Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

GENOME CHARACTERIZATION OF NOVEL *BACILLUS CEREUS*-GROUP INFECTING BACTERIOPHAGES

By

Cheng-Han Chung

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Biological Sciences

Idaho State University

Summer 2015

Committee Approval

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Cheng-Han Chung find it satisfactory and recommend that it be accepted.

_____

Michael A. Thomas, Ph.D.
Advisor

_____

Shu-Chuan (Grace) Chen, Ph.D.
Advisor

_____

Vern Winston, Ph.D.
Committee Member

_____

Peter Sheridan, Ph.D.
Committee Member

_____

Caryn Evilia, Ph.D.
Graduate Faculty Representative

DEDICATION

I dedicate this thesis to my parents (Ruei-Chuan Chung and Ruei-Wen Chang) who have always been unwavering supportive of me and my studies throughout my life.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr. Michael A. Thomas for his mentorship, patience and understanding during my graduate studies at Idaho State University. He has given me the opportunity to develop my own research topics while training my independent thinking. He is a role model as a scientist who taught me the critical thinking and time management. He is also a friend who told me how to enjoy the life during the graduate study.

I especially extend my heartfelt gratitude to my supervisor and advisor, Dr. Shu-Chuan (Grace) Chen. Her guidance and tremendous support allowed me to explore different topics in Statistics and Biology during my research journey. I am very lucky to have both advisors on board to lead me through my research journey.

I am grateful to have Dr. Michael H. Walter to be the principle investigator of this project. He is the lead researcher that brought me into the 'phage world'. His guidance and knowledge in bacteriophage allows me to grasp the concept of this project from start to finish. I am also grateful for his generosity to share all the materials and techniques in this project.

I would like to thank my committee – Dr. Vern Winston, Dr. Caryn Evilia and Dr. Peter P. Sheridan. I want to give my gratitude to Dr. Vern Winston for providing the space and supplies in his laboratory for my research project. He also inspired me by sharing his knowledge and experience in Virology to initiate this project. I am also thankful to Dr. Caryn Evilia who troubleshoots the experimental steps of phage DNA extraction and purification.

I wish to express my sincere thanks to Dr. Luobin Yang. I wouldn't have been able to accomplish the computational analysis without his assistance and maintenance of servers and clusters on GPU2 and Galaxy. I am thankful to the staffs in MRCF for helping me in getting the sequencing data. I am also thankful to Graham F. Hatfull's lab for sharing the sequencing data of nine bacteriophages. I want to show my gratitude to my colleagues, Dr. Gaurav Kaushik and Roger Long for helping me and learning with me during my research. I would like to thank the graduate students from the department of Biology who ever helped me in experiments, research and thesis writing.

I take this opportunity to express my gratitude to my wife, Jing-Huei Huang, for giving me support and faith to tackle challenges head on during my research studies. She is always there for me, and she is the one who makes me feel like home when we are far away from home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BJC: novel *Bacillus anthracis* phage name in this study

