A Thesis for the Degree of Master of Science

Insights into the genomic characterization and pathogenicity

유전체 특성 및 병원성에 대한 이해

2017년 8월

서울대학교 대학원

농생명공학부

구 현 진

Insights into the genomic characterization and pathogenicity

유전체 특성 및 병원성에 대한 이해

지도교수 김 희 발

이 논문을 석사학위 논문으로 제출함

2017년 6월

서울대학교 대학원

농생명공학부

구 현 진

구 현 진 의 농학 석사 학위논문으로 인준함

2017년 6월

위 원 장 한 재 용 (인)

부위원장 김 희 발 (인)

위 원 윤 숙 희 (인)

Abstract

Insights into the genomic characterization and pathogenicity

Hyunjin Koo

Department of Agricultural Biotechnology

Seoul National University

In these studies, genome-wide studies were conducted to have insights into the genomic characterization and pathogenesis. Although many researchers have tried to focus on the pathogenesis interaction with living organisms, little genomic research has been performed into the mechanism of pathogenesis leading to the disease. Therefore, to extend comprehensive understanding at the genomic level, I conducted comparative analysis and genome-wide association study.

In chapter 2, I performed a comparative genomic analysis with the complete genomes of *B. cereus* strains which is well known as a gastrointestinal pathogen that induces food poisoning. Based on the result of comparative analysis, I could identify the virulence factors playing a crucial role in

pathogenesis, positively selected genes and strain-specific genes of FORC_013.

In chapter 3, pan-genome analysis was conducted using 20 *Shigella* complete genomes available on the NCBI genome database. I found not only the overall genetic contents in *Shigella* spp. but also newly obtained genes through horizontal gene transfer. And these species are defined as monophyletic groups using the phylogenetic tree based on the core gene clustering and ANI analysis. Genes under positive selection account for very low rate in the core genome.

In chapter 4, genome-wide association studies on fusarium wilt resistance in radish (*Raphanus sativus*) was conducted to identify the relationship between phenotypic and genotypic data. I found the significant loci responsible for leading fusarium disease.

Through these studies, I could not only understand about genetic features in living organisms but also provide a comprehensive insight into further studies regarding pathogenesis mechanism at genomic level.

Key words: comparative genomic analysis, pan-genome analysis, genome-wide association study, pathogenicity

Student number: 2015-23125

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 Comparative genomic analysis

Sequencing refers to the process of deciphering DNA sequences in the genome. The first modern sequencing technology, commonly known as the first generation sequencing method, was developed by Sanger in 1977 (Sanger et al. 1977). Sanger sequencing technology has been a great help for many studies in multiple bioinformatics fields, but it was too expensive. This initial approach also produced short read length and very small amounts of data. To complement it, a technology called Next Generation Sequencing (NGS) has been developed. This technology is characterized by its ability to generate large amounts of data in a short time at a lower cost than Sanger sequencing technology. Following development of sequencing technology, lots of species have been completely sequenced. Using these genomic data, many researches were performed via bioinformatics method. Among of a variety of species, comparative genomic analysis using multiple microbiome was conducted to get more insight into its genetic characteristic at genomic level. There are various methods in comparative genomic analysis such as phylogenetic analysis, dN/dS analysis and pan-genome analysis. Comparative genomic approach has been considered as one of the major method for functional annotation of genomes and evolutionary inference (Miller et al. 2004).

It's a really important tool for identifying the evolutionary history of organisms at the molecular level via phylogenetic analysis. This analysis provides more understanding about the mechanism of transition of polymorphic sites (Nei 1996). Up to date, divergent typing methods have been used for prokaryotic species classification such as multi locus sequence typing (MLST), sequencing of 16S rRNA genes, sequencing of 23S rRNA genes, pulsed field gel electrophoresis (PFGE) and multi locus variable-number of tandem-repeat analysis (MLVA). It's rather time consuming and rarely sufficient to standardize using PFGE and MLVA methods (Malorny et al. 2011). Whereas 16S rRNA and MLST methods which depend on one or some common genes are able to classify at the species level, these methods have been shown to fail to define between nearby strains. In this respect, it is better to use more genomic components than to use only one or some genes (Chan et al. 2012; Harris et al. 2010; Lukjancenko et al. 2010). Instead of above methods, there are more reliable methods include a core genome phylogenetic tree and a dendrogram based on ANI value. ANI approach provides results that match with other phylogenetic classifications. Core genome phylogenetic method and ANI analysis are suitable way for prokaryote species delineation, in which bacteria are considered as closely related strains as well as a diverse group of species (Chan et al. 2012).

Evolution means that genetic elements of a group have changed over many generations. As a result of the evolutionary process, biodiversity such as species differentiation was created. The similarity of DNA sequences and the excavation of fossils are evidence that they evolved from common ancestors. Tracing the evolutionary history with genomic data has been considered as meaningful to identify the genetic differences based on various traits. Furthermore, understanding evolution is most relevant to changing over time and can help solve problems with evolutionary patterns. Among various methods which detect the evolutionary meaning, dN/dS analysis has been used to detect the selection pressure and molecular adaptation. dN/dS ratio was measured by comparing the rate of substitutions at synonymous sites (dS), which are regarded as neutral, to the rate of substitutions at non-synonymous sites (dN), which are assumed experience selection. The dN/dS ratio was designed for the analysis of genome sequences in phylogenetic lineage divergence (Kryazhimskiy and Plotkin 2008). If dN/dS ratio was lower than 1, it suggests that genes have undergone purifying selection connected with stabilization. In contrast, if the dN/dS ratio was higher than 1, it suggests positive selection, in which variants increase in frequency until they fix in the population. Especially, branch model and branch-site model is used for testing positive selection. Branch model is likelihood ratio tests among branches and

branch-site model has been used for dN/dS analysis on individual codons along specific lineages (Yang and Nielsen 1998; Zhang et al. 2005). In the case of branch model, the rate of false positive is quite high, so the branch-site approach has been used more often in detecting evolutionary meaning (Pond et al. 2011).

Since sequencing technology has been developed, data on various bacterial genomes have begun to gather dramatically. Studies on genus level as well as species level have been actively conducted with data available in this database. Among a lots of studies, the term "pan-genome" has been used for analysis of genome content focusing on genus level and even at the class, phylum or kingdom level. Pan-genome concept gives a foundation for predicting overall genetic characteristic of the dataset and facilitates guessing adaptation in the lifestyles of organisms. Several terms have been used in the pan-genome analysis: pan-genome (supra-genome), core genes, dispensable genes (distributed genes) and unique genes (strain-specific genes) (Hiller et al. 2007; Medini et al. 2005). The Pan genome literally means the entire genes involved in all strains studied. Core genome refers to genes that are common to all strains in the system. The genes found in the core genome not only implies the biological basic characteristics of the lines studied, but also has a major influence on determining their major phenotypic characteristics.

Dispensable genomes represent genes that are present in more than one strain, and strain-specific genes represent genes that are specifically present in only one strain. Dispensable and strain-specific genes are often referred to as accessory genomes. This not only has a great impact on biodiversity, but also benefits the selective advantage of adapting to other environments and surviving the new host. Accessory genomes have been considered as newly introduced genes via horizontal gene transfer. Also if the species has a large number of strain-specific genes among other species in the genus, this suggests that this species is easy to acquire genes from external environment more easily than other species. In the results of pan-genome analysis, It's able to divide bacteria species into open pan-genome and closed pan-genome using the pan-genome analysis. The open pan-genome, which was formed by species that live in diverse environments, has a very large gene reservoir and implies the exchange of diverse genetic material through horizontal gene transfer. On the other hand, species that live in an isolated environment present a closed pan-genome and have a limited gene pool. When new strains were added, difference tendency determines whether open pan-genome or closed pan-genome (Medini et al. 2005; Tettelin et al. 2008). If the size of the pan-genome has grown up and the size of core genome has declined, this is called open pan-genome. If not, it's called as closed pan-genome.

1.2 Genome-wide association study

As the sequencing method has been developed, a large amount of sequence data has been accumulated at a low price. Using genome data, we could conduct further analysis via bioinformatics methods. Until now, genome wide association study (GWAS) is a research method that explores genetic features for diseases and various phenotypes (Atwell et al. 2010; Duerr et al. 2006; Easton et al. 2007; Satake et al. 2009; Scott et al. 2007). As the genomic information analysis technology develops, a large number of research results are reported. Based on the assumption that the diversity of the traits depends on the genetic polymorphism, GWAS is performed on quantitative traits such as the presence or absence of diseases, qualitative traits such as height, weight, and degree of disease and genetic variants present in DNA (Chatterjee et al. 2013; Cho et al. 2009; Heath et al. 2011). Especially, SNP is mainly used among various genetic variants such as CNV, Indel and SNP (Cooper et al. 2008). GWAS estimates odds ratios for SNP locus differences between predominantly diseased and normal individuals. If there are environmental factors in the comparison group, select appropriate models among the various statistical models to correct environmental factors such as sex and year designated as explanatory variables. Generalized linear model (GLM) approach is generally used for analyzing quantitative traits. For

example, if the environmental factors of an object are referred to as age and sex, a linear model for a quantitative trait is expressed as

$$y = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 SNP + \epsilon.$$

The genetic variant actually used here is a numerically coded value. And we focus on the regression coefficient for SNP and evaluate β_3 value estimated by least square method (Balding 2006). In addition, analysis of the genome data will lead to the problem of population structure. At least several generations of random mating should be repeated until several species are mixed and reach equilibrium, which is difficult in practice. To solve the problem of population stratification, the samples in the dataset are measured using STURTURE software (Falush et al. 2003). The result from this analysis can be used as a covariate or to exclude samples. If used this value as covariates in association study, these values can be used as adjusting for ancestry effects in the dataset. Also, principal component analysis (PCA) is able to be used for characterizing the pattern of each sample in the population (Price et al. 2006). Lots of false positive can be occurred in GWAS because multiple tests are conducted. Therefore, it's crucial to reduce false positive by correcting the result. Bonferroni correction and false discovery rate (FDR) have been widely used to correct this error (van den Oord 2008). The Bonferroni correction adjusts the α value as $0.05/n$ where n is the count of statistical tests performed. This method is considered as too strict approach.

FDR is used for adjusting the false positive rate as alternative (Benjamini and Hochberg 1995). This approach is a prediction of the rate of significant results that are false positives.

This chapter was published in *Gut Pathogens*
as a partial fulfillment of Hyunjin Koo's Master program.

Chapter 2. Comparative genomic analysis reveals genetic features related to the virulence of *Bacillus cereus* FORC_013

2.1 Abstract

Bacillus cereus is well known as a gastrointestinal pathogen that causes foodborne illness. In the present study, we sequenced the complete genome of *B. cereus* FORC_013 isolated from fried eel in South Korea. To extend our understanding of the genomic characteristics of FORC_013, we conducted a comparative analysis with the published genomes of other *B. cereus* strains.

We fully assembled the single circular chromosome (5,418,913 bp) and one plasmid (259,749 bp); 5,511 open reading frames (ORFs) and 283 ORFs were predicted for the chromosome and plasmid, respectively. Moreover, we detected that the enterotoxin (NHE, HBL, CytK) induces food-borne illness with diarrheal symptom, and that the pleiotropic regulator (*Plc*R), along with other virulence factors, play a role in surviving and biofilm formation. Through comparative analysis using the complete genome sequence of *B. cereus* FORC_013, we identified both positively selected genes related to virulence regulation and 224 strain-specific genes of FORC_013.

Through genome analysis of *B. cereus* FORC_013, we identified multiple virulence factors, that may contribute to pathogenicity. These results will provide insight into further studies regarding *B. cereus* pathogenesis mechanism at the genomic level.

2.2 Introduction

For several decades, food-borne illnesses caused by microorganisms have attracted political and media attention around the world (Newell et al. 2010), because outbreaks of these diseases have a strong association with public health problems such as hospitalizations and even deaths (Altekruse et al. 1997; Scallan et al. 2011). Preventing these illnesses is challenging, involving a complicated process rather than simple conventional methods (Tauxe 1997). Therefore, it is essential that we elucidate the phenotypic and genotypic features of agents that cause outbreaks. Among a variety of organisms that can cause food-borne illness, *Bacillus,* which is characterized by several features—a Gram-positive, rod-shaped, motile, spore-forming and aerobic-to-facultative—has been considered an important opportunistic pathogen and may be categorized as 1) gastrointestinal pathogens inducing emetic and diarrheal symptoms and 2) systemic and local infections associated with the respiratory tracts of immunologically-compromised patients and neonates. (Carretto et al. 2000; Rasko et al. 2005). There were food poisoning outbreaks from consuming food including meat, soups, vegetable dishes, dairy products and seafood that were contaminated with *B. cereus* (Johnson et al. 1982; Kotiranta et al. 2000). Especially, *B. cereus* spore, which survives within the small intestine of the host, has a connection with food-borne illness inducing diarrheal symptom (Granum 2005; Granum and Lund 1997). A recent study

reported that enterotoxins might cause diarrheal foodborne illness, including hemolysin BL (HBL), nonhemolytic enterotoxin (NHE) and cytotoxin K (CytK) (Guinebretière and Broussolle 2002; Kotiranta et al. 2000). Although many research efforts have focused on the foodborne diseases caused by *B. cereus*, little genomic research of this species has been conducted into the mechanism of toxicity leading to food poisoning.

In the present study, we sequenced the complete genome of *B. cereus* FORC_013, to better understand its pathogenesis at the genomic level, which was isolated from fried eel in South Korea. Using this sequence data, we assembled the complete genome of *B. cereus* FORC_013 and determined its genomic characteristics. Then, we conducted a comparative genome analysis of the FORC_013 with the complete sequences of 29 other *B. cereus* strains to gain more information of this strain. These results may be useful in elucidating the pathogenicity of *B. cereus* and its role in food poisoning.

2.3 Materials and Methods

Sample Collection, Strain Isolation and Whole Genome Sequencing

B. cereus FORC_013 was isolated from fried eel in South Korea and cultivated in Brain Heart Infusion (BHI; Difco, Detroit, MI, USA) medium. Genomic DNA was isolated and purified using the MoBio UltraClean Microbial DNA Isolation Kit (MoBio, Carlsbad, CA, USA) following the manufacturer's recommendations. Approximately 5μg of genomic DNA was cut into 8–12kb fragments using the Hydroshear system (Digilab, Marlborough, MA, USA). SMRTbell libraries were prepared for each sample using the DNA Template Prep kit 2.0 (3–10kb) for SMRT sequencing which was carried out with C2 chemistry on a PacBio RS II system (Pacific Bioscience, Menlo Park, CA, USA). The AMPure XP bead purification system (Beckman Coulter, Brea, CA, USA) was used to purify libraries by removing small fragments (<1.5 kb). An Agilent 12,000 DNA kit (Applied Biosystems, Santa Clara, CA, USA) was used to characterize the size distribution of sheared DNA templates. Sequencing primers were annealed to the templates and DNA polymerase enzyme C2 was added following the manufacturer's instructions. Loading the enzyme template-complexes and libraries onto 75,000 zero-mode waveguides (ZMWs) was conducted using DNA/Polymerase Binding kit P4 (Pacific Bioscience) according to the

manufacturer's instructions. SMRTbell library sequencing using a 120-min sequence capture protocol with PacBio RS II to maximize read length via the DNA sequencing kit Reagent 2.0 (Pacific Bioscience). The summary of sequencing result is contained in Table 2-1.

Genome Assembly and Annotation

Whole genome assembly was performed using the SMRT portal system. Sequencing reads from the PacBio RS II system were assembled using the HGAP assembly-3 algorithm with curation of the genome size parameter which was set to 3 Mb using the Compute Minimum Seed Read Length option while other parameters were set to default (Chin et al. 2013). Sequencing errors were removed and a polishing assembly process was repeatedly performed to reduce errors, such as indels in the draft assembly using Quiver until none genomic variants were detected. The genome sequences were assembled into contigs using the PacBio RS II system. The orientation and direction of the assembled sequence was defined via Basic Local Alignment Search Tool (BLAST) and Mummer analyses (Delcher et al. 2003). BioEdit software was used to curate the polished sequence based on alignment results (Hall 2011). We used Rapid Annotation of Prokaryotic Genomes (PROKKA) to predict open reading frame (ORF) count and the rRNA and tRNA contents of *B. cereus* FORC_013 (Seemann 2014). The Rapid Annotation using

Table 2-1. Summary of *B. cereus* FORC_013 genome.

