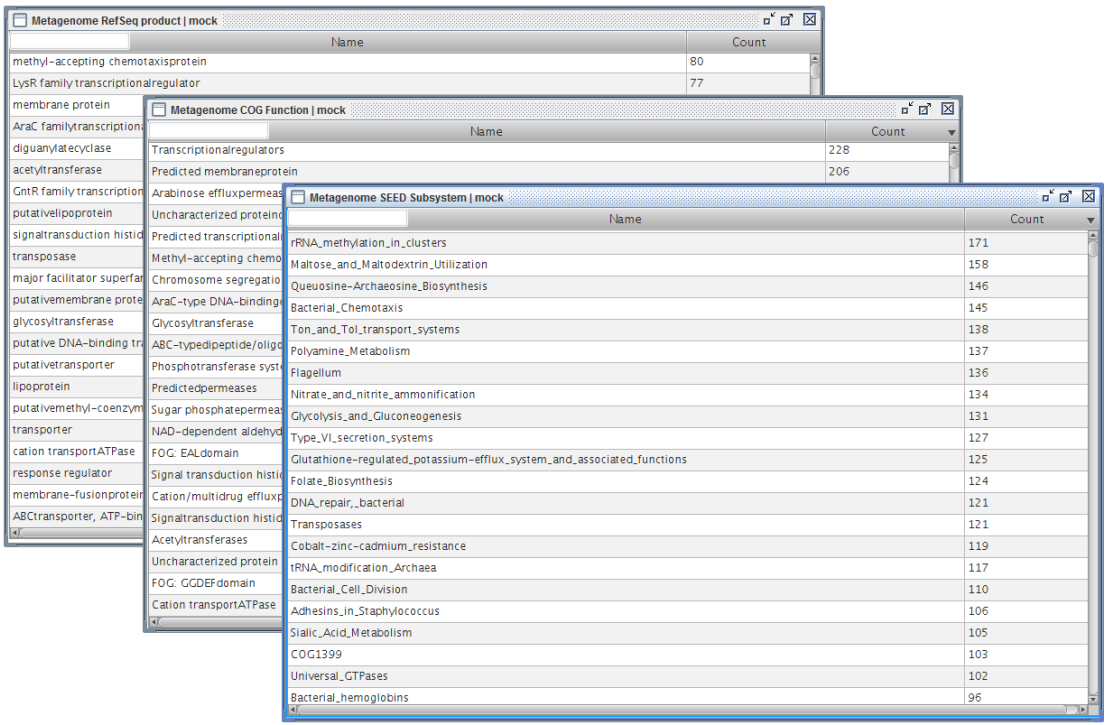


Figure 45. Visualization - Tables for annotated gene profiles refined from SEED subsystem, NCBI, and COG.

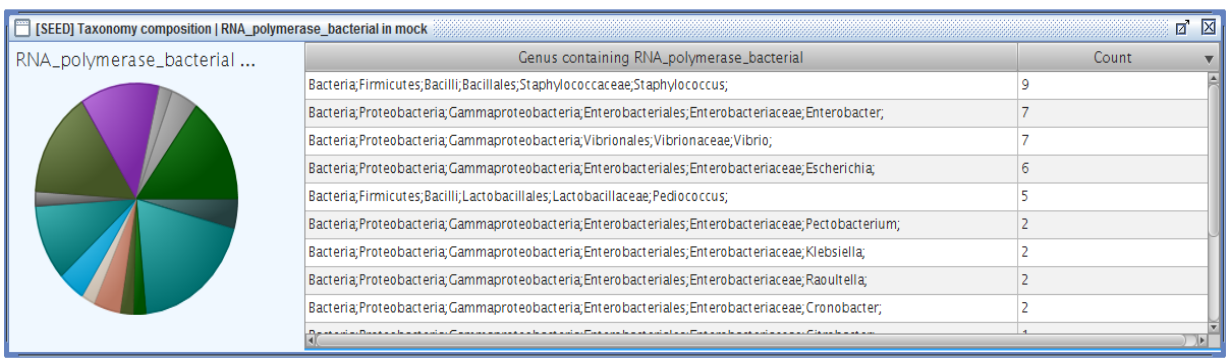


Figure 46. Visualization - Example of visualization of taxonomy composition of functional coding gene.

Application of Pipeline to an Experimental Sample

The pipeline was applied to the analysis of metagenome from an environmental soil and one human fecal sample and the results are summarized in Table 20.

Table 20 Summary of metagenome shotgun sequencing of a soil and fecal sample.

	Soil	Fecal
Total raw reads	6,065,964	8,227,067
Quality reads[(a)]	5,884,164	8,155,359
Forward nucleotides (bp)	1,376,179,304	1,638,015,900
Reverse nucleotides (bp)	1,388,776,581	1,642,306,245
Total nucleotides (bp)	2,764,955,885	3,280,322,145

[a] Average Quality > 25 and contains no ambiguous 'N' nucleotides.

Table 21. Estimation of sequencing coverage.

	Genome size	Coverage[(a)]	
		Soil	Fecal
10 Species	46,861,370	61X	70X
100 Species	468,613,700	6.1X	7X
1000 Species	4,686,137,000	0.61X	0.7X

[a] The coverage was estimated given nucleotide yield using the genome size of E. coli K12 DH10B strain, 4,686,137 bps.

Preprocessing is followed by paired end merging which resulted in 3,205,747 and 2,173,142 merged reads from soil and fecal samples respectively. With these merged reads and unmerged forward and reverse reads, taxonomic profiling was carried out against the HMM profile of ribosomal RNA sequences. Unlike the mock metagenome consisting of just 5 genome sequences billions of microbes are known to inhabit soil. Assuming that the 1,000 bacterial organisms are inhabit soil environments and the size of theirs genome is similar with Escherichia coli, at least 4,686,137,000 nucleotides should be obtained to capture the all the genomes in the environment at least one time, with a depth of 1 (Table 21). Thus the estimated sequencing depth was approximately less than 1X in both soil and fecal samples indicating that not all the genomic sequences were obtained. cf) coverage of sequencing of mock community is about 150X. 42 reference genomes took part in the mapping genome set to which

the raw reads were mapped in case of the soil while 75 genomes detected and comprised the mapping genome set. After mapping to reference genomes, 5,879,499 out of 5,884,164 reads remained unmapped indicating that very few reads from soil mapped to the selected reference genomes. In the interim, 1,792,535 reads accounting for about 21.9% were mapped to the reference genomes in fecal sample. It is because the sequence coverage was not enough to capture the metagenome contents as expected based on the estimated sequencing depth. Although, the sequencing coverage of the two samples was not very different under the assumption that there are 1,000 E. coli size genomes in the samples. However, in practice, it is known that the soil has more diverse microbial community than the fecal community. Therefore, it may be no use of mapping raw reads in the metagenome having complex microbial community unless sufficient sequencing coverage can't be obtained. With these unmapped read, de novo assembly was carried out and the result is summarized in Table 22. Like the result of the raw read mapping process, not only the number of assembled contigs but also all the other statistics indicate that the soil shotgun reads were poorly assembled.

Elapsed time and memory usage of both mapping and de novo assembly programs were measured (Fig. 47). IDBA took more time than the other two assembly programs in both soil and fecal samples. Running time of MetaVelvet was the sum of running time of velveth, velvetg and meta-velvetg. In mapping process, memory usage was higher in fecal sample than soil sample, because the size of the reference genome for fecal reads mapping is bigger than that for soil reads mapping.

Table 22. Result of reference mapping and de novo assembly of soil and fecal metagenome shotgun reads.

Statistics	Reference mapping		De novo assembly	
	Soil	Fecal	Soil	Fecal
No. of contigs	1,587	104,764	34,922	84,743
Max length (bp)	1,913	6,002	89,429	110,629
N50	293	411	1,420	1,291
Predicted CDSs	1,215	115,289	48,512	134,814