BLAST: Basic Local Alignment Search Tool

cryo-ET: cryo-electron tomography

DTR: Direct terminal repeat

dUTPase: dUTP diphosphatase

ICTV: The International Committee on Taxonomy of Viruses

HHblits: HMM-HMM-based lightning-fast iterative sequence search

HMM: Hidden Markov Model

NCBI: National Center for Biotechnology Information

NCR: Neighboring coverage ratio

NGS: Next Generation Sequencing

ORF: Open reading frame

QCM: Quartz crystal microbalance

RPB: Rotations per minute

rpsBLAST: Reversed Position Specific BLAST

SBP8a: Spore-binding phage 8a

SigF: RNA polymerase sigma 28 subunit F

SigG: RNA polymerase sigma 28 subunit G

TerL: Terminase large subunit

TerS: Terminase small subunit

TP: Terminal protein

TR: Transcription regulator

TSA: Tryptic Soy Agar

TSB: Tryptic Soy Broth

ABSTRACT

In this study, we sequenced 48 isolates of phages infecting *Bacillus anthracis* and analyzed Next Generation Sequencing (NGS) data to predict the terminus sequence of novel phages. These *Bacillus anthracis* phages were classified into four major clusters. Their terminal sequences were successfully identified using a method developed for this study, along with terminus validation by direct sequencing and inference of terminal types and DNA packaging strategies by phylogenetic analysis. Phage BJC (I48-like), SBP8a and QCM8 (Q8-like) were found to have exact terminal repeats up to 6,731 bp according to evident characteristics at the genome redundant region by analyzing NGS data, while QCM11 presented a linear sequence without genome redundancy. The terminus prediction of nine published phages using contig sequences and raw read data demonstrated that different types of DNA packaging mechanisms and terminal sequences could be identified by NGS data. BJC, SBP8a and QCM8 were predicted as *Myoviridae* with a DNA packaging strategy similar to SPO1. The genome characterization of QCM11 reveals that 12 putative proteins including the terminal protein are conservative between QCM11 and MG-B1. QCM11 is the first *Podoviridae* phage to be sequenced that infects *Bacillus anthracis*. This study provides a package to effectively identify terminal sequences based on read and contig characteristics of phage genome NGS data, suggesting that the complete genome configuration of a phage sequence should be undertaken for genome sequence documentation in databases and is crucial for subsequent analyses including gene annotation and comparative genomics. With the aid of high throughput sequencing along with the terminus prediction method, this study gives insights into characterization of genome configuration for novel bacteriophages.

CHAPTER 1: INTRODUCTION

This chapter explored the literature of bacteriophage genetics. Specifically, it focused on the phage structure genetics, physical maps, and associated mechanisms among tailed phages. Additionally, the experimental analyses and computer-based methods for identifying the phage genome terminus and underlying genome configuration in literature were reviewed. Finally, the primary research objectives of this study and approaches used were described.

1.1 Taxonomy of bacteriophages

The bacteriophage (phage) is an entity of viruses that infect bacteria. They possess high specificity and sensitivity to the target host, creating potential biomedical and scientific research applications. The International Committee on Taxonomy of Viruses (ICTV) uses 70 properties as criteria to classify the bacterial viruses and includes 14 officially accepted families along with an amount of unassigned bacteriophages to five potential families (Ackermann, 2009). To date, those identified phages that are infecting bacteria belong to either *Caudovirales* or *Tectiviridae* (International Committee on Taxonomy of Viruses and King, 2012). *Caudovirales* contains three tailed phage families: *Myoviridae*, *Siphoviridae* and *Podoviridae* (Ackermann *et al.*, 1994). All three phage families share geometric morphology on their various- sized polyhedral heads. The tail structure is the major distinguishing characteristic for their classification (International Committee on Taxonomy of Viruses and King, 2012). *Myoviridae* have contractile tails and small base plates. *Siphoviridae* and *Podoviridae* feature long and short non-contractile tails, respectively. The genetics of *Caudovirales* phages, the so-called tailed phages, was discussed in this study.

1.2 Phage genome configuration and associated DNA-packaging mechanisms

Linear, double-stranded DNA is the genome feature that all tailed bacteriophages share (Brussow and Hendrix, 2002). The genomes of phages that have been reported span a size range between 18 and 500 kbp (International Committee on Taxonomy of Viruses and King, 2012). Although tailed phages share a similar machinery for DNA packaging, the diversity of protein sequences and underlying mechanisms has been found to generate different forms of DNA recognition and cleavage reactions, which results in different types of genome terminal sequences (Casjens and Gilcrease, 2009; Fujisawa and Morita, 1997). Replication strategies and packaging mechanisms determine the different forms of genomic termini among phages (Casjens, 2011). With the exception of phi29-like and Mu-like phages that have monomeric DNA as packaging substrates, most tailed bacteriophages synthesize concatemeric DNA that contains up to 10 or more copies of the genome before DNA packaging begins (Casjens and Gilcrease, 2009). The concatemer is a linear DNA that has multiple copies of unit-length genome that is synthesized with a circular template multiple rounds without termination during the phage DNA replication process (Skalka, 1977; Takahashi, 1975). Phi29 only replicates monomeric linear DNA (Meijer *et al.*, 2001). Mu phage integrates their genome randomly into the host DNA chromosome and incises its genome with flanking host DNA during genome replication (Bukhari and Taylor, 1975; Bukhari and Zipser, 1972).