Property	Term
Finishing quality	Finished
Libraries used	PacBio SMRTbell™ library
Number of SMRT cells	2
Sequencing platforms	PacBio RS II sequencer
Assemblers	PacBio SMRT analysis 3.0
Gene calling method	PROKKA, and RAST
Number of reads	109,880 (PacBio_20K)
N50 read length	7,373
Average genome coverage	71.39x
Contigs no.	2 (chromosome) 1 (plasmid)
Scaffolds no.	1 (chromosome) 1 (plasmid)
N50 contig length	2,948,960
Chromosome length (bp)	5,418,913 bps (chromosome) 259,749 bps (plasmid)
ORFs	5,424
Locus Tag	FORC13
Genbank ID	CP011145, CP011146
BIOPROJECT	PRJNA279901
Source Material Identifier	FORC_013

Subsystem Technology (RAST) (Aziz et al. 2008) was used for SEED annotation with the default settings. Cluster of Orthologous Groups (COG) annotation was conducted using WebMGA (Wu et al. 2011) and DNAPlotter (Carver et al. 2009) was employed to generate an annotation map.

Genome Accession Numbers

To study the comparative genomics of *B. cereus* strains, 29 complete genome sequences were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/genome/genomes/157). The accession numbers for these 29 *B. cereus* sequences are CP009318.1 (03BB102), CP009641.1 (03BB108), CP009941.1 (03BB87), CP009596.1 (3a), CP015727.1 (A1), CP001177.1 (AH187), CP001283.1 (AH820), AE017194.1 (ATCC10987), AE016877.1 (ATCC14579), CP009628.1 (ATCC4342), CP001176.1 (B4264), CP001746.1 (CI), CP011153.1 (CMCCP0011), CP011151.1 (CMCCP0021), CP009300.1 (D17), CP009968.1 (E33L), CP003187.1 (F837/76), CP009369.1 (FM1), CP009686.1 (FORC_005), CP012691.1 (FORC_024), CP003747.1 (FRI-35), CP008712.1 (FT9), CP009590.1 (G9241), CP001186.1 (G9842), CP011155.1 (HN001), AP007209.1 (NC7401), CP012483.1 (NJ-W), CP000227.1 (Q1) and CP009605.1 (S2-8).

Analysis of Virulence Factors

To investigate the virulence factor encoding genes of FORC_013 strain, we downloaded full DNA sequences from the virulence factor database (VFDB). For virulence gene identification, we used BLASTn method against VFDB (identity ≥ 0.95).

Phylogenetic and Comparative genome analysis

JSpecies software was employed to compute Average nucleotide identity (ANI) values of all 30 strains (Goris et al. 2007). The MESTORTHO algorithm (Kim et al. 2008) was used to build an orthologous gene set for the 30 complete genomes (identity ≥ 0.95; coverage ≥ 0.8). Multiple sequence alignment of each orthologous gene was conducted using PRANK (Löytynoja and Goldman 2005). After removing the poorly aligned positions using Gblocks (Talavera and Castresana 2007), orthologous sequences were joined into one sequence to build a phylogenetic tree. The neighbor-joining method was used to construct a tree using MEGA7 (Kumar et al. 2016). The Codeml program based on PAML4 (phylogenetic analysis by maximum likelihood) (Yang 2007) was used to detect the genes which were under selective pressure by estimating dN (rate of non-synonymous substitution) and dS (rate of synonymous substitution) based on the branch and branch-site models. Prior to pan-genome analysis, 30 complete genome sequences of *B. cereus* were

annotated using the PROKKA annotation tool (Seemann 2014). After annotation, GFF files output from PROKKA were used as the input files for creating the pan-genome with Roary software (Page et al. 2015).

Quality Assurance

The 16s rRNA gene was identified from the assembled sequence using PROKKA. Pairwise distances were calculated by comparison of FORC_013 with published *B. cereus* genomes using ANI values.

2.4 Results and discussion

Genome Features of *Bacillus Cereus* FORC_013

The *B. cereus* FORC_013 genome consists of a circular DNA chromosome and a single circular plasmid. The whole genome sequence comprises 5,418,913 bp with a GC content of 35.3%. The *B. cereus* FORC_013 plasmid contains 259,749 bp with a GC content of 33% and a total of 259 predicted ORFs. The FORC_013 genome contains 5,424 ORFs, 107 tRNA sequences and 42 rRNA sequences. Among the predicted ORFs, 3,750 (69%) were predicted based on annotated genes and 1,674 (31%) were hypothetical and unknown proteins (Figure 2-1). Figure 2-2 presents the categorization of estimated functional genes based on SEED subsystem categories and COG functional categories; 3,525 genes were classified into 26 SEED subsystem categories. Of these, 286 ORFs were categorized into the cell wall and capsule subsystem, which includes pathogenicity; 128 ORFs were responsible for virulence, disease and defense, which may be related to toxin production. In total, 82 ORFs were related to motility and chemotaxis due to the formation of biofilms, which affect the persistence of the pathogen. Functional annotation based on COG categorization using WebMGA identified 3,537 ORFs. Excluding ORFs that were related to the General function prediction only and Function unknown categories (26.4%), 1,235 ORFs, accounting for

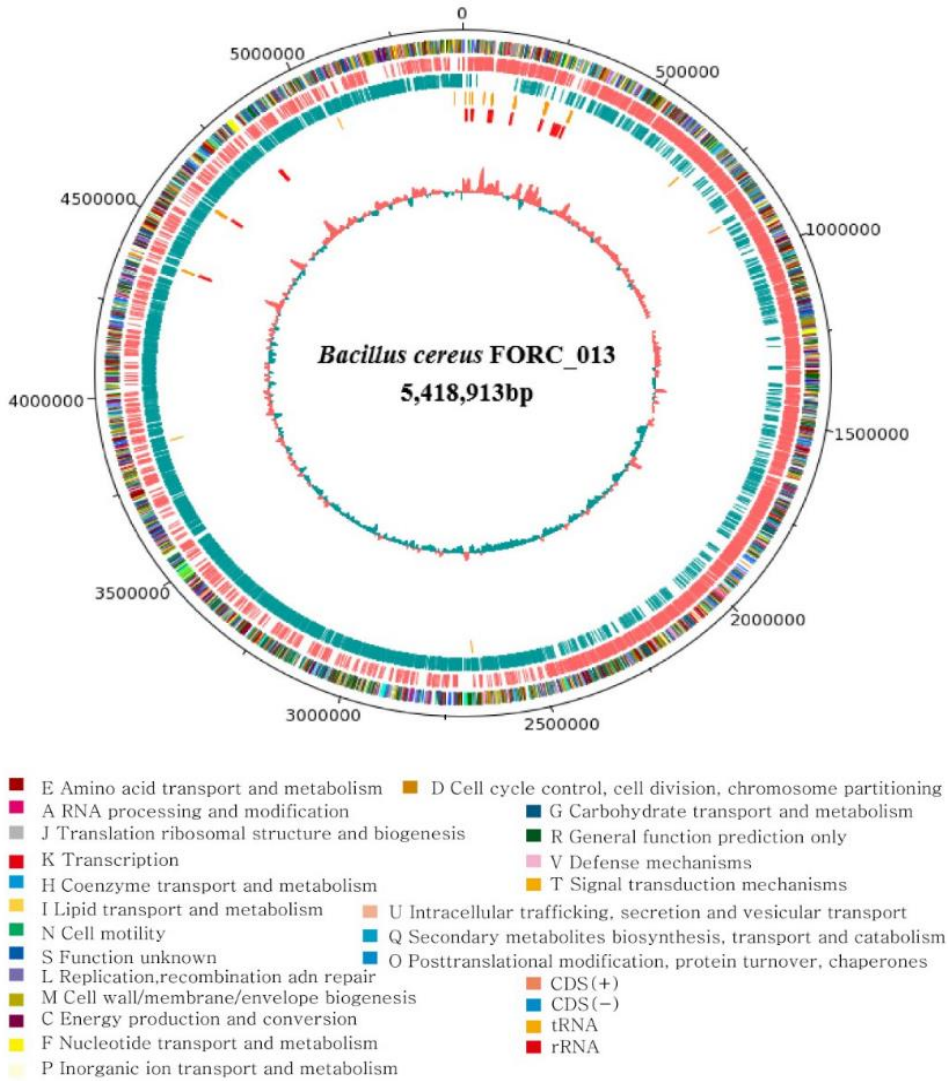

Figure 2-1. Circular genome map of the *B. cereus* FORC_013 chromosome. Circles, from outer to inner, represent the Cluster of Orthologous Groups (COG) distribution, with protein coding sequences (CDS) in the leading strand, CDS in the lagging strand, tRNA, rRNA, and the GC content. Functional genes are labeled around the outer circle as evolutionarily selected genes.

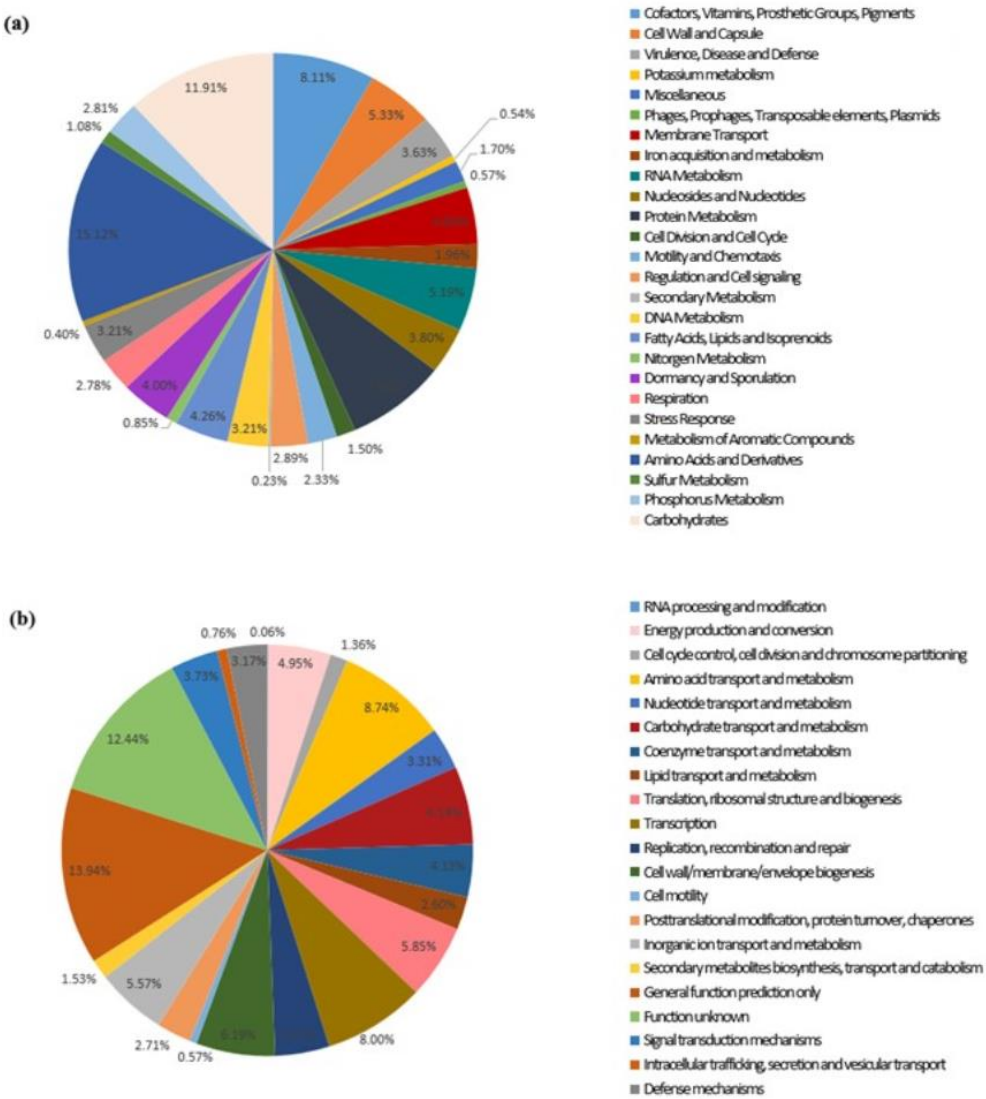

Figure 2-2. Functional categorization of all estimated open reading frames (ORFs) in the *B. cereus* FORC_013 genome based on the (a) SEED and (b) COG databases.

more than one-third of the COG-assigned ORFs, were classified into five major COG categories: 309 ORFs in category E (amino acid transport and metabolism), 283 ORFs in category K (transcription), 219 ORFs in category M (cell wall/membrane/envelope biogenesis), 217 ORFs in category G (carbohydrate transport and metabolism) and 207 ORFs in category J (translation, ribosomal structure and biogenesis).

Virulence Factors

The 20 virulence genes of FORC_013 were identified via BLASTn method against VFDB (Table 2-2). The virulence factors of FORC_013 were classified into six categories: host immune evasion, lipase, protease, regulation, toxin and others. As previous studies reported, the diarrheal symptom is well known for having a close relationship with the enterotoxin, such as hemolytic enterotoxin HBL, non-hemolytic enterotoxin NHE and cytotoxin K (Guinebretière and Broussolle 2002; Kotiranta et al. 2000). The genome of FORC_013 has all of these enterotoxins; *Cyt*K gene, HBL gene cluster (*hbl*A, *hbl*B, *hbl*C and *hbl*D) and NHE gene cluster (*nhe*A, *nhe*B and *nhe*C), suggesting that these genes are responsible for pathogenicity of FORC_013. In the protease category, immune inhibitor A metalloprotease (*inh*A) was detected; this gene assists in surviving the macrophage environment, which is an important factor of the host immune system

Virulence factor	Annotation	Locus tag
Host immune evasion		
-	Polysaccharide capsule	FORC13_5198, FORC13_5217
Lipase		
plcA	Phosphatidylcholine-preferring phospholipase C (PC-PLC)	FORC13_4514
piplc	Phosphatidylinositol-specific phospholipase C (PI-PLC)	FORC13_1400
Protease		
inhA	Immune inhibitor A metalloprotease	FORC13_4518
-	Immune inhibitor A metalloprotease	FORC13_3892
Regulation		
plcR	PlcR	FORC13_5291
Toxin		
-	Anthrolysin O	FORC13_5042
cytK	Cytotoxin K	FORC13_4071
hlyII	Hemolysin II	FORC13_1637
hlyIII	Hemolysin III	FORC13_3034
-	Hemolysin III homolog	FORC13_5388
hblC, hblD, hblB, hblA	Hemolytic enterotoxin HBL	FORC13_2078~ FORC13_2081
nheC, nheB, nheA	Nonhemolytic enterotoxin NHE	FORC13_3362~ FORC13_3364
Others		
-	Internalin-like	FORC13_4633
-		FORC13_3846

Table 2-2. Virulence factors of *B. cereus* FORC_013

(Guillemet et al. 2010). Further, this supports that *inh*A in FORC_013 may contribute to retain living in the macrophage intracellular system. The FORC_013 strain has hemolysin II (*hyl* II) and hemolysin III (*hyl* III) that form the pores by adapting under the harsh environment (Andreeva-Kovalevskaya et al. 2008; Baida and Kuzmin 1996). We also identified an regulation protein, *Plc*R, which is a well-known pleiotropic regulator of genes related to pathogenicity (Salamitou et al. 2000). This gene plays a role in the biofilm formation, which may induce the sporulation of bacteria (Hsueh et al. 2006; Ryu and Beuchat 2005). Biofilm formation facilitate generating adhesive spores and contribute to high resistance (Majed et al. 2016). Detection of *Plc*R indicated that the FORC_013 may take advantage of both biofilm formation and virulence gene regulation. Based on the results, it is reasonable to assume that these virulence factors contribute to pathogenicity of FORC_013. Additionally, we conducted a lactate dehydrogenase (LDH) release assay to identify cytotoxicity, which indicated that FORC_013 has pathogenic activity (Figure 2-3).

Phylogenetic and Comparative genome analysis

An ANI tree and a phylogenetic dendrogram based on orthologous genes were built for a comparative analysis of the FORC_013 strain. Both trees were generated using 29 complete genome sequences acquired from the

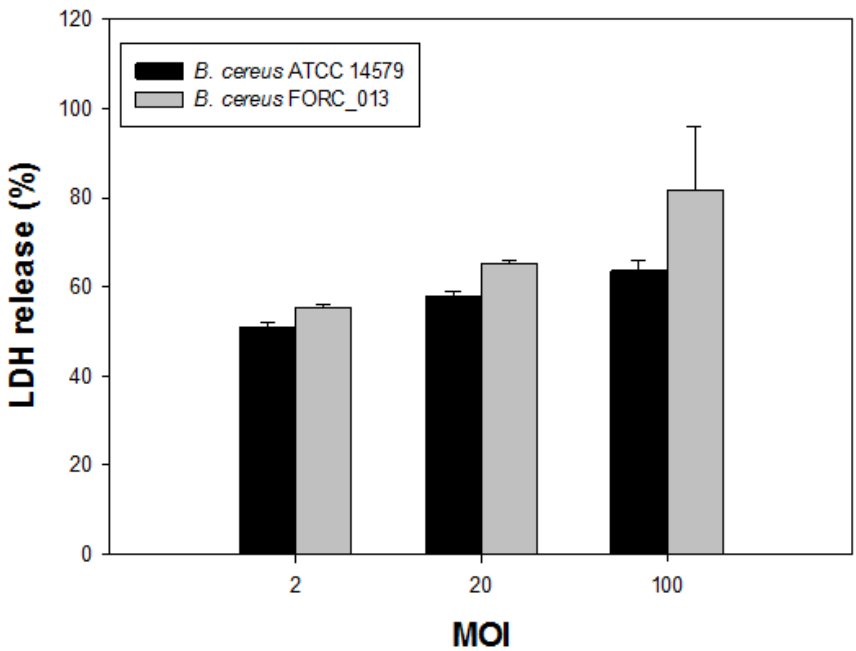

Figure 2-3. Cytotoxicity analysis for two strains of *B. cereus* The cytotoxicity analyses of FORC_013 strain were compared with ATCC 14579 strain by measuring the activity of cytoplasmic lactate dehydrogenase (LDH). INT-407 cells were infected with FORC_013 or ATCC 14579 at various multiplicities of infection (MOIs) for 3 h. Cytotoxicity was determined as the percentage of LDH leakage using the amount of LDH from the cells that were completely lysed by 2% Triton X-100. Error bars represent the standard errors of the means (SEM).

