이 학 석 사 학 위 논 문

Understanding the genome of living organism based on the second generation sequencing

차세대 염기서열 분석방법을 이용한

생물 유전체의 이해

2014 년 2 월

서울대학교 대학원

생물정보협동과정 생물정보학전공

곽 우 리

차세대 염기서열 분석방법을 이용한 생물 유전체의 이해

지도교수 김 희 발

이 논문을 이학석사 학위논문으로 제출함

2013 년 12 월

서울대학교 대학원

생물정보협동과정 생물정보학전공

곽 우 리

곽우리의 이학석사 학위논문을 인준함

2013 년 12 월

위 원 장 김 선 (인)

부위원장 김 희 발 (인)

위 원 조 서 애 (인)

Abstract

Understanding the genome of living organism based on the second generation sequencing

Woori Kwak

Interdisciplinary Program in Bioinformatics

Seoul National University

These studies are mainly about the rebuilding genome sequence of living organism using *de novo* assembly based on the second generation sequencing technologies and understanding the gene level features of organisms. Even though the next generation sequencing, especially the second generation sequencing, make the genome project can be conducted in reasonable price, assembling the short read from the second generation sequencing is challenging. Various programs which have its unique characteristics are available but one program or pipeline cannot be the best choice at any times. Therefore, researchers who want to rebuild the genome sequence using *de novo* assembly have to choose the best combination of programs and pipeline for specific data.

In chapter 2, I make the efficient combination of programs for the *de novo* assembly of microbes and the finished level genome assembly of the probiotic candidates had been conducted using short reads from two sequencing technologies. Based on the result of assembly, I found the potential risk as a useful probiotic strain.

In chapter 3, minke whale genome assembly had been conducted using low coverage re-sequencing data. I found the efficient genome assembly pipeline using various open source programs which showed better performance than the assembly result of the expensive commercial program. And contig extension and bridging were conducted to combine the result of assembly from different samples.

In chapter 4, assembly of unaligned reads from short read alignment to the reference genome was conducted to identify the unique sequence and gene contents of Korean Native Chicken (KNC) samples. Based on the unaligned reads assembly and gene prediction, KNC specific genes and sequences were identified for further analysis.

Through these studies, I trained making some efficient genome assembly pipelines suitable for specific data and learned the way to understand the characteristics of living organisms based on the assembly and gene level features.

Key words: Next generation sequencing, second generation sequencing, *de novo* assembly, genome assembly

Student number : 2012-20410

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 The sequencing technologies

1.1.1 The first generation sequencing.

The process which decodes the DNA sequence of genome is called sequencing. Even though first modern sequencing technology had been developed by Maxam and Gilbert in 1977(Maxam and Gilbert 1977), Sanger sequencing method(Sanger, Nicklen et al. 1977) is known as the first generation sequencing method or the conventional sequencing method. This method used ddNTP(dideoxyribo nucleotides triphosphate) which don't have OH in 3' carbon of center sugar. The oxygen in OH residue of 3' carbon provided the energy which can continue the chain reaction of DNA synthesis. So ddNTP which have 3'-H residue made the chain reaction terminated. The Sanger sequencing methods used these characteristic of ddNTP. Researchers made numerous fragments of DNAs which have 1 base pair length difference and conducted electrophoresis for ordering the base sequence of DNA fragment. Decoding of the order of DNA sequence had been conducted manually. The early stage Sanger sequencing has 100bp read length and small output of data generation. In 1986, Applied Biosystems(ABI) introduced automated DNA sequencing which uses different fluorescently end-labelled primer for each ddNTP sequencing reaction. Using the fluorescent spectrum of each ddNTP in combined electrophoresis gel with data analysis using computer(Smith, Sanders et al. 1986), sequence decoding was more easily and quickly

conducted compared to manually decoding. DNA sequencing became truly automated with the ABI Prism 310 which used capillary electrophoresis in 1996

1.1.2 The second generation sequencing

When the information which can be used as the base of bioinformatics was accumulated from Sanger sequencing based research, new sequencing technology named "Next Generation Sequencing" began to appear. These new sequencing technology had very low cost for data generation compared to Sanger sequencing method and it was rapidly used for various researches and research fields.(Metzker 2009) Nowadays, in 2013, two major sequencing technologies are ecumenically used in the market. One is Illumina's Hiseq which uses the reversible terminator and the other is the Roche's GS FLX which uses pyrosequencing.

Pyrosequencing is known as the first commercialized NGS technology and is was utilized by Jonathan Rothberg(Rothberg and Leamon 2008). Pyrosequencing technology detects the pyrophosphate release on nucleotide incorporation. Released PPis quantitatively are converted to ATPs by ATP sulfurylase and generated ATP provides energy to the luciferate-mediated conversion of luciferin to oxyluciferin. Oxyluciferin generates visible light for detecting the DNA synthesis

and this light is detected by camera. In pyrosequencing, there is no different detection signal between 4 dNTPs. So each dNTPs (A,T,G,C) are used one at a time and then we figure out the each base specific signal. In case of homopolymer, repeats of same base sequence, DNA synthesis is conducted until the end of homopolymer at a time. The amount of signal is different in proportion to the number of bases elongated at a time. However, it is not exactly proportional to the number of elongated bases and the variation of detection signal is increased with the length of homopolymer increasing. For this reason pyrosequencing frequently generated InDel sequencing error in homopolymer region compared to conventional Sanger sequencing and Illumina sequencing technology. Pyrosequencing read the fragmented DNA sequence using single direction method and paired-end read can be generated using mate-pair library. Read length is 600 base pair in average in GS FLX system of Roche and it is almost 6 times longer than Illumina Hiseq platforms.

Illumina's sequencing platform represented by Hiseq2000 is the most widely used sequencing platform in 2013. It uses SBS(sequencing by synthesis) sequencing principle with reversible terminator. Reversible terminator is blocked 3'-end for nucleotide incorporation like ddNTP used in conventional Sanger sequencing. However, as the name might

suggest, reversible terminator can be recovered its 3'-OH for nucleotide incorporation. Nucleotide which is blocked 3'-OH is incorporated to the primer sequence and the process of DNA synthesis is stop. Each reversible terminator has fluorescence dye and it can be detected by camera. And the 3'-end recover it's OH residue. These three steps (nucleotide incorporation, detecting fluorescence, recover 3'-OH) consist 1 cycle which is the basic unit of Illumina sequencing. Only one nucleotide can be detected in 1 cycle, so Illumina's sequencing platform have advantages in InDel sequencing error compared to pyrosequencing method. Illumina's sequencing platform read the fragmented DNA sequence using paired-end read method and read length is 101bp in Hiseq2000 and 150 in Hiseq 2500. The read length of illumina platforms is shorter than GS FLX system using pyrosequencing but it is a lot cheaper to generated sequencing data. And Illumina's sequencing platform have lowest sequencing error rate compared to other next generation sequencing platforms.

Recent NGS technology is classified under two categories, second generation technology and third generation technology, and these two major sequencing platforms described above are classed as second generation sequencing. First generation sequencing technology means Sanger sequencing technology. Compared to Sanger sequencing technology, second generation technology has some characters. First,

the read length of second generation sequencing is shorter than Sanger sequencing. The read length of Sanger sequencing is almost 1kbp and it is 10 times longer than Illumina's Hiseq 2000. Second is the data generation throughput and time. Second generation sequencing generates high throughput sequencing data in short time compared to Sanger sequencing. For example, one Hiseq2500 device generate 10X coverage genome data of 20 peoples in almost one day and it is tens of thousand times bigger than Sanger sequencing device. Next is the low cost. It is expected that sequencing the genome of one individual will be cost under $100 in the near future. Forth, sequencing reactions are conducted in the miniaturized device compared to Sanger sequencing. Fifth, error rate is higher than Sanger sequencing method. The error rate of Illumina platform and pyrosequencing is known as 0.26% and 1.07%. It is higher than the error rate of Sanger sequencing method (about 0.1%)

1.1.3 Development of sequencing technology

Illumina Hiseq and Roche GS FLX sequencing systems are what we call second generation sequencing platform and these platforms still have some common limitations. First, it uses fluorescence detection system. Illumina system uses nucleotides which have the specific fluorescence color for each base and the laser make the fluorescence molecule light. GS FLX system also detects the light as signal of the

nucleotide incorporation. This type of technology must have camera system for detecting the signal and error can be generated and accumulated. For example, illumine sequencing system using reversible terminator which have fluorescence molecule. In each cycle of sequencing, fluorescence molecules in cluster have to be removed for next nucleotide incorporation. However, all fluorescence molecules cannot be removed perfectly. Remained fluorescence molecules are accumulated as the sequencing proceeding and this make the weak fluorescence signal. This is the reason why the result of illumina sequence has lower quality scores in end of the read. Imaging system using camera also can be a hurdle for miniaturizing sequencer. Second, the second generation sequencing use PCR for preparing sequencing library. GC contents which mean the proportion of G and C nucleotide in total nucleotides can affect the PCR result. Normal PCR method shows that all region of whole genome cannot be amplified evenly and high or low GC regions are more difficult to amplify using PCR. Therefore, sequencing results based on PCR library preparation are necessarily biased. Third, the error rate of these sequencing technologies is still higher than Sanger sequencing method.

To overcome the limitation, various sequencing platforms and technologies are developing. Two commercialized sequencing

platforms, ion torrent from life science and RS system from Pacific Biosystem, have unique sequencing principle.

Ion torrent(Rothberg, Hinz et al. 2011) is similar to GS FLX system which uses the byproduct of nucleotide incorporation like pyrophosphate. Instead of pyrophosphate, ion torrent system detects the hydrogen ion which is also byproduct of nucleotide incorporation using PH value. This method conducts sequencing without fluorescence molecule and imaging device. And sequencing process is processed on small semiconductor chip. So the sequencing device can be efficiently miniaturized. Error rate of InDels, the major error type of pyrosequencing, can be reduced. Sequencing speed is increased. However ion torrent still amplify DNA fragment using emulsion PCR for library preparation. Sequencing result is not GC bias free because the amplification efficiency of DNA fragment vary according to the GC ratio of DNA fragment. DNA fragments with high or low GC ratio cannot be amplified efficiently.

Pacbio RS system(English, Richards et al. 2012) fixed the DNA polymerase in the bottom of well and the DNA synthesis process is conducted in fixed point of location. It uses fluorescence molecules and imaging device like second generation sequencing and it has same limitations like second generation sequencing technologies. However,

pacbio RS system adopt single molecular sequencing technology. It means that RS system doesn't amplify the DNA fragment for sequencing process like other sequencing platforms (Illumina, Roche and Ion torrent). Sequencing without amplification of DNA fragments using PCR has benefits like GC bias free which is useful for genome sequencing and accurate expression profile which is useful for RNA-seq analysis. Rs system also generates long read length sequencing data compared to other sequencing platforms. In spite of these advantages of pacbio RS system, higher error rate (almost 15%) is a big problem of RS system. To solve this weakness of RS system, the method called CCS system (circular consensus sequencing system)(Travers, Chin et al. 2010) has been developed. Hairpin structure adaptor attaches to the end of DNA fragment and this adaptor structure make the sequencing process repeatedly conducted for the DNA fragment. DNA fragment is sequenced at least 3 times and the consensus base call can efficiently reduce the error rate of RS system using independent sequencing reactions.

Illumina also make up for the weak points. First, read length is improved. Miseq V2 which is the most recent sequencing platform of Illumina produces 300bp pair-end data. This is three times longer than Hiseq2000 which is the most world widely used sequencing platforms

in these days. And almost 600bp of single read can be generated for metagenome community analysis using overlapping library. Second, GC bias problem can be solved using PCR-free library preparation kit. PCR amplification and Gel electrophoresis has been used for making sequencing library for Illumina. PCR-free Library preparation(Kozarewa, Ning et al. 2009) kit does not use gel electrophoresis for size selection and magnetic bead base for DNA isolation in library preparation protocol. Because it doesn't amplify DNA using PCR, genome coverage of sequencing can be increased in high or low GC contents region and dispersion of sequencing overage is greatly reduced. This can make the genome project more efficiently conducted. Third one is the molecule technology for long read data generation. This technology is developed base on the genome assembly research of Botryllus schlosseri(Voskoboynik, Neff et al. 2013). Main concept of this technology is partitioning specific size of DNA molecule and multiplexing technology using index sequences. Partitioned DNA molecules marked with index sequence and short read assembly reconstruct limited number of DNA molecules. This technology is expected to be supplied to the market within 2014.

Nanopore sequencing(Branton, Deamer et al. 2008) is the third generation sequencing platform in the true sense of the word. It doesn't

use fluorescence molecules and imaging devices. It doesn’t amplify DNA fragments and it conducts sequencing on a single molecule of DNA fragment. The prototype device of oxford nonopore is very small (palm size) and it can conduct sequencing just connect USB 3.0 cable with laptop computer. Even though the accuracy and the throughput of oxford nanopore have to be improved, this shows the blue print of future sequencing technology. The changes of sequencing technologies are extremely fast and researchers who want to analysis NGS data have to understand the unique characteristics and principle of sequencing technologies.

1.2 Genome assembly using short read NGS data

1.2.1 Conventional genome assembly

Genome assembly means that reconstruct the original sequence of DNA using fragment DNA sequence from sequencing machine. This is necessary to figure out the genome sequence of organism because sequencing machine cannot read the complete genome sequence at one time. Genome assembly is often likened to jig saw puzzle. Assembly process relies on the assumption that two fragment of sequenced DNA reads share a same string of letters in specific location of the original DNA sequence. Using this overlapped sequence, two sequence reads can be connected and extended to the longer sequence. Various programs like Celera Assembler(Myers, Sutton et al. 2000), ARACHNE(Batzoglou, Jaffe et al. 2002),PCAP(Huang, Wang et al. 2003) were widely used for assembly of whole genome shotgun sequencing data. However, assembly methods based on the overlapping sequence information need to compare all sequence one by one and required memory size of computing server increase proportionally to the number of input sequence. Whole genome sequence data using second generation sequencing consist of huge number of short reads

and comparing all the pair of short read sequences is almost impossible in these days computing system resources.

1.2.2 Genome assembly using NGS data

There are several characteristics of NGS data which make the genome assembly difficult. First, regardless of the sequencing technology, sequencing results is shorter than conventional sequencing. The read length of Sanger sequencing which generate longer reads than NGS sequencer is not also fully enough for genome assembly. The read length of NGS data is even much shorter than Sanger sequencing and NGS attempts to overcome this limitation through high throughput data generation. Second one is huge amounts of data produced by NGS. NGS experiments produce a massive amount data which make the required memory resources of server increase drastically. Most of NGS data assemblers manage these kinds of large sequencing data through use of K-mers. A K-mer means a series of contiguous base of length K and K is positive integer. Converting reads to the set of K-mers reduces the total amount of data efficiently. Searching shared K-mers is easier than searching overlaps of each read. However, if positive integer K is small, the unique information of original reads is lost. And opposite, if K value is large enough, the information of original read remained but the amount of huge data cannot be reduced efficiently. Third, error rate

is higher than conventional sequencing. This can induce assembly errors which make incorrect assembly or shortened contigs., Many statistical approaches used To correct the errors from sequencing machines. Forth, repeat sequences make the assembly more difficult. Genome regions which share perfect repeats or the length of repeated sequence longer than read length of NGS data cannot be distinguishable. Paired-end and mate-pair libraries can be a help for solving this problem. Last one is the sequencing coverage of different genome regions is not uniform. This induce the gaps in the assembly result and coverage variability.

Because of these characteristics of NGS data mentioned above, the new assembly methods are necessary. Assemblers using NGS data are classified under 3 categories including greedy assemblers, overlap-layout-consensus assemblers, and de Bruijm graph assemblers. Even though the details of these three kinds of assembler are different between each other, these programs uses graphs based techniques. Greedy assemblers use the simple principle. It iteratively extends a read or contig by adding reads based on sequence overlaps. This is repeated until the read or contig cannot be extended more. The choice of reads for extension is based on the number of matching bases in the overlapping sequence. Assemblers using OLC (overlap-layout-

consensus) method conduct the assembly in three stages. First stage is overlap discovery. In the first stage, read overlaps are calculated. All pair-wise comparison of reads is conducted based on the precomputed K-mer contents. Based on the overlaps in the first stage, the overlap graph is built and optimized. Last stage is consensus stage using multiple sequence alignment. De bruijn graph assembler is the most widely used assembler for short read NGS data like Illumina. This method has the advantages for dealing with large quantities of data. De bruijin graph is based on K-mers and calculating the all pair-wise read overlapping is unnecessary. Individual reads don't need to be stored and redundant sequences are collapsed.