The genome packaging mechanism, a packaging cleavage site, called *cos*, was first identified in lambda phage (Feiss *et al.*, 1983). The terminase gpNu1 recognizes the *cosN* site and generates cohesive termini of DNA to initiate the DNA packaging until it

identifies another *cos* site on the concatemer. The packaging machinery of phages T3 and T7 (direct terminal repeats) binds to a packaging recognition site and cuts at the cleavage site with a unit length of virion genome from concatemeric DNA (Chung and Hinkle, 1990; Hashimoto and Fujisawa, 1992). There is only one copy of the terminal repeat between the unit-length genome in head-to-tail concatemers. It is believed that the terminal redundancy is duplicated at the rear end of each packaged genome after the cleavage of the unit-length genome. (Dunn and Studier, 1983; Pajunen *et al.*, 2002). Previously, the terminal duplication models were proposed for unraveling the mechanism of DNA processing and maturation during DNA packaging in T7 (short direct terminal repeat) phages (Chung and Hinkle, 1990; Fujisawa and Morita, 1997; Zhang and Studier, 2004). The mechanism by which long terminal redundancies are synthesized however remains unknown. Terminase of circularly permuted phage P22 also recognizes the *pac* site to initiate DNA packaging (Wu *et al.*, 2002). Phage P22 does not cleave at the second *pac* site identified on the concatemer; rather, it packages between 102% and 110% of genome sequences from concatemers (Casjens and Hayden, 1988). The cleavage site on the concatemer is the start site of the next genome to be packaged into the next prohead. Thus, the genome sequences have up to 10% genome redundancy and yet have different genome termini among packaged DNA.

The protein components of the DNA packaging process were revealed in the 1970's (Kaiser *et al.*, 1975; Luftig *et al.*, 1971). Packaging of tailed-bacteriophage dsDNA involves a capsid-formed prohead and DNA translocating machinery. The prohead is a preformed shell consisting of capsid protein. The terminase contains two subunits: a terminase large subunit (TerL) and a terminase small subunit (TerS). TerL

possesses prohead binding activity (Sun *et al.*, 2008). The N-terminal domain of TerL catalyzes DNA translocating activity powered by ATPase (Duffy and Feiss, 2002; Kondabagil *et al.*, 2006). The C-terminal domain of the TerL carries out the DNA cleavage activity (Hwang *et al.*, 2000; Smits *et al.*, 2009). The TerS is responsible for packaging site recognition and DNA binding activity (Catalano *et al.*, 1995; de Beer *et al.*, 2002). To generate a mature DNA genome, the TerS binds to concatermeric DNA (Casjens, 2011; Rao and Feiss, 2008). The TerS then forms a complex with the TerL that loads genomic DNA into proheads through portal proteins. Finally, TerL recognizes the packaging cleavage site and cuts one copy off of the concatemer by endonuclease activity.

1.3 Identification of genomic termini and complete genome sequences

Several packaging strategies and corresponding configurations of packaged DNA were classified based on different genomic termini (Casjens and Gilcrease, 2009). A collection of diverse types of genome termini has been described in the literature (Born *et al.*, 2011; Casjens *et al.*, 2005). A straightforward experiment to detect the type of genome terminus is to perform restriction enzyme digestion under different conditions. Phages lambda and P2 have 5' cohesive ends (Catalano *et al.*, 1995; Murray and Murray, 1973). Mycobacteriophage L5 and D29 (Donnelly-Wu *et al.*, 1993; Ford *et al.*, 1998), *Bacillus subtilis* phage phi105 (Ellis and Dean, 1985), and *E. coli* phage HK97 have 3' extensions (Juhala *et al.*, 2000). The type of cohesive terminus and corresponding sequences can be identified by restriction enzyme digesting patterns in agarose electrophoresis gels with different cooling rate treatments. The two terminal fragments that contain cohesive ends can be detected by comparing the restriction patterns between slowly cooled samples and rapidly cooled ones. Heating the restriction enzyme digested

DNA to a temperature between 75-80 degrees Celsius can separate cohesive ends but not double-stranded DNA. The cohesive fragments are annealed together by gradually decreasing the temperature of the restriction digested sample, and forming a combined size band rather than two separated and smaller bands which appeared by rapidly cooling the sample down. The different patterns between the two cooling conditions suggest whether the genomic DNA have cohesive terminus or not.