NCBI database and the FORC_013 genome sequence (Figure 2-4). The neighbor-joining method was used to construct an ANI tree with pairwise distance matrix and a phylogenetic tree with orthologous genes. The ATCC14579 strain was shown to contain pathogenicity-related genes in a previous study (Ivanova et al. 2003). In both of our tree analyses, FORC_013 clustered closely with ATCC14579. The high ANI value (98.6%) indicates that FORC_013 may have genes affecting virulence, similar to ATCC14579.

To identify the positively selected genes, we calculated the positive selection sites for the orthologous gene using the branch and branch-site models. In the branch model, 12 genes were revealed as being selected: YtxC-like family protein, post-transcriptional regulator ComN, cytochrome c-550, HTH-type transcriptional regulator NorG, putative HTH-type transcriptional regulator, flagellar hook-basal body complex protein FliE, putative murein peptide carboxypeptidase, HTH-type transcriptional regulator GltC, AT synthase subunit a, regulatory protein YeiL, superoxide dismutase (Mn)2, and oligoendopeptidase F (Table 2-3). In the branch-site model, two putative genes were detected with their own functions: mycinamicin III 3"-O-methyltransferase and putative efflux system component YknX (Table 2-4). In the methyltransferase gene, amino acid 190 was changed to asparagine from aspartic acid in FORC_013. Asparagine and aspartic acid are categorized in the carboxamide group and the negatively charged group,

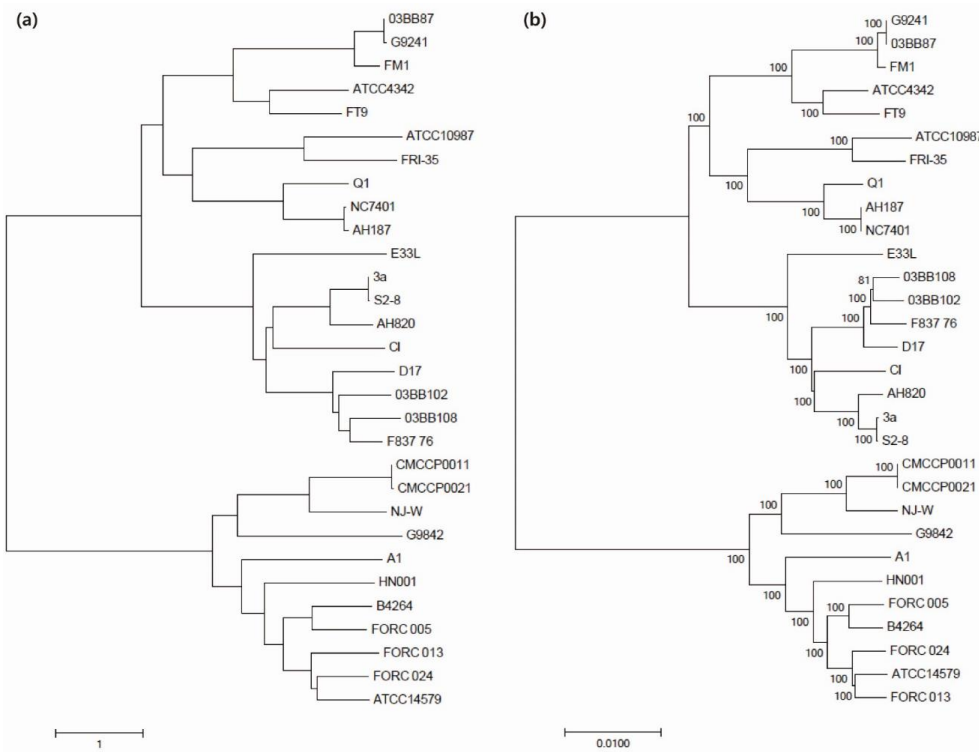

Figure 2-4. Average nucleotide identity (ANI) tree and phylogenetic tree based on (a) ANI value and (b) orthologous genes, respectively.

Locus	Function	p-value	$\omega_{F.G.}$	$\omega_{B.G.}$
FORC13_0586	YtxC-like family protein	2.04E-02	1.03927	0.04993
FORC13_0734	Post-transcriptional regulator ComN	4.62E-02	1.0206	0.05055
FORC13_0851	Cytochrome c550	2.09E-05	1.00565	0.07063
FORC13_1405	HTH-type transcriptional regulator NorG	3.91E-11	1.09935	0.05774
FORC13_3326	Putative HTH-type transcriptional regulator	5.13E-03	1.11237	0.03878
FORC13_3534	Flagellar hook-basal body complex protein FliE	2.93E-02	1.0579	0.10196
FORC13_3811	putative murein peptide carboxypeptidase	7.82E-03	1.21275	0.11958
FORC13_5156	HTH-type transcriptional regulator GltC	4.76E-21	1.18227	0.0696
FORC13_5251	ATP synthase subunit a	1.41E-02	1.05517	0.12087
FORC13_5350	Regulatory protein YeiL	2.27E-04	2.89351	0.13594
FORC13_5384	Superoxide dismutase (Mn) 2	3.19E-03	1.18661	0.04302
FORC13_5392	Oligoendopeptidase F	1.52E-03	1.28818	0.07277

Table 2-3. Positively selected genes predicted in the branch model and related data for *B. cereus* FORC_013

Function	Mycinamicin III 3"-O-methyltransferase	Putative efflux system component YknX
Locus position	FORC13_3977	FORC13_5192
Peptide length	258	395
p-Value	8.77E-20	2.04E-02
ω2 F.G.	999.00	74.42
ω2a B.G.	0.0009	0.0185
Proportion of site class 2a	0.893	0.934
Bayes Empirical Bayes	0.967*	0.975*
Position	190	153
FORC_013	N	I
Other strains	D	V

Table 2-4. Positively selected genes predicted in the branch-site model and related data for *B. cereus* FORC_013

respectively. In the gene identified as a putative efflux system component, amino acid 153 in FORC_013 was changed to isoleucine from valine. Isoleucine is in the hydrophobic group, while valine is categorized as nonpolar. Through evolutionary analysis, we detected some positively selected genes related to virulence. The Superoxide dismutase (Mn) 2 plays a crucial role in protecting cells from the oxidative stress (Duport et al. 2016; Wang et al. 2011). Toxin from FORC_013 can survive low gastric pH condition in the presence of the Superoxide dismutase (Mn) 2. The YknX gene encodes an ABC transporter, which is contributed to the export of virulence factor (Butcher and Helmann 2006; Davidson et al. 2008). These results suggest that the positively selected genes identified in the FORC_013 strain may have an influence on pathogenicity.

Furthermore, Pan-genome analysis of 30 strains revealed 25,247 genes comprising the supra-genome based on the Roary pipeline (Figure 2-5a). The relation between the number of genomes (x) and the pan-genome size (y) was $y = 7520.62x^{0.37} - 2066.4$ ($R^2 = 0.999926$). Also, the relationship between the core genome size and the genome number was calculated as $n = 7276.95e^{-0.82m} + 2284.85$ ($R^2 = 0.960822$). The size of the *B. cereus* pan-genome has grown, while the scale of core genome has decreased with the addition of new strains (Figure 2-5b). Based on this result, we can consider this pan-genome to be an open pan-genome, providing evidence that this species dwells under

conditions that encourage the transfer of genetic material through pathways such as horizontal gene transfer (Muzzi and Donati 2011; Polz et al. 2013). Above all, we examined the unique genes of FORC_013 to elucidate the strain's specific biological characteristics. We detected that the unique genes of the FORC_013 strain comprise 224 genes, including 130 hypothetical proteins. The proportion of unique genes of FORC_013 was 4.16% (Figure 2-5c). Furthermore, we could detect strain-specific genes of FORC_013 associated with virulence through Pan-genome analysis. Here, we identified A-type flagellin and flagellin genes that are involved in biofilm formation (Houry et al. 2010), which are candidates for assisting the activation of FORC_013's pathogenicity.

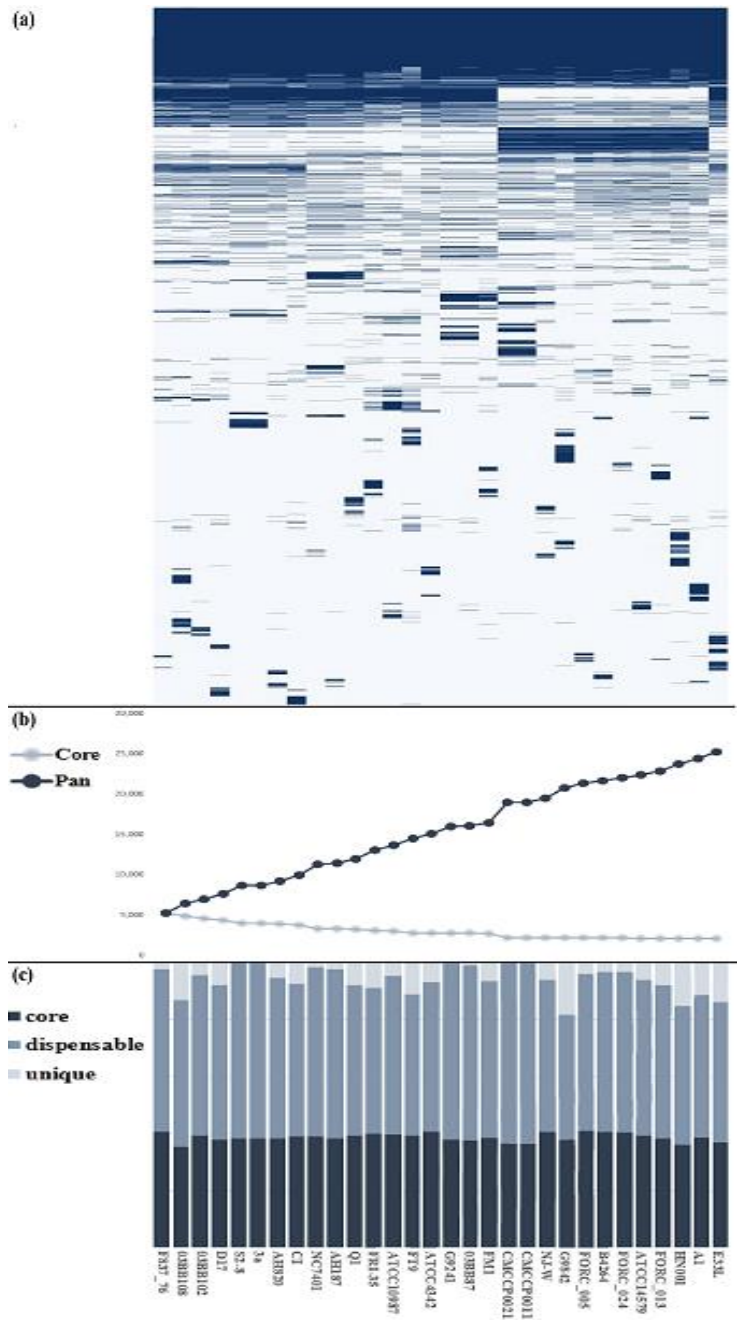

Figure 2-5. Pan-genome- each strain is represented by a vertical line (a) pan-genome structure for 30 *B. cereus* species (b) core genome and pan genome (c) proportion of core, dispensable, and unique genes.

2.5 Conclusions

In this study, we sequenced the genome of *B. cereus* FORC_013, which is an opportunistic pathogen that occurs food borne illness, and performed comparative analysis with 29 published strains. As a result, we detected the virulence factors of this strain that can assist its pathogenicity. We also identified positively selected genes and unique genes of FORC_013. This study advances our understanding of the genetic characteristics of FORC_013. In addition, these findings will provide useful information for further research related to the virulence mechanisms used by this pathogen.

This chapter will be published in elsewhere
as a partial fulfillment of Hyunjin Koo's Master program.

Chapter 3. Pan-genome analysis revealed the core genome of *Shigella* spp.

3.1 Abstract

The *Shigella* spp. is a known cause for the zoonotic disease called shigellosis, which is an intestinal disease and characterized by symptoms such as diarrheal episode, fever, cramps, and mucous bloody stools. To study about the genetic information of the bacteria and its evolutionary background, we conducted pan-genome and evolutionary analysis using 20 complete genome sequences of *Shigella*. The results revealed that 10,124 orthologous clusters comprising the pan-genome of *Shigella* spp. These clusters include 2,105 in the core genome, 4,268 in the dispensable genome, and 3,751 in the strain-specific genes. In the COG analysis, we could detect the categories that account for significant proportion of COGs, which is responsible for protein metabolism as the conserved genes. We could obtain a congruent result between phylogenetic tree, based on core gene clusters, and dendrogram using ANI values; this supports the monophyletic state within the genus. Finally, we performed dN/dS analysis using the core genome. In this analysis, the proportion of genes taking the positive selection occupies a bit lower rate in the core gene clusters set, suggesting they are relatively conservative. This study might offer some clues for the major phenotypic traits and provide insight into the functional evolution of *Shigella* spp.

3.2 Introduction

Shigellosis is an important zoonosis caused by the *Shigella* spp., a gram-negative and rod-shaped bacteria. As an intestinal disease, the shigellosis is characterized by symptoms such as moderate-to-severe diarrhoeal episode, dysentery with fever, cramps, and mucous bloody stools (Schroeder and Hilbi 2008). This disease has been estimated to affect about 164 million patients and lead to 1.1 million deaths each year, including young children under 5 years accounting for 60% (JingYuan et al. 2011; Kotloff et al. 1999). The four species of *Shigella*, *Shigella flexneri*, *Shigella boydii*, *Shigella sonnei* and *Shigella dysenteriae*, belong to the genus *Shigella* and are responsible for the disease. It is frequently transmitted through fecal-oral contact and highly contagious due to low infectious inoculum (Niyogi 2005). To investigate such bacteria, one of the most recent study performed a comparative genomic analysis using a multitude of genomes to reveal the genetic diversity among prokaryotic species (Abby and Daubin 2007).

Among a variety of comparative analysis methods, the pan-genome concept, which comprise of the ‘core genome’ containing genes common in all strains, the ‘dispensable genome’ containing genes common in two or more strains, and ‘unique genes’ specific to single strain was proposed a decade ago. This analysis has been frequently used to provide the overall genetic feature of the

microbial genomes (Tettelin et al. 2005). The genes found in the core genome play a crucial role in housekeeping functions and major phenotypic traits, whereas the dispensable genome renders selective advantages including the ability to colonize new hosts and niche adaptation (Medini et al. 2005; Tettelin et al. 2005; Tettelin et al. 2008; Vernikos et al. 2015). Using the pan-genome analysis, we can divide bacteria species into open pan-genome and closed pan-genome. The open pan-genome presents the species that live in a variety of environments, have an unlimited large gene repertoire, and have a multitude of methods to exchange genetic material; while, the closed pan-genome presents the species that colonize isolated niches and have a limited gene pool (Medini et al. 2005). Although multiple complete genomes in *Shigella* spp. have been sequenced, the pan-genome analysis of *Shigella* is yet to be proposed.

Herein, to facilitate comprehensive understanding of *Shigella* at the genomic level, we present the result of pan-genome analysis of 20 *Shigella* complete genomes available on the NCBI genome database, including the genetic components information. Through this study, we could identify the overall genetic contents in this genus and provide insights into the core conservation and genomic diversity of *Shigella* spp.

3.3 Materials and methods

Sequence retrieval from a public database

The 20 complete genomes of *Shigella* were obtained from the NCBI GenBank and used for the pan-genome and evolutionary analyses. These genomes were clustered in 4 different species such as *boydii, flexneri, sonnei,* and *dysenteriae*. The retrieved data are summarized in Table 3-1.