1.2.3 The representative assemblers for NGS data

SSAKE(Warren, Sutton et al. 2007) is known as the first short read assembler and it is designed to assemble unpaired reads. This assembler uses the greedy algorithm and index reads by their prefixes. Reads of which prefix overlaps over minimum length are searched iteratively. SHARCGS(Dohm, Lottaz et al. 2007), VCAKE(Jeck, Reinhardt et al. 2007) and QSRA(Bryant, Wong et al. 2009) use the greedy algorithm like SSAKE.

Newbler(Margulies, Egholm et al. 2005) is a widely used OLC-based assembler specially designed for GS FLX sequence platform of Roche. Using long read length compared to Illumina, First release of Newbler conducts whole genome assembly based on unpaired reads of approximately 100 bp but now it can handle much longer reads. It is usually used for small size genome like bacterial genome. Newbler is an exclusively designed assembler for GS FLX system and it can use short read data from Illumina sequencer to support GS FLX sequencing data.

CABOG(Miller, Delcher et al. 2008) is the revised version of Celera Assembler originally designed for Sanger reads and the pipeline is designed for 454 data. CABOG collapses the homopolymer to single bases to overcome the InDel error of pyrosequencing. Each read in the set of overlapping reads is compared for error correction and errors can be inferred where bases are contradicted by overlapped bases.

Euler(Pevzner, Tang et al. 2001) is developed originally for Sanger reads and modified to operate on various data, 454 pyrosequencing reads(Chaisson, Pevzner et al. 2004), single-end Illumina reads(Chaisson and Pevzner 2008), and paired-end Illumina reads(Chaisson, Brinza et al. 2009). Like other assembler, Euler also conducts error correction before building de Bruijn graph. Based on K-

mer frequency distribution, it detects base-call errors by identifying K-mers with low frequency. Because K-mers resulting from base-call errors show much lower frequency while most true K-mers are repeated over many times. K-mers with frequency below threshold are removed or corrected from input sequence data. While error correction step before assembly is important, this modification can mask true polymorphism or true K-mers which show low frequency by chance. When reads are converted to K-mers, there is some information lost compared to using reads directly. Read-threading step by laying entire read onto its graph is conducted to recover this information.

Velvet(Zerbino and Birney 2008) is a very popular assembler among de bruijn assemblers. It performs graph simplification which collapse simple paths into single nodes and this makes the graphs much simpler. Three parameters used in Velvet affect the result of assembly. First one is K value used in K-mers, which have to be an odd integer. Second is the minimum expected frequency threshold of K-mers for error correction step. Finally, the expected genome coverage controls spurious connection breaking.

AllPaths(Butler, MacCallum et al. 2008) and Allpaths-LG(Gnerre, MacCallum et al. 2011) are de bruijn graph based assemblers for large genome assembly. Allpaths begin with error correction preprocess

similar to Euler's and it uses the base quality score for this process. Filtered reads may be retained if the substitution of two low-quality base makes its K-mers trusted or the read is essential for building a path between pair-end reads. AllPaths-LG adds improvements to the AllPaths algorithm like better error correction for remaining true SNVs while filtering as many sequencing error as possible, more efficient gap filling and scaffolding, and graph simplification which can show better result in eukaryotic genome assembly. However, AllPaths-LG require high memory resources compared to other de bruijn assemblers and at least one more specially designed overlap paired-end library is essential for staring assembly.

SOAPdenovo(Li, Zhu et al. 2010) and SOAPdenovo2(Luo, Liu et al. 2012) is freely available large genome assemblers using de bruijn graph and small memory resource requirement is a characteristics of these programs. Especially SOAPdenovo2 uses sparse graph which can reduce the memory requirement more effectively while the results of assembly are maintained. SOAPdenovo can conduct contig assembly and scaffolding independently. Gapcloser, the inner module of SOAPdenovo, shows good performance for gap filling process and it can be conducted independently with SOAPdenovo.

Process of genome assembly using NGS data

De novo assembly using NGS data is conducted through 6 steps.

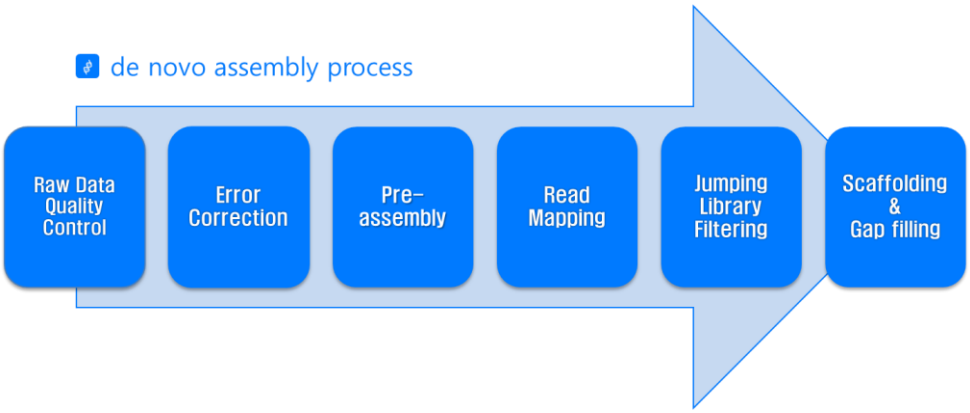

Figure 1. Typical *de novo* assembly process

First is raw data quality control step. There is no golden standard but here are some typically used conditions for raw data quality control.

- Read with N bases more than 10%.
- Read with low quality bases more than half of the read length.(quality score <5)
- Read contain more than 10bp adapter or primer sequence.
- Overlap paired-end read (except for Allpaths-LG).
- Two reads in each paired-end read are completely identical.

After raw data quality control, error correction process is conducted. Most of assemblers have its own error correction module, but some independent error correction module like Quake(Kelley, Schatz et al. 2010) and error correction module of AllPaths-LG are frequently used. Next, pre-assembly using only paired-end reads is conducted for checking and filtering the jumping library. Jumping libraries of which the insert size is longer than 2kbp are mapped to the assembled scaffolds and insert size distribution is calculated. Based on the result of insert size distribution, researchers decide the use of particular jumping library. Contigs assembly and scaffolding are conducted in order of insert size and the gap filling process follows. The best result is chosen among the result using various programs and libraries.

Chapter 2. *De novo* assembly and comparative analysis of the *Enterococcus faecalis* genome (KACC 91532) from a Korean neonate

2.1 Introduction

Ever since Ilya Mechnikov first proposed lactic acid from fermented milk as a secret to longevity, lactic acid bacteria (LAB) have been an area of considerable interest (Mercenier, Pavan et al. 2003). LAB have been shown to provide numerous health benefits including prevention and treatment of diarrhea, immunological activation against pathogens and cancer, prevention of allergies, improved gastrointestinal function, improved lactose tolerance, and treatment of high blood pressure (Sarkar 2008). Because of these wide-ranging benefits, significant research has been aimed at isolating LAB from feces for food and medicinal use (Martín, Jiménez et al. 2006). Based on existing probiotics selection criteria, we hypothesized that the most numerous LAB in feces would exhibit superior proliferation and adaptability in the intestine of neonates (Penders, Thijs et al. 2006), and could therefore be used as a probiotic strain for infants and children. Accordingly, we collected feces samples from newborns from all areas of South Korea. From these samples, Enterococcus faecalis isolate KACC 91532 exhibited the fastest growth rate, and was chosen as a potential new probiotic bacterial strain (Kim., Choi. et al. 2011).

E. faecalis is the first LAB to colonize the intestines of infants, and is generally considered a non-pathogenic commensal of the mammalian gastrointestinal tract (Ryan and Ray 2010). However, recent studies have established *E. faecalis* as the major cause of enterococcal infection (Hancock, Gilmore et al. 2006); other virulence factors in *E. faecalis* and their role in pathogenicity have also been established (Chow, Thal et al. 1993, Schlievert, Gahr et al. 1998). Many studies have shown that the presence or absence of specific virulence factors is important for enterococcal infections, and that these virulence factors can enhance the ability of *E. faecalis* to cause disease (Eaton and Gasson 2001, Lempiäinen, Kinnunen et al. 2005, Cariolato, Andrighetto et al. 2008). As bacteria undergo significant horizontal gene transfer and gene loss compared to eukaryotic species, whole genome *de novo* assembly is essential to fully understand the genetic composition of newly isolated bacteria (Ochman and Moran 2001). For example, a previous study (Feil, Feil et al. 2005) demonstrated large differences in gene contents between fully sequenced pathovars of *Pseudomonas syringae*. Such divergence between known reference genomes and newly isolated bacteria makes it difficult to obtain complete genome sequences for newly isolated bacteria using resequencing or reference-based assembly (Salzberg, Sommer et al.

2008). Therefore, we sought to characterize the gene content and virulence factors of *E. faecalis* KACC 91532 using whole genome *de novo* assembly.

In this study, we identified acid, heat, and antibiotic resistance in *E. faecalis* isolate KACC 91532, and conducted whole genome *de novo* assembly using a newly constructed *de novo* assembly pipeline. Based on this assembled genome sequence, we identified the gene contents and virulence factors of *E. faecalis* KACC 91532, and compared them to four available *E. faecalis* reference genomes. We were also able to identify evolutionarily accelerated genes and variation in *E. faecalis* KACC 91532 using dN/dS analysis. We establish the potential risk of *E. faecalis* KACC 91532 as a probiotic strain, and present a newly constructed *de novo* assembly pipeline which can be used for performing *de novo* assembly of other microorganisms.

2.2 Material and methods

LAB Isolation

Fecal samples were collected from 25 neonates (16 male, 9 female) born over the course of 5 days across 6 regions of South Korea (Seoul, Incheon, Gang-

won, Chungcheong, Jeolla, and Gyeongsang). Samples were stored under anaerobic conditions at 4℃.

Fecal samples were plated on BCP(Bromocresol Purple) agar (Eiken, Japan) and incubated at 37℃ for 48 h. LAB were quantified by manually counting all yellow colonies, which then were subcultured in MRS broth (Difco, USA) and screened on TOS agar (Eiken, Japan). Then, isolated colonies were cultured on MRS agar (Difco, USA) under anaerobic incubation, and preserved in cryovials (Key Scientific Products, USA) at -70℃ for further study.

To identify the LAB isolate, we performed 16S rRNA gene sequence analysis according to the method of Pavlova et al. (Pavlova, Kilic et al. 2002). To amplify 16S rDNA, we used universal primers corresponding to six conserved regions of the *Escherichia coli* numbering system. Chromosomal DNA was isolated using a genomic DNA extraction kit (Qiagen, Germany). PCR was performed in a 50 μL reaction mixture containing primers (50 pmol), template DNA (50 ng), 5 μL 10×*Taq* DNA polymerase buffer, 4 μL dNTP at 2.5 mM, and 1 U *Taq* DNA polymerase (Takara, Japan). The PCR amplification product was purified using a QIAquick gel extraction kit (Qiagen), ligated into a pSTBlue-1 vector (Novagen, USA), and transformed into *E. coli* DH5α competent cells. The recombinant plasmids were purified using a DNA purification kit (Qiagen) and digested with *Eco*RI to confirm the insert. The nucleotide sequence of the insert was determined using a BigDye™-

terminator sequencing kit and ABI PRISM 377 sequencer (Perkin-Elmer, USA), according to the manufacturer's instructions. The 16S rDNA sequences were subjected to a similarity search of the GenBank database. The strain exhibiting the highest growth rate in MRS broth from 26 XR7 strain *E. faecalis* was donated to the Rural Development Administration (RDA)-Genebank Information Center, Republic of Korea.

Testing physiologic features

Antimicrobial susceptibility to erythromycin, gentamicin, oxacillin, tylosin, and vancomycin was performed by disk diffusion in accordance with Clinical Laboratory Standard Institute guidelines(Wikler 2006). For heat challenge and survival measurement, growth phase cultures were heat treated at 95℃ for 30 s, 1 min, and 2 min, respectively. Heat-treated culture (100 μL) was spread on MRS agar (Difco, USA) and incubated at 37℃ for 24 h. And growth phase culture (100 μL) was spread on MRS agar (Difco, USA) adjusted to pH 4.8, 5.0, and 5.5 with 1.0N HCl, and incubated at 37℃ for 24 h for acid tolerance

Genomic Sequencing, assembly and annotation

Roche 454 pyrosequencing reads (shotgun and 8 kb mate pair) and Illumina Hiseq 2000 sequencing reads were generated by the National

Instrumentation Center for Environmental Management at Seoul National University. The details of the sequence data are provided in Supplementary Table 2.1 and Figure 2.1.

Raw read data (sff) were modified for de Bruijn assembly (SOAPdenovo (Li 2009), Allpaths-LG (Gnerre, MacCallum et al. 2011)). Sff file format reads were converted into fastq using sff2fasta (http:// github.com/indraniel/sff2fastq). Linker sequences were removed from 454 8 Kb mate pair reads and converted into a paired-end library using in-house software. Quality control and trimming of 454 shotgun reads was performed using FastQC(Andrews 2012) and FastX-toolkit(Gordon and Hannon 2010). Then the reads were converted into overlapped paired-end fastq for Allpaths-LG input.

The overall process of *de novo* assembly used in this paper is described in Figure 2.2. Newbler assembly software (Chaisson and Pevzner 2008) (gsAssembler 2.8) was used to perform *de novo* genome assembly using 454-FLX sequence data. Independent *de novo* assembly was performed three times to generate contigs using SOAPdenovo and Allpaths-LG. First, Illumina reads were error-corrected using the Allpaths-LG error-correction module, and assembled using SOAPdenovo. Second, converted 454 reads were edited for SOAPdenovo assembly using the Allpath-LG error-correction module.

Last, modified 454 reads were converted into paired-end reads for assembly using Allpaths-LG. Gap filling was conducted by Gapcloser (Li 2009) and Gapfiller (Nadalin, Vezzi et al. 2012); Illumina and 454 reads were error-corrected using the Allpaths-LG module, and the combined contigs from the de Bruijn assembly were used for Gapcloser. For Gapfiller, error-corrected Illumina reads were used as input data. Gapcloser and Gapfiller were reiterated until no change in N base number was seen.

BLAST was used to identify scaffolds that could be connected to each other. For each scaffold, 200 bp were cut from both ends, and the resulting collection of 200-bp sequences was used as query sequences. Contigs from SOAPdenovo, Allpaths-LG, and Newbler were used as the BLAST database. Before connecting two scaffolds, the match of base pair in the connection end of each scaffold were manually checked using Bioedit (Hall 2005)(Figure 2.3).

Enterococcus faecalis reference genome sequences from four strains (EF62, D32, V583, OG1RF) were obtained from the NCBI database. Genome sequences were annotated using the RAST (Aziz, Bartels et al. 2008) server pipeline. Assembly of *E. faecalis* KACC 91532 was compared to reference sequences using a BLAST dotplot with the RAST SEED viewer. Sequences and functions of genes mentioned in

this paper were manually confirmed using BLASTp; Cn3D (Wang, Geer et al. 2000) was used to identify the location of variants in the 3D protein structure.

dn/dS Analysis

To conduct dN/dS analysis using the branch model, we gathered nucleotide and amino acid sequences with the same FIGfam IDs as our RAST annotation results. The orthologous gene sets were aligned by PRANK (Löytynoja and Goldman 2005) using the default settings; poorly aligned sites were eliminated using Gblocks (Castresana 2000). To build a standard phylogenetic tree, we performed bootstrap analysis on the combined data set sequences using PHYLIP (Seqboot) (PLOTREE and PLOTGRAM 1989). We used TREE-PUZZLE (Schmidt, Strimmer et al. 2002) to estimate the Ts/Tv ratio and calculated the distance of each strain using a Kimura 2-parameter model. A consensus tree was built using the neighbor joining method. The maximum likelihood method (codeml of PAML 4) (Yang 2007) was used to estimate the dN (the rate of non-synonymous substitution), dS (the rate of synonymous substitution), and ω (the ratio of non-synonymous substitutions to the rate of synonymous substitutions) with F3X4 codon frequencies under the branch model (model=2, NS sites=0)

and basic model (model=0, NS sites=0). Orthologs with dS > 3 or ω > 5 were filtered (Castillo-Davis, Kondrashov et al. 2004, Peacock, Seeger et al. 2007). The overall process of dN/dS analysis is provided in Figure 2.4.