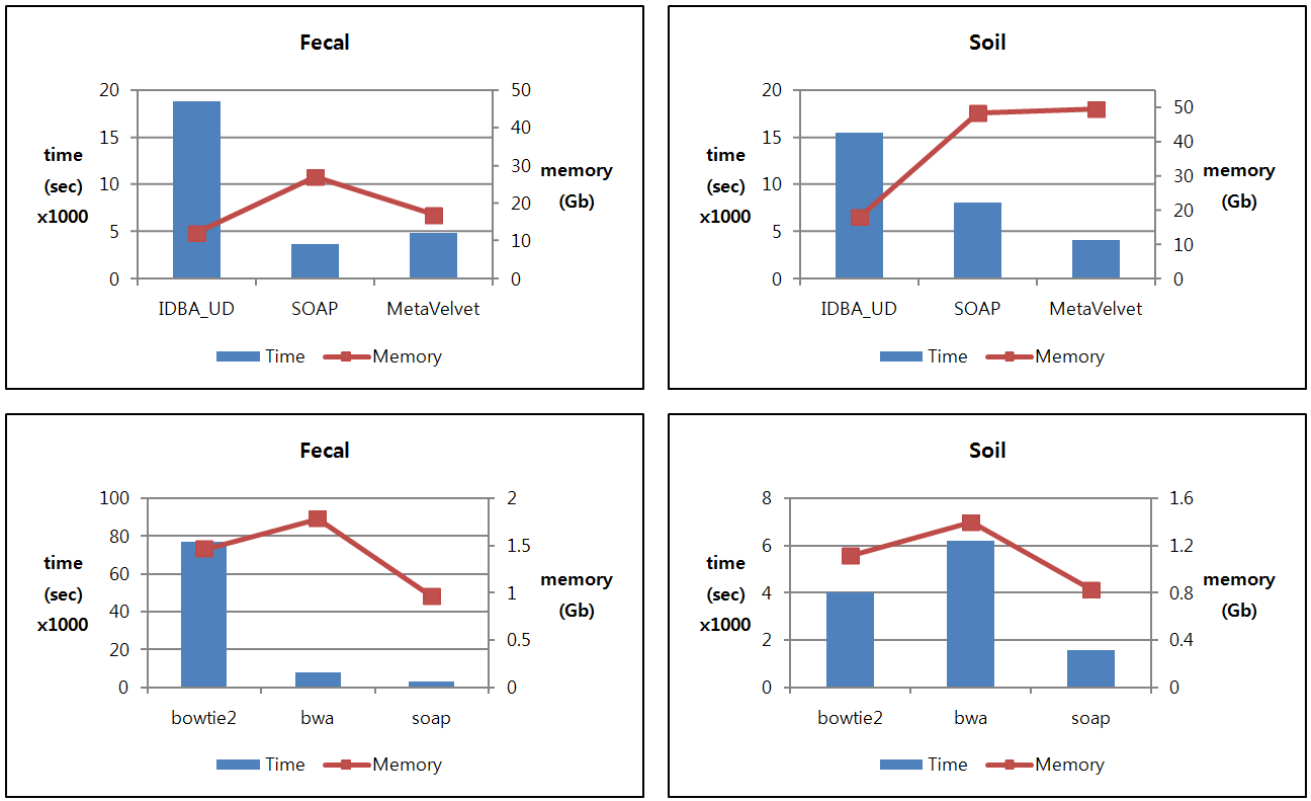


Figure 47. Time and memory usage of mapping and assembly programs for each sample.

3.4 Summary and Discussion

In this study, a bioinformatic pipeline for the random shotgun metagenome analysis was established for short paired end reads generated from Illumina MiSeq machine. In addition to the analysis pipeline, an application for the visualization of the metagenome was also developed to provide graphical representation of metagenome. Major obstacle in treating the metagenome shotgun reads was the length of the reads because short reads do not provide enough information to get assigned properly. To overcome the difficulty incurred by the short read length, the pipeline was dedicated to make longer contigs. Thus, three steps were devised and incorporated into the pipeline to make contiguous sequences. Firstly the paired end reads were merged by aligning the shared overlap region to generate a read whose length at least approached 400bp. Secondly, short reads were mapped to the reference genomes inferred from taxonomic profile. Lastly, the remaining unmapped reads got into the de novo assembly. With these longer contig sequences, gene prediction followed by functional annotation process was carried out. Since the purpose of metagenome analysis is not only to reveal the taxonomical and functional profile but also to correlate these two features, an annotation database was compiled containing all coding sequences in DB carrying taxonomic information as well as functional information. Even though coding sequences of EzGenome has a meticulously structured taxonomy for

bacterial genomes, Pfam's eukaryote and virus taxonomy are poorly structured so the taxonomy presentation of the eukaryote and virus was limited to a certain taxonomy level.

With the constructed pipeline, the mock metagenome consisting of 5 known strains was analyzed and compared to the result obtained from MG-RAST metagenome portal which have been cited the most so far. The comparison showed that the Chunlab pipeline performed better in terms of taxonomical and functional profiling accuracy. By means of the positive effect of the lengthened sequence, the bias in both taxonomical and functional profiling was thought to be reduced as compared to MG-RAST, but there still existed a bias. As for the visualization, considering the purpose of metagenome analysis, both taxonomic structure and functional profile including SEED, COG, and GO were presented in a form of intuitive and interactive graphical chart. Also, taxonomic composition indicated by each functional category of SEED and COG were presented in relation to the functions and taxonomies.

Environmental soil and fecal metagenome samples were analyzed using the pipeline. Unlike the known mock metagenome consisting of only 5 strains, soil and fecal have complex microbial community hence the estimated sequencing coverage was not enough to capture the whole content of all bacterial genomes in metagenome. Therefore the reference genome database could not be set up with a sufficient number of genomes and so very little number of reads were mapped to the reference genomes resulting in fewer number of contigs. Thus a relatively large number of reads were taken for the de novo assembly which requires larger

computational resources and has higher complexity leading to the chimeric contigs. Between soil and feces, the number of uncultured strains in soil was higher than that of feces, and these uncultured genome sequences were not in the genome database. This lack of reference genome of soil samples resulted in poor raw read mapping and de novo assembly result. However, rapid growth of genome sequence database with the development of sequencing technology is expected to cover the uncultured genome soon.

Chapter 4 EzEditor: A versatile Molecular Sequence Editor for Both Ribosomal RNA and Protein Coding Genes

4.1 Overview

General procedure for molecular phylogenetic analysis consists of two major steps, namely multiple sequence alignment and phylogenetic treeing. The former is regarded important as it can affect the accuracy of all downstream analyses. Many computer programs are available for multiple alignment of DNA or protein sequences with CLUSTAL-series programs being most popular (Thompson *et al.*, 2002). However, computer-generated sequence alignment is often required to be improved by considering biological knowledge such as secondary structure of RNA and reading frame of protein coding genes. This task can be fulfilled using molecular sequence editing softwares called sequence alignment editors by which computer-generated alignment can be viewed and adjusted manually by adding or deleting gaps. Several sequence alignment editors are available for the general usage, including SEAVIEW (Galtier *et al.*, 1996), BioEdit (Hall, 1999), DNAAlignEditor (Sanchez-Villeda *et al.*, 2008), INTERALIGN (Pible *et al.*, 2005), and JalView (Waterhouse *et al.*, 2009). In case of ribosomal RNA genes which are most widely used phylogenetic markers, special sequence editing programs are developed, including ARB (Ludwig *et al.*, 2004) and jPhydit (Jeon *et al.*, 2005). These software tools allow users to consider secondary structure information while manually editing the sequences. Conserved protein-coding genes, such as *rpo*B, *rec*A and *gyr*B, are also being widely used in molecular phylogenetic studies

(Case *et al.*, 2007; Feng *et al.*, 1997). Unlike rRNA genes, protein-coding genes can be utilized either as DNA or translated protein sequences in phylogenetic analysis. Because of the degenerate nature of the genetic code, single amino acid can be encoded by multiple codons, and DNA sequence in coding region contains position-specific information as a component of codon, depending on reading frame. Therefore, in case of protein sequence, the original DNA sequences coding for the proteins are better aligned by codon-based alignment (Goldman *et al.*, 1994) in which the protein sequences are first aligned and their DNA sequences are then rearranged by inserting gaps, following the previously aligned protein sequences. In this scheme, one gap in protein sequence alignment is translated into three consecutive gaps in DNA sequence alignment. There are few software tools and web services (Bininda-Emonds, 2005; Suyama *et al.*, 2006; Wernersson *et al.*, 2003) to achieve this task. However, to our knowledge, there is no sequence editor that allows codon-based DNA alignment and manual editing at the same time. In this study, new sequence editor, named EzEditor, is introduced for simultaneous codon-based editing of protein and DNA sequence alignment. Since it is the descendent of jPhydit program (Jeon *et al.*, 2005), it provides all functionality for editing rRNA alignment using secondary structure information.

4.2 Features of EzEditor

Graphical User Interface

Graphical User Interface (GUI) is one of the most important features of EzEditor. While majority of the biological software is command line interface, sequence editor programs including EzEditor must be GUI for the manual alignment editing. As EzEditor is a JAVA-based application, the GUI of EzEditor is based on JAVA swing library. Figure 48 shows the main work space of EzEditor.

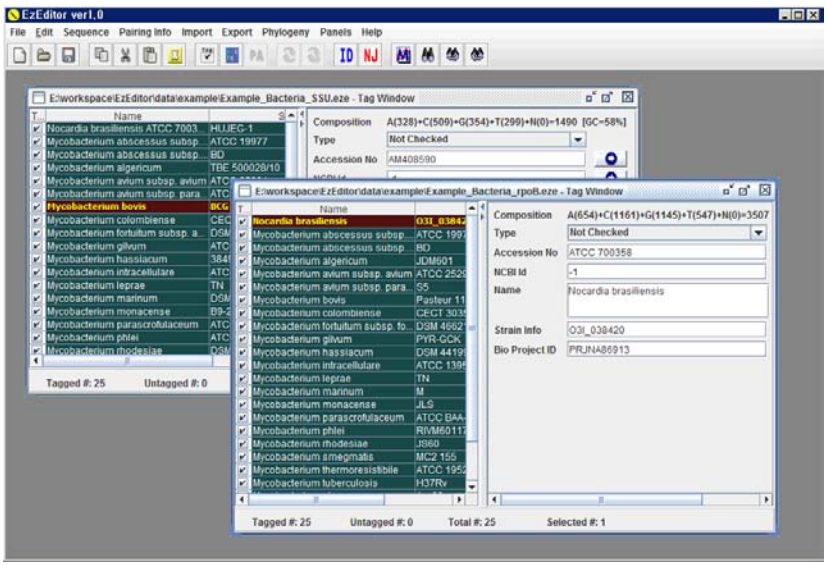


Figure 48. Main work space of EzEditor. The work space can be loaded with multiple data files at once.