Pan-genome and core genome analysis

For the pan-genome analysis of the 20 *Shigella* species, we used the PGAP software (Zhao et al. 2012). Here, we applied the MultiParanoid (MP) method for searching the ortholog clusters using the default parameters (score:40; evalue:1e-10; coverage:0.5; identity:0.5). Then, the pan-genome and core genome profile were constructed. We classified the clusters of orthologous groups (COGs) using COGNIZER (Bose et al. 2015).

Phylogenetic and Positive selection analysis

JSpecies was used for computing average nucleotide identity (ANI) values of all 20 strains (Goris et al. 2007). Using the ANI value, we constructed a

Strain	Genome size (Mb)	GC (%)	Gene	Protein	Assembly
Shigella flexneri 2a str.301	4.82882	50.67	4788	4313	GCA_000006925.2
Shigella flexneri 2a str.2457T	4.59935	50.9	4906	4362	GCA_000007405.1
Shigella flexneri 5 str.8401	4.57428	50.9	5168	4482	GCA_000013585.1
Shigella flexneri 2002017	4.89449	50.66	5273	4686	GCA_000022245.1
Shigella flexneri 2003036	4.59581	50.9	5259	4534	GCA_000743955.1
Shigella flexneri Shi06HN006	4.6209	50.9	4947	4404	GCA_000743995.1
Shigella flexneri NCTC1	4.52658	50.9	5144	4464	GCA_000953035.1
Shigella flexneri G1663	4.81731	50.67	5192	4659	GCA_001021855.1
Shigella flexneri 1a	4.85334	50.77	5206	4635	GCA_001578125.1
Shigella flexneri 4c	5.19607	50.38	5650	5073	GCA_001579965.1
Shigella flexneri 2a	4.89175	50.73	5289	4714	GCA_001580175.1
Shigella boydii Sb227	4.64652	51.1	5073	4523	GCA_000012025.1
Shigella boydii CDC3083-94	4.87466	51	5915	5046	GCA_000020185.1
Shigella boydii ATCC9210	4.57425	51.2	5370	4704	GCA_001027225.1
Shigella sonnei Ss046	5.05532	50.76	5839	5057	GCA_000092525.1
Shigella sonnei 53G	5.22047	50.73	6030	5286	GCA_000283715.1
Shigella sonnei FORC_011	5.13271	50.75	5948	5151	GCA_001518855.1
Shigella sonnei FDAARGOS_90	4.98824	51	5752	4953	GCA_001558295.1
Shigella dysenteriae Sd197	4.56091	50.92	4834	4294	GCA_000012005.1
Shigella dysenteriae 1617	4.4802	50.95	6503	6409	GCA_000497505.1

Table 3-1. Characteristics of 20 Shigella spp. strains used in this study

dendrogram with the pairwise distance using MEGA7 (Kumar et al. 2016). All core genes from 20 *Shigella* genomes were used for constructing the consensus tree. Multiple alignments for each core gene were conducted using PRANK (Löytynoja and Goldman 2005) and then the poorly aligned positions were removed using Gblocks (Talavera and Castresana 2007). To build a phylogenetic tree, we joined all orthologous core gene set into a single sequence. A phylogenetic tree was generated by the neighbor-joining method with the bootstrap replications of 1,000 using MEGA7. For the positive selection analysis, protein sequences of each cluster were aligned by MAFFT (Katoh and Toh 2010). These alignment results were converted into the corresponding codon-based nucleotide sequences. Then, PGAP software scans the protein and converted nucleotide sequences, and outputs non-synonymous mutation and synonymous mutation. Finally, to assess the selection pressure, dN/dS was calculated using this software.

3.4 Results and Discussion

Pan-genome and core genome

20 complete genome sequences of *Shigella* species downloaded from the NCBI genbank were used in this analysis. A total of 10,124 orthologous clusters were revealed as comprising the pan-genome. We calculated the relations of the pan-genome size and the core genome size respectively against the number of genome using PGAP. The relation between the number of genomes (n) and the pan-genome size (m) was $m = 2147.8n^{0.444} + 1941.5$ ($R^2 = 0.999926$). Also, the relationship between the core genome size and the genome number was calculated as $y = 2291.5e^{-0.322x} + 2208.3$ ($R^2 = 0.960822$). Based on these results, we could identify that the size of the *Shigella* pan-genome has grown, whereas the size of core genome has declined when new strains were added (Figure 3-1A). Through this result, this pan-genome can be considered as an open pan-genome, suggesting that *Shigella* spp. have a variety of genome contents and inhabited diverse habitats that enhance exchanging the genetic material by horizontal gene transfer (Mann et al. 2013; Medini et al. 2005; Tettelin et al. 2008). As visualized in Figure 3-1B, the core genome is composed of 2,105 orthologous clusters. These genes represent the major biological characteristics of *Shigella*. Also, we could detect 3,751 strain-specific genes in *Shigella* spp. In this result, *S. dysenteriae* 1617 had

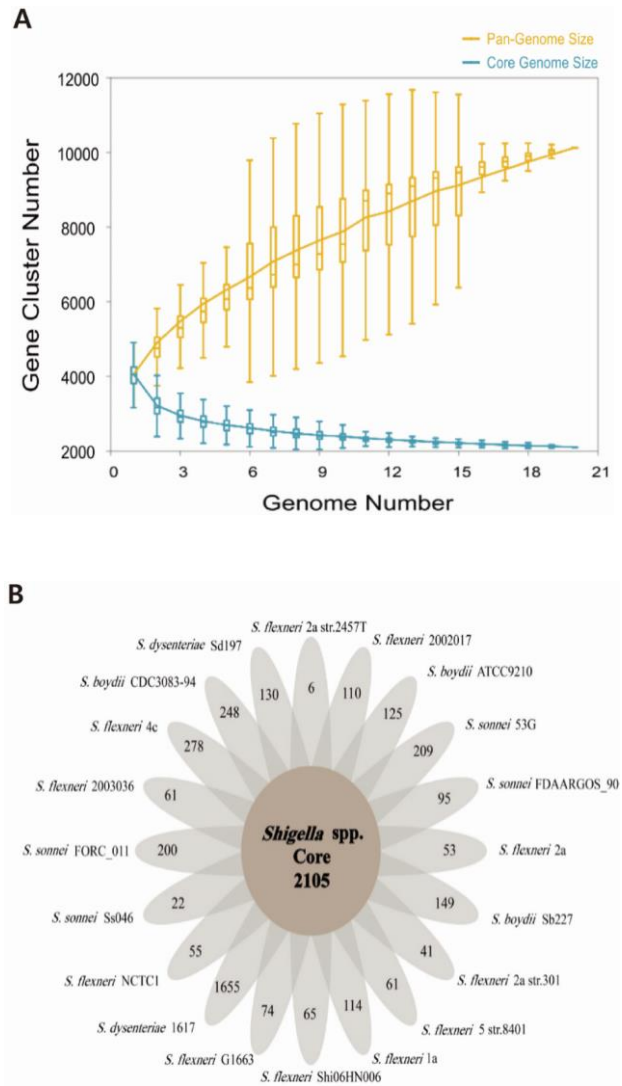

Figure 3-1. Pan-genome analysis of *Shigella* spp.

(A) Characteristics of pan-genome and core genome. As the number of genomes increased, the pattern of genes in the pan-genome and core genome are plotted. **(B)** The side parts represent the number of strain-specific genes. The center represents the number of core genome.

the largest unique genes (1,655) whereas *S. flexneri* 2a str.2457T had the fewest strain-specific genes (6). These genes have been considered as newly obtained genes through horizontal gene transfer (Boto 2010). *S. dysenteriae* shows the largest number of unique genes among the four *Shigella* species; this suggests that this species facilitate to acquire genes from external environment more easily than the others. Through this analysis, we could obtain an insight into not only the overall pattern about *Shigella* spp. genomes, but also observe the genetic material exchanges in the species.

Functional characterization of Ortholog clusters

Functional characterization based on COG categorization was examined using COGNIZER. Excluding Function unknown and general function prediction, three major COG categories were detected as accounting for more than one-fourth of the COG category in the pan-genome: carbohydrate transport and metabolism (588 gene clusters), amino acid transport, and metabolism (547) and transcription (473). In the core genome, the most frequent COG category was the amino acid transport and metabolism (253). The next frequent COGs were followed by 'translation, ribosomal structure and biogenesis' (190) and 'inorganic ion transport and metabolism' (164). After that, we calculated the proportion of each group (core, dispensable and

unique) to identify the distribution in each category (Figure 3-2). Among somewhat abundant COGs as mentioned above, we could observe that a considerably high proportion in core genome was ‘amino acid and metabolism’ and ‘translation, ribosomal structure and biogenesis’. This further supports the analysis of Liu (Liu et al. 2012), Ouzounis and Kyrpides (Ouzounis and Kyrpides 1996), who proved that genetic procedure such as translation is conserved and related to the original form. Synthesis of proteins is held on ribosomes complexes including large RNA molecules and some proteins. In addition, a previous study supports that the ribosome is not only a survivor of RNA, which plays the fundamental role in protein synthesis, but also abundant in information on very early stage in evolution (Berg et al. 2002). This result signifies the importance of regulating a variety of biochemical processes in charge of the synthesis of proteins, and the collapse of proteins or other large molecules as the conserved genes. Not in the range of abundant COGs, we found four COG categories occupying more than 50% as the core genome except RNA processing and modification, which had low count of two gene clusters: ‘cell cycle control, cell division, chromosome partitioning’, ‘coenzyme transport and metabolism’, ‘nucleotide transport and metabolism’ and ‘posttranslational modification, protein turnover, chaperones’. These might have a crucial role in major phenotypes in *Shigella* spp. On the other hand, some category has tendency to have lower proportion

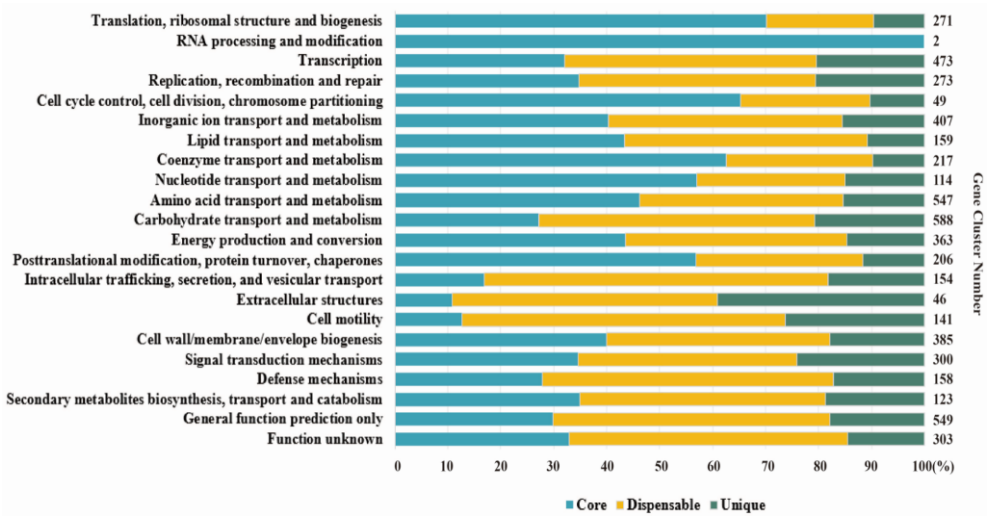

Figure 3-2. Functional characterization of orthologous group based on COG category. These bar represent the proportion of genomes such as core, dispensable and unique in each category.

of core genome, considering the large size of gene clusters (i.e. ‘carbohydrate transport and metabolism’), suggesting that they are related to adaptation of *Shigella* in their special environmental niche.

Phylogenetic tree

As the previous study on the combination of ANI and core genome dendrogram can be a plausible method to analyze a diverse group of bacteria delineation, in the case of defining the species as monophyletic groups, over 95% pair-wise ANI values is used as threshold (Chan et al. 2012). We observed that all strains exhibited the ANI values higher than 97%; thus, this is appropriate to describe the phylogenetic pattern of *Shigella*. The dendrogram based on ANI represents four lineages according *felxneri, boydii, sonnei,* and *dysenteriae* (Figure 3-3A). This dendrogram supports that each lineage is in monophyletic state within the genus, without any exceptions. We could identify that two lineages*, boydii* and *sonnei,* are a sister group. In the previous study, four *shigella* species was classified with four serogoups: *dysenteriae* (serogroup A); *flexneri* (serogroup B); *boydii* (serogroup C); and *sonnei* (serogroup D). From the physiological point of view, A, B, and C were similar, and group D is able to be differentiated on the biochemical reaction (Abby and Daubin 2007). Unlike the physiological point of view, in this study

it can suppose that the C and D groups were differentiated from the most recent common ancestor on the genomic level. Thus, it is important to perform a validation process, such as morphological and cytological studies, for a more precise interpretation of the portion of C and D bound to the sister group. In addition, a consensus tree was generated from the orthologous sequence of all 2,105 *Shigella* core gene clusters (Figure 3-3B). In the result from 1,000 iterations, most of the bootstrap value in the branch are near 100, supporting good confidence. The phylogenetic tree based on core gene clusters is congruent with the dendrogram based on ANI values. No misclassification was found using the phylogeny base on the core gene clustering and ANI analysis. This supports that phylogenetic classification based on core genome is compatible with that of ANI-based approach.

Positive selection analysis based on core gene clusters

To identify the selection pressures, many studies used the dN/dS value on protein coding sequences (Kryazhimskiy and Plotkin 2008; Spielman and Wilke 2015); in contrast, we analyzed the positive selection focusing on the core genome in this study (Leekitcharoenphon et al. 2012). If the dN/dS value was lower than 1, it means that genes have undergone purifying selection related to the stabilization. On the other hand, if the value was higher than 1,

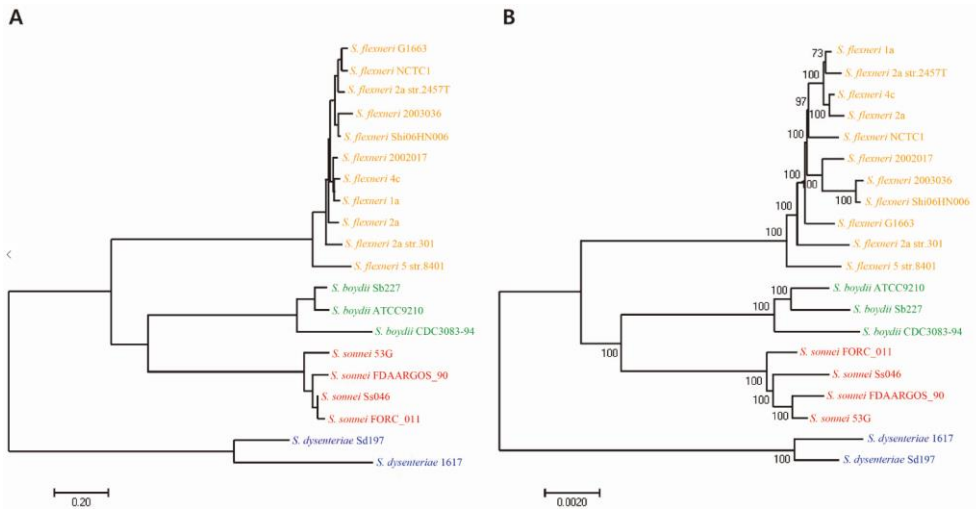

Figure 3-3. (A) Average nucleotide identity (ANI) tree and **(B)** phylogenetic tree based on orthologous core genome.

it suggests positive selection. We tested the positive selection analysis using 2060 clusters comprising the core genome, except dS ratio= ‘0’. 299 genes were detected as positively selected. COG proportion in positive selection genes was visualized in Figure 3-4. We identified that mainly 2 COG categories—‘carbohydrate transport and metabolism’ and ‘Transcription’—account for the positively selected region . In this analysis, genes under positive selection occupied very low rate in the core genome, suggesting that core genes were significantly conservative.

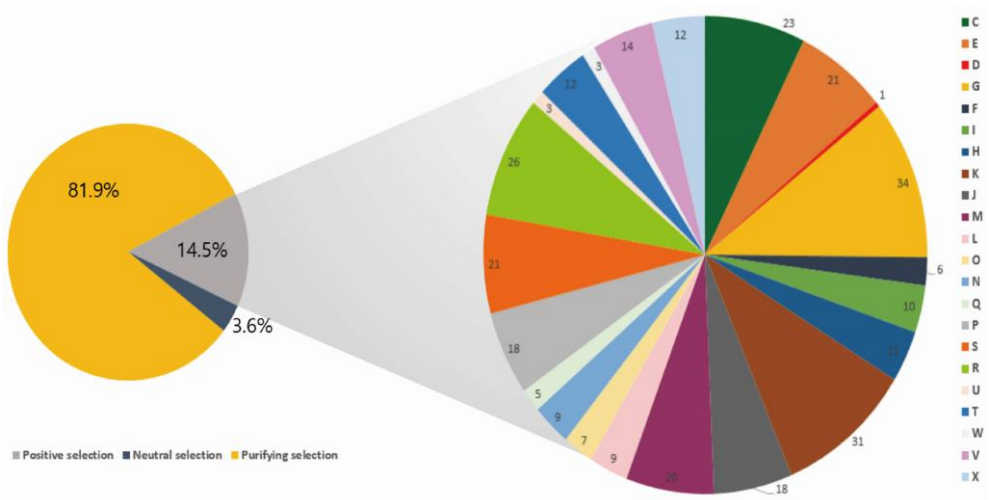

Figure 3-4. The proportion of dN/dS results and the distribution of COGs under positive selection in the core genome.