2.3 Results

Physiologic features of *E. faecalis* KACC 91532

The details of test results pertaining to acid resistance, heat resistance, and antibiotic resistance are provided in Table 2.2. *E. faecalis* KACC 91532 was able to survive at pH 4.8, and was resistant to gentamicin. It was also able to withstand 2 min heat treatment at 95℃.

De novo assembly and gene contents of *E. faecalis* KACC 91532

The first assembly using Newbler (gsAssembler 2.8) produced 5 scaffolds with an N50 of 2,966,033 and a total length was 3,120,175 bp, with 63,577 N bases (2.037%). Newbler provide hybrid assembly using illumina read or contigs. However the results of hybrid assembly had more fragmented scaffolds with more N bases than assembled genome which used 454 raw read only. For example, when we added contigs from soap denovo and allpaths-LG to Newbler, the assembly result had

6 scaffolds with 63,785 N bases. After gap filling using GapCloser process based on the first assembly using 454 raw data only, numbers of N bases were reduced from 63,577 to 1,046. Additional gap filling using GapCloser did not close gap anymore and continued gap filling process using Gapfiller increased the number of N bases to 1,117. However, Gapfiller helped GapCloser to close more gaps and we could reduce the number of N bases to 71. A BLAST search using 200 bp ends from each of the five scaffolds revealed the longest and shortest scaffolds to be reverse complementary to each other; these two scaffolds were subsequently joined to form a single scaffold. After all iterated gap filling and bridging, the final assembly consisted of four scaffolds (2,97 Mb, 70526 bp, 51709 bp, 28528 bp) with a total length of 3,123,166 bp and 71 N bases. A comparison of the *E. faecalis* KACC 91532 assembly sequence to four reference genome sequences using a BLAST dotplot is provided in Figure 2.5. BLAST dotplot shows a bidirectional comparison of two genome sequences and the closer genome sequence shows a more diagonal line. The E.faecalis KACC 91532 assembly shows the least disconnection in a diagonal line and the highest query cover rate (94%) with the EF62 strain isolated from Norwegian infants. The *E. faecalis* KACC 91532 assembly was the

most similar to the EF62 strain isolated from Norwegian infants (Solheim, Aakra et al. 2009).

A comparison of four *E. faecalis* reference genomes and *E. faecalis* KACC 91532 using RAST is provided in Table 2.3. *E. faecalis* KACC 91532 draft genome had 346 subsystems, 3,061 CDS, and 67 structural RNAs. After removing hypothetical proteins, a total of 2286 CDSs remained. The four reference genomes (EF62, D32, V583, OG1RF) had 2887, 2919, 3172, and 2548 CDSs each. We found numerous genes associated with resistance to antibiotics and toxic compounds including the bile hydrolysis gene for survival in the gastrointestinal tract, the tetracycline-resistance gene, beta lactamase (which confer resistance to beta-lactam antibiotics such as penicillin), topoisomerase IV genes (which confer resistance to fluoroquinolone antibiotics), and multidrug resistance efflux pump genes all within the *E. faecalis* KACC 91532 assembly. There are 11 KACC 91532-specific genes compared to EF62 strain, 37 genes compared to D32 strain, 25 genes compared to V583 strain, and 39 genes compared to OG1RF strain. There are nine KACC 91532-specific genes compared to four references and two of them (Exodeoxynuclease V Beta and Cadmium-transporting ATPase) could be assigned with an enzyme commission (EC) number. All KACC 91532-specific genes are provided in Table 2.4.

dN/dS analysis

The dN/dS analysis revealed 18 evolutionarily accelerated genes with p-value < 0.05 and FDR < 0.2; the gene list is provided in Table 2.5. Among 18 evolutionarily accelerated genes, 7 (Arginine deiminase [EC 3.5.3.6], O-succinylbenzoic acid-COA ligase [EC 6.2.1.26], Probable L-ascorbate-6-phosphate lactonase UlaG [EC 3.1.1.-], Two-component sensor histidine kinase, malate [EC 2.7.3.-], tRNA nucleotidyltransferase [EC 2.7.7.21,25], Predicted PTS system, galactosamine-specific IIA component [EC2.7.1.69], and UDP-galactosephosphotransferase [EC 2.7.8.6]) had been assigned with an enzyme commission number. With the exception of Probable L-ascorbate-6-phosphate lactonase UlaG (EC 3.1.1.-), sequences and annotations of the remaining six genes were confirmed using BLASTp. A summary comparing the nucleotide and amino acid sequences of these six genes is provided in Table 2.6. There are three genes (tRNA nucleotidyltransferase, arginine deiminase, UDP-galactosephosphotransferase) whose non-synonymous sites were found only in *E. faecalis* KACC 91532. In tRNA nucleotidyltransferase, glutamic acid and serine had been changed to lysine and asparagines; for arginine deiminase, arginine had been changed to histidine. UDP-

galactosephosphotransferase had many variable positions near the end of its amino acid sequence region. The amino acid sequence changes in these three genes are provided in Figure 2.6A. Arginine deiminase had an available protein 3D structure in NCBI and we identified the location of the variant in 3D protein structure using Cn3D. The variant in arginine deiminase was estimated to be located in a chain between an α-helix and β-strand sheet of 1LXYA (Figure 2.6B).

2.4 Discussion

454 GS FLX+(Roche) using pyrosequencing chemistry is routinely used for microorganism genome sequencing due to its long read length. However, the high error rate in homo polymers and high reagent costs for pyrosequencing are significant obstacles to whole genome sequencing (Loman, Constantinidou et al. 2012). The addition of Illumina sequencing reads (sequencing by synthesis) to pyrosequencing data provides a useful and cost effective way to get more complete genome assemblies. Accordingly, many studies have employed a combination of these two different methods for whole genome sequencing. Through testing various combinations of read data sets and various *de novo* assembly programs employing different algorithms, we

have developed a newly constructed *de novo* assembly pipeline. Using Newbler under default conditions and only 454 reads showed better performance than using both 454 reads and Illumina reads at the same time. We think that this result stems from the characteristics of the *E. faecalis* genome, including small genome size and low repeat content. Allpaths-LG is a *de novo* assembler that uses Illumina reads, and was not designed to perform error correction for raw read data. However, it has been used to generate input data for other *de novo* assembly programs (Salzberg, Phillippy et al. 2012). An independent error correction module (Errorcorrectread.pl) from Allpaths-LG (http://www.broadinstitute.org/software/allpaths-lg/blog/) can be used for error correction of raw read data more easily. The result of gap filling showed that the combinational use of GapCloser and Gapfiller effectively closed gaps existed in the scaffolds. We think that this result comes from the algorithm of Gapfiller which remove the low quality edge and extend gap edges with k-mers. The *de novo* assembly pipeline described in this paper is an effective pipeline that uses independent and suitable programs for each step. This *de novo* assembly pipeline can reduce costs for iterated Sanger sequencing to generate complete genomes and provide more accurate information about microorganisms.

No known gentamicin-resistance genes were identified in *E. faecalis* KACC 91532 to account for the observed gentamicin-resistant phenotype. However, *E. faecalis* KACC 91532 does possess multi-drug resistance (MDR) efflux pump genes. As drug efflux systems are a well-established mechanism of antibiotic resistance (Lomovskaya and Watkins 2001) (Alekshun and Levy 2007), we believe that the gentamycin resistance of *E. faecalis* KACC 91532 is related to this system. *E. faecalis* KACC 91532 exhibited thermal resistance in a heat-resistance assay. Chaperone protein DnaK, one of the heat-shock proteins, was found in *E. faecalis* KACC 91532 and may be related to its thermo tolerance. In terms of virulence factors, *E. faecalis* KACC 91532 had sex pheromone-related genes (aggregation substance Asa1/PrgB) present in only one of four reference strains (V583). Aggregation substance Asa1/ProgB encodes for cell-surface protein Asc10. This protein is involved in cell aggregation, and can lead to the horizontal transfer of antibiotics-resistance genes, such as pheromone-inducible tetracycline-resistance plasmid pCF10 (Chung, Bensing et al. 1995), to other bacteria. Hence, *E. faecalis* KACC 91532 should only be used as a probiotic strain for Korean infants after careful consideration. This analysis demonstrates that by understanding gene

contents using *de novo* assembly, we can build upon existing probiotics selection criteria to produce safer probiotics and health supplements.

Evolutionarily accelerated genes and KACC 91532-specific variants

tRNA nucleotidyltransferase performs 3'-terminal-CCA tRNA sequence repair in conjunction with poly(A) polymerase I and polynucleotide phosphorylase; this process is essential for the growth of the bacteria.(Reuven, Zhou et al. 1997) O-succynylbenzoic acid-COA ligage (EC 6.2.1.26) is used in vitamin K production, which can in turn stimulate microbial growth(Baronets 2003). As *E. faecalis* KACC 91532 is the fastest-growing strain among all enterococci isolates examined, the presence of these two evolutionarily accelerated genes may account for the elevated growth rate of this strain.

Arginine deiminase (ADI) is the first enzyme in the arginine deiminase pathway, and is commonly found in acid-resistant LAB. This pathway produces ammonia by converting L-arginine into L-citrulline; the resulting ammonia helps to buffer the organism under acidic conditions. Acid resistance of *E. faecalis* KACC 91532 was above average compared to other *E. faecalis* isolates. More research is needed to

establish a direct correlation between variation in arginine deiminase and acid resistance in *E. faecalis* KACC 91532; gene-specific population analysis can be used to determine the effect of these variants

Table 2 - 1. Summary of raw read data

Type of Read	Number of Reads	Encoding	GC ratio	Read length
Illumina paired end A	6504978	Sanger/Illumina 1.9	37	101
Illumina paired end B	6504978	Sanger/Illumina 1.10	37	101
454 shotgun	833495	Sff	37	53–1200
454 8kb mate pair	620901	Sff	38	57–1200

Table 2 - 2. Summary of antibiotics resistance, acid resistance, and heat resistance test (+ : resistance, O : survival)

Erythomycin	Gentamicin	Oxacillin	Tylosin	Vancomycin		PH 4.8	PH 5.0	PH 5.5		30 Sec	1 Min	2 Min
-	+	-	-	-		O	O	O		O	O	O

Table 2 - 3. RAST annotation summary of four *E. faecalis* reference genomes and KACC 91532.

Genome	Genome Size	Subsystem	CDS	Structural RNAs	Hypothetical CDS Removed
E62	2,988,673 bp	349	2887	67	2214
D32	2,987,450 bp	344	2919	74	2187
V583	3,218,031 bp	355	3172	80	2327
OG1RF	2,739,625 bp	332	2548	71	2021
KACC 91532	3,123,166 bp	346	3061	67	2286

Table 2 - 4. Summary of *E. faecalis* KACC91532-specific gene lists compared to four reference genomes

(a) EF62

FIGfam, ref ID	Function
FIG01289107	rRNA methylase (FIG011178)
FIG00000595	DNA recombination protein RmuC
FIG00043030	Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
FIG01304124	D-hydantoinase (EC 3.5.2.2)
FIG01304225	Phage tail fibers
FIG00631844	Pheromone binding protein TraC/PrgZ
FIG00628727	Pheromone shutdown protein TraB/PrgY
FIG00133384	Putative pheromone precursor lipoprotein
FIG00008572	Cadmium efflux system accessory protein
FIG00503691	Cadmium-transporting ATPase (EC 3.6.3.3)

(b) D32

FIGfam, ref ID	Function
ref|ZP_04434451.1|	Branched-chain phosphotransacylase
FIG00001401	Gluconate dehydratase (EC 4.2.1.39)
FIG00003174	Alpha-xylosidase (EC 3.2.1.-)
FIG00019456	Xylose isomerase (EC 5.3.1.5)
FIG00000793	Xylulose kinase (EC 2.7.1.17)
FIG00009450	Peptidoglycan N-acetylglucosamine deacetylase (EC 3.5.1.-)
FIG00035627	Glycerol-3-phosphate cytidylyltransferase (EC 2.7.7.39)
FIG01955844	CDP-glycerol:poly(glycerophosphate) glycerophosphotransferase (EC 2.7.8.12)
ref|ZP_19425819.1|	ATP-binding protein

FIG01289107	rRNA methylase (FIG011178)
FIG00000364	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)
FIG00008304	Aspartate 1-decarboxylase (EC 4.1.1.11)
FIG00000440	Pantoate--beta-alanine ligase (EC 6.3.2.1)
FIG00000595	DNA recombination protein RmuC
FIG00003334	DNA-cytosine methyltransferase (EC 2.1.1.37)
FIG00002504	DNA-damage-inducible protein J
FIG00043030	Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
ref|NP_816141.1|	Undecaprenyl diphosphate synthase (EC 2.5.1.31)
FIG00340292	Duplicated ATPase component BL0693 of energizing module of predicted ECF transporter
FIG00013534	Substrate-specific component BL0695 of predicted ECF transporter
FIG00010442	Transmembrane component BL0694 of energizing module of predicted ECF transporter
FIG00004764	Ammonium transporter
FIG01304124	D-hydantoinase (EC 3.5.2.2)
FIG00003520	Phage tail fiber protein
FIG01304225	Phage tail fibers
FIG01304517	Putative GTPases (G3E family) (COG0523)
FIG00628006	Aggregation substance Asa1/PrgB
FIG00631844	Pheromone binding protein TraC/PrgZ
FIG00628727	Pheromone shutdown protein TraB/PrgY
FIG00630030	Surface exclusion protein Sea1/PrgA
FIG00001676	Ferredoxin
FIG00446664	Retron-type RNA-directed DNA polymerase (EC 2.7.7.49)
FIG00008572	Cadmium efflux system accessory protein
FIG00503691	Cadmium-transporting ATPase (EC 3.6.3.3)
FIG00017458	Sensor histidine kinase VncS
FIG00016235	Two-component response regulator VncR
FIG01303728	Tetracycline resistance protein TetM

(c) OG1RF

FIGfam, ref ID	Function
ref|ZP_04434451.1|	Branched-chain phosphotransacylase
FIG00003507	6-phospho-beta-galactosidase (EC 3.2.1.85)
FIG00001401	Gluconate dehydratase (EC 4.2.1.39)
FIG00003174	Alpha-xylosidase (EC 3.2.1.-)
FIG00019456	Xylose isomerase (EC 5.3.1.5)
FIG00000793	Xylulose kinase (EC 2.7.1.17)
FIG00077849	Ethanolamine utilization protein similar to PduU
FIG01955844	CDP-glycerol:poly(glycerophosphate) glycerophosphotransferase (EC 2.7.8.12)
FIG01289107	rRNA methylase (FIG011178)
FIG00011382	Membrane proteins related to metalloendopeptidases
FIG00000364	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)
FIG00008304	Aspartate 1-decarboxylase (EC 4.1.1.11)
FIG00000440	Pantoate-beta-alanine ligase (EC 6.3.2.1)
FIG00000595	DNA recombination protein RmuC
FIG00002504	DNA-damage-inducible protein J
FIG00002410	Recombinational DNA repair protein RecT (prophage associated)
FIG00043030	Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
ref|NP_816141.1|	Undecaprenyl diphosphate synthase (EC 2.5.1.31)
FIG00340292	Duplicated ATPase component BL0693 of energizing module of predicted ECF transporter
FIG00013534	Substrate-specific component BL0695 of predicted ECF transporter
FIG00010442	Transmembrane component BL0694 of energizing module of predicted ECF transporter
FIG00004764	Ammonium transporter
FIG01283764	Phage DNA binding protein
FIG00003520	Phage tail fiber protein
FIG01304225	Phage tail fibers
FIG00011468	Phage tail length tape-measure protein
FIG01304517	Putative GTPases (G3E family) (COG0523)

FIG00628006	Aggregation substance Asa1/PrgB
FIG00631844	Pheromone binding protein TraC/PrgZ
FIG00628727	Pheromone shutdown protein TraB/PrgY
FIG00133384	Putative pheromone precursor lipoprotein
FIG00630030	Surface exclusion protein Sea1/PrgA
FIG00446664	Retron-type RNA-directed DNA polymerase (EC 2.7.7.49)
FIG00009563	Choloylglycine hydrolase (EC 3.5.1.24)
FIG00008572	Cadmium efflux system accessory protein
FIG00503691	Cadmium-transporting ATPase (EC 3.6.3.3)
FIG00017458	Sensor histidine kinase VncS
FIG00016235	Two-component response regulator VncR
FIG01303728	Tetracycline resistance protein TetM

(d) V583.