Many data files can be opened at the same time for comparison of the different gene trees of the same organisms. SQLite (http://www.sqlite.org/), small scale database file appropriate for applications, format file named as EZE is the data file of EzEditor. Data fields in the data file is listed in Table 23. Accession number is the unique key of the sequence in the table and random number will be assigned to the sequence with no accession number.

Table 23. Data field of EZE file and example data.

Data Field	Example
Name	Mycobacterium bovis
Accession	AM408590
Taxonomy	*Bacteria;Actinobacteria;Actinobacteria_c;Corynebacteriales;Mycobacteriaceae;Mycobacterium;Mycobacterium bovis*
Is type	Yes/No
Sequence	AGAGTTTGATCCTGGCTCA--GACGAACGCT-G-..
BioProjectID	PRJEA18059
GenbankID	121491530
Strain	BCG Pasteur 1173P2

Select Panel and Align Panel are the main panels displaying data in an EZE file. Select Panel (Fig. 49) is the meta data manager panel. All sequences are listed in the left panel and the meta information of the selected sequence appears in the right panel. Summary of the data file is in the bottom of the Select Panel.

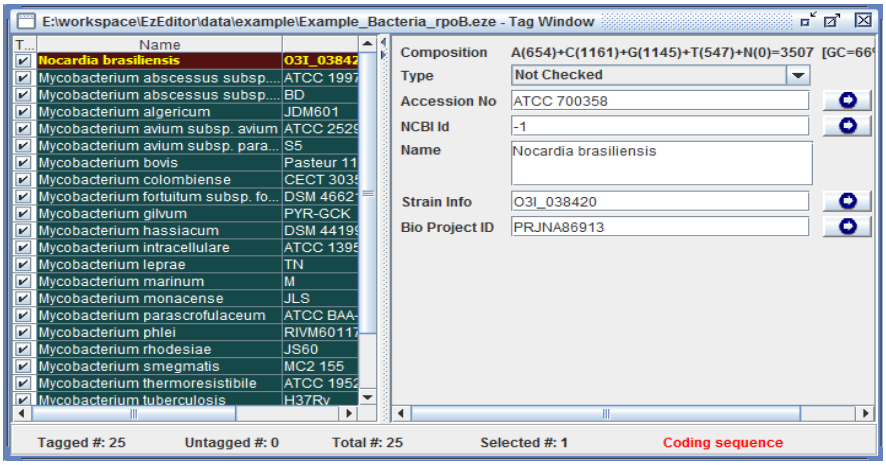


Figure 49. All information except DNA sequences is shown in SelectPanel. Meta data of selected sequence in the left panel is shown in the right panel.

There are four different sequence data types in EzEditor. 1) Functional coding sequence, 2) other non-coding gene sequence, 3) bacterial 16S rRNA sequence and 4) archaeal 16S rRNA sequence. These data types can be categorized into two data types i.e., RNA and non-RNA sequence. EzEditor recognizes different sequence data types to show different AlignPanel. AlignPanel of ribosomal RNA sequence composed of

DNA sequence alignment panel and secondary pairing information panel is at the top of the DNA sequence alignment. Figure 50 is the AlignPanel of bacterial 16S rRNA sequence. Secondary pairing information of current selected nucleotide is shown in the panel at the bottom. In Figure 50, the selected strain is Mycobacterium bovis and the cursor is located at the 6th column of 'T' in the black square box. The secondary pairing nucleotide of the selected 'T' is 'G' in yellow square box which is located at the 14th columns shown in light gray box. The reference position of secondary structure pairing is Escherichia coli in case of bacteria and Methanocaldococcus jannaschii in case of archaea.

Figure 50. Align Panel of 16S ribosomal RNA. Secondary structure pairing information of selected sequence in DNA alignment panel is shown.

The AlignPanel (Fig. 51) of functional coding sequence is different

from that of rRNA's in that the translated protein sequence is shown instead of secondary structure pairing information. Whenever manual DNA sequence editing is done, insertion or deletion occurs and the protein sequence reflects the changes in DNA sequence and the changed translated protein sequence is shown instantly so that users can take advantage of the translated sequence for the accurate sequence alignment. Protein sequence is under different selective pressure than the DNA sequence- information of DNA substitutions, synonymous and non-synonymous, could be used for DNA sequence alignment of functional coding gene (Yang *et al.*, 2000).

Figure 51. Align Panel of functional coding sequence. Protein sequence alignment is shown in the panel below DNA alignment panel.

4.2.1 Algorithms and Models Implemented in EzEditor

Sequence Alignment

The ultimate goal of molecular phylogenetics is to infer a reliable phylogenetic tree. Practically, *in silico* analysis of molecular phylogenetics is conducted in following sequential manner. 1. Sequence alignment, 2. manual editing of sequence alignment, 3. phylogenetic tree inference, and 4. evaluation of inferred tree. Because the phylogenetic tree is inferred from sequence alignment, the tree's accuracy is dependent upon the accuracy of the sequence alignment. Both pairwise and multiple sequence alignment functions are contained in EzEditor. The semi-automated pairwise alignment is to align an unaligned sequence to the existing multiple alignment. 'Optimal Linear Space' algorithm (Myers *et al.*, 1988), one of the Dynamic Programming, was applied to the semi-automated pairwise alignment. ClustalW2 was integrated into the EzEditor as a binary form.

Phylogenetic tree inference.

There are two different types of phylogenetic tree inference approaches namely., Character-based approach and Distance-based approach. Character-based approach uses the sequence alignment directly while distance-based approach requires transformation of sequence alignment into distance matrix using evolutionary models. Character based approach includes Maximum parsimony (Fitch, 1971) and Maximum likelihood (Guindon *et al.*, 2003) algorithms and distance based approaches

include Minimum Evolution (Rzhetsky *et al.*, 1992), Least Square (Felsenstein, 1997), UPGMA (Yap *et al.*, 1996) and Neighbor Joining method (Saitou *et al.*, 1987). In EzEditor, Neighbor Joining and UPGMA were implemented.

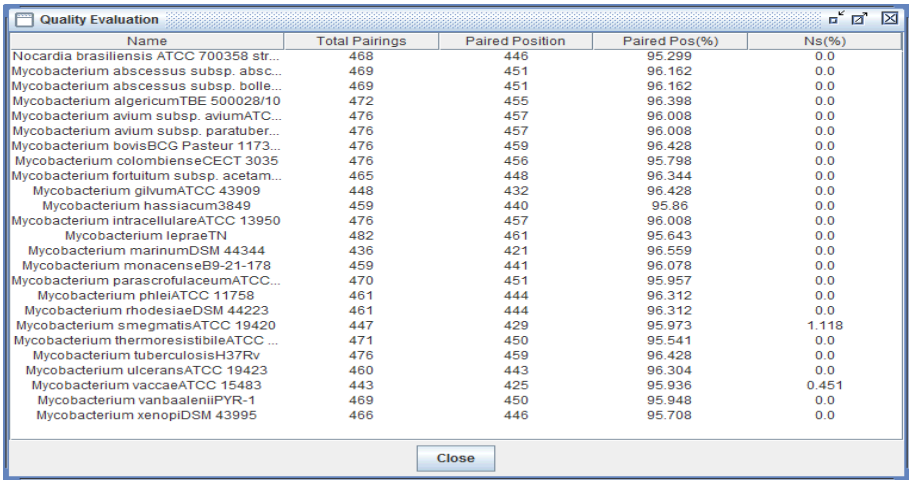
Evolutionary model

For these two distance based approaches (NJ and UPGMA) in phylogenetic tree inference, mathematical evolutionary model is required to transform given sequence alignment into distance matrices. Widely used two types of model for distance matrix, Juke and Cantor's model and Kimura 2 Parameter model were implemented in EzEditor.

4.2.2 Miscellaneous Functions

Sequence alignment and phylogenetic tree inference are the core of molecular phylogenetics. However, some miscellaneous functions are also required for the efficiency and accuracy of the analysis. EzEditor has a function for alignment quality evaluation which is based on the secondary structure pairing information. As studies using rRNA sequence incorporate secondary structure pairing (Cole *et al.*, 2009; Schloss, 2009a) data to get robust sequence alignment, multiple alignment must be followed by manual alignment editing to improve sequence alignment accuracy in many cases. Secondary structure pairing information plays the important role as a guide for manual editing. Figure 52 shows the result of alignment quality

evaluation.