This chapter will be published in elsewhere
as a partial fulfillment of Hyunjin Koo's Master program.

Chapter 4. Genome-wide association studies on Fusarium wilt resistance in *Raphanus sativus* using genotyping-by-sequencing approach

4.1 Abstract

Fusarium wilt (FW) is a fungal disease that causes severe yield losses in radish. The most effective method to control the FW is thought to develop and use of resistant varieties in cultivation. The identification of marker loci linked to FW resistance is expected to facilitate breeding for disease-resistant radish. In the present study, we applied an integrated framework of genome-wide association studies using genotyping-by-sequencing (GBS) to identify FW resistance loci in a panel of 227 radish accessions including 58 elite breeding lines. Phenotyping was conducted by manual inoculation of the FW pathogen to 10 days-old seedlings, and scoring for disease index at 3 weeks after inoculation in two constitutive years. Genome-wide association analysis identified 13 single nucleotide polymorphism (SNP) markers and 8 putative candidate genes significantly associated with FW resistance. Four of the total 13 SNPs lied in a region syntenic to a quantitative trait locus for FW resistance in radish. In synteny analysis using linked markers, three SNPs on chromosome 1 and 7 (R1_10850192, R1_10850198, R7_7288076) were mapped to the *A. thaliana* genomic region, which contains *RFO1* contribute to immunity against *F. oxisporum* infection in the root vascular cylinder. These markers will be valuable for breeding programs using marker-assisted selection to develop FW resistant varieties.

4.2 Introduction

Radish (*Raphanus sativus* L., 2n=18) belongs to the family *Brassicaceae* and is an economically important crop used as vegetable, animal fodder and oil production in worldwide. The swollen taproot, young leaves, fresh sprouts and immature siliques are eaten as vegetables in many ways: raw, pickled, dried, and simmered. All types of radishes are cultivated for forages, and oil radishes are used as a source for seed oil. Radish contains valuable nutrients and phytochemicals such as sugars, minerals, glucosinolates and flavonoids (Lim et al. 2016). Radish is estimated about 7 million tons/year for world production and represented about 2% of total world vegetable production (Kopta and Pokluda 2013). It is of major important vegetable in East Asia, especially in China, Japan and Korea. Production areas of radish is estimated about 1,200 thousand hectares in China, 33 thousand hectares in Japan, and 20 thousand hectares in Korea (China; Annual report on Food, Agriculture and Rural Area in Japan 2015; Kostat 2015).

The genus *Raphanus* is originated from coastal region of Mediterranean (Kitamura 1958). Radish was introduced into China along the Silk Road about 2,000 years ago, to Japan and Korea about 1,300 years ago (Kaneko et al. 2007; Kitamura 1958; Li 1988; Park HG 2008). Most scholars considered that the cultivated radish (*Raphanus sativus*) originated from wild radishes (*R.*

raphanistrum L.), while others considered *R. sativus* could be a progeny derived from hybridizat(Kitamura 1958)ion of *R. maritimus* and *R. landra* which both are subspecies of *R. raphanistrum* (Kitamura 1958). While, in the study of chloroplast DNA variations of radishes, it was revealed that *R. raphanistrum* is not the maternal ancestor of the cultivated radish, and East Asian wild radish has played a role to the establishment of the East Asian cultivated radish (Yamane et al. 2005). Recently, high-depth resequencing analysis revealed that Asian cultivated radish types were closely related to the Asian wild accessions, but were distinct from European cultivated radishes (Kim et al. 2016). Another wild *Raphanus* species, *R. sativus* var. *raphanistroides* Makino, was naturally grown on the coastal areas of China, Japan and Korea (Kitamura 1958; Lü et al. 2008)

Radish cultivars have been categorized into five varieties based on the morphological traits (Kitamura 1958; Lü et al. 2008; Yamane et al. 2009), which are *Raphanus sativus* var. *sativus* L. (syn. var. *radicular* Pers.) (European small radish), var. *hortensis* Becker (East Asian big long radish), var. *niger* Kerner (black Spanish radish), var. *chinensis* Gallizioli (Chinese oil radish) and var. *caudatus* Hooker & Anderson (rat tail radish or feed radish).

Fusarium wilt (FW) of radish is a soil-borne disease caused by the fungal pathogen *Fusarium oxysporum* f. sp. *raphani*. FW causes severe yield losses in radish production in the continuous cropping field (Armstrong and

Armstrong 1976; Bosland and Williams 1987; Garibaldi et al. 2006; Yu et al. 2013). Disease symptoms start with yellowing and dropping of the young leaves, vascular discoloration, severe stunting, and infected plants eventually wilted and died as the disease progressed (Kendrick and Snyder 1942). The pathogen of FW can survive for long periods in the soil without a suitable host plant. Biotic and abiotic dispersal mechanisms, and agronomic control were not sufficiently efficacious for the disease prevention. Therefore, the most preferred method for control of the disease is the development and use of radish cultivars resistant to FW.

Plant resistance to disease can be regulated by major genes, multiple additive genes, race specific genes, or host-pathogen recognition genes (Knepper and Day 2010). FW of radish was reported to be controlled by quantitative multiple genes anonymous in the dominant or recessive interaction (Ashizawa et al. 1979; Kaneko et al. 2007; Peterson and Pound 1960). The QTL for FW was detected in the BC_1S_1 families, OPJ14 of RAPD locus on LG1, which explained 60.4% of phenotypic variation (Kaneko et al. 2007). In the result of the comparative map between *R. sativus* and *A. thaliana*, this locus was predicted to correspond to the long arm of *A. thaliana* Chr 1, which contained TIR-NBS genes implicated in disease resistance (Shirasawa et al. 2011). In mapping of QTL using a RIL population, two QTLs for FW resistance and one QTL for FW susceptibility were detected on LG1 and LG7,

respectively (Kaneko et al. 2007). Another QTL analysis using F2 population, a total of 8 QTLs for FW resistance were identified on LG 2, 3, 6 and 7 (Yu et al. 2013). In the synteny analysis, the QTL on LG 3 with high LOD value showed homology to *A. thaliana* Chr 3, which contains disease-resistance gene clusters.

Until now, genomic evidence is not yet available to elucidate the correlation with the fusarium FW resistance of *R. sativus*. In spite of lots of QTL mapping research, the genetic features related to multiple agronomic traits is still poorly understood. QTL mapping is conducted using dividing bi-parental populations. This approach has a limitation of allelic diversity and the occurring amount of recombination in the RIL population (Korte and Farlow 2013). Recently, a genome-wide association study (GWAS) has been thought to be a powerful method to detect the candidate markers related to complex trait by screening a large number of individuals (Li et al. 2013; Zhao et al. 2011). A GWAS approach has been successfully clarified the genetic basis of agronomic traits and candidate genes in rice, maize, wheat and soybean (Li et al. 2013; Wu et al. 2015).

In this study, a collection consisting 227 accessions was used to perform GWAS based on the Genotyping-By-Sequencing (GBS) approach to detect the candidate markers to predict the loci related to fusarium resistance in *R. sativus*. From this study, we could identify the loci responsible for inducing

fusarium disease. This result may provide us a comprehensive insight on the genetic characteristic associated with FW resistance of *R. sativus*.

4.3 Materials and Methods

Plant materials

A panel containing 227 radish accessions was used to evaluate FW resistance. A total of 227 radish collection was comprised of 126 accessions provided by National Agrobiodiversity Center (NAC)-RDA in Korea, 43 accessions provided by Genebank-NARO in Japan, and 58 elite breeding lines. The 126 accessions from NAC-RDA were selected from 818 accessions to maximize the phenotypic diversity (based on the genetic diversity of morphological characteristics), and self-pollinated by single-seed descent for three generations. The 43 accessions from Genebank-NARO were also selected from 419 accessions based on the morphological characteristic, such as root shape and root skin color, and self-pollinated for one generation. In addition, 58 elite lines used for commercial radish breeding programs are F8-10 progenies generated by single seed descent. Seedlings were grown in the greenhouse and used to inoculate with FW pathogen.

Pathogen inoculation

A panel of 227 radish accessions were evaluated for resistance to FW in September of 2015 and 2016, using the method described by Yu et al. (2013) with some modification. Thirty seedlings from each accessions were grown

for 7 days and inoculated with FW pathogen using the root-cut dipping methods (Smith et al. 1981). Seven-day-old seedlings were lifted and dropped soil from roots with tapping fingers, and their root-rip was cut to 2-3 cm in length. They were then immediately immersed in a paper cup contained 10 ml of fusarium spore suspension solution adjusted to 10^6 per mm^2 concentration for 10 min. Twenty-five seedlings were transplanted into plug trays containing commercial potting soil (Barokeo, Seoul Bio, Chungbuk, Korea). The temperature of glasshouse was maintained between 26 and 29 ℃. Two and three weeks after inoculation, disease symptoms were scored on a scale of 1, 3, 5, 7 and 9 of disease index (DI), with DI 1 being most resistant and DI 9 being most susceptible to FW (Figure 4-1). The severity of disease symptoms as follows: DI 1, healthy and no symptoms; DI 3, chlorosis of lower leaves and slightly dwarfed; DI 5, well developed symptoms of chlorosis, necrosis and dwarfed; DI 7, severe symptoms of chlorosis, necrosis, and defoliation; DI 9, plants died.

Preparation of genotyping-by-sequencing libraries

Total genomic DNA was extracted from 0.1g of leaf tissue using DNeasy Plant Minikit (Qiagen, Hilden, Germany) following the manufacturer's instruction. The amount of DNA was quantified using the standard procedure

Figure 4-1. Symptoms of Fusarium wilt of *Raphanus sativus*. Plants were visually assessed for development of symptoms 2 and 3 weeks after inoculation with the pathogen. Symptom severity score was rated on a 5-point scale: 1 = healthy and no symptoms, 3 = chlorosis of lower leaves and slightly dwarfed, 5 = well developed symptoms of chlorosis, necrosis and dwarfed, 7 = severe symptoms of chlorosis, necrosis, and defoliation, 9 = plants died.

of Quant-iT PicoGreen dsDNA Assay Kit (Molecular Probes, Eugene, OR, USA) with Synergy HTX Multi-Mode Reader (Biotek, Winooski, VT, USA) and normalized to 20 ng·ul^{-1}. DNA (200ng) was digested with 8U of High-fidelity *Apek*I (New England BioLabs, Ipswich, MA, USA) at 75 ˚C for 2 h.

Preparation of genotyping-by-sequencing libraries

DNA libraries for GBS were constructed according to the protocols as described previously (De Donato et al. 2013; Elshire et al. 2011) with minor modifications. The restriction digestion of DNA with ApekI was followed by ligation with adapters. The adapters included a set of 96 different barcode-containing adapters for tagging individual samples and a common adapter for all samples. The ligation was performed using 200 cohesive end units of T4 DNA ligase (New England Biolabs) at 22 ˚C for 2 h and the ligase was inactivated by holding at 65 ˚C for 20 min. The sets of 87 ligations from Korean populations and 18 ligations from the other populations were pooled into one sample respectively, and purified using QIAquick PCR Purification Kit (Qiagen). The pooled ligations (5ul) were amplified in 50ul reaction by multiplexing PCR using AccuPower Pfu PCR Premix (Bioneer, Daejeon, South Korea) and 25 pmol of each primer. PCR cycles consisted of 98 ˚C for 5 min followed by 18 cycles of 98 ˚C for 10 s, 65 ˚C for 5 s, and 72 ˚C for 5 s,

with a final extension step at 72 ˚C for 5 min. The PCR products were also purified using QIAquick PCR Purification Kit (Qiagen) and then evaluated the distribution of fragment sizes with BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). The GBS libraries were sequenced in the Illumina NextSeq500 (Illumina, San Diego, CA, USA) with the length of 150 bp single-end reads.

Sequencing and genotyping

Single-end sequencing was performed using Illumina Hiseq2000 sequencer for a total of 227 DNA samples. Raw data was demultiplexed to sort samples using the GBSX tool (Herten et al. 2015). The chromosome pseudomolecule level genome data (Rs 1.0 Chromosome) from the Radish Genome Database (http://radish-genome.org) was used as a reference for radish (Jeong et al. 2016). After performing demultiplexing, single-end sequence reads were mapped to the radish reference genome using Bowtie2 (Langmead and Salzberg 2012). For calling variants, we used software packages Genome Analysis Toolkit (GATK) and Picard tools (McKenna et al. 2010). We conducted local realignment of reads to correct misalignment caused by the presence of insertion and deletion using GATK 'RealignerTargetCreator' and 'IndelRealigner' sequence data processing tools. Subsequently, GATK

'HaplotypeCaller' and 'SelectVariants' instructions were used to call variants.

Imputation and SNP Filtering

A total of 211,499 "raw" SNPs were identified after calling variants. Of these 211,499 SNPs, we used 188,813 SNPs mapped to assembled chromosomes excluding markers (22,686) mapped to scaffolds that unassigned to a chromosome. SNPs with minor allele frequency(MAF) (<0.05) and call rate (<80%) were discarded using PLINK (Purcell et al. 2007). After filtering, missing genotypes were imputed using BEAGLE v4.1 with default settings (Browning and Browning 2016).

Population structure and linkage disequilibrium

STRUCTURE version 2.3.4 (http://pritchardlab.stanford.edu/structure.html) was used to assume the pattern of population stratification through a model-based approach (Pritchard et al. 2000). Five runs were performed with 32,778 markers for k-values from 2 to 14. To achieve a more accurate result, we used 5,000 burn-in period followed by 50,000 MCMC (Markov Chain Monte Carlo) iteration parameter. The most likely subgroups were determined by measuring the estimated likelihood values delta K acquired from this result.

Linkage disequilibrium (LD) between marker loci on each chromosome was assessed with the squared allele frequency correlation using TASSEL v.5.0 standalone (Bradbury et al. 2007).

Genome-wide association analysis

Association analysis was conducted in TASSEL v.5.0 standalone. Kinship (K) matrix was estimated familial relatedness between lines as an identical-by-state (IBS) matrix. To solve the skewed problem of trait distribution, we conducted transformations using the box-cox function in the R package (Juliana et al. 2015). Then, we fitted a mixed linear model (MLM) with correction for both population structure and relatedness : QK model (Bastien et al. 2014). The critical *P*-values ($-\log_{10}P \geq 3$) was used as detecting the significant marker of fusarium resistance (Visioni et al. 2013). The loci of the significant SNPs were used for predicting the candidate product via annotation using the Radish Genome Database.

4.4 Results

Population structure and linkage disequilibrium with Genotypic Data

A total of 32,778 markers were detected with the GBS approach after filtering options. Population structure was assessed by calculating the ad hoc statistic (delta K) to identify the subpopulations within the 227 individuals. K values were ranged from 2 to 14 on the entire dataset using filtered SNP markers. Based on changing delta K, a significant increase was observed both from 2 to 3 and from 8 to 9. The population was sorted into three sub-population (K=3) (Figure 4-2). The 227 radish accessions were divided into three clusters based on the geographical origins: (1) the south Chinese and Japanese radishes, (2) the north Chinese and Korean radishes, and (3) the European radishes. For controlling false positive problem, we used population structure covariate in association analysis. LD decay was shown as the scatter plots of squared allele frequency correlation (r^2) between the SNP loci in Figure 4-3. Overall the genotypes, the distribution of r^2 rapidly declined with increasing the physical distance. The LD varied all chromosomes decayed to <0.2 at about 900kb.