FIGfam, ref ID	Function
FIG01304276	Cystathionine beta-lyase (EC 4.4.1.8)
FIG00003507	6-phospho-beta-galactosidase (EC 3.2.1.85)
FIG00745599	Alpha-glucosidase (EC 3.2.1.20)
FIG00001401	Gluconate dehydratase (EC 4.2.1.39)
FIG00077849	Ethanolamine utilization protein similar to PduU
FIG00000350	D-alanine--D-alanine ligase (EC 6.3.2.4)
FIG01289107	rRNA methylase (FIG011178)
FIG00003190	BH1670 unknown conserved protein in B. subtilis
FIG00000595	DNA recombination protein RmuC
FIG01303866	DNA repair exonuclease family protein YhaO

FIG00043030	Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
FIG00008752	ComF operon protein C
FIG00001450	Hydroxymethylglutaryl-CoA reductase (EC 1.1.1.34)
FIG00133144	Muconate cycloisomerase (EC 5.5.1.1)
FIG01304124	D-hydantoinase (EC 3.5.2.2)
FIG01283764	Phage DNA binding protein
FIG01304225	Phage tail fibers
FIG01304517	COG0523: Putative GTPases (G3E family)
FIG00631844	Pheromone-binding protein TraC/PrgZ
FIG00631844	Pheromone-binding protein TraC/PrgZ
FIG00628727	Pheromone shutdown protein TraB/PrgY
FIG00133384	Putative pheromone precursor lipoprotein
FIG00008572	Cadmium efflux system accessory protein
FIG00503691	Cadmium-transporting ATPase (EC 3.6.3.3)
FIG01303728	Tetracycline resistance protein TetM

Table 2 - 5. Evolutionarily accelerated genes in *E. faecalis* KACC 91532 ($p < 0.05$, FDR < 0.2)

Figfam ID	Function	p-value
FIG00000165	Translation elongation factor LepA	0.03784
FIG00000184	Putative deoxyribonuclease YcfH	0.02807
FIG00000289	tRNA nucleotidyltransferase (EC 2.7.7.21) (EC 2.7.7.25)	0.04455
FIG00000298	GTP-binding protein EngB	0.00505
FIG00000860	O-succinylbenzoic acid--CoA ligase (EC 6.2.1.26)	0.02428
FIG00001104	Arginine deiminase (EC 3.5.3.6)	0.03786
FIG00001384	"Catalyzes the cleavage of p-aminobenzoyl-glutamate to p-aminobenzoate and glutamate, subunit A"	0.02479
FIG00001663	GTP-sensing transcriptional pleiotropic repressor codY	0.04811
FIG00002133	"Two-component sensor histidine kinase, malate (EC 2.7.3.-)"	0.03056
FIG00002968	Undecaprenyl-phosphate galactosephosphotransferase (EC2.7.8.6)	0
FIG00004373	Putative membrane protein YeiH	0.01028
FIG00018224	Substrate-specific component MtsA of methionine-regulated ECF transporter	0.02596
FIG00051666	Lon-like protease with PDZ domain	0.0133
FIG00134695	Predicted PTS system, galactosamine-specific IIA component (EC 2.7.1.69)	0.04318
FIG00139589	Probable L-ascorbate-6-phosphate lactonase UlaG (EC 3.1.1.-) (L-ascorbate utilization protein G)	0.03166
FIG01115339	ABC transporter substrate-binding protein	0.03338
FIG01321817	Quaternary ammonium compound-resistance protein sugE	0.01516
FIG01333098	NtrC family Transcriptional regulator, ATPase domain	0.0176

Table 2 - 6. Number of substitution sites and non-synonymous sites in evolutionarily accelerated enzymes.

(S.site : Substitution site, K-S.site : KACC91532 Substitution site, NS.site : Non-Synonymous substitution site, K-NS.site : KACC91532 Non-Sysnonymous site)

Figfam ID	Function	S.site	K-S.site	NS.site	K-NS.site
FIG00000289	tRNA nucleotidyltransferase	28	10	2	2
FIG00000860	O-succinylbenzoic acid-COA ligase	21	1	7	1
FIG00001104	Arginine deiminase	18	3	1	1
FIG00002133	Two-component sensor histidine kinase, malate	14	2	5	1
FIG00002968	Undecaprenyl-phosphate galactosephosphotransferase	51	32	11	11
FIG00134695	Predicted PTS system, galactosamine-specific IIA component	10	3	3	2

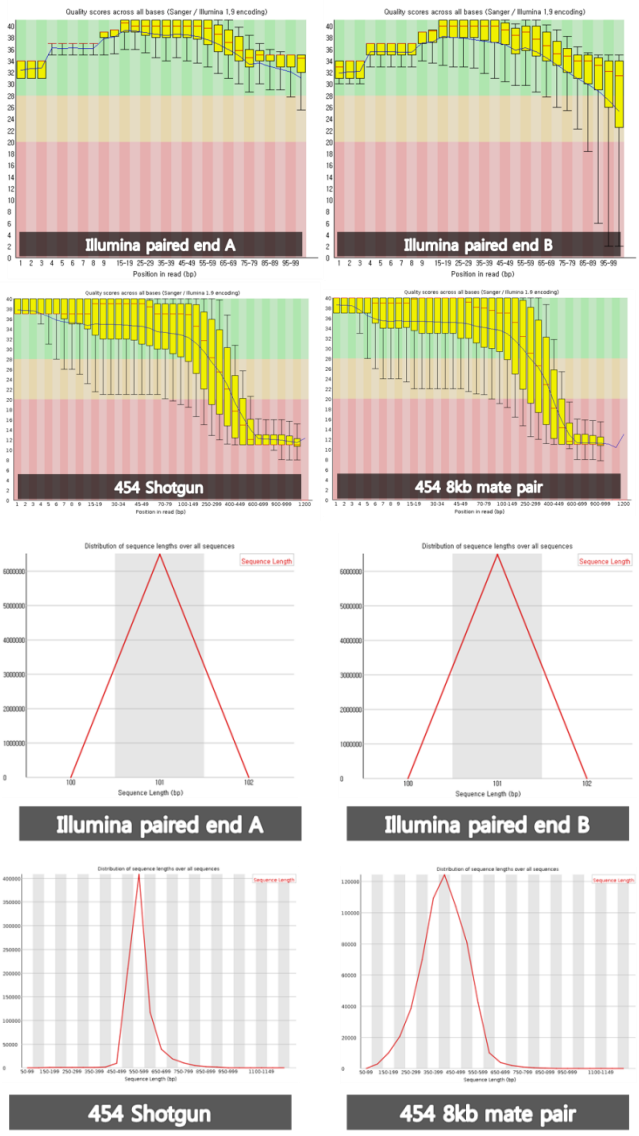

Figure 2 - 1 FastQC results of raw data

(a. Per base quality score of four raw reads, b. Length distribution of four raw reads.)

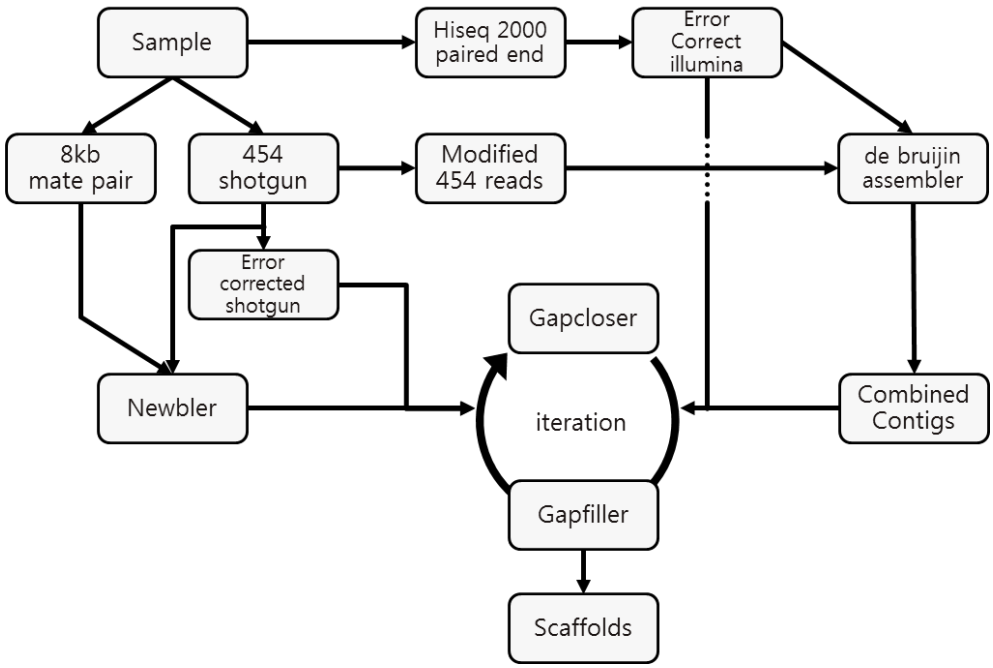

Figure 2 - 2. *De novo* assembly pipeline used for E. faecalis KACC91532 genome assembly.

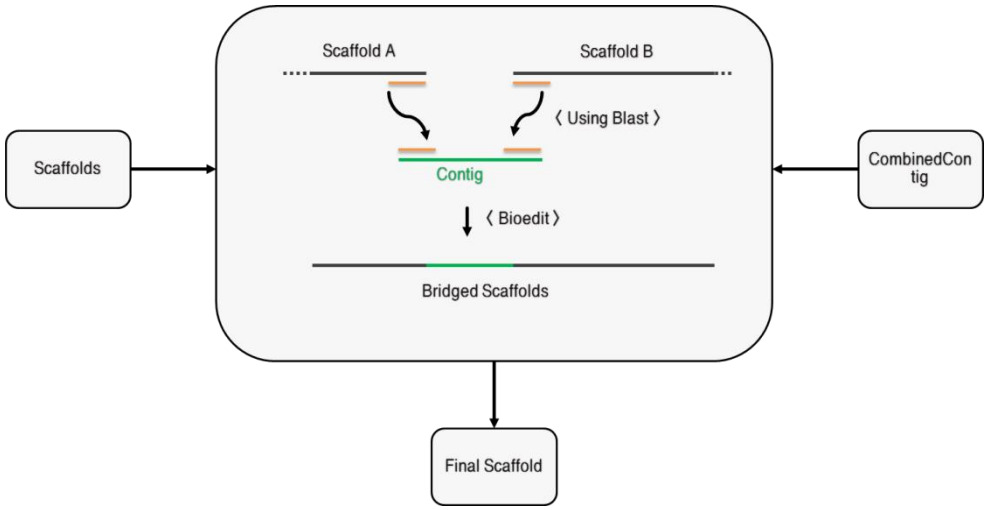

Figure 2 - 3. Summary of scaffold-bridging process

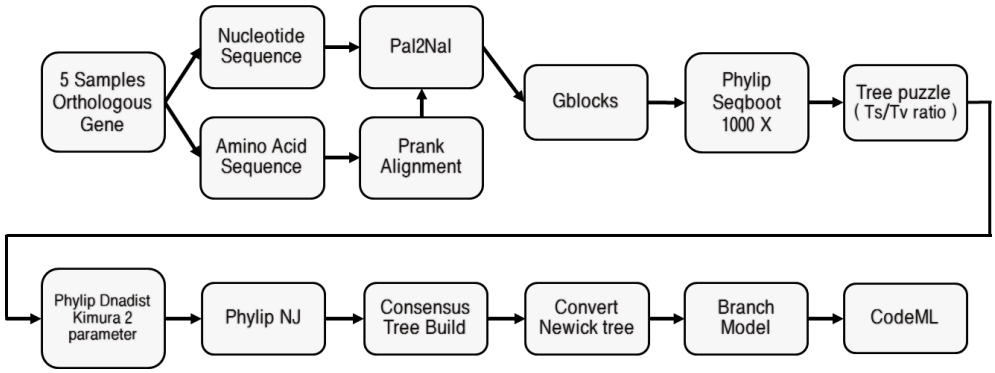

Figure 2 - 4. Overall process of dN/dS analysis used in this work.

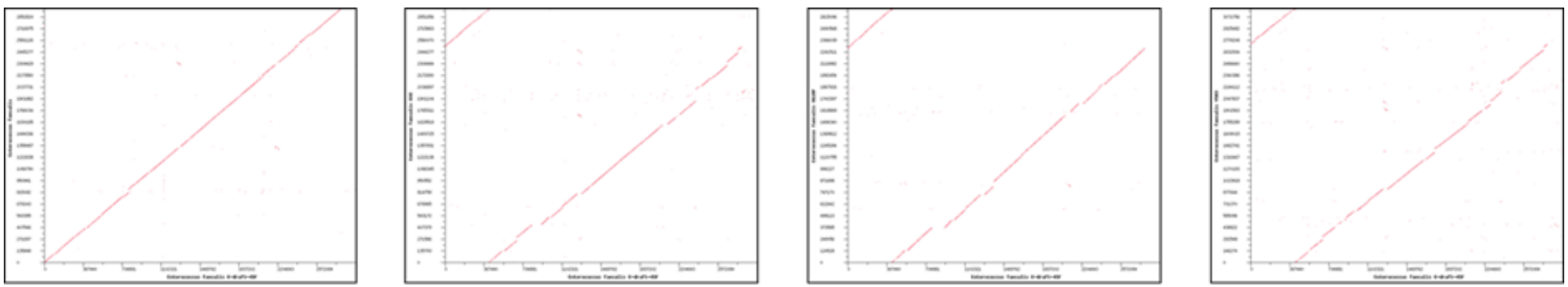

Figure 2 - 5. BLAST dotplot of *E. faecalis* KACC91532 assembly with four reference genomes (X-axis: KACC91532, Y-axis: reference genome).

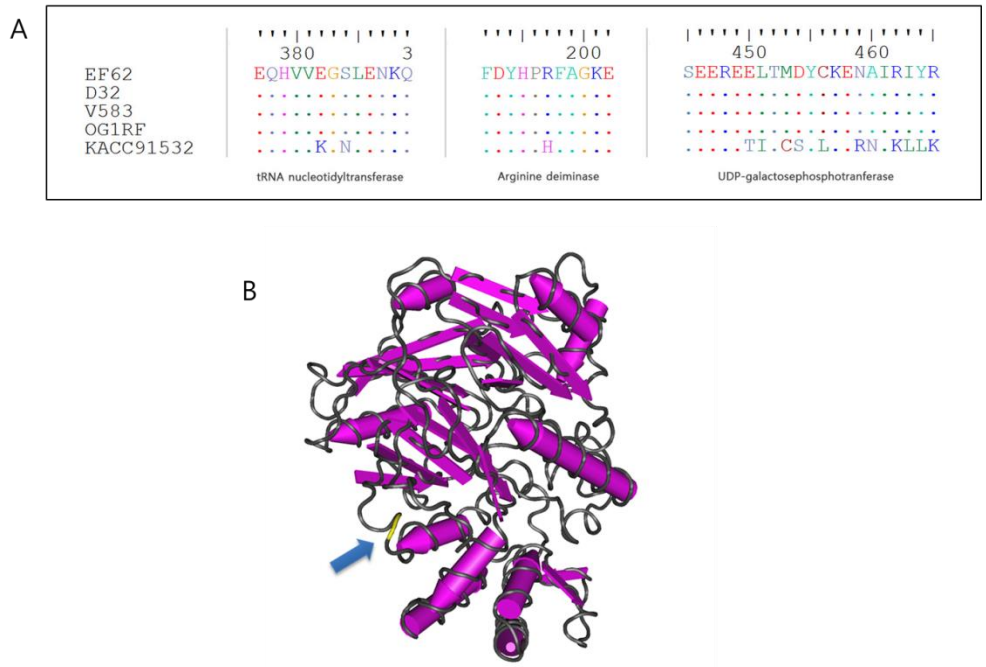

Figure 2 - 6. E. faecalis KACC 91532-specific amino acid variants in evolutionarily accelerated genes.