Circularly permuted headful packaging systems are used by phages such as P22, SPP1 and T4 (Ratcliff *et al.*, 1979; Rhoades *et al.*, 1968; Streisinger *et al.*, 1964). If the cleavage sites of a headful packaging phage are greatly varied among virions, the length of some fragments will be so variable that they form a smear-like background stain on the gels. Since the permuted genome has a different length and varied termini of genome as compared to a genome with precise length and physical ends, the restriction pattern of a headful packaging phage will have one wider band at the position where the fragments approach the end of the genome rather than a single band width of an exact size on electrophoresis gels. This condition can be seen on a restriction pattern of phage Sf6 and ES18 (Casjens *et al.*, 2004; Casjens *et al.*, 2005).

Phages T3 and T7 have non-permuted and relatively short direct terminal repeats with exact length in every virion genome (Dunn and Studier, 1983; Pajunen *et al.*, 2002). T5 and SPO1 possess long terminal repeats with length of 10,139 bp and 13,185 bp, respectively (Stewart *et al.*, 2009; Wang *et al.*, 2005). N4-like phages feature dynamic length of terminal repeats (Ohmori *et al.*, 1988). In this type of genome configuration, different cooling conditions will not change patterns as cohesive end phages do.

Therefore, the identification of genome redundancy relies on a reference sequence of nucleotide or direct sequencing analysis 'walking' toward both ends of the genome.

While most of the tailed phages generate concatemeric DNA as substrates for genomic DNA packaging, some phages only generate monomeric DNA. Phi29-like phages have covalently bound terminal proteins (Ito, 1978; Salas *et al.*, 1978). Headful packaging is evident in Mu-like phages with host DNA sequences flanking the integrating position of host chromosomes (Bukhari and Taylor, 1975; George and Bukhari, 1981; Groenen and van de Putte, 1985). Bukhari and Zipser (1972) demonstrated that the Mu phage can randomly integrate its genome into the *E. coli* host chromosome. The transposable genome contains adjacent host DNA at the integrated site and presents non-specific host DNA in its progeny (Ljungquist and Bukhari, 1977).

The restriction digests pattern could help one determine the approximate terminal region, in cases such as cohesive-end phages and circularly permuted phages. However, the determination of exact terminal sequence depends on direct sequencing reactions with the aid of known restriction mapping on a reference sequence beforehand. It would be more difficult to identify the terminal region by restriction digests if the test phage has blunt terminal ends and genome redundancy when the sequence is not available. Furthermore, a phage that was isolated independently from samples is very likely to have a novel genome that has never been sequenced. Therefore, a contig sequence assembled by shotgun sequencing or high throughput sequencing becomes the very first step of the identification of phage genome configuration.

1.4 High throughput sequencing of phage genomes

Next generation sequencing (NGS) of DNA has greatly enhanced various genetic research, including studies of bacteriophage genomics. The high throughput volume of NGS allows one to acquire high sequencing coverage of phage genomes and to reveal genome characteristics such as terminal redundancy and cleavage sites at genome ends. Three major characteristics of phage genome NGS are discussed in the following.

1.4.1 Circularity of phage sequence assembly

While NGS was widely used for phage genome projects, the circularity of a sequence assembly was reported when a phage contained a duplication of direct repeats (Casjens and Gilcrease, 2009). Genome redundancy of *Bacillus subtilis* phage SPO1 was characterized in 1978 (Cregg and Stewart, 1978). The circularity of the SPO1 assembly was found after a library of sheared SPO1 DNA was incorporated into plasmids of *E. coli*, sequenced with direct sequencing individually and assembled afterward (Stewart *et al.*, 2009). Circular contigs of three *Bacillus cereus* phages were revealed after genome assembly of Roche 454 pyrosequencing reads (Grose *et al.*, 2014). However, this study assigned the first base pair t at the non-coding region upstream of terminase large subunit gene on the genome. The lengths of genome redundancy were then determined by the PAUSE raw sequencing data processing method (http://cpt.tamu.edu/pause/) (Grose *et al.*, 2014). For the genome sequencing of 63 novel Mycobacteriophages, 11 assemblies had circular contigs which are seemingly circularly permuted termini (Hatfull *et al.*, 2013).