Genome-wide association study (GWAS)

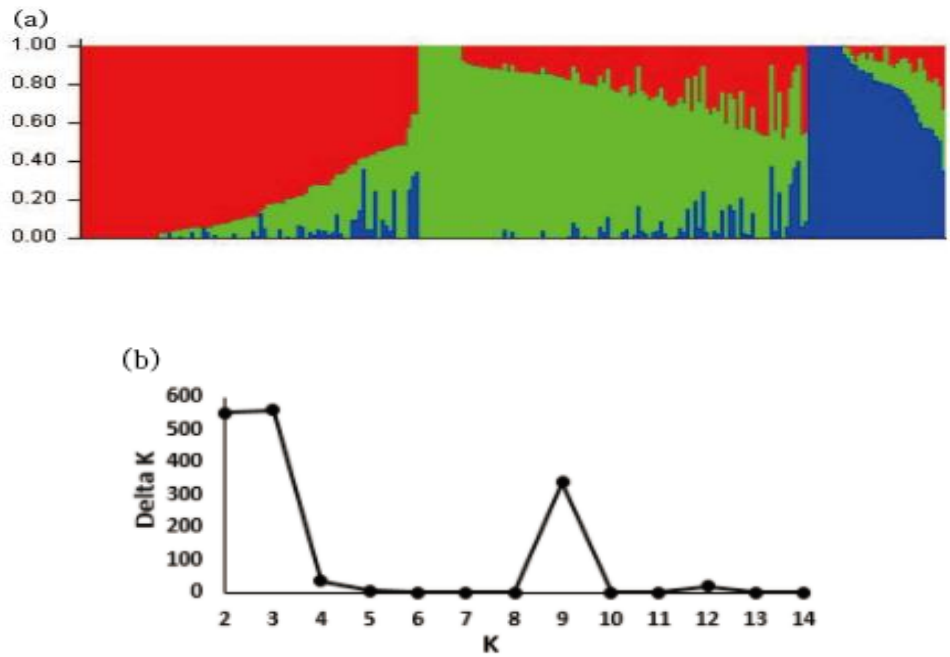

Figure 4-2. Genetic diversity and population structure (a) Population structure of Radish cultivars in Korea, each accession is represented by a single vertical line and one cluster is represented by color (b) estimated delta K ranging from 2 to 14

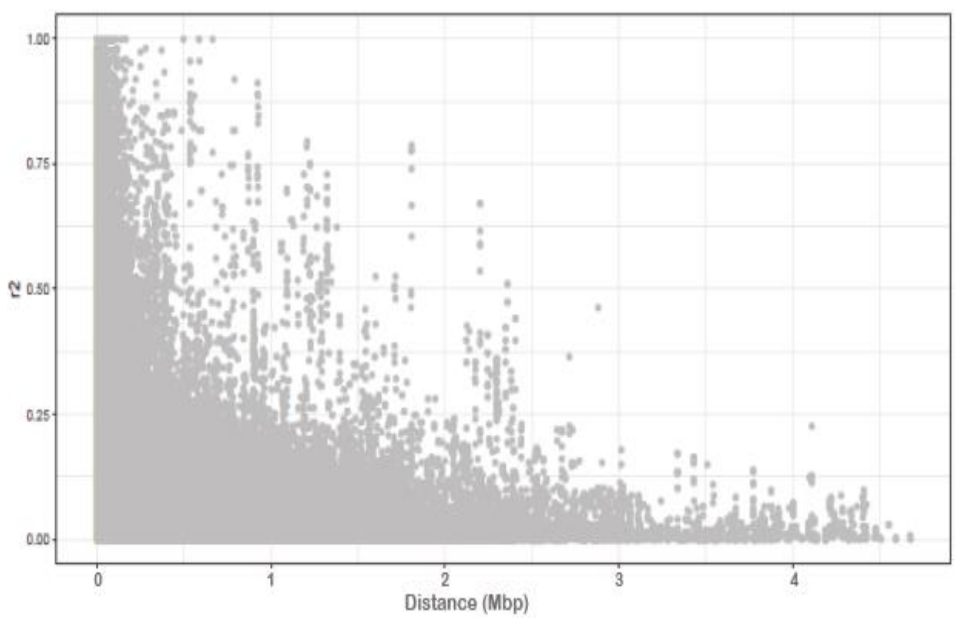

Figure 4-3. Estimated linkage disequilibrium decay Scatter plot showing the linkage disequilibrium (LD) decay across the chromosomes of radish

GWAS was performed using mixed linear model (MLM) with corrections for population structure, relatedness and year, witch identified significant 13 SNPs which were matched to 8 candidate genes distributed on chromosome 1, 2, 3, 4, 5, 6 and 7. Theses SNPs represent a minimum allele frequency (MAF) ranging from 0.05 and with a highest value of 9.9E-04. On chromosome1, two SNPs were significantly associated with Rs019110. This gene was encoded as formin 8 protein. Three SNP markers mapping on chromosome 2 were detected and located in two genes, Rs065080 and Rs045840, which encoded WD40 repeat like superfamily protein and Sec23/Sec24 protein transport family protein, respectively. On chromosome 3, one marker was identified as having significantly associated with one candidate gene, Rs116800, which was annotated as pentatricopeptide repeat (PPR) superfamily protein. One significant SNP locating in chromosome 4 was detected and included in Rs191190. This gene was encoded as ABL integrator like protein. A marker located in chromosome 5 was encoded as protein kinase family protein, Rs229140. The last chromosome be shown as significantly associated with FW trait was chromosome 7. On chromosome 7, three markers were detected and located in 2 genes, Rs382210 and Rs392970. We could identify these genes related to PLC-like phosphodiesterase superfamily protein and movement protein binding protein 2C, respectively. We arbitrary checked 30 Kb upstream and downstream region of the

significant SNP loci (Sakiroglu and Brummer 2016). We could detect 80 candidate genes (Table 4-1), including the genes encoding disease resistance-responsive proteins, such as LRR-RLK and Zinc finger protein.

Synteny analysis

We conducted synteny analysis for the *R.sativus* genome in the present study with the another *R.sativus* in previous QTL study, and the *A.thaliana* genome (Yu et al. 2013). Several genomic regions harboring common FW resistance were identified (Figure 4-4). Comparative genomic analysis revealed extensive synteny between *R.sativus* genomes in the present study and previous QTL study. The syntenic regions are Rs045840 on chromosome 2 and *qFW6* in LG6, Rs116800 on chromosome 3 and *qFW4* in LG3, and Rs229140 on chromosome 5 and *qFW3* in LG3 (Yu et al. 2013). Five genomic regions on chromosome 1, 2, 4 and 7 were distinctly detected in this study.

Table 4-1. Potential candidate genes identified 30Kbp upstream and downstream of the 13 SNP loci associated with Fusarium wilt in radish

No.	GeneID	SNP	Function
1	Rs065070	30112064	high affinity nitrate transporter [BLAST2GO] | Best-Hit: high affinity nitrate transporter 2.7 [AT5G14570]
2	Rs191150	29844872	tetratricopeptide repeat domain-containing protein [BLAST2GO] | Tetratricopeptide repeat (TPR)-like superfamily protein [AT5G41950]
3	Rs045900	40678277 40678250	electron carrier electron transporter iron ion binding protein [BLAST2GO] | 2Fe-2S ferredoxin-like superfamily protein [AT4G32590]
4	Rs065130	30112064	abhydrolase domain-containing protein fam108b1-like [BLAST2GO] | alpha/beta-Hydrolases superfamily protein [AT5G14390]
5	Rs191210	29844872	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT5G42070]
6	Rs392980	7288076	transcription factor bim1 [BLAST2GO] | basic helix-loop-helix (bHLH) DNA-binding superfamily protein [AT5G08130]
7	Rs137390	23325880 23325887	protein transport protein sec61 subunit beta [BLAST2GO] | Preprotein translocase SecSec61-beta subunit protein [AT5G60460]
8	Rs045800	40678277 40678250	Unknown protein
9	Rs065050	30112064	hypothetical protein 23.t00068 [Brassica oleracea] [ABD65627]
10	Rs110050	2765088	aspartic proteinase [BLAST2GO] | Saposin-like aspartyl protease family protein [AT4G04460]
11	Rs191230	29844872	p-glycoprotein 2 [BLAST2GO] | P-glycoprotein 2 [AT4G25960]
12	Rs191130	29844872	scarecrow-like protein 23-like [BLAST2GO] | GRAS family transcription factor [AT5G41920]
13	Rs392920	7288076	rpa70-kda subunit b [BLAST2GO] | Best-Hit: RPA70-kDa subunit B [AT5G08020]
14	Rs045790	40678277 40678250	protein kinase ame3 [BLAST2GO] | Protein kinase superfamily protein [AT4G32660]
15	Rs045820	40678277 40678250	potassium channel kat3 [BLAST2GO] | potassium channel in Arabidopsis thaliana 3 [AT4G32650]
16	Rs229090	5906417	Unknown protein

17	Rs137450	23325880 23325887	Best-Hit: FUNCTIONS IN: molecular_function unknown [AT5G60630]
18	Rs045860	40678277 40678250	enhancer of polycomb-like transcription factor protein [BLAST2GO] | Best-Hit: Enhancer of polycomb-like transcription factor protein [AT4G32620]
19	Rs393000	7288076	suppressor of phytochrome b 5 [BLAST2GO] | Best-Hit: suppressor of phytochrome b 5 [AT5G08150]
20	Rs137380	23325880 23325887	elongation factor 1-alpha [BLAST2GO] | GTP binding Elongation factor Tu family protein [AT5G60390]
21	Rs019090	10850192 10850198	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT1G24160]
22	Rs045840	40678277 40678250	sec24-like transport protein [BLAST2GO] | Best-Hit: Sec23/Sec24 protein transport family protein [AT4G32640]
23	Rs229110	5906417	feruloyl ortho-hydroxylase 1 [BLAST2GO] | 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein [AT3G13610]
24	Rs137410	23325880 23325887	fasciclin-like arabinogalactan protein 11-like [BLAST2GO] | FASCICLIN-like arabinogalactan-protein 12 [AT5G60490]
25	Rs191250	29844872	translation initiation factor eif3 subunit [BLAST2GO] | Translation initiation factor eIF3 subunit [AT5G37475]
26	Rs229130	5906417	disease resistance-responsive (dirigent-like protein) family protein [BLAST2GO] | Disease resistance-responsive (dirigent-like protein) family protein [AT3G13650]
27	Rs229150	5906417	kinase family protein [BLAST2GO] | Protein kinase family protein [AT3G13670]
28	Rs065080	30112064	wd repeat-containing protein 82-b-like [BLAST2GO] | Best-Hit: Transducin/WD40 repeat-like superfamily protein [AT5G14530]
29	Rs116800	2765088	pentatricopeptide repeat-containing protein chloroplastic-like [BLAST2GO] | Pentatricopeptide repeat (PPR) superfamily protein [AT3G48250]
30	Rs191190	29844872	abl interactor-like protein 2 [BLAST2GO] | ABL interactor-like protein 4 [AT5G42030]
31	Rs382210	15034105 15034182	glycerophosphoryl diester phosphodiesterase [BLAST2GO] | PLC-like phosphodiesterases superfamily protein [AT1G74210]
32	Rs019110	10850192 10850198	formin-like protein [BLAST2GO] | Best-Hit: formin 8 [AT1G70140]
33	Rs382180	15034105 15034182	Unknown protein

34	Rs045890	40678277 40678250	e3 ubiquitin-protein ligase at4g11680-like [BLAST2GO] | RING/U-box superfamily protein [AT4G32600]
35	Rs116820	2765088	ribosomal protein l7ae l30e s12e gadd45 family protein [BLAST2GO] | Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein [AT5G20160]
36	Rs229190	5906417	3-epi-6-deoxocathasterone 23-monooxygenase-like [BLAST2GO] | cytochrome P450 family 90 subfamily D polypeptide 1 [AT3G13730]
37	Rs392990	7288076	ring u-box domain-containing protein [BLAST2GO] | RING/U-box superfamily protein [AT5G08139]
38	Rs019130	10850192 10850198	rna-binding ras-gap sh3 binding protein [BLAST2GO]
39	Rs392890	7288076	glucan endo- -beta-glucosidase-like protein 3 [BLAST2GO] | glucan endo-13-beta-glucosidase-like protein 3 [AT5G08000]
40	Rs065100	30112064	surfeit locus protein 2 [BLAST2GO] | Surfeit locus protein 2 (SURF2) [AT5G14440]
41	Rs191200	29844872	dcd (development and cell death) domain protein [BLAST2GO] | DCD (Development and Cell Death) domain protein [AT5G42050]
42	Rs392970	7288076	movement protein binding protein 2c [BLAST2GO] | movement protein binding protein 2C [AT5G08120]
43	Rs382200	15034105 15034182	probable lrr receptor-like serine threonine-protein kinase at1g74360-like [BLAST2GO] | Leucine-rich repeat protein kinase family protein [AT1G74360]
44	Rs191140	29844872	tbc domain-containing protein [BLAST2GO] | Ypt/Rab-GAP domain of gyp1p superfamily protein [AT5G41940]
45	Rs382190	15034105 15034182	zinc-finger protein 1 [BLAST2GO] | zinc-finger protein 1 [AT5G67450]
46	Rs065040	30112064	rna helicase drh1 [BLAST2GO] | Best-Hit: DEAD box RNA helicase family protein [AT5G14610]
47	Rs191170	29844872	upf0160 protein mitochondrial-like [BLAST2GO] | Best-Hit: Metal-dependent protein hydrolase [AT5G41970]
48	Rs137460	23325880 23325887	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT5G60650]
49	Rs191220	29844872	dynamin-related protein 1a [BLAST2GO] | dynamin-like protein [AT5G42080]
50	Rs045810	40678277 40678250	potassium channel kat3 [BLAST2GO] | potassium channel in Arabidopsis thaliana 3 [AT4G32650]
51	Rs065060	30112064	rna helicase drh1 [BLAST2GO] | Best-Hit: DEAD box RNA helicase family protein [AT5G14610]

52	Rs191160	29844872	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT5G41960]
53	Rs191260	29844872	dna-binding storekeeper transcriptional regulator [BLAST2GO] | DNA-binding storekeeper protein-related transcriptional regulator [AT1G11510]
54	Rs045850	40678277 40678250	ap -like zinc finger domain-containing protein [BLAST2GO] | ArfGap/RecO-like zinc finger domain-containing protein [AT4G32630]
55	Rs229080	5906417	Unknown protein
56	Rs191120	29844872	alpha beta fold family protein [BLAST2GO] | alpha/beta-Hydrolases superfamily protein [AT5G41900]
57	Rs045780	40678277 40678250	e3 ubiquitin-protein ligase march6-like [BLAST2GO] | RING/FYVE/PHD zinc finger superfamily protein [AT4G32670]
58	Rs392940	7288076	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT5G08060]
59	Rs045830	40678277 40678250	potassium channel kat3 [BLAST2GO] | potassium channel in Arabidopsis thaliana 3 [AT4G32650]
60	Rs137400	23325880 23325887	transcription factor hb29-like [BLAST2GO] | homeobox protein 26 [AT5G60480]
61	Rs229120	5906417	probable polyamine transporter at3g13620-like [BLAST2GO] | Amino acid permease family protein [AT3G13620]
62	Rs045870	40678277 40678250	copper ion binding protein [BLAST2GO] | copper ion binding [AT4G32610]
63	Rs109080	2765088	Unknown protein
64	Rs137420	23325880 23325887	late embryogenesis abundant protein [BLAST2GO] | late embryogenesis abundant protein-related / LEA protein-related [AT5G60530]
65	Rs019080	10850192 10850198	paired amphipathic helix repeat-containing protein [BLAST2GO] | Best-Hit: Paired amphipathic helix (PAH2) superfamily protein [AT1G27240]
66	Rs191180	29844872	nuclease harbi1-like [BLAST2GO] | Best-Hit: CONTAINS InterPro DOMAIN/s: Putative harbinger transposase-derived nuclease (InterPro:IPR006912) [AT5G41980]

67	Rs229100	5906417	calmodulin binding [BLAST2GO] | calmodulin-binding family protein [AT3G13600]
68	Rs392900	7288076	zinc ion binding protein [BLAST2GO] | Best-Hit: zinc ion binding [AT5G61110]
69	Rs392910	7288076	uncharacterized protein [BLAST2GO] | Best-Hit: unknown protein [AT5G08010]
70	Rs116810	2765088	mutant tfiif-alpha [BLAST2GO] | Best-Hit: transcription activators [AT4G12610]
71	Rs229170	5906417	rna recognition motif-containing protein [BLAST2GO] | RNA-binding (RRM/RBD/RNP motifs) family protein [AT3G13700]
72	Rs065090	30112064	beta- -n-acetylglucosaminyltransferase family protein [BLAST2GO] | beta-14-N-acetylglucosaminyltransferase family protein [AT5G14480]
73	Rs045880	40678277 40678250	mitochondrial acidic protein mam33-like [BLAST2GO] | Mitochondrial glycoprotein family protein [AT4G32605]
74	Rs229140	5906417	kinase family protein [BLAST2GO] | Protein kinase family protein [AT3G13670]
75	Rs392930	7288076	glycerophosphoryl diester phosphodiesterase [BLAST2GO] | PLC-like phosphodiesterases superfamily protein [AT5G08030]
76	Rs392960	7288076	low quality protein: atp-dependent helicase hrq1-like [BLAST2GO] | ATP-dependent helicases [AT5G08110]
77	Rs065110	30112064	e3 ubiquitin-protein ligase rglg2 [BLAST2GO] | RING domain ligase2 [AT5G14420]
78	Rs229180	5906417	prenylated rab acceptor family protein [BLAST2GO] | PRA1 (Prenylated rab acceptor) family protein [AT3G13720]
79	Rs065030	30112064	dna (cytosine-5)-methyltransferase drm1 [BLAST2GO] | domains rearranged methyltransferase 2 [AT5G14620]
80	Rs229160	5906417	kinase family protein [BLAST2GO] | Protein kinase protein with adenine nucleotide alpha hydrolases-like domain [AT3G13690]