(A) Amino acid variant in tRNA-nucleotidyltransferase, arginine deiminase, and UDP-galactosephosphotransferase. (B) Location of the variant in arginine deiminase.

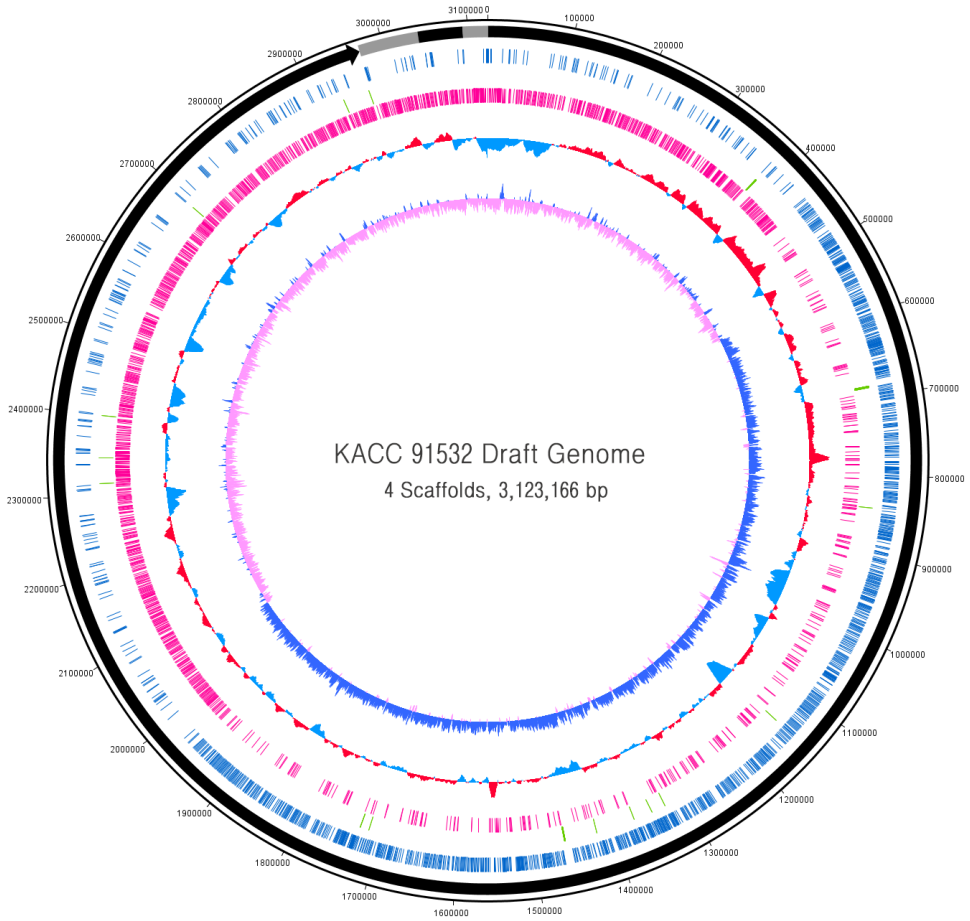

Figure 2 - 7. E. faecalis KACC 91532 genome assembly map generated using DNAPlotter [5]. From outside to inside, tracks describe (i) four scaffolds, (ii) CDS on forward strand in blue, (iii) RNA genes in green, (iv) CDS on reverse strands in pink, (v) GC content, and (vi) GC skew.

Chapter 3. Minke whale genome assembly using low coverage whole genome re-sequencing data.

3.1 Introduction

Cetaceans are a group of marine mammals with a history of transition from land to ocean. This group includes whales, dolphins and porpoises which have unique adaptations to the aquatic environment. While the complete picture of the Cetacean evolutionary history is unknown, it is thought that the transition to sea occured about 50 MYA. The order Cetacea includes close to 85 species divided into two suborders of Mysticeti, the Baleen Whales, and the Odontoceti, the Toothed Whales. The common minke whale is the second smallest Baleen whale and is widely distributed in both Southern and Northern hemispheres. It is predicted that in order for common minke whales to survive in the aquatic environment, it has characteristic evolutionary traits. For example, while the minke whale lives in the ocean it breathes through the lungs. With this one breath, the minke whale can sustain itself for much longer than terrestrial mammals. To sustain itself in a hypoxic state, the minke whale is thought to have developed a mechanism to deal with problems that arises with such a condition. Another interesting characteristic of the whale is the exceptionally low cancer rates in comparison to humans or mice despite the larger body size. This suggests that cancer defense mechanisms are well-developed in the whales. By understanding this adaptation in the whales, it would be

possible to gain insights into cancer defense to aid cancer treatment in humans. There are a number of different ways to study these mechanisms, however, there are difficulties with using a biological experimentation approach. First, experiments are not only time consuming but also requires a lot of resources. Second, there are limitations in obtaining specimens as whaling is currently banned for minke whales. With these difficulties, molecular genetics research using genomic analysis is an effective tool for studying traits specific to minke whales. In this study, for the minke whale, which lacks a reference genome, we performed a *de novo* assembly using the genomic data generated by NGS. Then, by combining the assembly results of multiple individuals, a reference assembly was successfully created. Using the reference assembly, gene prediction analysis was conducted to present basic information for future research including comparative genomics.

3.2 Material and Methods

From the National Fisheries Research & Development Institute (NFRDI), Korea, four muscle samples of minke whales were obtained. These samples were donated to NFRDI after being accidentally caught by fishing net in Hypo, Ganggu, Pohang and off the coast of Korea.

DNA was extracted from the each sample of 4 whales using the Illumina whole genome sequencing protocol. Then, paired-end libraries were constructed for each sample with insert sizes of 270 bp and 480 bp. Illumina Hiseq 101 cycle paired-end sequencing was performed to obtain genomic sequences. Figure 3 – 1 shows the overall process of assembly and gene prediction. To check the quality of the obtained read data, FastQC was used and to exclude sequencing errors, the error-correction module of Allpaths-LG was used. Reads of each sample were merged into one shuffled-form fasta file with the N bases filtered. These reads were assembled using IDBA_UD(Peng, Leung et al. 2012) with the following options: pre-correction and kmin=40. The resulting gaps in the sequences were filled using Gapcloser with parameter k value of 31. Then, the genome assembly for the S30 sample was performed using CLC for the comparing results. Assembly with the following parameters: minimum contig lengths of 2000, similarity 0.85, length fraction 0.5, insert cost 3, deletion cost 3, and mismatch cost 2. RepeatMasker(Tarailo-Graovac and Chen 2009) with the options mammal species, no-low and no-is, the repeats in the sequence were screened. Gene prediction was carried out using Augustus(Stanke, Diekhans et al. 2008) and the results were used as input for a BLASTP(Altschul, Madden et al. 1997) search. The BLASTP search

was filtered with the following options: peptide length of more than 100 amino acids and gene coverage of over 70% without gaps. The assembly results from each sample were combined to maximize the gene contents and make a representative assemble of the minke whale genome. The predicted gene sequences that passed the filtering step were combined and designed at the BLASTP DB. The gene sequences of each sample were checked against the BLASTP DB and the results were filtered with the following parameters: identity > 95%, 70% < q.cov and s.cov < 130. The contigs from each sample were grouped in to one of the following: 1. contigs without genes, 2. contigs with sample-specific genes, 3. contigs with only one gene, and 4. contigs with multiple genes. The Sample S30 showed the best assembly results and so using that as a basis, the contigs of the sample-specific genes from the other three samples. Using ClustalW(Thompson, Gibson et al. 2002), multiple sequence alignments were conducted for each group. Also, the contig extension and bridging were based on the S30 contigs or in the case of absence of S30 contig, the consensus sequence. Figure 3 – 2 shows the contig extension and bridging process. Then, following the same process as above, the combined genome sequence was masked and gene prediction was carried out. The short reads were

mapped to the combined assembled genome using Bowtie2(Langmead and Salzberg 2012) with the default option.

3.3 Results

The details of raw read data is described in Table 3 – 1. Average depth was about 15 X compared to the approximately estimated mink whale genome size. N contents was 2.13% n average and S35 sample had high GC ratio (42.4%) compared to other samples. The result of assembly of each sample was described in Table 3 – 2. Input reads were error corrected using the in dependent module of Allpath-LG. The sequences shorter than 2,000bp were filtered. S30 had small number of contigs(262,747) and maximum contig length(105,339). N50 length was 10,321bp for S30, 5,359bp for S34, 4,236bp for S35 and 7,810bp for S37. The result of S30 assembly showed longest N50 length. Total residue count was showed deviation between samples. Total residue count of S30 was 2,010,222,571bp and it was almost two times bigger than S35 which showed the smallest total residue counts (1,126,905,396). Total residue count of S30 accounted for 67% of estimated minke whale genome size (3Gbp). Genome assembly of S30 showed better assembly result among 4 samples based on the Maximum contig length, N50 length, N contents and Coverage. Table 3

– 3 shows the comparing result of the genome assembly of S30 based on combination of open source programs and commercial program named clc assembly cell. The assembly result based on the combination of open source program had longer sequence length, less N bases and more coverage than the result of commercial program. Especially in N contents, open source based assembly show much smaller N contents compared to CLC assembly cell. Using RepeatMasker, various repeat elements identified such as Sine, Line and etc. The details of the RepeatMasker are in Table 3 - 4. Identified numbers of repeat element of S35 was smaller than other samples. Line element had the largest proportion in every sample and the proportion of Sine and LTR element was similar. On the whole, the composition of repeat element using RepeatMasker showed similar pattern in three samples (S30, S34 and S37). The gene prediction results using Augustus from masked genome sequence of each sample is in Table 3 - 5. S30 sample showed the largest number of predicted genes. Based on the result of gene prediction, we classified assembled contigs of each sample under 4 categories. The result of classification is described in Table 3 – 6. After merging, extension and bridging process based on S30 genome assembly with 3 other samples, the combined genome assembly of minke whale was build. The combined genome assembly showed same

maximum length with S30 genome assembly. N50 length and average length were slightly increased to 10,400bp and 7,727bp. The genome coverage was increased from 67.0% to 73.7% based on estimated minke whale genome size (3Gb). Summary statistic of combined minke whale genome assembly is in Table 3 – 7 and repeat elements of combined genome are described in Table 3 - 8. The number of repeat elements in combined assembly was generally increased with increasing number of contigs. Figure 3 - 3 shows the comparing result of comparing 3 different version of assembly.

Table 3 - 9 shows the results of short read mapping using Bowtie2 to the combined assembly and Figure 3 – 4 shows the comparing result of short read mapping to 3 different version of assembly. Combined genome assembly showed almost 89% mapping rate in every sample and combination of open source programs showed higher alignment rate compared to assembly of clc assembly cell.

3.4 Discussion

Combined minke whale genome assembly using combination of various open source program showed better results than assembled genome using clc assembly cell in the maximum length, N50 length,

mapping the raw read and Gaps. It shows the nature of *de novo* assembly using NGS short read data. There are numerous assemblers available for NGS data, but the results of assembly are very different between each other. The result of assembly showed in this study is one of many assembly trials using combination of various programs and one specific assembly pipeline or program doesn't always show the best result. Because of this reason, researchers have to find proper combination of pipeline for each NGS data. Combining the individual assembly base on the gene contents using contig extension and bridging showed that it can be one of efficient way to merge the assembly result. S35 sample showed the different pattern in identified repeat elements compared to other samples. To identify the reason of this result, read mapping coverage of S35 sample to combined minke whale genome can be an answer to identify the reason of this.

Table 3 - 1. Sequencing result of 4 Mink Whale Samples.

Sample Name	Insert Size	Total Base (bp)	Depth (X)	Read Count	N (%)	GC (%)	Q20 ratio(%) /depth(X)	Q30 ratio(%) /depth(X)
S30	270bp	51,959,636,648	17.32	514,451,848	1.98	39.89	92 / 15.9	87 / 15.0
S34	270bp	40,610,199,180	13.54	402,081,180	2.54	39.72	92 / 12.3	87 / 11.8
S35	480bp	47,488,854,074	15.83	470,186,674	1.82	42.4	90 / 14.2	83 / 13.1
S37	480bp	40,666,301,650	13.56	402,636,650	2.17	40.98	89 / 12.1	82 / 11.1
Total		180,724,991,552	60.25	1,789,356,352	2.13	40.75	90 / 54.5	85 / 51.0

* Estimated mink whale genome size : 3Gb

* Fastq Quality Encoding : Sanger Quality (ASCII Character Code = Phred Quality Value + 33

Table 3 - 2. The result summary of minke whale genome assembly.

Sample name	S30	S34	S35	S37
Number of contig	262,747	313,490	282,736	294,439
Sequence length				
Minimum length	2,000	2,000	2,000	2,000
Maximum length	105,339	61,954	53,070	68,378
Average length	7,651	4,719	3,985	6,286
N50 length	10,321	5,359	4,236	7,810
Residue information				
Total residue count(bp)	2,010,222,571	1,479,651,607	1,126,905,396	1,851,132,035
N contents	17,875	303,142	4,484,941	2,988,329
Closed N by Gapcloser	15,243	239,309	727,324	1,298,712
GC content (%)	40.51	38.58	43.21	40.59

* Minimum cut off contig length = 2000bp.

Table 3 - 3. Result summary of S30 minke whale genome assembly.

Statistic	IDBA_UD	CLC
Number of Contig	262,747	266,806
Sequence Length		
Minimum length	2,000	2,000
Maximum length	105,339	92,838
Average length	7,651	6,782
N50 length	10,321	8,650
Residue information		
Total residue count (bp)	2,010,222,571	1,809,558,425
N content	17,875(0.00%)	14,452,650 (0.8%)
GC content (%)	40.49%	40.51%

* Minimum cut off contig length = 2000bp, OS – open source program, bowtie2 conducted with default option.

Table 3 - 4. The RepeatMasker result summary of 4 samples

Sample name	S30		S34		S35		S37	
Elements	No.	Length	No.	Length	No.	Length	No.	Length
Sine	1,015,281	150,800,833 (6.77%)	742,339	105,477,583 (6.31%)	123,993	17,666,179 (1.66%)	861,733	125,495,334 (6.01%)
Line	1,280,241	456,915,337 (20.52%)	1,017,906	344,636,222 (20.61%)	186,869	59,040,424 (5.54%)	1,203,118	415,185,547 (19.87%)
LTR elements	455,514	144,478,462 (6.49%)	369,818	110,307,049 (6.60%)	74,671	24,543,226 (2.30%)	435,092	139,417,009 (6.67%)
DNA elements	369,942	78,295,071 (3.52%)	292,067	60,486,743 (3.62%)	56,988	11,114,767 (1.04%)	358,544	74,122,957 (3.55%)
Unclassified	5,055	956,831 (0.04%)	4,024	751,816 (0.04%)	771	134,971 (0.01%)	4,968	927,640 (0.04%)
Small RNA	5,272	575,413 (0.03%)	3,936	429,088 (0.03%)	787	84,733 (0.01%)	4,931	547,234 (0.03%)
Satellites	183,837	62,416,759 (2.80%)	140,168	46,115,494 (2.76%)	29,359	10,183,045 (0.96%)	169,308	55,462,818 (2.65%)

Table 3 - 5. Summary of gene prediction results using Augustus.

	S30	**S34**	**S35**	**S37**
Number of genes	41,098	28,348	30,383	37,422
Number of Exons	131,164	62,167	73,950	113,794
Total gene length	229,649,018	82,573,850	84,286,704	178,879,019
Average length	5587.84	2912.86	2774.14	4780.05

* Over 100 peptide length and 70% gene coverage.

Table 3 - 6. The contig classification result of four samples.