Name	Total Pairings	Paired Position	Paired Pos(%)	Ns(%)
Nocardia brasiliensis ATCC 700358 str...	468	446	95.299	0.0
Mycobacterium abscessus subsp. absc...	469	451	96.162	0.0
Mycobacterium abscessus subsp. bolle...	469	451	96.162	0.0
Mycobacterium algericumTBE 500028/10	472	455	96.398	0.0
Mycobacterium avium subsp. aviumATC...	476	457	96.008	0.0
Mycobacterium avium subsp. paratuber...	476	457	96.008	0.0
Mycobacterium bovisBCG Pasteur 1173...	476	459	96.428	0.0
Mycobacterium colombienseCECT 3035	476	456	95.798	0.0
Mycobacterium fortuitum subsp. acetam...	465	448	96.344	0.0
Mycobacterium gilvumATCC 43909	448	432	96.428	0.0
Mycobacterium hassiacum3849	459	440	95.86	0.0
Mycobacterium intracellulareATCC 13950	476	457	96.008	0.0
Mycobacterium lepraeTN	482	461	95.643	0.0
Mycobacterium marinumDSM 44344	436	421	96.559	0.0
Mycobacterium monacenseB9-21-178	459	441	96.078	0.0
Mycobacterium parascrofulaceumATCC...	470	451	95.957	0.0
Mycobacterium phleiATCC 11758	461	444	96.312	0.0
Mycobacterium rhodesiaeDSM 44223	461	444	96.312	0.0
Mycobacterium smegmatisATCC 19420	447	429	95.973	1.118
Mycobacterium thermoresistibileATCC ...	471	450	95.541	0.0
Mycobacterium tuberculosisH37Rv	476	459	96.428	0.0
Mycobacterium ulceransATCC 19423	460	443	96.304	0.0
Mycobacterium vaccaeATCC 15483	443	425	95.936	0.451
Mycobacterium vanbaaleniiPYR-1	469	450	95.948	0.0
Mycobacterium xenopiDSM 43995	466	446	95.708	0.0

Figure 52. Secondary structure pairing information can be used to assess the robustness of the sequence alignment.

Another useful functional module of EzEditor is sequence similarity statistics. When a new unaligned sequence is imported, provided that the unaligned sequence is aligned to the existing multiple alignment, EzEditor can identify the closest sequence in the dataset without inferring a tree. Figure 53 is the similarity table of a selected sequence to all the other sequences.

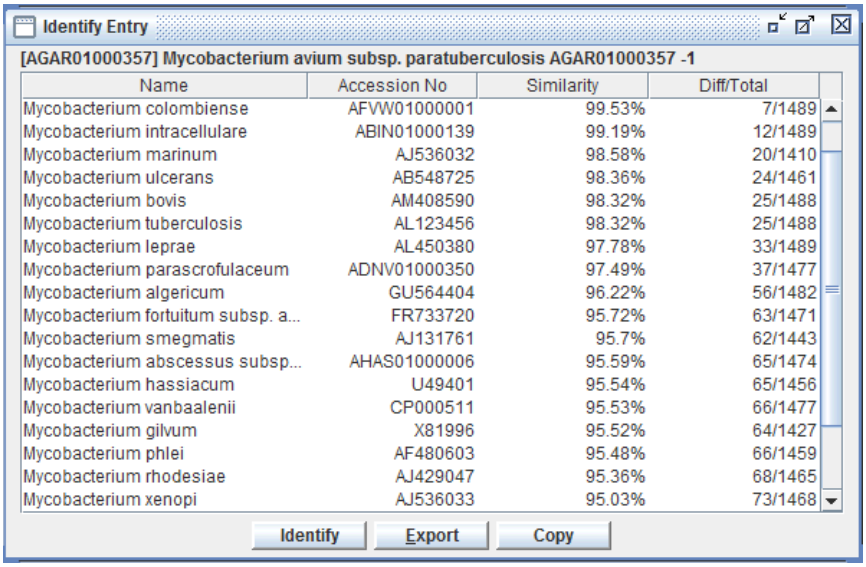


Figure 53. Sequence similarity to all the other sequences in the dataset is shown.

Other supplementary functional modules such as importing/exporting sequences and batch alignment editing etc., were built in EzEditor to provide a convenient tools for molecular phylogenetics.

4.3 Summary and Discussion

EzEditor provides researchers with powerful and efficient environment for molecular phylogenetics. From sequence alignment, sequence alignment editing to phylogenetic tree inference, every required *in silico* analysis could be conducted in this integrated application. In addition, to the secondary structure pairing information of ribosomal rRNA sequence, EzEditor featured translated protein sequence information of functional coding sequence to make use of the protein sequence for the robust DNA sequence alignment. With this protein sequence module, other genetic marker sequences can be analyzed in EzEditor. As genome scale large datasets are becoming available with the advent of Next Generation Sequencing technology, many approaches have been addressed to reveal evolutionary relationship among organisms with this genome scale sequence data (Chan *et al.*, 2013). However, this genome scale phylogenetics, called phylogenomics, through typical method still remains infeasible due to many factors including incomplete genome sequence, error-prone NGS reads, and limited computational resources to deal with multiple genome sequences. Moreover, genomic sequence properties such as genome rearrangement, gene-fusion and deletion, lateral genetic transfer, and transcript variation, make phylogenomics more complicated. Rather, drawback of lack of sequence resolution with only single gene could be overcome by using multiple genetic markers or by

creating a shared gene tree. Thus using more than a single genetic marker could provide more robust phylogeny, and the functional coding sequence alignment editing of EzEditor could come up as a useful function.

Not only as a phylogenetic analysis environment, EzEditor can also be used as a basic sequence alignment editing and management tool. As NGS has replaced the traditional Sanger sequencing, a large quantity of the sequence are being treated in batch mode. Therefore, direct sequence examination may be impossible or not required anymore. However, sequence editor still plays important part of various analyses where molecular sequence is the main framework of the analysis. Therefore, EzEditor is a useful application for phylogenetics as well as basic sequence analysis tool for various biological projects.

Conclusions

Development of sequencing technology provides the biological disciplines with the opportunity to scale up the research and these scaled up researches again promote the development of the bioinformatics. Of course the application of bioinformatics is not limited only to the NGS related biological research, however recent development of bioinformatics seemed to be spurred by the development of the sequencing technology. And this trend would last for the time being because NGS technology continues to step toward the next generation.

Metagenomics, one of the beneficiary research field of the improved sequencing technology, which is becoming the indispensable choice for understanding our environment and human health as they would not have gathered the limelight without the aid of the bioinformatics. In this study, bioinformatics played a crucial role in analyzing the NGS data obtained from metagenome. For the PCR based amplicon target sequencing for the bacterial community analysis, two different types of analysis systems corresponding to the 454 pyrosequencing and Illumina MiSeq platforms were constructed. Both two sequencing platforms have different type of errors which could lead to a biased result. Thus the pipelines are focusing on detecting, correcting and/or removing the errors. For the 454 pyrosequencing pipeline, CDenoiser, homopolymeric error handling program was developed and showed better performance than other programs. Suitable program and database for chimera handling step were

selected through evaluation by means of known mock sequence so as to minimize the effect of sequencing errors. In Illumina MiSeq pipeline, the errors in the 3' end could be corrected by merging the paired end. Further, errors in other regions could be corrected using iterative consensus clustering approach. 99.5% of resultant consensus reads showed over 95% similarity to their template sequence indicating that Illumina MiSeq amplicon analysis using this pipeline could generate more accurate result. Further, swine fecal sample was analyzed using both platforms and the result showed that the more diverse bacterial community was recovered from the Illumina MiSeq paired end data indicating that this sequencing platform is a potential alternative platform to 454 pyrosequencing for amplicon based metagenomics.

Random shotgun metagenome analysis pipeline applicable for the Illumina MiSeq paired end data was also developed. Because the short read length is the major hindrance causing bias (Wommack *et al.*, 2008), development of the pipeline focused on creating longer contig sequences either by raw read mapping to the reference genomes or by de novo assembly of short reads. The feasibility of the pipeline was evaluated with known mock metagenome by comparing the result with the MG-RAST results. The pipeline captured the taxonomic and functional profile more closely to those of mock metagenome than the MG-RAST result. When the pipeline was applied to the environmental soil and fecal samples, results showed that the mapping strategy could not successfully make the longer contigs in the soil sample while in the fecal sample, as many as 21% of the raw reads were mapped to the reference genome sequences creating longer

contig sequences. During the *de novo* assembly step, the longer contigs were created in the fecal sample analysis while smaller contigs were created in the soil sample.

The observed low performance of the raw read mapping step in soil sample is partly due to both, the lack of sequencing coverage and relatively high ratio of uncultured microorganisms in soil sample (Tyson, 2008). The low sequencing coverage implies that the ribosomal RNA operons of microbes in the metagenome were not sequenced. In addition, because the ribosomal RNA sequence profile of Hidden Markov Model was inferred from the multiple alignment of rRNA sequence of the existing culturable microorganisms, the rRNA sequence of uncultured bacteria could not participate in the multiple alignment. Therefore, the uncultured bacteria were not detected by rRNA profile of HMM even when the rRNA sequences of the uncultured bacteria were captured by random shotgun sequencing.

De novo assembly of soil samples could be directly affected by the low sequencing coverage. Since the algorithm of the assemblers use adjacent sequences, the missing flanking sequences of Kmer substring in de novo assembly resulted in many short broken contigs or unassembled raw reads. Further, the mapping process failed to reduce the number of reads and the complexity of the reads remained so high that the assembled contigs were more likely to be of chimeric origin.

As a consequence, even though the mapping and de novo assembly strategies worked relatively well for the metagenomes having low microbial diversity, metagenome with highly complex bacterial community

could not take advantage of the mapping and de novo assembly without the sufficient sequencing coverage or mapping genome sequences. However NGS technologies are currently under development toward larger sequencing output so, those uncultured bacterial genomes are to be sequenced in the near future. Hence the mapping and de novo strategies are expected to be useful for the complex metagenomes provided that the sufficient sequencing coverage and assembled mapping genome data are available.

Ezeditor, new sequence alignment editor, provides the codon based alignment in addition to the secondary structure based ribosomal RNA alignment. As a phylogenetic analysis tool, phylogeny of the functional coding sequences obtained from the metagenome analysis can be inferred by using the function in EzEditor.