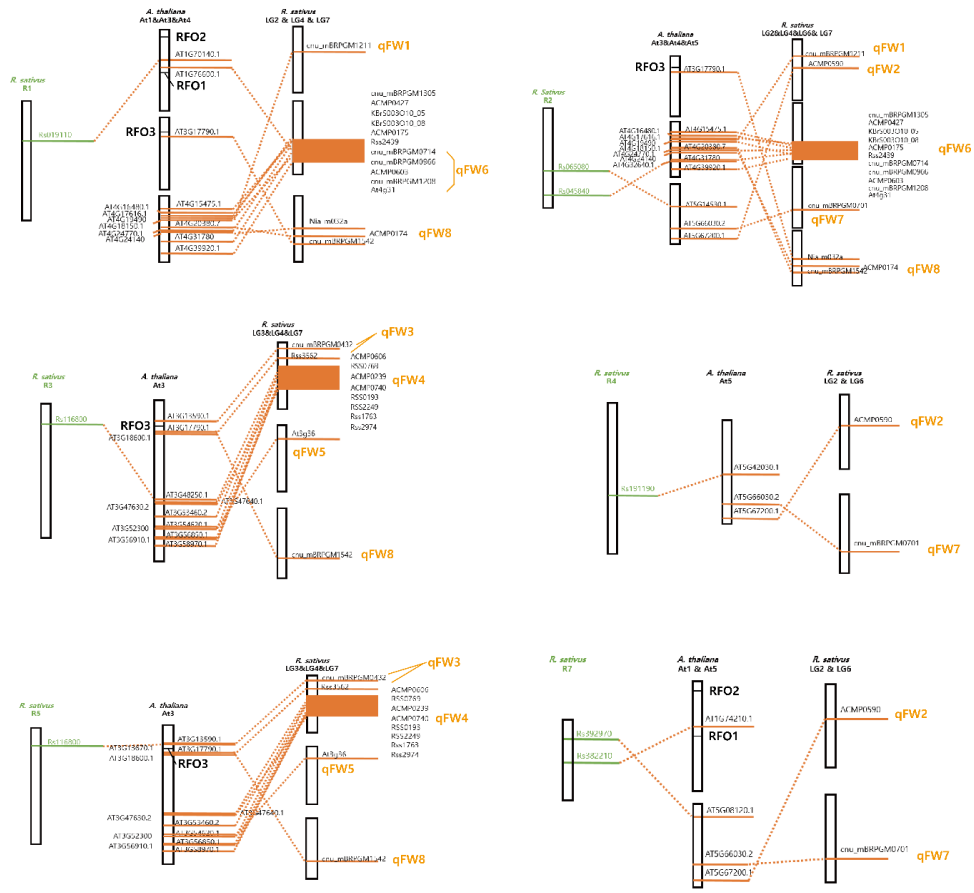

Figure 4-4. Genetic map and distribution of the position of Fsarium wilt resistance gene in *R. sativus*. *R. sativus* genome in the present study (left), *A. thaliana* genome (middle), and *R. sativus* genome in the previous QTL study (Yu et al., 2013, right). Locus designations are provided on the right side of the chromosome. QTL positions are indicated by orange colored-letter.

4.5 Discussion

Genome-wide association study has been used to detect putative functional markers and candidate genes related to complex traits of various plant species (Iwata et al. 2010; Pasam et al. 2012; Zhao et al. 2011). Genotyping-by-sequencing (GBS), high throughput and cost-effective genotyping methods, can support GWAS even in species with limited genomic information (Byrne et al. 2013). In this study, we reported the results of GBS-GWAS in a germplasm panel of 227 radish accessions to identify single nucleotide polymorphism (SNP) and candidate genes associated with Fusarium wilt resistance.

In the analysis of population structure, three clusters were confirmed (K=3) (Figure 4-2). East Asian cultivated radishes were divided into two groups, and are distinct from European cultivated radishes. This is supported with that Asian cultivated radish types are closely related to wild Asian accessions, but are distinct from European/American cultivated radishes (Kim et al. 2016). For linkage disequilibrium (LD), it was affected by many different factors, such as natural selection, domestication, founding events, genetic diversity, and population stratification (Bastien et al. 2014; Hyten et al. 2010; Vuong et al. 2015). In maize, the average decline of LD distance was 5 – 100kb from the SNP analysis of genetic diversity using 447 germplasm

(Xing et al. 2011; Yan et al. 2009). Long LD decay was observed in self-pollinated crops in soybean (125–600 kb), which has very narrow genetic diversity compared to other cultivated crops. We used LD across 32,778 SNPs with minor allele frequency (MAF) > 0.05 to determine the structure of the entire 227 lines. LD decay distance for r^2 greater than 0.2 was 900 kb in radish, which means LD broke down significantly slower than maize, rice (75–150 kb) and soybean (Figure 4-3). Longer LD was explained that the crop has experienced less recombination and contained more common alleles compared to shorter LD decay (Xing et al. 2011; Yan et al. 2009). It was also implied that genetic bottle-neck had increased LD block-size resulting into longer GWAS regions being associated with the phenotypes (Vuong et al. 2015).

In multiple previous association analysis study, due to leading false positive problem of GLM, MLM was thought to be more acceptable than GLM method (Huang et al. 2010; Zhang et al. 2010). A GWAS identified 13 genomic regions strongly associated with FW resistance using the MLM (QKmodel) (Figure 4-5, Table 4-2). Two SNPs are in a gene (Rs01911) that encoding formin 8 proteins which are essential for the creation of actin-based structure in polarized plant cell growth (Vidali et al. 2009). One SNP on chromosome 2 was detected in WD-40 repeat-like superfamily proteins which have been recognized as regulator of plant-specific developmental

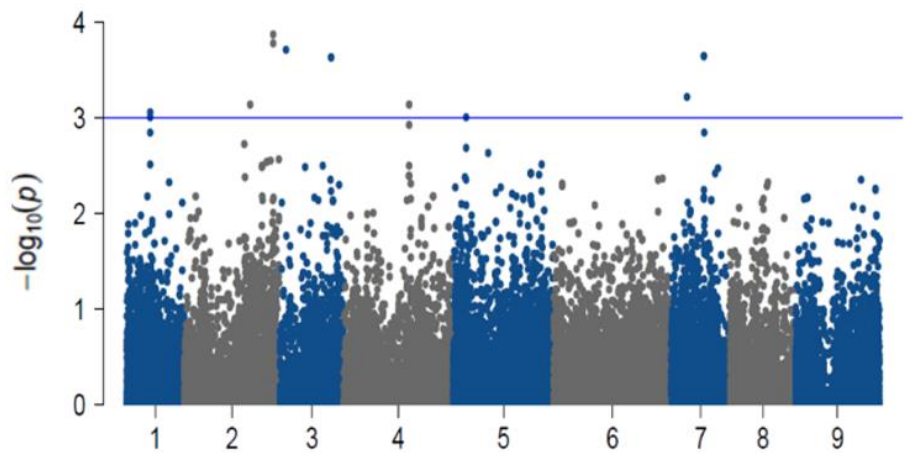

Figure 4-5. Manhattan plots of Radish SNP markers mapped to 9 chromosomes of *R. sativus* using mixed model (Q+K)

Table 4-2. Potential candidate genes detected by GWAS

GeneID	SNP	Encoding	*p* Value
Rs019110	R1_10850192	formin 8	8.69E-04, 9.91E-04
	R1_10850198		
Rs065080	R2_30112064	WD40 repeat-like superfamily protein	7.33E-04
Rs045840	R2_40678277	Sec23/Sec24 protein transport family protein	1.33E-04, 1.67E-04
	R2_40678250		
Rs116800	R3_2765088	Pentatricopeptide repeat (PPR) superfamily protein	1.96E-04
	R3_23325880		2.30E-04
	R3_23325887		2.30E-04
Rs191190	R4_29844872	ABL interactor-like protein 4	7.25E-04
Rs229140	R5_5906417	kinase family protein	9.92E-04
Rs382210	R7_15034105	PLC-like phosphodiesterases superfamily protein	2.24E-04
	R7_15034182		
Rs392970	R7_7288076	movement protein binding protein 2C	6.02E-04

process (Van Nocker and Ludwig 2003). Two SNPs were located in Rs045840 encoded Sec23/Sec24 proteins which are the cytosolic proteins to consist the COPII coat core machinery (Pahuja et al. 2015). Sec24 is responsible for binding to membrane cargo proteins at the endoplasmic reticulum (ER). One of the candidate genes, Rs116800 on chromosome 7, encoded pentatricopeptide repeat (PPR) superfamily proteins localized to the mitochondrial, chloroplast intracellular space, cytosol or nucleus, and have been known to play important roles in plant developmental processes and responses to environmental stresses (Jiang et al. 2015). The candidate Rs382210 encoded PLC-like phosphodiesterase superfamily proteins which integrate signaling cascades by being involved in hydrolytic process (Sedzielewska Toro and Brachmann 2016). The candidate Rs392970 encoded movement protein binding protein 2C (MPB2C) belongs to the group of microtubule-associated proteins, which thought to mediate the versatility and architecture of diverse microtubular assemblies (Ruggenthaler et al. 2009). Further study of gene expression with these SNPs is needed to offer an effective marker-assisted selection (MAS) tool to accelerate radish breeding.

In addition, we arbitrarily searched potential candidate genes in 30Kbp upstream and downstream of the 13 SNP loci. Total 80 genes were identified and some of them are considered to associate with disease resistance, such as LRR and zinc finger proteins (Table S2). Leucine-rich repeat receptor-like

kinase (LRR-RLK) implies to play a crucial role in plant disease resistance pathways (Bent and Mackey 2007). Multiple LRR-RLKs are well known controlling disease interacting or tolerance, having relation with development and growth functions and characterizing in associated with both biotic and abiotic stress conditions (Lee and Choi 2013; Xing et al. 2011). In addition, zinc finger protein is known to play a role as encoding a variety of NBS-LRR type of proteins. Especially, since it has a crucial role in host-pathogen interaction to combine zinc finger protein with NBS-LRR domain, these genes may have a strong connection to the FW in radish (Gupta et al. 2012).

Synteny analysis using the linked markers to the candidate gens, three genomic regions were common with the QTL regions detected in the previous study (Yu et al. 2013). Rs116800 on chromosome 3 showed homology with *qFW4* in LG3 which indicated higher LOD value and percentage of phenotypic variation. *Foc-Bol* region of *B. oleracea* was designated to be involved in *qFW4* , which region is corresponding to chromosome 3 of *A. thaliana* (Pu et al. 2012; Yu et al. 2013). Two candidate genes, Rs019110 and Rs382210, located closely with *RFO1* region of *A. thaliana*. In synteny analysis using linked markers, three SNPs on chromosome 1 and 7 (R1_10850192, R1_10850198, R7_7288076) were mapped to genomic region of the *A. thaliana* genome, which contains RFO1, encodes a member of the wall-associated kinase family of receptor-like kinases (RLKs) and

contribute to immunity against *F. oxisporum* infection in the root vascular cylinder (Diener 2012; Diener and Ausubel 2005). While one candidate gene, Rs229140, positioned closely to *RFO3* region of *A. thaliana*. In previous study, RFO1 and RFO3 encode RLKs were reported that the diversity in RLK genes could be a major source of FW resistance in A. thaliana (Cole and Diener 2013). These candidate genes could be a clue to understand FW resistance in radish.

Fusarium wilt is a severe threat to the radish production in the continuous cropping field with limited prevention for disease. It was reported that resistance to Fusarium wilt is controlled by quantitative multiple genes (Ashizawa et al. 1979; Kaneko et al. 2007). Here, we were able to generate a large number of genome-wide markers by using GBS, which made GWAS possible to identify 8 candidate genes and 13 SNPs significantly associated with FW resistance. The identification of these significant markers associated with FW resistance would provide new insight for further study related to FW mechanism in *R. sativus* and offer a fundamental knowledge for breeding FW resistant radish.

Reference

Abby S, Daubin V (2007) Comparative genomics and the evolution of prokaryotes. Trends in microbiology 15:135-141

Altekruse S, Cohen M, Swerdlow D (1997) Emerging foodborne diseases. Emerging infectious diseases 3:285

Andreeva-Kovalevskaya ZI, Solonin A, Sineva E, Ternovsky V (2008) Pore-forming proteins and adaptation of living organisms to environmental conditions. Biochemistry (Moscow) 73:1473-1492

Armstrong G, Armstrong JK (1976) Common hosts for Fusarium oxysporum formae speciales spinaciae and betae. Phytopathology 66:542-545

Ashizawa M, Hida K-i, Yoshikawa H (1979) Studies on the breeding of Fusarium resistance in radish. I. Screening of radish varieties for Fusarium resistance. Yasai Shikenjo hokoku= Bulletin of the Vegetable and Ornamental Crops Research Station Series A

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627-631

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M (2008) The RAST Server: rapid annotations using subsystems technology. BMC genomics 9:1

Baida GE, Kuzmin NP (1996) Mechanism of action of hemolysin III from Bacillus cereus. Biochimica et Biophysica Acta (BBA)-Biomembranes 1284:122-124

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7:781-791

Bastien M, Sonah H, Belzile F (2014) Genome wide association mapping of resistance in soybean with a genotyping-by-sequencing approach. The Plant Genome 7

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological):289-300

Bent AF, Mackey D (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. Annu Rev Phytopathol 45:399-436

Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry. 5th. New York: WH Freeman 38:76

Bose T, Haque MM, Reddy C, Mande SS (2015) COGNIZER: a framework for functional annotation of metagenomic datasets. PloS one 10:e0142102

Bosland PW, Williams PH (1987) An evaluation of Fusarium oxysporum from crucifers based on pathogenicity, isozyme polymorphism, vegetative compatibility, and geographic origin. Canadian Journal of Botany 65:2067-2073

Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society of London B: Biological Sciences 277:819-827

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. The American Journal of Human Genetics 98:116-126

Butcher BG, Helmann JD (2006) Identification of Bacillus subtilis σW-dependent genes that provide intrinsic resistance to antimicrobial compounds produced by Bacilli. Molecular microbiology 60:765-782

Byrne S, Czaban A, Studer B, Panitz F, Bendixen C, Asp T (2013) Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. PLoS One 8:e57438

Carretto E, Barbarini D, Poletti F, Capra FM, Emmi V, Marone P (2000) Bacillus cereus fatal bacteremia and apparent association with nosocomial transmission in an intensive care unit. Scandinavian journal of infectious diseases 32:98-100

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25:119-120

Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ (2012) Defining bacterial species in the genomic era: insights from the genus Acinetobacter. BMC microbiology 12:302

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature genetics 45:400-405

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A,

Copeland A, Huddleston J, Eichler EE (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods 10:563-569

Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban H-J, Yoon D, Lee MH, Kim D-J, Park M (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature genetics 41:527-534

Cole SJ, Diener AC (2013) Diversity in receptor-like kinase genes is a major determinant of quantitative resistance to Fusarium oxysporum f. sp. matthioli. New Phytologist 200:172-184

Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI, Ritchie MD, Stein CM, Roden DM, Smith JD (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112:1022-1027

Davidson AL, Dassa E, Orelle C, Chen J (2008) Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiology and Molecular Biology Reviews 72:317-364

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One 8:e62137

Delcher AL, Salzberg SL, Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. Current Protocols in Bioinformatics:10.13. 11-10.13. 18

Diener A (2012) Visualizing and quantifying Fusarium oxysporum in the plant host. Molecular Plant-Microbe Interactions 25:1531-1541

Diener AC, Ausubel FM (2005) RESISTANCE TO FUSARIUM OXYSPORUM 1, a dominant Arabidopsis disease-resistance gene, is not race specific. Genetics 171:305-321

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. science 314:1461-1463

Duport C, Jobin M, Schmitt P (2016) Adaptation in Bacillus cereus: From Stress to Disease. Frontiers in microbiology 7

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-1093

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6:e19379

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587

Garibaldi A, Gilardi G, Gullino ML (2006) Evidence for an expanded host range ofFusarium oxysporum f. sp. raphani. Phytoparasitica 34:115-121

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International journal of systematic and evolutionary microbiology 57:81-91

Granum PE (2005) Bacillus cereus. Foodborne pathogens: microbiology and molecular biology:409-414