Source	30	35	34	37
Number of contig	628,081	938,541	994,603	668,188
Number of gene	41,098	30,383	28,348	37,422
Contig with gene	27,635	18,268	14,849	24,506
Contig with one gene	14,285	11,065	8,057	14,296
Contig with multi gene	600	542	212	637
Contig with sample specific gene	12,750	6,661	6,580	9,573
Contig without gene	600,446	920,273	979,754	643,682

Table 3 - 7. The result summary of combined minke whale genome assembly

Source	Value
Number of Contig	286,129
Sequence Length	
Minimum length	2,000
Maximum length	105,339
Average length	7,727
N50 length	10,400
Residue information	
Total residue count (bp)	2,211,014,055
N content	524,763(0.02%)
GC content (%)	40.92%

* Minimum cut off contig length = 2000bp

Table 3 - 8. The result summary of repeat masking using RepeatMask.

Source	Number of elements	Length occupied (%)
Total Length	-	2,226,669,555
Sine	1,015,281	150,800,833 (6.77%)
Line	1,280,241	456,915,337 (20.52%)
LTR elements	455,514	144,478,462 (6.49%)
DNA elements	369,942	78,295,071 (3.52%)
Unclassified	5,055	956,831 (0.04%)
Small RNA	5,272	575,413 (0.03%)
Satellites	183,837	62,416,759 (2.80%)
Total Masked		893,922,896(40.15%)

Table 3 - 9. The result summary of read mapping using Bowtie2

Categories	Samples			
	30	**34**	**35**	**37**
Total number of reads	514,451,848 (100%)	402,081,180 (100%)	470,186,674 (100%)	402,636,650 (100%)
Concordantly 1 time	290,495,048 (56.4%)	211,464,096 (52.5%)	236,048,842 (50.2%)	162,187,448 (40.2%)
Concordantly > 1 time	85,699,662 (16.6%)	76,771,800 (19.0%)	79,939,174 (17.0%)	34,242,644 (8.5%)
Discordantly 1 time	14,322,176 (2.7%)	6,252,922 (1.5%)	19,396,744 (4.1%)	55,572,032 (13.8%)
1 time in mixed mode (single reads)	40,957,298 (7.9%)	33,725,380 (8.3%)	33,216,429 (7.0%)	49,889,813 (12.3%)
> 1 time in mixed mode (single reads)	31,097,230 (6.0%)	25,804,959 (6.4%)	54,233,679 (11.5%)	52,303,191 (12.9%)
Total include singletone	462,571,414 (89.9%)	354,019,157 (88.0%)	422,834,868 (89.9%)	354,195,128 (87.9%)

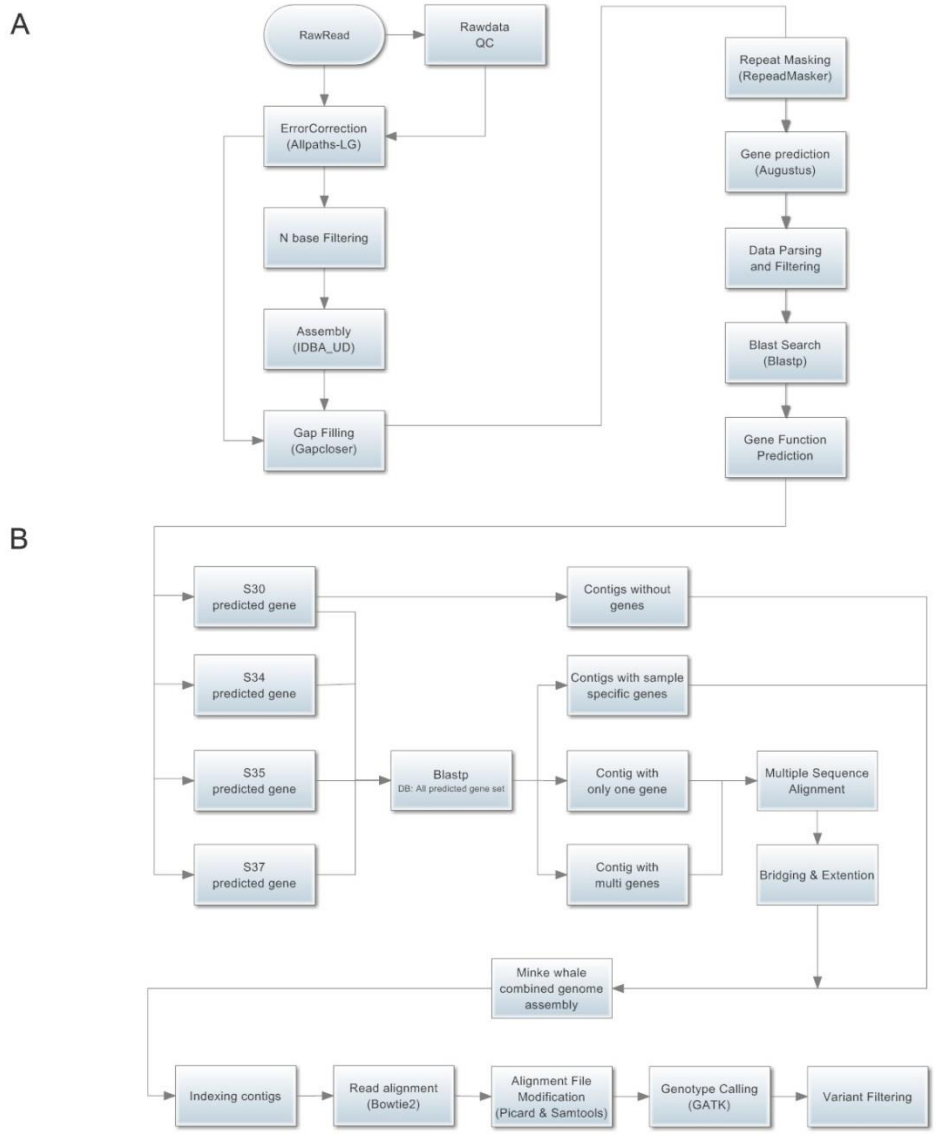

Figure 3 - 1. Overall Process of assembly, gene prediction and variant calling.

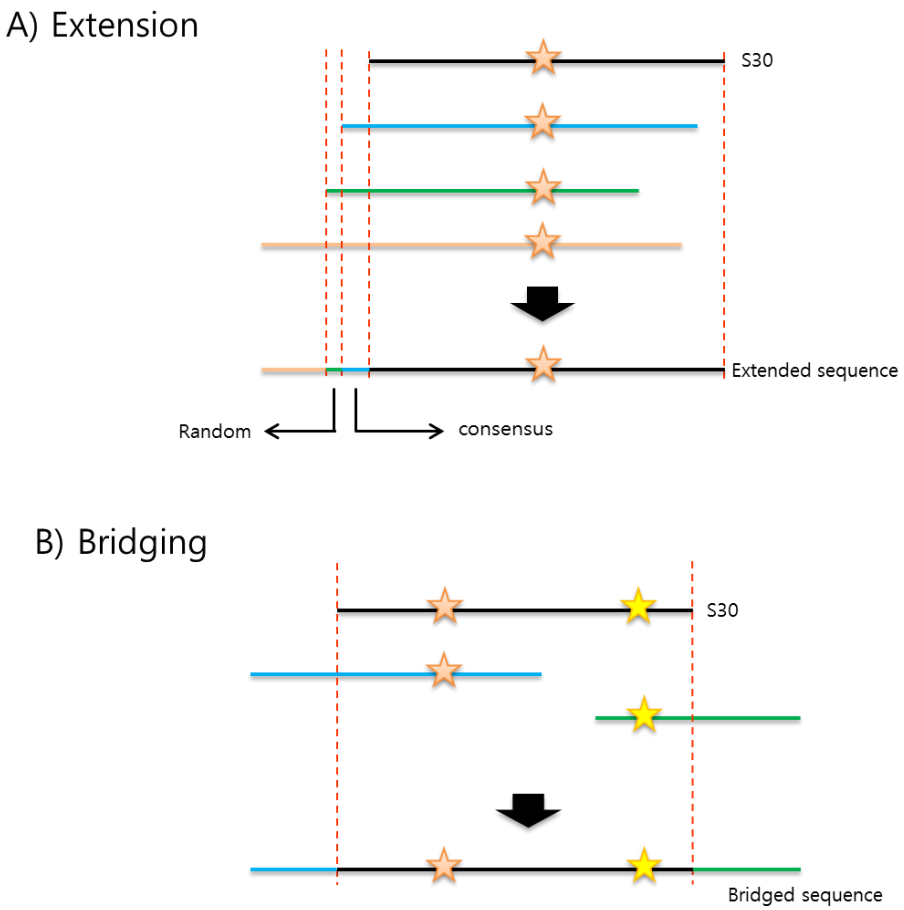

Figure 3 - 2. Process of contig extention and bridging.

A. Process of Extension. Star shows same gene in the cluster.

B. Process of bridging.

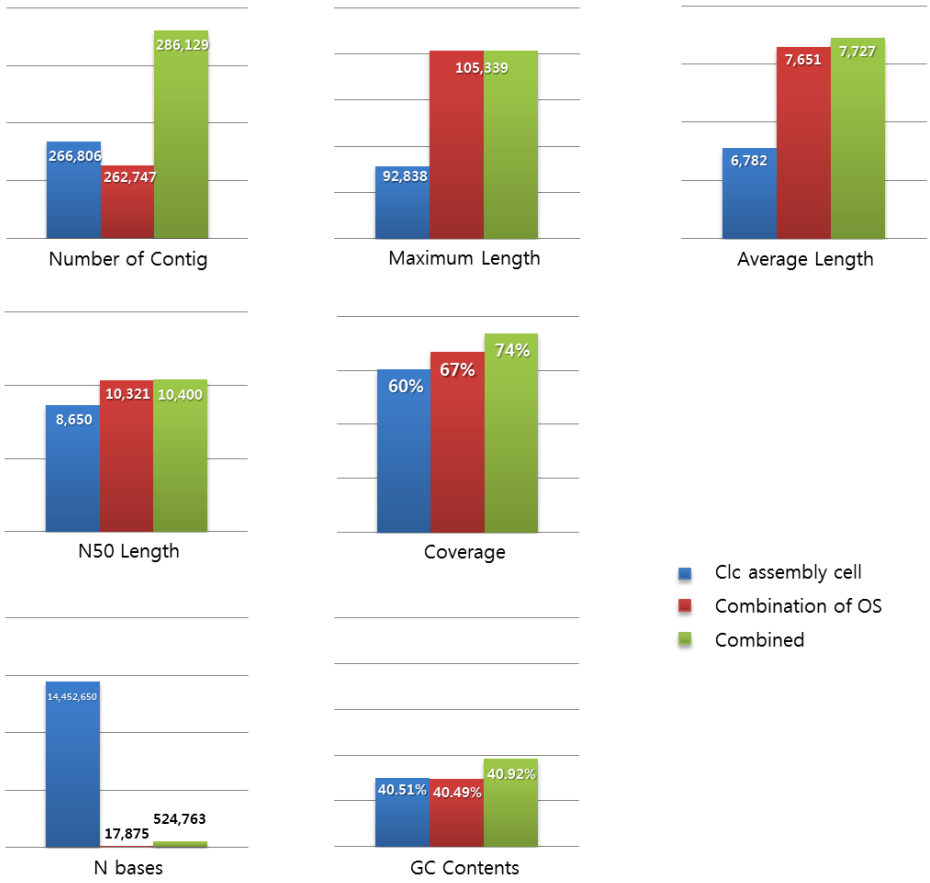

Figure 3 - 3. Comparing result of 3 assembly (Clc assembly cell, Combination OS, Combined)

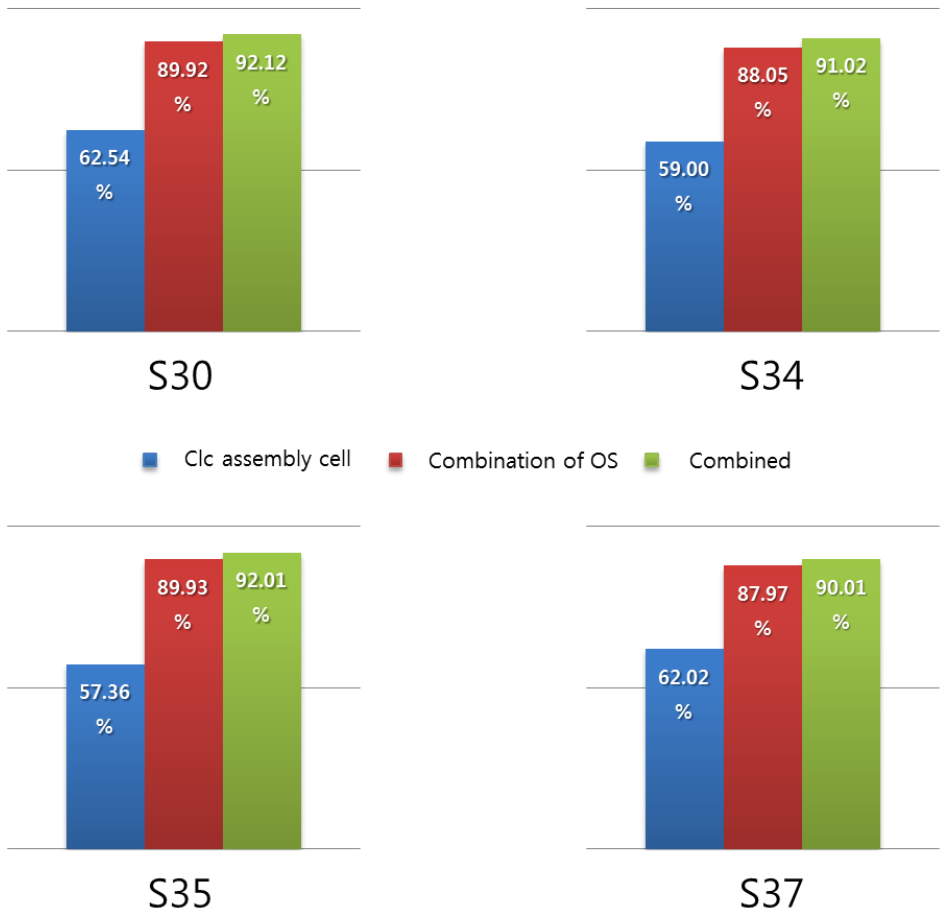

Figure 3 - 4. The result summary of read mapping to three assembled genome using Bowtie2.

Chapter 4. Korean Native Chicken genome assembly based on unaligned reads of whole genome re-sequencing

4.1 Introduction

For a country that imports large amounts of biological resources, the Korea Native Chicken (KNC) is an important genetic resource. Though KNC grows slower than the commercial broiler breeds, it is preferred by Korean people for tis taste and nutrient content.(Liu, Jayasena et al. 2012) This breed underwent severe decline in genetic diversity during the second World War and the Korean War. Furthermore, many different foreign chicken breeds were imported into Korea during the Japanese occupation and the recovery period. The resulting hybridization has further endangered the survival of KNC. The National Institute of Animal Science of Korea has recently conducted a KNC restoration project to recover the lost endemic genetic resources. The restored KNC consists of five strains identified by its plumage color of red, yellow, black, grey, and white. To understand the genetic traits of KCN, its genome was sequenced using NGS techniques. Then, through re-sequencing process based on the published reference genome (*Gallus gallus* 4), SNVs, InDel, and various SVs were obtained. However, re-sequencing has limitations due to its dependence on the comparison with the reference genome. For example, there may be regions in the newly sequenced genome that is very different to the reference genome, or the region may not exist in the reference genome.

Therefore, it is helpful to take the reads that do not map to the reference genome and perform a *de novo* assembly to find KNC specific genes or sequences. A problem that arises from this approach is that the assembly obtained from unaligned reads is not from a library designed for *de novo* assembly and has insufficient coverage. In this study, we focused on efficiently creating an assembly with low coverage unaligned reads post re-sequencing and identifying KCN specific genes.

4.2 Material and methods

Five different Korean native chicken strains, each with a different plumage color, was used in this study. From five male Korean native chickens of different strains, blood samples were collected. DNA was extracted from the samples following the Illumina's whole genome sequencing protocol. Genomic DNA library was constructed for each sample using Truseq Kit of Illumina. HiSeq 2000 101 cycle paired-end sequencing was performed on each sample and the resulting reads were checked for quality using FastQC. Using the error correction module in the program Allpaths-LG, sequencing errors were excluded to yield error corrected pair-end reads for each sample. Each read of samples was merged into a single shuffled form fasta file and N bases in the reads were filtered. The error-corrected reads were assembled using

IDBA_UD using pre_correction option and using Gapcloser, all the gaps in the assembled sequences were filled with parameter k value of 31.