References

Acinas, S. G., R. Sarma-Rupavtarm, V. Klepac-Ceraj and M. F. Polz (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied And Environmental Microbiology* **71**, 8966-8969.

Angiuoli, S. V., M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White & other authors (2011). CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC bioinformatics* **12**, 356.

Ansorge, W., B. S. Sproat, J. Stegemann and C. Schwager (1986). A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods* **13**, 315-323.

Ansorge, W., B. Sproat, J. Stegemann, C. Schwager and M. Zenke (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic acids research* **15**, 4593-4602.

Arthur Brady and S. L. Salzberg (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*.

Arumugam, M., E. D. Harrington, K. U. Foerstner, J. Raes and P. Bork

(2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26**, 2977-2978.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight & other authors (2000). Gene Ontology: tool for the unification of biology. *Nature genetics* **25**, 25-29.

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones and A. J. Weightman (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied And Environmental Microbiology* **71**, 7724-7736.

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones and A. J. Weightman (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Applied And Environmental Microbiology* **72**, 5734-5741.

Aydogdu, H., A. Asan and M. T. Otkun (2010). Indoor and outdoor airborne bacteria in child day-care centers in Edirne City (Turkey), seasonal distribution and influence of meteorological factors. *Environmental Monitoring and Assessment* **164**, 53-66.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass & other authors (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Barker, J. and S. F. Bloomfield (2000). Survival of *Salmonella* in bathrooms and toilets in domestic homes following salmonellosis.

Journal of Applied Microbiology **89**, 137-144.

Bartram, A. K., M. D. Lynch, J. C. Stearns, G. Moreno-Hagelsieb and J. D. Neufeld (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl Environ Microbiol* **77**, 3846-3852.

Bateman, A. and J. Quackenbush (2009). Bioinformatics for next generation sequencing. *Bioinformatics* **25**, 429.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall & other authors (2002). The Pfam protein families database. *Nucleic Acids Research* **30**, 276-280.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon & other authors (2004). The Pfam protein families database. *Nucleic Acids Research* **32**, D138-D141.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16**, 545-552.

Berg, R. D. (1996). The indigenous gastrointestinal microflora. *Trends in microbiology* **4**, 430-435.

Bininda-Emonds, O. R. (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC bioinformatics* **6**, 156.

Binnewies, T. T., Y. Motro, P. F. Hallin, O. Lund, D. Dunn, T. La, D. J.

Hampson, M. Bellgard, T. M. Wassenaar & other authors (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & integrative genomics* **6**, 165-185.

Bohnebeck, U., T. Lombardot, R. Kottmann and F. O. Glöckner (2008). MetaMine–A tool to detect and analyse gene patterns in their environmental context. *BMC bioinformatics* **9**, 459.

Bowers, J., J. Mitchell, E. Beer, P. R. Buzby, M. Causey, J. W. Efcavitch, M. Jarosz, E. Krzymanska-Olejnik, L. Kung & other authors (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593-595.

Bragg, L., G. Stone, M. Imelfort, P. Hugenholtz and G. W. Tyson (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nature methods* **9**, 425-426.

Buchholz, U., H. Bernard, D. Werber, M. M. Bohmer, C. Remschmidt, H. Wilking, Y. Delere, M. an der Heiden, C. Adlhoch & other authors (2011). German Outbreak of *Escherichia coli* O104:H4 Associated with Sprouts. *New England Journal of Medicine* **365**, 1763-1770.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer and R. Knight (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108**, 4516-4522.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D.

Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich & other authors (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336.

Carpentier, B., E. Lagendijk, D. Chassaing, P. Rosset, E. Morelli and V. Noel (2012). Factors impacting microbial load of food refrigeration equipment. *Food Control* **25**, 254-259.

Case, R. J., Y. Boucher, I. Dahllöf, C. Holmström, W. F. Doolittle and S. Kjelleberg (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied And Environmental Microbiology* **73**, 278-288.

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

Chan, C. X. and M. A. Ragan (2013). Next-generation phylogenomics. *Biology direct* **8**, 1-6.

Chatterji, S., I. Yamazaki, Z. Bai and J. A. Eisen (2008). CompostBin : A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. *Recomb*.

Check, H. E. (2009). Genome sequencing: the third generation. *Nature* **457**, 768.

Claesson, M. J., Q. O. Wang, O. O'Sullivan, R. Greene-Diniz, J. R. Cole, R. P. Ross and P. W. O'Toole (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA

gene regions. *Nucleic Acids Research* **38**.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. Kulam-Syed-Mohideen, D. McGarrell, T. Marsh & other authors (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37**, D141-D145.

Collins, F., E. Lander, J. Rogers, R. Waterston and I. Conso (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

Colwell, R. K. (1997). *EstimateS*. Robert K. Colwell.

Degnan, P. H. and H. Ochman (2012). Illumina-based analysis of microbial community diversity. *ISME J* **6**, 183-194.

Desai, N., D. Antonopoulos, J. A. Gilbert, E. M. Glass and F. Meye (2012). From genomics to metagenomics. *Curr Opin Biotechnol*.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu & other authors (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied And Environmental Microbiology* **72**, 5069-5072.

Do, C. B. and S. Batzoglou (2008). What is the expectation maximization algorithm? *Nature biotechnology* **26**, 897-899.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**, e105-e105.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS*

Computational Biology **7**.

Edgar (2011). UCHIME improves sensitivity and speed of chimera detection. *bioinformatics* **27**.

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.

Ehrlich, S. D. (2011). MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In *Metagenomics of the Human Body*, pp. 307-316: Springer.

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

Engelbrektson, A., V. Kunin, K. C. Wrighton, N. Zvenigorodsky, F. Chen, H. Ochman and P. Hugenholtz (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**, 642-647.

Engle, M. L. and C. Burks (1994). GenFrag 2.1: new features for more robust fragment assembly benchmarks. *Computer applications in the biosciences: CABIOS* **10**, 567-568.

Evans, J. A., S. L. Russell, C. James and J. E. L. Corry (2004). Microbial contamination of food refrigeration equipment. *Journal of Food Engineering* **62**, 225-232.

Fan, L., K. McElroy and T. Thomas (2012). Reconstruction of Ribosomal

RNA Genes from Metagenomic Data. *Plos One*.

Feldmeyer, B., C. W. Wheat, N. Krezdorn, B. Rotter and M. Pfenninger (2011). Short read illumina data for the de novo assembly of a non-model snail species transcriptome and a comparison of assembler performance. *BMC Genomics*.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic biology* **46**, 101-111.

Feng, D.-F., G. Cho and R. F. Doolittle (1997). Determining divergence times with a protein clock: update and reevaluation. *Proceedings of the National Academy of Sciences* **94**, 13028-13033.

Fey, P. D. and M. E. Olson (2010). Current concepts in biofilm formation of *Staphylococcus epidermidis*. *Future Microbiology* **5**, 917-933.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic biology* **20**, 406-416.

Flores, G. E., S. T. Bates, D. Knights, C. L. Lauber, J. Stombaugh, R. Knight and N. Fierer (2011). Microbial Biogeography of Public Restroom Surfaces. *PLoS One* **6**, e28132. doi:28110.21371/journal.pone.0028132.

Flores, G. E., S. T. Bates, J. G. Caporaso, C. L. Lauber, J. W. Leff, R. Knight and N. Fierer (2013). Diversity, distribution and sources of bacteria in residential kitchens. *Environ Microbiol* **15**, 588-596.

Galtier, N., M. Gouy and C. Gautier (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer applications in the biosciences:*

CABIOS **12**, 543-548.

Gilbert, J. A. and C. L. Dupont (2011). Microbial metagenomics: beyond the genome. *Annual Review of Marine Science* **3**, 347-371.

Gilbert, J. A., F. Meyer, J. Jansson, J. Gordon, N. Pace, J. Tiedje, R. Ley, N. Fierer, D. Field & other authors (2010). The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010. *Standards in genomic sciences* **3**, 249.

Gilles, A., E. Meglécz, N. Pech, S. Ferreira, T. Malausa and J.-F. Martin (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *Bmc Genomics* **12**, 245.

Glass, E. M., J. Wilkening, A. Wilke, D. Antonopoulos and F. Meyer (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* **2010**, pdb. prot5368.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz & other authors (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences* **103**, 11240-11245.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution* **11**, 725-736.

Goll, J., D. B. Rusch, D. M. Tanenbaum, M. Thiagarajan, K. Li, B. A.

Methé and S. Yooseph (2010). METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics* **26**, 2631-2632.

Guindon, S. and O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52**, 696-704.

Haas, B. J., D. Gervers and A. M. Earl (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*.

Haft, D. H., J. D. Selengut and O. White (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research* **31**, 371-373.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series*, pp. 95-98.

Hamady, M., C. Lozupone and R. Knight (2010). Fast UniFrac - facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *nature ismej* **4**.

Hamady, M., J. J. Walker, J. K. Harris, N. J. Gold and R. Knight (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature methods* **5**, 235-237.

Handelsman J, R. M., Brady SF, Clardy J and Goodman RM (1998). Molecular biological access to the chemistry of unknown soil microbes : a new frontier for natural products. *Chemistry and Biology*.

Harris, T. D., P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo & other authors (2008). Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109.