Granum PE, Lund T (1997) Bacillus cereus and its food poisoning toxins. FEMS microbiology letters 157:223-228

Guillemet E, Cadot C, Tran S-L, Guinebretière M-H, Lereclus D, Ramarao N (2010) The InhA metalloproteases of Bacillus cereus contribute concomitantly to virulence. Journal of bacteriology 192:286-294

Guinebretière M-H, Broussolle V (2002) Enterotoxigenic profiles of food-poisoning and food-borne Bacillus cereus strains. Journal of Clinical Microbiology 40:3053-3056

Gupta SK, Rai AK, Kanwar SS, Sharma TR (2012) Comparative analysis of zinc finger proteins involved in plant disease resistance. PLoS One 7:e42578

Hall T (2011) BioEdit: an important software for molecular biology. GERF Bull Biosci 2:6

Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469-474

Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA,

McEvoy BP, Schrage AJ, Grant JD, Chou Y-L (2011) A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. Biological psychiatry 70:513-518

Herten K, Hestand MS, Vermeesch JR, Van Houdt JK (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC bioinformatics 16:1

Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J (2007) Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. Journal of bacteriology 189:8186-8195

Houry A, Briandet R, Aymerich S, Gohar M (2010) Involvement of motility and flagella in Bacillus cereus biofilm formation. Microbiology 156:1009-1018

Hsueh Y-H, Somers EB, Lereclus D, Wong ACL (2006) Biofilm formation by Bacillus cereus is influenced by PlcR, a pleiotropic regulator. Applied and environmental microbiology 72:5089-5092

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics 42:961-967

Hyten DL, Cannon SB, Song Q, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC genomics 11:38

Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikhailova N, Lapidus A (2003) Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 423:87-91

Iwata H, Ebana K, Uga Y, Hayashi T, Jannink J-L (2010) Genome-wide association study of grain shape variation among Oryza sativa L. germplasms based on elliptic Fourier analysis. Molecular Breeding 25:203-215

Jeong Y-M, Kim N, Ahn BO, Oh M, Chung W-H, Chung H, Jeong S, Lim K-B, Hwang Y-J, Kim G-B (2016) Elucidating the triplicated ancestral genome structure of radish based on chromosome-level comparison with the Brassica genomes. Theoretical and Applied Genetics:1-16

Jiang SC, Mei C, Liang S, Yu YT, Lu K, Wu Z, Wang XF, Zhang DP (2015)

Crucial roles of the pentatricopeptide repeat protein SOAR1 in Arabidopsis response to drought, salt and cold stresses. Plant molecular biology 88:369-385

JingYuan Z, GuangCai D, HaiYan Y, QingTang F, YuanLin X (2011) Multi-drug resistance and characteristic of integrons in Shigella spp. isolated from China. Biomedical and Environmental Sciences 24:56-61

Johnson K, Nelson C, Busta F (1982) Germination and Heat Resistance of Bacillus cereus Spores from Strains Associated with Diarrheal and Emetic Food-Borne Illnesses. Journal of Food Science 47:1268-1271

Juliana P, Rutkoski JE, Poland JA, Singh RP, Murugasamy S, Natesan S, Barbier H, Sorrells ME (2015) Genome-wide association mapping for leaf tip necrosis and pseudo-black chaff in relation to durable rust resistance in wheat. The Plant Genome 8

Kaneko Y, Kimizuka-Takagi C, Bang SW, Matsuzawa Y (2007) Radish. Vegetables. Springer, pp 141-160

Katoh K, Toh H (2010) Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26:1899-1900

Kendrick J, Snyder W (1942) Fusarium wilt of radish. Phytopathology 32:1-1033

Kim KM, Sung S, Caetano-Anollés G, Han JY, Kim H (2008) An approach of orthology detection from homologous sequences under minimum evolution. Nucleic acids research 36:e110-e110

Kim N, Jeong Y-M, Jeong S, Kim G-B, Baek S, Kwon Y-E, Cho A, Choi S-B, Kim J, Lim W-J (2016) Identification of candidate domestication regions in the radish genome based on high-depth resequencing analysis of 17 genotypes. Theoretical and Applied Genetics 129:1797-1814

Kitamura S (1958) Varieties and transitions of radish. Japanese Radish, Jpn Sci Soc, Tokyo:1-19

Knepper C, Day B (2010) From perception to activation: the molecular-genetic and biochemical landscape of disease resistance signaling in plants. The Arabidopsis Book:e012

Kopta T, Pokluda R (2013) Yields, quality and nutritional parameters of radish (Raphanus sativus) cultivars when grown in organically in Czech Republic. Hort Sci 40:16-21

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9:1

Kotiranta A, Lounatmaa K, Haapasalo M (2000) Epidemiology and pathogenesis of Bacillus cereus infections. Microbes and Infection 2:189-198

Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, Adak G, Levine M (1999) Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. Bulletin of the World Health Organization 77:651-666

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet 4:e1000304

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular biology and evolution:msw054

Lü N, Yamane K, Ohnishi O (2008) Genetic diversity of cultivated and wild radish and phylogenetic relationships among Raphanus and Brassica species revealed by the analysis of trnK/matK sequence. Breeding Science 58:15-22

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National academy of sciences of the United States of America 102:10557-10562

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9:357-359

Lee S, Choi D (2013) Comparative transcriptome analysis of pepper (Capsicum annuum) revealed common regulons in multiple stress conditions and hormone treatments. Plant cell reports 32:1351-1359

Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW (2012) Genomic variation in Salmonella enterica core genes for epidemiological typing. BMC genomics 13:88

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nature Genetics 45:43-50

Li S (1988) The origin and resources of vegetable crops in China. International symposium on horticultural germplasm, cultivated and wild, Beijing, China, pp 197-202

Lim S-H, Song J-H, Kim D-H, Kim JK, Lee J-Y, Kim Y-M, Ha S-H (2016) Activation of anthocyanin biosynthesis by expression of the radish R2R3-MYB transcription factor gene RsMYB1. Plant cell reports 35:641-653

Liu W, Fang L, Li M, Li S, Guo S, Luo R, Feng Z, Li B, Zhou Z, Shao G (2012) Comparative genomics of Mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. PLoS One 7:e35698

Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced Escherichia coli genomes. Microbial ecology 60:708-720

Majed R, Faille C, Kallassy M, Gohar M (2016) Bacillus cereus Biofilms—Same, Only Different. Frontiers in Microbiology 7

Malorny B, Hauser E, Dieckmann R (2011) New approaches in subspecies-level Salmonella classification. Salmonella From Genome to Function:1-23

Mann RA, Smits TH, Bühlmann A, Blom J, Goesmann A, Frey JE, Plummer KM, Beer SV, Luck J, Duffy B (2013) Comparative genomics of 12 strains of Erwinia amylovora identifies a pan-genome with a large conserved core. PLoS One 8:e55644

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20:1297-1303

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Current opinion in genetics & development 15:589-594

Miller W, Makova KD, Nekrutenko A, Hardison RC (2004) Comparative genomics. Annu Rev Genomics Hum Genet 5:15-56

Muzzi A, Donati C (2011) Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. International Journal of Medical Microbiology 301:619-622

Nei M (1996) Phylogenetic analysis in molecular evolutionary genetics. Annual review of genetics 30:371-403

Newell DG, Koopmans M, Verhoef L, Duizer E, Aidara-Kane A, Sprong H, Opsteegh M, Langelaar M, Threfall J, Scheutz F (2010) Food-borne diseases—the challenges of 20years ago still persist while new ones continue to emerge. International journal of food microbiology 139:S3-S15

Niyogi SK (2005) Shigellosis. Journal of microbiology (Seoul, Korea) 43:133-143

Ouzounis C, Kyrpides N (1996) The emergence of major cellular processes in evolution. FEBS letters 390:119-123

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691-3693

Pahuja KB, Wang J, Blagoveshchenskaya A, Lim L, Madhusudhan M, Mayinger P, Schekman R (2015) Phosphoregulatory protein 14-3-3 facilitates SAC1 transport from the endoplasmic reticulum. Proceedings of the National Academy of Sciences 112:E3199-E3206

Park HG KO, Kim HT, Na JH, Park Y, Park JY, Park CS, Song KH, Yang DH, Om YH, Yoo IW, Yoon J-Y, Lee Bs, Sug HD, Jeong SY, Oh D-G, Cheong JW, Cho YH, Cho Y-S, Cho YC (2008) The recent history of vegetable seed industry in Korea. Seoul National Univ Press, Korea

Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, Graner A (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC plant biology 12:16

Peterson JL, Pound G (1960) Studies on resistance in Radish to Fusarium oxysporum f. eonglutinans. Phytopathology 50

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29:170-175

Pond SLK, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K (2011) A random effects branch-site model for detecting episodic diversifying selection. Molecular biology and evolution:msr125

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38:904-909

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Pu Z-j, Shimizu M, Zhang Y-j, Nagaoka T, Hayashi T, Hori H, Matsumoto S, Fujimoto R, Okazaki K (2012) Genetic mapping of a fusarium wilt resistance gene

in Brassica oleracea. Molecular breeding 30:809-818

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81:559-575

Rasko DA, Altherr MR, Han CS, Ravel J (2005) Genomics of the Bacillus cereus group of organisms. FEMS microbiology reviews 29:303-329

Ruggenthaler P, Fichtenbauer D, Krasensky J, Jonak C, Waigmann E (2009) Microtubule-associated protein AtMPB2C plays a role in organization of cortical microtubules, stomata patterning, and tobamovirus infectivity. Plant physiology 149:1354-1365

Ryu J-H, Beuchat LR (2005) Biofilm formation and sporulation by Bacillus cereus on a stainless steel surface and subsequent resistance of vegetative cells and spores to chlorine, chlorine dioxide, and a peroxyacetic acid–based sanitizer. Journal of Food Protection® 68:2614-2622

Sakiroglu M, Brummer EC (2016) Identification of loci controlling forage yield and nutritive value in diploid alfalfa using GBS-GWAS. Theoretical and Applied Genetics:1-8

Salamitou S, Ramisse F, Brehélin M, Bourguet D, Gilois N, Gominet M, Hernandez E, Lereclus D (2000) The plcR regulon is involved in the opportunistic properties of Bacillus thuringiensis and Bacillus cereus in mice and insects. Microbiology 146:2825-2832

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences 74:5463-5467

Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, Kubo M, Kawaguchi T, Tsunoda T, Watanabe M, Takeda A (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nature genetics 41:1303-1307

Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M-A, Roy SL, Jones JL, Griffin PM (2011) Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17

Schroeder GN, Hilbi H (2008) Molecular pathogenesis of Shigella spp.:

controlling host cell signaling, invasion, and death by type III secretion. Clinical microbiology reviews 21:134-156

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. science 316:1341-1345

Sedzielewska Toro K, Brachmann A (2016) The effector candidate repertoire of the arbuscular mycorrhizal fungus Rhizophagus clarus. BMC genomics 17:101

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics:btu153

Shirasawa K, Oyama M, Hirakawa H, Sato S, Tabata S, Fujioka T, Kimizuka-Takagi C, Sasamoto S, Watanabe A, Kato M (2011) An EST-SSR linkage map of Raphanus sativus and comparative genomics of the Brassicaceae. DNA research 18:221-232

Smith S, Ebbels D, Garber R, Kappelman A (1981) Fusarium wilt of cotton. Fusarium: diseases, biology, and taxonomy Pennsylvania State University, University Park:29-38

Spielman SJ, Wilke CO (2015) The relationship between dN/dS and scaled selection coefficients. Molecular biology and evolution:msv003

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology 56:564-577

Tauxe RV (1997) Emerging foodborne diseases: an evolving public health challenge. Emerging infectious diseases 3:425

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America 102:13950-13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Current opinion in microbiology 11:472-477

van den Oord EJ (2008) Controlling false discoveries in genetic studies. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 147:637-

644

Van Nocker S, Ludwig P (2003) The WD-repeat protein superfamily in Arabidopsis: conservation and divergence in structure and function. BMC genomics 4:50

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Current opinion in microbiology 23:148-154

Vidali L, van Gisbergen PA, Guerin C, Franco P, Li M, Burkart GM, Augustine RC, Blanchoin L, Bezanilla M (2009) Rapid formin-mediated actin-filament elongation is essential for polarized plant cell growth. Proceedings of the National Academy of Sciences of the United States of America 106:13341-13346

Visioni A, Tondelli A, Francia E, Pswarayi A, Malosetti M, Russell J, Thomas W, Waugh R, Pecchioni N, Romagosa I (2013) Genome-wide association mapping of frost tolerance in barley (Hordeum vulgare L.). BMC genomics 14:1

Vuong TD, Sonah H, Meinhardt C, Deshmukh R, Kadam S, Nelson R, Shannon J, Nguyen H (2015) Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC genomics 16:593

Wang Y, Mo X, Zhang L, Wang Q (2011) Four superoxide dismutase (isozymes) genes of Bacillus cereus. Annals of Microbiology 61:355-360

Wu J, Feng F, Lian X, Teng X, Wei H, Yu H, Xie W, Yan M, Fan P, Li Y (2015) Genome-wide Association Study (GWAS) of mesocotyl elongation based on re-sequencing approach in rice. BMC plant biology 15:218

Wu S, Zhu Z, Fu L, Niu B, Li W (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC genomics 12:1

Xing HT, Guo P, Xia XL, Yin WL (2011) PdERECTA, a leucine-rich repeat receptor-like kinase of poplar, confers enhanced water use efficiency in Arabidopsis. Planta 234:229-241

Yamane K, Lü N, Ohnishi O (2005) Chloroplast DNA variations of cultivated radish and its wild relatives. Plant Science 168:627-634

Yamane K, Lü N, Ohnishi O (2009) Multiple origins and high genetic diversity of cultivated radish inferred from polymorphism in chloroplast simple sequence repeats. Breeding Sci 59:55-65

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize

collection using SNP markers. PloS one 4:e8451

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 24:1586-1591

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of molecular evolution 46:409-418

Yu X, Choi SR, Ramchiary N, Miao X, Lee SH, Sun HJ, Kim S, Ahn CH, Lim YP (2013) Comparative mapping of Raphanus sativus genome using Brassica markers and quantitative trait loci analysis for the Fusarium wilt resistance trait. Theoretical and applied genetics 126:2553-2562

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular biology and evolution 22:2472-2479

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM (2010) Mixed linear model approach adapted for genome-wide association studies. Nature genetics 42:355-360

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature communications 2:467

Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics 28:416-418

국문초록

유전체 특성 및 병원성에 대한 이해

구현진

농생명공학부

서울대학교 대학원 농업생명과학대학

이번 유전체 전체에 대한 연구는 유전체 특성 및 병인에 대한 통찰력을 얻기 위해 수행되었다. 많은 연구자들이 살아있는 유기체와 병원체의 상호 작용에 초점을 맞추어 연구를 하려고 노력하였지만, 질병에 이르는 병인의 메커니즘에 대해서는 유전적 수준에서 연구가 수행되지 않았다. 따라서 유전체 수준에서 포괄적인 이해를 넓히기 위해 비교 분석과 유전체 관련 연구를 실시했다.

2 장에서 나는 식중독을 유발하는 위장병 병원균으로 잘 알려진 *B. cereus* 균주의 완전한 유전체로 비교 유전체 분석을 수행했다. 비교 분석 결과를 토대로 병인 발생에 중요한 역할을 하는 병독성 인자, 선택 압력을 받은 유전자 및 FORC_013 의 strain-specific genes 을 확인할 수 있었다.

3 장에서 NCBI 유전체 데이터베이스에 20 개의 완전 유전체를 수집하여 범 유전체 분석을 수행하였다. 나는 *Shigella* spp. 에서

전반적인 유전적 내용뿐만 아니라 수평적 유전자 전달을 통해 새롭게 획득된 유전자들도 확인하였다. 그리고 이 종들은 핵심 유전자 클러스터링과 ANI 값에 기초한 계통수 분석을 사용하여 단일 계통군으로 나뉘는 것을 알 수 있었다. 선택 압력을 받는 유전자는 핵심 유전체에서 매우 낮은 비율을 차지하여 핵심 유전체 보존이 잘 되어 있다는 사실도 알 수 있었다.

제 4 장에서는 표현형 데이터와 유전형 데이터 사이의 관계를 규명하기 위해 무 (*Raphanus sativus*)의 fusarium wilt 저항성에 대한 전장 유전체 분석이 수행되었다. 이 연구를 통해 fusarium disease를 일으키는 중요한 좌들을 발견하였다.

이러한 연구를 통해 나는 살아있는 생물체의 유전적 특징을 이해할 수 있었을 뿐만 아니라, 유전체 수준의 병인 기작에 관한 앞으로의 연구에 좀 더 포괄적인 통찰력을 제공할 수 있었다.

주요어: 비교 유전체 분석, 범 유전체 분석, 전장 유전체 분석, 병원성

학 번: 2015-23125