To detect unique sequences for each strain of the Korean Native Chicken, the unaligned reads from the Bowtie2 mapping step were mapped to the assembled scaffolds. For gene prediction, scaffolds with lengths longer than 2000 bp and average read depth of over 10 were used. RepeatMasker was used on the scaffolds with the following options: nolow and no_is. On the repeat masked sequence, Augustus was implemented and the resulting output was used as input for Blastp search with the NR database. The blast search result was filtered excluding peptides of less than 1000 amino acids and a gene coverage less than 70%.

4.3 Results

Table 4 – 1 shows the result of short reads mapping of each sample using bowtie2. Average overall alignment rate was over 97%. Almost 87% of paired-end read properly mapped to the reference genome with estimated insert size. Figure 4-2 shows the insert size distribution of 5 KNC samples. The average insert size of 5 KNC was 268.1 and the

average standard deviation was 22.4. Discordantly mapped reads which can be inferred to structure variation was under 3%. The proportions of mapped single tone in each sample were also under approximately 2%. Table 4 – 2 shows the assembly result of each sample using IDBA_UD. Same assembly pipeline used in minke whale genome assembly was also used for KNC. KNC_G(53) sample showed the longest maximum contig length, average contig length and N50 length among samples. Assembled total residue of samples was approximately 0.95G which covers approximately 90% of the reference genome size. KNC_L(40) showed the smallest number of N bases(29,649). GC content ratio of each sample was approximately 41% which is smaller than the reference genome (41.9%). Unaligned read mapping to the assembled genome is described in Table 4 – 3. Under 30% of unaligned read were mapped to the assembled contigs. Figure 4 – 3 shows the scatter plots of the length of assembled contigs and the number of read mapped to each contigs before and after filtering with an average read depth coverage of 10. Gene prediction using the filtered contigs which have sufficient mapped read coverage was conducted and the number of predicted genes for each KNC sample is as follows: 74 for KNC_R(16), 82 for KNC_Y(24), 105 for KNC_W(3), 91 for KNC_L(40) and 112 for KNC_G(53). Among these predicted genes, 8 genes were common

and Table 4 – 4 showed the list of commonly predicted genes in KNC. 3 genes (maltase-glucoamylase, carbohydrate response element binding protein variant, and beta-keratin-realeted protein) are related to sugar metabolism.

4.4 Discussion.

Overall alignment rates of 5 KNC samples were almost 98% and this is higher than ordinary alignment rate using other species. Especially the proportions of reads classified as concordantly > 1 time, discordantly 1 time, and single tone were low and these reads are generally related to the repeat region or structural variation in the reference genome. For the assembly of unaligned reads, I used not unaligned read but whole raw data because of the nature of error correction process conducted before de novo assembly. Many de novo assembly programs conduct error correction based on the information of K-mer frequency distribution. If only unaligned reads were used for error correction, the K-mer frequency distribution can be biased and reads coverage cannot be enough. Therefore, I conducted assembly based on the raw data and mapped unaligned read to the assembled contigs to find the contigs using unaligned read. However, the contigs with low coverage or gaps

had to be filtered to identify the contigs of unaligned reads. Generally the result based on the raw data shows longer assembly statistics compared to the assembly of exacted unaligned reads. Further analysis is necessary for common genes, contigs and sample specific contigs to identify the unique genome features of KNC breed.

Table 4 - 1. The result summary of read mapping using Bowtie2.

Categories	Samples				
	KNC_W	KNC_R	KNC_Y	KNC_L	KNC_G
Total number of reads	367,190,708	350,530,416	369,509,968	354,548,474	375,424,656
Concordantly 1 time	318,853,898 (86.84%)	306,401,382 (87.41%)	324,791,532 (87.90%)	305,782,626 (86.25%)	329,086,008 (87.66%)
Concordantly > 1 time	23,125,556 (6.30%)	20,814,642 (5.94%)	23,372,912 (6.33%)	22,478,014 (6.34%)	23,615,846 (6.29%)
Discordantly 1 time	9,448,684 (2.57%)	7,738,304 (2.21%)	5,992,216 (1.62%)	9,898,092 (2.79%)	4,431,962 (1.18%)
aligned exactly 1 time (single tone)	4,688,679 (1.28%)	4,719,392 (1.35%)	4,379,557 (1.19%)	4,972,690 (1.40%)	5,355,669 (1.43%)
aligned > 1 time (single tone)	2,221,074 (0.60%)	1,910,112 (0.54%)	1,815,259 (0.49%)	2,196,368 (0.62%)	1,857,673 (0.49%)
Overall alignment rate	358,341,410 (97.59%)	341,591,890 (97.45%)	360,346,120 (97.52%)	345,330,212 (97.40%)	364,349,628 (97.05%)

Table 4 - 2. The result summary of Korean Native Chicken genome assembly using IDBA_UD.

Sample name	KNC_R(16)	KNC_Y(24)	KNC_W(3)	KNC_L(40)	KNC_G(53)
Number of contig	59,399	61,337	57,632	61,413	55,873
Sequence length					
Minimum length	2,000	2,000	2,000	2,000	2,000
Maximum length	287,034	279,299	374,701	366,767	403,774
Average length	16,034	15,526	16,668	15,570	17,240
N50 length	27,072	25,802	28,281	25,807	29,589
Residue information					
Total residue(bp)	952,410,824	952,354,786	960,613,709	956,228,021	963,288,513
N contents	56,857	38,874	41,275	29,649	44,072
Closed N by Gapcloser	140,521	100,621	103,203	65,194	106,207
GC content (%)	40.90%	40.89%	41.08	40.99%	41.13%

Table 4 - 3. The result summary of remapping unaligned reads to assembled genome using bowtie2

Categories	Samples				
	KNC_R(16)	**KNC_Y(24)**	**KNC_W(3)**	**KNC_L(40)**	**KNC_G(53)**
Total number of reads	11,657,196	10,272,762	12,605,627	13,143,917	11,361,401
Concordantly 1 time	1,953,151 (16.75%)	2,134,380 (20.00%)	2,089,660 (16.58%)	1,950,721 (14.84%)	2,315,587 (20.38%)
Concordantly > 1 time	741,837 (6.36%)	919,732 (8.62%)	865,203 (6.86%)	822,836 (6.26%)	1,017,708 (8.96%)
Discordantly 1 time	109,080 (1.22%)	83,377 (1.09%)	92,372 (0,.96%)	108,375 (1.05%)	132,314 (1.65%)
Overall alignment rate	24.05%	29.40%	24.17%	21.93%	30.50%

Table 4 - 4. The list of commonly predicted genes from assembled scaffolds among 5 Korean Native Chicken.

Predicted gene ID	Predicted Gene Name
gi|118095337|ref|XP_422811.2|	PREDICTED: maltase-glucoamylase, intestinal [Gallus gallus]
gi|157783877|gb|ABV72703.1|	carbohydrate response element binding protein variant 2 [Gallus gallus]
gi|326931845|ref|XP_003212034.1|	PREDICTED: thyrotropin-releasing hormone receptor-like [Meleagris gallopavo]
gi|363728185|ref|XP_416447.3|	PREDICTED: heat shock transcription factor, X-linked-like [Gallus gallus]
gi|363736543|ref|XP_003641728.1|	PREDICTED: zinc finger CCCH domain-containing protein 11A-like [Gallus gallus]
gi|363746570|ref|XP_425998.3|	"PREDICTED: PHD finger protein 7-like, partial [Gallus gallus]"
gi|449280403|gb|EMC87722.1|	"hypothetical protein A306_03536, partial [Columba livia]"
gi|45382325|ref|NP_990177.1|	beta-keratin-related protein [Gallus gallus]

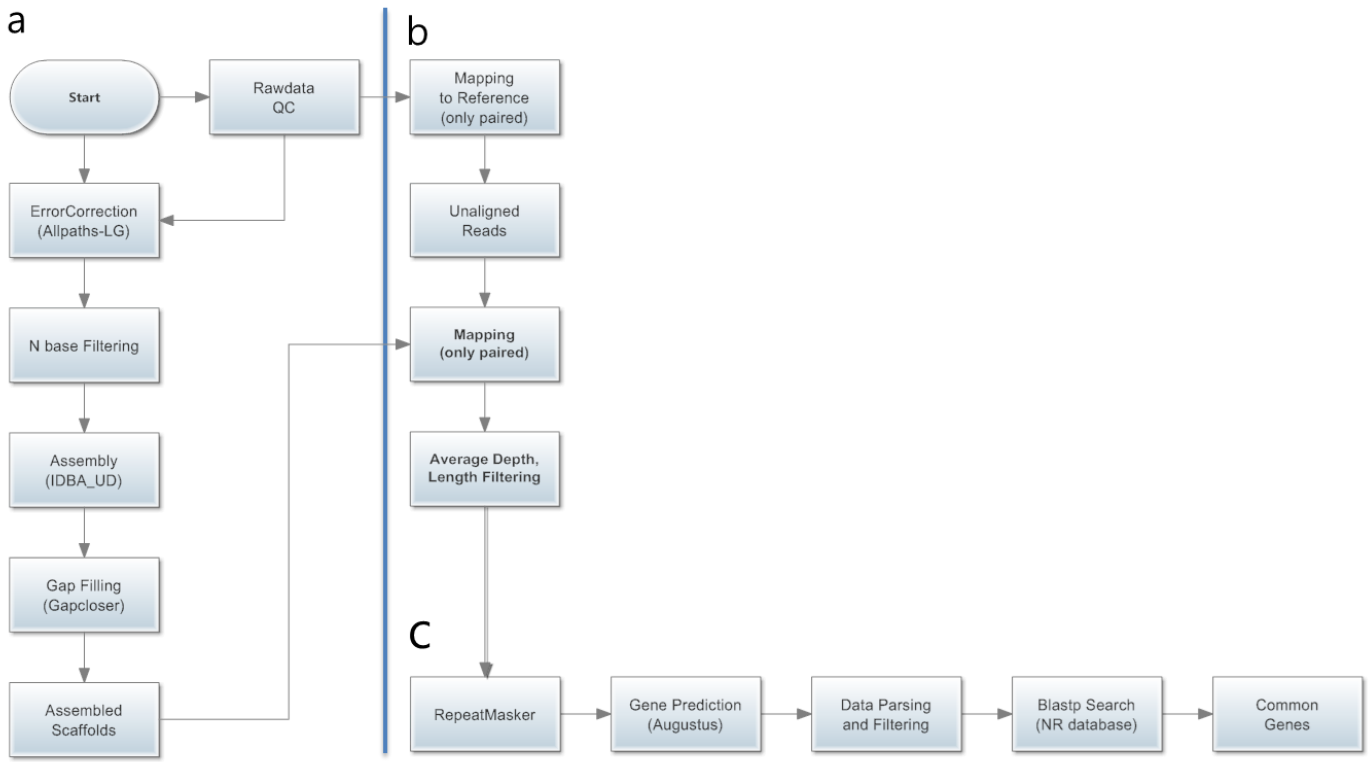

Figure 4 - 1. Overall process of genome assembly and gene prediction.

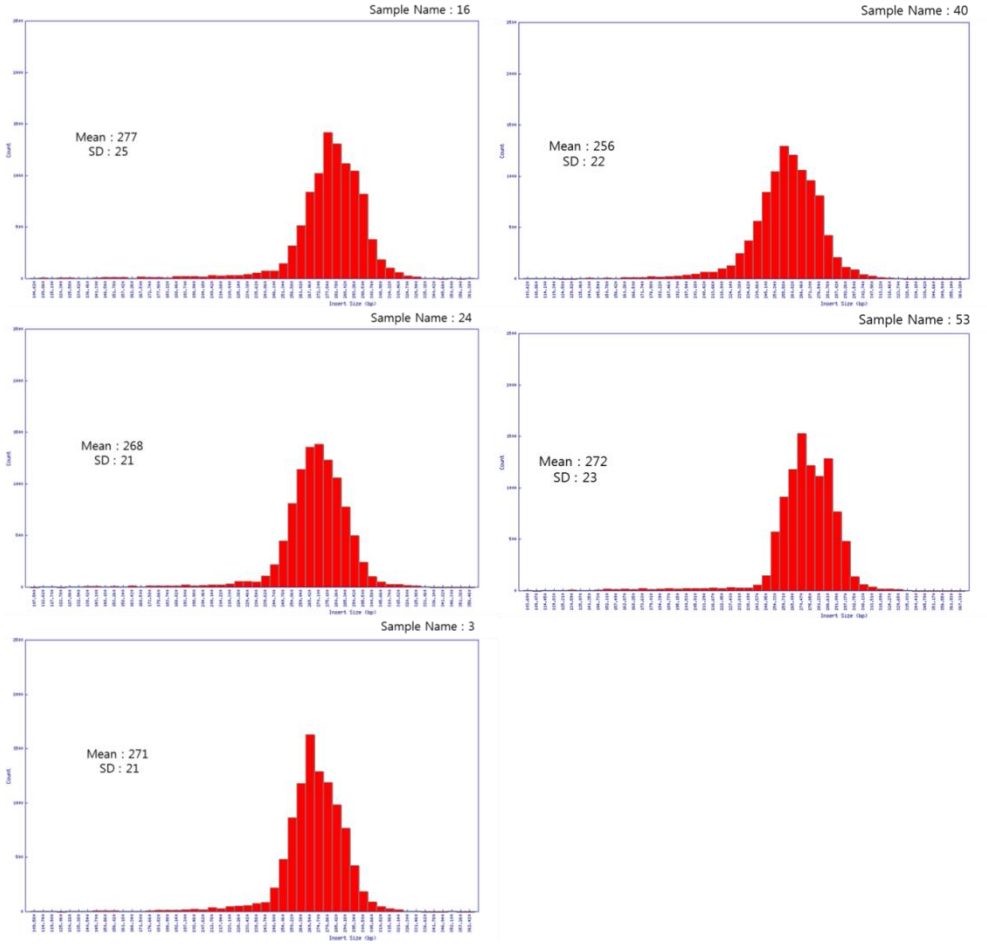

Figure 4 - 2. Insert size distributions of 5 KNC samples.

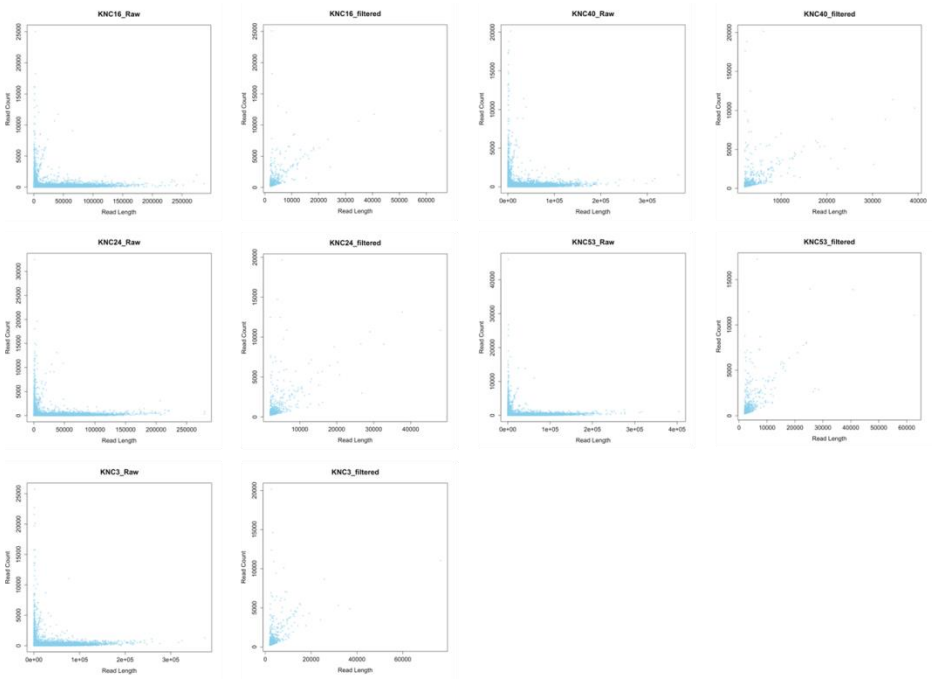

Figure 4 - 3. Scatter plot using scaffold length and read count number, before and after average read coverage filtering

References

Alekshun, M. N. and S. B. Levy (2007). "Molecular mechanisms of antibacterial multidrug resistance." Cell **128**(6): 1037-1050.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**(17): 3389-3402.