Hewitt, K. M., C. P. Gerba, S. L. Maxwell and S. T. Kelley (2012). Office Space Bacterial Abundance and Diversity in Three Metropolitan Areas. *PLoS One* **7**, e37849. doi:37810.31371/journal.pone.0037849.

Hoff, K. J., T. Lingner, P. Meinicke and M. Tech (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research* **37**.

Hoff, K. J., M. Tech, T. Lingner, R. Daniel, B. Morgenstern and P. Meinicke (2008). Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC bioinformatics* **9**, 217.

Hogeweg, P. and B. Hesper (1978). Interactive instruction on population interactions. *Computers in biology and medicine* **8**, 319-327.

Horner, D. S., G. Pavesi, T. Castrignanò, P. D. O. De Meo, S. Liuni, M. Sammeth, E. Picardi and G. Pesole (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in bioinformatics* **11**, 181-197.

Huang, W., L. Li, J. R. Myers and G. T. Marth (2011). ART: a next-generation sequencing read simulator. *Bioinformatics*.

Huang, Y., P. Gilna and W. Li (2009). Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*.

Huber, J. A., D. B. M. Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D.

A. Butterfield and M. L. Sogin (2007). Microbial population structures in the deep marine biosphere. *Science* **318**, 97-100.

Huber, T., G. Faulkner and P. Hugenholtz (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319.

Hudson, D. H., A. Auch, J. Qi and S. C. Schuster (2007). MEGAN analysis of metagenomic data. *Genome Research* **17**.

Huerta, M., G. Downing, F. Haseltine, B. Seto and Y. Liu (2000). NIH working definition of bioinformatics and computational biology. *US National Institute of Health*.

Hugenholtz, P. and T. Huber (2003). Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International journal of systematic and evolutionary microbiology* **53**, 289-293.

Hur, M., Y. Kim, H. R. Song, J. M. Kim, Y. I. Choi and H. Yi (2011). Effect of Genetically Modified Poplars on Soil Microbial Communities during the Phytoremediation of Waste Mine Tailings. *Appl Environ Microb* **77**, 7611-7619.

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.

Huse, S. M., L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman and M. L. Sogin (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genetics* **4**, e1000255.

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

Hyman, E. D. (1988). A new method of sequencing DNA. *Analytical biochemistry* **174**, 423-436.

Jackson, V., I. S. Blair, D. A. McDowell, J. Kennedy and D. J. Bolton (2007). The incidence of significant foodborne pathogens in domestic refrigerators. *Food Control* **18**, 346-351.

Jeon, Y.-S., J. Chun and B.-S. Kim (2013). Identification of Household Bacterial Community and Analysis of Species Shared with Human Microbiome. *Current microbiology*, 1-7.

Jeon, Y.-S., H. Chung, S. Park, I. Hur, J.-H. Lee and J. Chun (2005). jPHYDIT: a JAVA-based integrated environment for molecular phylogeny of ribosomal RNA sequences. *Bioinformatics* **21**, 3171-3173.

Keahey, K. (2010). Cloud Computing for Science.

Keegan, K. P., W. L. Trimble, J. Wilkening, A. Wilke, T. Harrison, M. D'Souza and F. Meyer (2012). A platform-independent method for detecting errors in metagenomic sequencing data: drisee. *PLoS computational biology* **8**, e1002541.

Kelley, D. R., B. Liu, A. L. Delcher, M. Pop and S. L. Salzberg (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research* **40**, e9-e9.

Kembel, S. W., E. Jones, J. Kline, D. Northcutt, J. Stenson, A. M.

Womack, B. J. M. Bohannan, G. Z. Brown and J. L. Green (2012). Architectural design influences the diversity and structure of the built environment microbiome. *The ISME journal* **6**, 1469-1479.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research* **12**, 656-664.

Kim, O.-S., Y.-J. Cho, K. Lee, S.-H. Yoon, M. Kim, H. Na, S.-C. Park, Y. S. Jeon, J.-H. Lee & other authors (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International journal of systematic and evolutionary microbiology* **62**, 716-721.

Kim, S., L. Liao, J.-F. Tomb, M. Zaki, H. Toivonen and J. Wang (2001). A probabilistic approach to sequence assembly validation. In *BIOKDD*, pp. 38-43: Citeseer.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied And Environmental Microbiology*.

Kumar, P. S., M. R. Brooker, S. E. Dowd and T. Camerlengo (2011a). Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS One* **6**, e20956.

Kumar, V. and J. Sivaraman (2011b). Structural characterization of BVU_3255, a methyltransferase from human intestine antibiotic

resistant pathogen Bacteroides vulgatus. *J Struct Biol* **176**, 409-413.

Kunin, V., A. Copeland and A. L. Konstantinos (2008). A Bioinformaticians Guide to Metagenomics. *Microbiol Mol Biol Rev*.

Kunin, V., A. Engelbrektson, H. Ochman and P. Hugenholtz (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental microbiology* **12**, 118-123.

Lai, B., R. Ding, Y. Li, L. Duan and H. Zhu (2012). A de novo assembly program for shotgun DNA reads. *Bioinformatics* **28**, 1455-1462.

Laserson, J., V. Jojic and D. Koller (2010). Genovo: De novo assembly for metagenomes. *Lecture Notes in Computer Science* **6044**, 341-356.

Lee, J.-H., H. Yi and J. Chun (2011). rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology* **49**, 689-691.

Li, H. and R. Durbin (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, W. and A. Godzik (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Lippmann, R. (1987). An introduction to computing with neural nets. *ASSP Magazine, IEEE* **4**, 4-22.

Liu, B., J. Yuan, S.-M. Yiu, Z. Li, Y. Xie, Y. Chen, Y. Shi, H. Zhang, Y.

Li & other authors (2012a). COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics* **28**, 2870-2874.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012b). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* **2012**.

Liu, Z., T. Z. DeSantis, G. L. Andersen and R. Knight (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research* **36**, e120-e120.

Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman and R. Knight (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* **35**, e120.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, A. Buchner, T. Lai, S. Steppi, G. Jobb & other authors (2004). ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363-1371.

Lukashin, A. V. and M. Borodovsky (1998). GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Research* **26**, 1107-1115.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan & other authors (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18.

Luscombe, N. M., D. Greenbaum and M. Gerstein (2001). What is bioinformatics? A proposed definition and overview of the field.

Methods of information in medicine **40**, 346-358.

Magoč, T. and S. L. Salzberg (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker Jr, P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt & other authors (2001). The RDP-II (ribosomal database project). *Nucleic Acids Research* **29**, 173-174.

Margulies, M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature methods* **437**, 376.

Mariette, J., C. Noirot and C. Klopp (2011). Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC research notes* **4**, 149.

Marinelli, L. J., S. Fitz-Gibbon, C. Hayes, C. Bowman, M. Inkeles, A. Loncaric, D. A. Russell, D. Jacobs-Sera, S. Cokus & other authors (2012). *Propionibacterium acnes* bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *mBio* **3**, doi:10.1128/mBio.00279-00212.

Markowitz, V. M., K. Mavromatis, N. N. Ivanova, I.-M. A. Chen, K. Chu and N. C. Kyrpides (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**, 2271-2278.

Markowitz, V. M., N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I.-M. A. Chen, Y. Grechkin, I. Dubchak & other authors (2008). IMG/M: a data management and analysis system for

metagenomes. *Nucleic Acids Research* **36**, D534-D538.

Mavromatis, K., N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski & other authors (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* **4**, 495-500.

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

McHardy, A. C., H. c. G. a. Martı´n, A. Tsirigos, P. Hugenholtz and I. Rigoutsos (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*.

Methe, B. A., K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker & other authors (2012). A framework for human microbiome research. *Nature* **486**, 215-221.

Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nature Reviews Genetics* **11**, 31-46.

Michaels, B., T. Ayers, M. Celis and V. Gangar (2001). Inactivation of refrigerator biofilm bacteria for application in the food service environment. *Food Service Technology* **1**, 169-179.

Miller, C. S., B. J. Baker, B. C. Thomas, S. W. Singer and J. F. Banfield (2011). EMIRGE- reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*.

Miller, J. R., S. Koren and G. Sutton (2010). Assembly Algorithms for

Next-Generation Sequencing Data. *Genomics*.

Mohammed, M. H., T. S. Ghosh, D. Komanduri and S. S. Mande (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25**.

Mohammed, M. H., T. S. Ghosh, N. K. Singh and S. S. Mande (2011a). SPHINX-an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* **27**.

Mohammed, M. H., T. S. Ghosh, S. Chadaram and S. S. Mande (2011b). i-rDNA- alignment free algorithm for rapid in silico detection of ribosomal gene fragments from metagenomic sequence data sets. *BMC Genomics*.

Moore, M., A. Dhingra, P. Soltis, R. Shaw, W. Farmerie, K. Folta and D. Soltis (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology* **6**, 17.

Myers, E. W. and W. Miller (1988). Optimal alignments in linear space. *Computer applications in the biosciences: CABIOS* **4**, 11-17.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai & other authors (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39**.

Namiki, T., T. Hachiya, H. Tanaka and Y. Sakakibara (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*.

Ni, J., Q. Yan and Y. Yu (2013). How much metagenomic sequencing is enough to achieve a given goal? *Scientific reports* **3**.

Noguchi, H., T. Taniguchi and T. Itoh (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA research* **15**, 387-396.