Andrews, S. (2012). "FastQC. A quality control tool for high throughput sequence data." http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc.

Aziz, R., D. Bartels, A. Best, M. DeJongh, T. Disz, R. Edwards, K. Formsma, S. Gerdes, E. Glass and M. Kubal (2008). "The RAST Server: rapid annotations using subsystems technology." BMC genomics **9**(1): 75.

Baronets, N. (2003). "Vitamin K as a stimulator of microbial growth." Zhurnal mikrobiologii, epidemiologii, i immunobiologii **4**: 104-105.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov and E. S. Lander (2002). "ARACHNE: a whole-genome shotgun assembler." Genome research **12**(1): 177-189.

Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs and X. Huang (2008). "The potential and challenges of nanopore sequencing." Nature biotechnology **26**(10): 1146-1153.

Bryant, D. W., W.-K. Wong and T. C. Mockler (2009). "QSRA–a quality-value guided de novo short read assembler." BMC bioinformatics **10**(1): 69.

Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum and D. B. Jaffe (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads." Genome research **18**(5): 810-820.

Cariolato, D., C. Andrighetto and A. Lombardi (2008). "Occurrence of virulence factors and antibiotic resistances in *Enterococcus faecalis* and *Enterococcus faecium* collected from dairy and human samples in North Italy." Food Control **19**(9): 886-892.

Castillo-Davis, C., F. Kondrashov, D. Hartl and R. Kulathinal (2004). "The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint." Genome research **14**(5): 802-811.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular Biology and Evolution **17**(4): 540.

Chaisson, M., P. Pevzner and H. Tang (2004). "Fragment assembly with short reads." Bioinformatics **20**(13): 2067-2074.

Chaisson, M. J., D. Brinza and P. A. Pevzner (2009). "De novo fragment assembly with short mate-paired reads: Does the read length matter?" Genome research **19**(2): 336-346.

Chaisson, M. J. and P. A. Pevzner (2008). "Short read fragment assembly of bacterial genomes." Genome research **18**(2): 324-330.

Chow, J., L. Thal, M. Perri, J. Vazquez, S. Donabedian, D. Clewell and M. Zervos (1993). "Plasmid-associated hemolysin and aggregation substance production contribute to virulence in experimental enterococcal endocarditis." Antimicrobial agents and chemotherapy **37**(11): 2474-2477.

Chung, J. W., B. A. Bensing and G. M. Dunny (1995). "Genetic analysis of a region of the Enterococcus faecalis plasmid pCF10 involved in positive regulation of conjugative transfer functions." Journal of bacteriology **177**(8): 2107-2117.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing." Genome research **17**(11): 1697-1706.

Eaton, T. J. and M. J. Gasson (2001). "Molecular Screening of EnterococcusVirulence Determinants and Potential for Genetic Exchange between Food and Medical Isolates." Applied and Environmental Microbiology **67**(4): 1628-1635.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid and K. C. Worley (2012). "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." PloS one **7**(11): e47768.

Feil, H., W. S. Feil, P. Chain, F. Larimer, G. DiBartolo, A. Copeland, A. Lykidis, S. Trong, M. Nolan and E. Goltsman (2005). "Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000." Proceedings of the National Academy of Sciences of the United States of America **102**(31): 11064-11069.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea and S. Sykes (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proceedings of the National Academy of Sciences **108**(4): 1513-1518.

Gordon, A. and G. Hannon (2010). "Fastx-toolkit : FASTQ/A short-reads pre-processing tools " http://hannonlab.cshl.edu/fastx_toolkit.

Hall, T. (2005). "Bioedit version 7.0. 4." Department of Microbiology, North Carolina State University.

Hancock, L. E., M. S. Gilmore, V. Fischetti, R. Novick, J. Ferretti, D. Portnoy and J. Rood (2006). "Pathogenicity of enterococci." Gram-positive pathogens(Ed. 2): 299-311.

Huang, X., J. Wang, S. Aluru, S.-P. Yang and L. Hillier (2003). "PCAP: a whole-genome assembly program." Genome research **13**(9): 2164-2170.

Jeck, W. R., J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl and C. D. Jones (2007). "Extending assembly of short DNA sequences to handle error." Bioinformatics **23**(21): 2942-2944.

Kelley, D. R., M. C. Schatz and S. L. Salzberg (2010). "Quake: quality-aware detection and correction of sequencing errors." Genome Biol **11**(11): R116.

Kim., M.-K., A. Choi., G.-S. Han., S.-G. Jeong., H.-S. Chae., A. Jang., K.-H. Seol., M.-H. Oh., D.-H. Kim. and J.-S. Ham. (2011). "Development of Probiotic Dairy Product for the Normalization of Microbial Flora in Korean Infants." Article : Development of Probiotic Dairy Product for the Normalization of Microbial Flora in Korean Infants **31**(2): 290-295.

Kozarewa, I., Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman and D. J. Turner (2009). "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes." Nature methods **6**(4): 291-295.

Löytynoja, A. and N. Goldman (2005). "An algorithm for progressive multiple alignment of sequences with insertions." Proceedings of the National Academy of Sciences of the United States of America **102**(30): 10557.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Lempiäinen, H., K. Kinnunen, A. Mertanen and A. Wright (2005). "Occurrence of virulence factors among human intestinal enterococcal isolates." Letters in applied microbiology **41**(4): 341-344.

Li, R. (2009). "Short Oligonucleotide Analysis Package: SOAPdenovo 1.03." Beijing Genomics Institute, Beijing.

Li, R., H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan and K. Kristiansen (2010). "De novo assembly of human genomes with massively parallel short read sequencing." Genome research **20**(2): 265-272.

Liu, X. D., D. D. Jayasena, Y. Jung, S. Jung, B. S. Kang, K. N. Heo, J. H. Lee and C. Jo (2012). "Differential Proteome Analysis of Breast and Thigh Muscles between Korean Native Chickens and Commercial Broilers." ASIAN-AUSTRALASIAN JOURNAL OF ANIMAL SCIENCES **25**(6): 895-902.

Loman, N. J., C. Constantinidou, J. Z. M. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson and M. J. Pallen (2012). "High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity." Nature Reviews Microbiology.

Lomovskaya, O. and W. Watkins (2001). "Inhibition of efflux pumps as a novel approach to combat drug resistance in bacteria." Journal of molecular microbiology and biotechnology **3**(2): 225-236.

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

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen and Z. Chen (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martín, R., E. Jiménez, M. Olivares, M. Marín, L. Fernández, J. Xaus and J. Rodríguez (2006). "*Lactobacillus salivarius* ECT 5713, a potential probiotic strain isolated from infant feces and breast milk of a mother–child pair." International journal of food microbiology **112**(1): 35-43.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proceedings of the National Academy of Sciences **74**(2): 560-564.

Mercenier, A., S. Pavan and B. Pot (2003). "Probiotics as biotherapeutic agents: present knowledge and future prospects." Current pharmaceutical design **9**(2): 175-191.

Metzker, M. L. (2009). "Sequencing technologies—the next generation." Nature Reviews Genetics **11**(1): 31-46.

Miller, J. R., A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry and G. Sutton (2008). "Aggressive assembly of pyrosequencing reads with mates." Bioinformatics **24**(24): 2818-2824.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert and K. A. Remington (2000). "A whole-genome assembly of Drosophila." Science **287**(5461): 2196-2204.

Nadalin, F., F. Vezzi and A. Policriti (2012). "GapFiller: a de novo assembly approach to fill the gap within paired reads." BMC Bioinformatics **13**(Suppl 14): S8.

Ochman, H. and N. A. Moran (2001). "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis." Science **292**(5519): 1096-1099.

Pavlova, S., A. Kilic, S. Kilic, J. S. So, M. Nader-Macias, J. Simoes and L. Tao (2002). "Genetic diversity of vaginal lactobacilli from women in different countries based on 16S rRNA gene sequences." Journal of applied microbiology **92**(3): 451-459.

Peacock, C. S., K. Seeger, D. Harris, L. Murphy, J. C. Ruiz, M. A. Quail, N. Peters, E. Adlem, A. Tivey and M. Aslett (2007). "Comparative genomic analysis of three Leishmania species that cause diverse human disease." Nature genetics **39**(7): 839-847.

Penders, J., C. Thijs, C. Vink, F. F. Stelma, B. Snijders, I. Kummeling, P. A. van den Brandt and E. E. Stobberingh (2006). "Factors influencing the composition of the intestinal microbiota in early infancy." Pediatrics **118**(2): 511-521.

Peng, Y., H. C. Leung, S.-M. Yiu and F. Y. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Pevzner, P. A., H. Tang and M. S. Waterman (2001). "An Eulerian path approach to DNA fragment assembly." Proceedings of the National Academy of Sciences **98**(17): 9748-9753.

PLOTREE, D. and D. PLOTGRAM (1989). "PHYLIP-phylogeny inference package (version 3.2)."

Reuven, N. B., Z. Zhou and M. P. Deutscher (1997). "Functional overlap of tRNA nucleotidyltransferase, poly (A) polymerase I, and polynucleotide phosphorylase." Journal of Biological Chemistry **272**(52): 33255-33259.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew and M. Edwards (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.

Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." Nature biotechnology **26**(10): 1117-1124.

Ryan, K. J. and C. G. Ray (2010). Sherris medical microbiology, McGraw Hill Medical.

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher and M. Roberts (2012). "GAGE: A critical evaluation of genome assemblies and assembly algorithms." Genome research **22**(3): 557-567.

Salzberg, S. L., D. D. Sommer, D. Puiu and V. T. Lee (2008). "Gene-boosted assembly of a novel bacterial genome from very short reads." PLoS computational biology **4**(9): e1000186.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences **74**(12): 5463-5467.

Sarkar, S. (2008). "Effect of probiotics on biotechnological characteristics of yoghurt: A review." British Food Journal **110**(7): 717-740.

Schlievert, P. M., P. J. Gahr, A. P. Assimacopoulos, M. M. Dinges, J. A. Stoehr, J. W. Harmala, H. Hirt and G. M. Dunny (1998). "Aggregation and binding substances enhance pathogenicity in rabbit models of Enterococcus faecalis endocarditis." Infection and immunity **66**(1): 218-223.

Schmidt, H. A., K. Strimmer, M. Vingron and A. Von Haeseler (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics **18**(3): 502-504.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood (1986). "Fluorescence detection in automated DNA sequence analysis."

Solheim, M., Å. Aakra, L. G. Snipen, D. A. Brede and I. F. Nes (2009). "Comparative genomics of Enterococcus faecalis from healthy Norwegian infants." BMC genomics **10**(1): 194.

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008). "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics **24**(5): 637-644.

Tarailo-Graovac, M. and N. Chen (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences." Current Protocols in Bioinformatics: 4.10. 11-14.10. 14.

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.

Travers, K. J., C.-S. Chin, D. R. Rank, J. S. Eid and S. W. Turner (2010). "A flexible and efficient template format for circular consensus sequencing and SNP detection." Nucleic Acids Research **38**(15): e159-e159.

Voskoboynik, A., N. F. Neff, D. Sahoo, A. M. Newman, D. Pushkarev, W. Koh, B. Passarelli, H. C. Fan, G. L. Mantalas and K. J. Palmeri (2013). "The genome sequence of the colonial chordate, Botryllus schlosseri." Elife **2**.

Wang, Y., L. Y. Geer, C. Chappey, J. A. Kans and S. H. Bryant (2000). "Cn3D: sequence and structure views for Entrez." Trends in biochemical sciences **25**(6): 300.

Warren, R. L., G. G. Sutton, S. J. Jones and R. A. Holt (2007). "Assembling millions of short DNA sequences using SSAKE." Bioinformatics **23**(4): 500-501.

Wikler, M. A. (2006). Performance standards for antimicrobial susceptibility testing: sixteenth informational supplement, Clinical and Laboratory Standards Institute.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

국문초록

차세대 염기서열 분석방법을 이용한 생물 유전체의 이해

곽우리

생물정보협동과정 생물정보학전공

서울대학교 대학원 자연과학대학

이 학위논문은 차세대 염기서열 분석기술 중 2 세대에 해당하는 Illumina 의 Hiseq2000 및 Roche GS FLX 시스템을 기반으로 생성된 염기서열 데이터를 이용하여 생물체의 유전체를 효과적으로 assembly 하는 방법과 특정 생물 종의 유전체를 재구성하여 이해하는 연구들로 구성되어있다. 수평적 유전자 이동으로 인해 de novo assembly 가 기본으로 수행되는 미생물의 유전체를 Sanger sequencing 없이 assembly gap 을 효과적으로 처리할 수 있는 pipeline 을 구성하고 이를 이용하여 finished genome 수준의 assembly 를 완성하였다. 원핵생물인 유산균뿐만 아니라 생물학적으로 중요한 밍크고래 및 국내 토종품종인 재래닭과 같은 진핵생물의 유전체를 low coverage 데이터를 활용하여 재구성하는 연구 등을 수행했다.

주요어 : 차세대 염기서열 분석, 유전체 assembly,

학 번 : 2012-20410

<감사의 글>

남들보다 늦은 나이에 대한 걱정, 그리고 이 분야를 내가 과연 잘 할 수 있을까 하는 의문으로 시작된 대학원 석사 생활은 돌이켜보면 함께 연구하고 생활했던 연구팀식구들이 있어 무사히 2 년간의 석사과정을 마칠 수 있었다는 생각이 듭니다. 김부장님, 윤숙희박사님, 이보영박사님, 삼선누나, 정례누나, 규원이형, 선진이형, 다정누나, 원철이형, 영섭이형, 수연이, 형민이, 태헌이, 동현이, 재민이, 현주, 영준이, 철이, 현수, 대원이, 세윤이, 현정씨, 서희, 소진이, 수인이까지 제 졸업논문의 이 작은 페이지를 빌어 평소 전하지 못한 제 마음 속 감사의 인사를 전하고 싶습니다. 비록 저와 함께한 연구, 시간 그리고 추억들은 개개인 마다 차이는 있겠지만 연구팀 식구들이 있어 지금까지 달려올 수 있었습니다. 감사합니다. 그리고 앞으로 저와 함께하는 시간이 저 뿐만 아니라 각자에게도 훗날 소중한 추억으로 기억될 수 있도록 노력하겠습니다.

선생님. 선생님께 대한 감사함을 글로 표현하기 위해서 아무리 고민하고 또 고민하고 생각해봐도 어떻게 표현할 길 없이 그저 가슴만 먹먹해지는 것 같습니다. 어떤 미사어구와 구차한 말로 이 마음을 표현한다 한들 그저 부질없는 일이 될 터, 제가 선생님의 부끄럽지 않은 제자가 될 수 있도록 행동으로 노력하겠습니다. 감사합니다. 그리고 사랑합니다. 선생님.

대표님. 작년 처음 회사를 만들고 이끌어 오기 시작하면서 대표님께서 감당하고 계신 많은 짐을 덜어드리지는 못하고 그저

과분한 사랑만을 받아 면목이 없을 뿐입니다. 저에게 주신 믿음과 기회에 꼭 보답할 수 있도록 열심히 분발하겠습니다.

또한 사랑하는 나의 가족들은 지금까지 저를 지탱하는 힘이었습니다. 철없는 못난 아들을 언제나 믿고 위해주는 엄마, 아빠, 누나, 매형, 조카들, 그리고 한결같이 옆을 지켜주는 서령이까지, 저는 우리 가족이 있어서 항상 힘을 낼 수 있고 엎어져도 일어설 수 있었습니다. 우리가족 아픈 사람 없이 오래오래 행복했으면 좋겠습니다.

제 지난 2 년 동안의 대학원 생활을 돌아보면 떠오르는 단어는 '즐거움' 입니다. 앞으로의 박사과정에서도 석사과정 동안의 즐거움을 안고 더욱 가치 있는 연구를 수행하며 주변 사람 및 제가 몸담고 있는 이 분야에 도움이 되는 사람이 될 수 있도록 더욱 매진하겠습니다.