Nyrén, P. and A. Lundin (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical biochemistry* **151**, 504-509.

Ojima, M., Y. Toshima, E. Koya, K. Ara, S. Kawai and N. Ueda (2002a). Bacterial contamination of Japanese households and related concern about sanitation. *International Journal of Environmental Health Research* **12**, 41-52.

Ojima, M., Y. Toshima, E. Koya, K. Ara, H. Tokuda, S. Kawai, F. Kasuga and N. Ueda (2002b). Hygiene measures considering actual distributions of microorganisms in Japanese households. *Journal of Applied Microbiology* **93**, 800-809.

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

Pati, A., N. N. Ivanova, N. Mikhailova, G. Ovchinnikova, S. D. Hooper, A. Lykidis and N. C. Kyrpides (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature*

Methods **7**, 455-457.

Peng, Y., H. C. M. Leung, S. M. Yiu and F. Y. L. Chin (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*.

Perry, A. and P. Lambert (2011). *Propionibacterium acnes*: infection beyond the skin. *Expert Review of Anti-Infective Therapy* **9**, 1149-1156.

Pevzner, P. a., H. Tang and M. S. Waterman (2001). An Eulerian path approach to DNA fragment assembly. *PNAS*.

Pible, O., G. Imbert and J.-L. Pellequer (2005). INTERALIGN: interactive alignment editor for distantly related protein sequences. *Bioinformatics* **21**, 3166-3167.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics* **10**, 354-366.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Glöckner (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188-7196.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric & other authors (2012). The Pfam protein families database. *Nucleic Acids Research* **40**, D290-D301.

Quince, C., A. Lanzen, R. J. Davenport and P. J. Turnbaugh (2011). Removing Noise From Pyrosequenced Amplicon. *BMC Bioinformatics*.

Quince, C., A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M.

Head, L. F. Read and W. T. Sloan (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods* **6**, 639-641.

Reeder, J. and R. Knight (2010). Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature methods* **7**, 668.

Rho, M., H. Tang and Y. Ye (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*.

Rintala, H., M. Pitkaeranta, M. Toivola, L. Paulin and A. Nevalainen (2008). Diversity and seasonal dynamics of bacterial community in indoor environment. *Bmc Microbiol* **8**, doi:10.1186/1471-2180-1188-1156.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26.

Roesch, L. F., R. R. Fulthorpe, A. Riva, G. Casella, A. K. Hadwin, A. D. Kent, S. H. Daroub, F. A. Camargo, W. G. Farmerie & other authors (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**, 283-290.

Rogers, K. L., P. D. Fey and M. E. Rupp (2009). Coagulase-Negative *Staphylococcal* Infections. *Infectious Disease Clinics of North America* **23**, 73-98.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén and P. Nyrén (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**, 84-89.

Rosen, M. J., B. J. Callahan, D. S. Fisher and S. P. Holmes (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics* **13**, 283.

Rouchka, E. C. and D. J. States (1998). Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis. In *ISMB*, pp. 140-147.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman & other authors (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *Plos Biology* **5**, e77.

Ruselervanembden, J. G. H., R. Vanderhelm and L. M. C. Vanlieshout (1989). Degradation of Intestinal Glycoproteins by *Bacteroides-Vulgatus*. *Fems Microbiology Letters* **58**, 37-41.

Rzhetsky, A. and M. Nei (1992). A simple method for estimating and testing minimum-evolution trees. *Mol Biol Evol* **9**, 945-967.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**, 406-425.

Sanchez-Villeda, H., S. Schroeder, S. Flint-Garcia, K. E. Guill, M. Yamasaki and M. D. McMullen (2008). DNAAlignEditor: DNA alignment editor tool. *BMC Bioinformatics* **9**, 154.

Sanger, F., S. Nicklen and A. R. Coulson (1977). DNA sequencing with chin-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463.

Schaller, R. R. (1997). Moore's law: past, present and future. *Spectrum, IEEE* **34**, 52-59.

Schloss, P. D. (2009a). A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One* **4**, e8230.

Schloss, P. D. (2009b). Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial community. *Applied And Environmental Microbiology* **75**, 7537.

Schloss, P. D., D. Gevers and S. L. Westcott (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *Plos One* **6**.

Schmidt, T. M., E. DeLong and N. Pace (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of bacteriology* **173**, 4371-4378.

Scholz, M. B., C.-C. Lo and P. S. Chain (2012). Next generation sequencing and bioinformatics bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*.

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature* **200**.

Scott, E. (1996). Foodborne disease and other hygiene issues in the home. *Journal of Applied Bacteriology* **80**, 5-9.

Scuderi, G., M. Fantasia, E. Filetici and M. P. Anastasio (1996). Foodborne outbreaks caused by *salmonella* in Italy, 1991-4. *Epidemiology and Infection* **116**, 257-265.

Searls, D. B. (2010). The roots of bioinformatics. *PLoS Computational*

Biology **6**, e1000809.

Seshadri, R., S. Kravitz, L. Smarr, P. Gilna and M. Frazier (2007). CAMERA: A Community Resource for Metagenomics. *Plos Biology*.

Sharma, V. K., N. Kumar, T. Prakash and T. D. Taylor (2012). Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin. *Plos One*.

Shendure, J. and H. Ji (2008). Next-generation DNA sequencing. *Nature biotechnology* **26**, 1135-1145.

Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra & other authors (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones and İ. Birol (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**, 1117-1123.

Sinclair, R. G. and C. P. Gerba (2011). Microbial contamination in kitchens and bathrooms of rural Cambodian village households. *Letters in Applied Microbiology* **52**, 144-149.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences* **103**, 12115-12120.

Staley, J. T. and A. Konopka (1985). Measurement of in situ activities of

nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology* **39**, 321-346.

Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya and E. F. DeLong (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of bacteriology* **178**, 591-599.

Sundquist, A., S. Bigdeli, R. Jalili, M. Druzin, S. Waller, K. Pullen, Y. El-Sayed, M. M. Taslimi, S. Batzoglou & other authors (2007). Bacterial flora-typing with targeted, chip-based Pyrosequencing. *BMC microbiology* **7**, 108.

Suyama, M., D. Torrents and P. Bork (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34**, W609-W612.

Tao, Y., Y. Liu, C. Friedman and Y. A. Lussier (2004). Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discovery Today: BIOSILICO* **2**, 237-245.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). A genomic perspective on protein families. *Science* **278**, 631-637.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova & other authors (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* **29**, 22-28.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov &

other authors (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41.

Tawfik, D. S. and A. D. Griffiths (1998). Man-made cell-like compartments for molecular evolution. *Nature biotechnology* **16**, 652-656.

Teeling, H., J. Waldmann, T. Lombardot, M. Bauer and F. O. Glöckner (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics* **5**, 163.

Temperton, B. and S. J. Giovannoni (2012). Metagenomics: microbial diversity through a scratched lens. *Current Opinion in Microbiology*.

Thompson, J. D., T. Gibson and D. G. Higgins (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, 2.3. 1-2.3. 22.

Treangen, T. J., S. Koren, I. Astrovskaya, D. Sommer, B. Liu and M. Pop (2011). MetAMOS: a metagenomic assembly and analysis pipeline for AMOS. *Genome Biology* **12**, 1-27.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight and J. I. Gordon (2007). The human microbiome project. *Nature* **449**, 804-810.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar & other authors (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*

428, 37-43.

Tyson, P. H. a. G. W. (2008). Metagenomics. *Nature Microbiology* **455**.

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa & other authors (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research* **18**, 1051-1063.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson & other authors (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Vergin, K. L., E. Urbach, J. L. Stein, E. F. DeLong, B. D. Lanoil and S. J. Giovannoni (1998). Screening of a Fosmid Library of Marine Environmental Genomic DNA Fragments Reveals Four Clones Related to Members of the OrderPlanctomycetales. *Applied and Environmental Microbiology* **64**, 3075-3078.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261-5267.

Wang, Z., M. Gerstein and M. Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp and G. J. Barton (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Wernersson, R. and A. G. Pedersen (2003). RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic acids research* **31**, 3537-3539.

Wilkening, J., A. Wilke, N. Desai and F. Meyer (2009). Using clouds for metagenomics: a case study. In *Cluster Computing and Workshops, 2009 CLUSTER'09 IEEE International Conference on*, pp. 1-6: IEEE.

Wintzing, F. v., U. B. Gobel and E. Stackbrandts (2006). Determination of microbial diversity in environmental samples_pitfalls of PCR-based rRNA analysis.

Woese, C. R. (1987). Bacterial evolution. *Microbiological reviews* **51**, 221.

Wommack, K. E., J. Bhavsar and J. Ravel (2008). Metagenomics : Read Length Matters. *Applied and Environmental Microbiology* **74**, 1453-1463.

Wooley , J. C. and Y. Ye (2010). Metagenomics: Facts and artifacts, and computational challenges. *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY* **25**.

Wright, E. S., L. S. Yilmaz and D. R. Noguera (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Applied And Environmental Microbiology* **78**, 717-725.

Wu, Y.-W. and Y. Ye (2008). A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples. *Recomb*.

Yang, X., S. P. Chockalingam and S. Aluru (2012). A survery of error-correction methods for next-generation sequencing. *Bioinformatics*.

Yang, Z., R. Nielsen, N. Goldman and A.-M. K. Pedersen (2000).

Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449.

Yap, I. and R. Nelson (1996). Winboot: a program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms. *International Rice Research Institute, Manila*, 1-22.

Zhang, J., K. Kobert, T. Flouri and A. Stamatakis (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, btt593.

APPENDIX I. Estimated Diversity Index of Household Microbiome

				CD-HIT			TBC			TDC-TBC		
Origin	Sample		Normalized reads	Observed OTUs	Estimated OTUs (Chao1)	Shannon index	Observed OTUs	Estimated OTUs (Chao1)	Shannon index	Observed OTUs	Estimated OTUs (Chao1)	Shannon index
Refrigerator	Culture independent	#1	2,000	57	85.11	2.15	284	383.09	4.57	75	177.50	2.45
		#2	2,000	103	218.00	3.19	308	543.77	4.04	76	142.11	2.80
		#3	2,000	343	835.77	3.98	557	900.70	5.00	340	820.31	3.48
		#4	2,000	145	244.53	3.48	361	674.89	4.60	108	190.88	2.90
		#5	1,477	226	391.15	3.97	340	433.25	5.01	267	446.45	4.15
		#6	1,946	278	415.04	4.25	497	616.94	5.30	329	497.30	4.36
		#7	2,000	580	1486.16	5.03	746	1,445.00	5.54	572	1,318.00	4.83
		#8	2,000	309	667.73	4.02	491	714.73	5.11	354	701.84	3.77
		#9	2,000	190	342.29	3.69	433	916.00	4.71	192	368.79	3.56
		#10	2,000	138	230.81	3.51	350	488.11	4.69	111	179.90	2.92
	Culture dependent	#1	2,000	43	54.00	2.56	300	380.73	4.55	57	78.00	2.37
		#2	2,000	105	138.07	3.39	332	442.01	4.42	79	157.11	2.46
		#3	1,401	100	145.11	3.78	290	409.02	4.88	79	91.83	2.71
		#4	1,028	76	108.50	3.17	187	255.25	4.17	61	149.00	2.26
		#5	2,000	76	89.13	2.90	296	359.01	4.42	80	113.21	2.62

		#6	2,000	41	63.00	2.07	180	193.23	3.39	44	57.91	1.20
		#7	2,000	158	246.50	3.70	397	553.14	4.70	116	201.55	2.43
		#8	2,000	123	246.50	3.79	295	343.09	4.72	144	215.55	3.47
		#9	2,000	181	246.00	3.97	395	572.03	4.78	161	258.33	3.00
		#11	2,000	169	206.94	4.05	487	720.63	5.20	109	179.50	2.67
Toilet	Culture independent	#1	2,000	153	224.87	3.25	528	1,089.99	5.10	143	221.55	2.93
		#3	1,409	196	237.62	4.21	324	377.31	5.02	201	251.40	4.07
		#8	1,150	158	243.69	3.80	307	502.15	4.67	137	242.06	3.72
		#10	2,000	313	422.17	4.79	497	559.40	5.57	286	357.84	4.63
		#12	2,000	157	312.08	3.75	346	431.12	4.77	128	258.20	3.07
		#13	2,000	126	165.00	3.47	426	581.00	4.84	88	138.21	1.56
	Culture dependent	#1	2,000	38	47.00	1.86	199	228.00	3.34	20	24.20	0.90
		#2	2,000	46	59.75	2.29	215	316.86	3.29	21	24.50	0.97
		#3	2,000	43	46.00	2.36	203	216.68	3.64	34	40.00	1.70
		#4	2,000	83	85.77	3.43	407	588.12	4.95	52	61.07	1.75
		#5	2,000	54	92.50	2.29	293	399.97	3.91	34	47.60	0.80
		#6	2,000	68	83.11	2.75	340	469.09	4.36	36	59.75	1.53
		#7	2,000	124	165.35	3.44	285	312.28	4.52	116	184.90	2.77

국문 초록 (Abstract in Korean)

메타지놈은 환경시료에서 직접 추출한 전체 DNA를 의미하며, 메타지놈 연구의 목적은 시료 내에 존재하는 세균들의 분류학적 구성과 기능적 측면을 알아내는 데 있다. 메타지놈 분석 방법으로는 크게 amplicon 기반의 분석 방법과 random shotgun 기반의 분석 방법이 존재한다. 이 두 방법은 대량의 sequencing 데이터를 필요로 한다는 공통점이 있고, 전통적인 Sanger sequencing 방법으로는 충분한 데이터를 얻을 수 없었다. 하지만, 차세대 염기 서열 분석 방법 (Next Generation Sequencing, NGS)의 발전은 낮은 가격에 대량의 sequencing data 생산을 가능하게 하였고, NGS를 통해 생산되는 복잡한 대용량의 데이터를 처리하기 위한 생물정보학 기술이 발전 함에 따라, 지난 수년 동안 메타지놈 연구가 활발하게 진행 되어왔으며, 메타지놈 분석은 미생물학 및 미생물 생태학 등 관련 연구에 필수적인 연구가 분야가 되었다. 하지만, NGS의 단점인 짧은 리드 (read)

길이, 상대적으로 높은 시퀀싱 에러 율 및 적합한 분석 시스템의 부족으로 인하여 메타지놈은 아직까지 완전하게 분석이 되지 못하고 있는 상황이며, 따라서 현재 사용되고 있는 여러 가지 에러처리 알고리즘에 대한 객관적인 평가와 효율적인 알고리즘 및 효과적인 데이터 처리 시스템의 개발이 필요하다.

본 연구에서는, NGS를 활용한 amplicon 및 random shotgun 기반의 메타지놈 분석을 수행하는데 필요한 분석 시스템을 개발하였으며, 생물정보학 기술을 활용하여 NGS 에러 및 짧은 리드 길이로 인해 발생하는 오류를 최소화하여 정확성을 향상시키는데 주안점을 두었다. Amplicon 분석을 위해, 454 pyrosequencing data 및 Illumina MiSeq paired end 데이터를 처리하는 두 종류의 분석 시스템을 개발 하였으며, 454 pyrosequencing 분석 시스템에서는 homo-polymer 및 PCR 에러를 처리하기 위한 새로운 알고리즘을 개발하였으며, 분석 시스템을 활용한 시범연구로, 일반 가정의 화장실 및 냉장고에 존재하는 미생물의 군집과 사람의 장내 미생물 군집을 비교 분석하였다. Illumina MiSeq

데이터를 이용한 amplicon 분석 시스템 개발 과정에서는, 최적의 sequencing 조건 및 sequencing region을 찾았으며, sequencing error를 처리하기 위해, paired end reads 병합 프로그램과 iterative consensus clustering 방법을 개발하여 분석의 정확도를 향상 시켰다.

Random shotgun 분석 시스템은 Illumina MiSeq paired end 데이터를 이용하였다. Amplicon 분석과 달리 대부분의 shotgun sequencing paired end reads는 병합 되지 않으며, 결과적으로 짧은 리드를 분석에 사용하게 되어 분석 과정에서 오류가 만들어질 가능성이 높다. 따라서, 짧은 리드길이로 발생하는 오류를 최소화하기 위한 방법으로 긴 contig sequence 를 만드는 raw read mapping 방법과 de novo assembly를 수행하도록 시스템을 개발하였다. Mapping 단계에서는, 동적으로 생성된 mapping genome database를 이용, raw read mapping을 통해 긴 contig sequences를 생성하였으며, mapping에 참여하지 않은 reads들은 de novo assembly를 통해 contig sequence를 생성하도록 하였다. 개발한 분석 파이프라인을 이용하여 토양과 사람의 장내에서 채취한 샘플을

분석 하였으며, 장내 샘플 분석에서는 mapping과 de novo assembly를 통해 긴 contig sequences를 생성하였으나, 토양 샘플 분석에서는 성공적으로 contig sequences를 생성하지 못했다. 이는, 토양 샘플은 장내 샘플 보다 상대적으로 복잡한 미생물 군집 구조를 가지기 때문에 샘플에 존재하는 미생물을 충분히 포함할 수 있을 정도의 많은 양의 sequencing 데이터를 얻는 것이 불가능하여, rRNA profile을 이용한 mapping genome database가 충분히 구성될 수 없었기 때문이며, 또한 알려지지 않은 난배양성 미생물이 존재하기 때문에 mapping 과정에 활용되는 reference genome이 존재하지 않을 수 있기 때문이다. 분석 시스템 개발에 더해, 분석 과정을 거쳐 생성된 결과를 해석하기 위해 군집 구조와 annotation 정보를 시각화하는 프로그램을 개발하였다.

Java 기반의 sequence alignment 에디터 프로그램인 EzEditor를 개발하였다. 분자 계통 분류 연구에 16S rRNA 뿐만 아니라, conserved coding sequence가 이용되면서, codon 기반의 sequence alignment editing 프로그램이 필요하며, EzEditor는 rRNA의

이차구조에 더해 functional coding sequence의 protein sequence를 이용하여 DNA alignment를 에디팅할 수 있도록 다양한 기능을 포함하고 있다. EzEditor는 분자 계통 분류학에 필요한 기능 뿐만 아니라, 염기서열 및 단백질의 아미노산 서열 등을 이용하는 연구에 사용할 수 있는 다양한 기능을 포함하고 있다.

주요어: 생물정보학, 차세대염기서열분석, 메타지놈, 분석 파이프라인, 미생물 군집분석, 이지에디터, 계통분류학

학 번: 2007-30126