1.4.2 Frequency of read edge position

NGS read frequencies from the 5’ end revealed that T4-like phage IME08, which has a circular permutation, has a sequence preference at the genome terminus rather than a random cleavage site during headful packaging (Jiang *et al.*, 2011). This implied that high frequency reads represented the sequence-specific terminus for circularly permuted phage IME08. Li *et al.* scrutinized raw sequencing reads from NGS data and further supported this finding (Li *et al.*, 2014). The terminus-tagged reads directed the exact sequence of first and last base pair of T3 genome, whereas untagged reads with the highest frequency appeared at the same position as tagged reads indicated (Li *et al.*, 2014). A number of newly isolated phages also retained the characteristic of high frequency reads at the terminal position of phage genome (Li *et al.*, 2014). However, this study did not validate the terminal sequence by experimental analysis such as direct sequencing.

1.4.3 Coverage build-up on a sequence redundant region of an assembly

A sequence repeat region in a genome results in higher coverage in NGS data, which would be linearly correlated to the number of repeat copies in the genome. The short raw reads from the repeat regions then overlap on the read mapping file. The overlapping assembly results in a non-redundant contig and a high coverage within repeat regions. This characteristic could be applied to phage genome sequencing to define whether a phage has genome redundancy. Yee *et al.* (2011) identified coverage build-up in the middle of a coverage map in the whole genome sequence of SPO1-similar phage

SP10. Yee *et al.* defined the region of higher read depth as genome terminal redundancy of approximately 12 kb. However, Yee *et al.* did not describe the characteristics of the terminal sequence. In 2012, Gill *et al.* characterized five *Caulobacter crescentus* phages that are closely related to phiCbK. These phages showed 10 to 17kb terminal redundancy based on a striking build-up of coverage over the assembled contig (Gill *et al.*, 2012). The genomic terminal redundancies were further confirmed by tagging the genome terminus with short nucleotide fragments as markers before high throughput sequencing (Gill *et al.*, 2012). Li *et al.* (2014) also tagged genome termini of phage T3 with ligated adaptors for high throughput sequencing to locate the terminus. Compared to the untagged T3 genome, Li *et al.* (2014) demonstrated that the high coverage region in the contig represented the genomic redundancy of T3 phage (Li *et al.*, 2014).

1.5 *Bacillus cereus*-group bacteria and susceptible bacteriophages

The *Bacillus cereus* group consists of *B. cereus, B. thuringiensis, B.weihenstephanensis, B. mycoides,* and *B. anthracis* (Tourasse *et al.*, 2006). These Gram-positive, spore-forming bacteria were frequently isolated from soils (Lee *et al.*, 2011). These firmicute bacteria tend to live in spore form in soil, and tend to infect target animal and plant hosts when in the vegetative state (Vilain *et al.*, 2006). *Bacillus anthracis* is the etiological agent of anthrax, which was the first pathogenic organism that Robert Koch used to demonstrate the linkage of a bacterium to a specific disease (Blevins and Bronze, 2010). The virulent elements are known to anchor on plasmids pXO1 and pXO2 in *B. anthracis*. pXO1 encodes toxin subunits *pag, lef, cya,* and pXO2 encodes encapsulated protein *capA, capB, capC* and *dep* (Fouet and Mock, 1996). *B. thuringiensis*

is pathogenic to insects by forming insecticidal crystal proteins during sporulation, known as Cry protein encoded by *cry* gene that is located on plasmids (Ibrahim *et al.*, 2010). *B. cereus* is now known as an opportunistic bacterium which causes serious infections either intestinal or non-intestinal in humans (Tourasse *et al.*, 2006). The pathogenicity of *B. cereus* is reported to secret four toxins associated with tissue-destructive/reactive activities (Granum, 1994).

The well-known phage strain pathogenic to *Bacillus cereus* is phage γ, which was first reported in 1955 as a variant of phage ω (Brown and Cherry, 1955). The ligand was identified as sortase A that targeted GamR on *Bacillus cereus* (Davison *et al.*, 2005). Bacteriophage CP-51 was also reported to attach the spore form of *Bacillus cereus* and express its genes after induction of endospore germination (Cohen *et al.*, 1973). It is worthwhile mentioning that most of the phages isolated infecting *B. anthracis* were also able to infect *B. cereus*, due to close genetic and morphological relatedness between members of *B. cereus* groups (Gillis and Mahillon, 2014). Furthermore, spore-forming *Bacillus* species in *B. cereus* groups are dependent on sporulation in adverse environmental conditions. (Koehler, 2009). Some phages are also known to infect the spore form of *B. cereus* (El-Arabi *et al.*, 2013; Lee *et al.*, 2011). However, few phages of *B. cereus* have been isolated and characterized in detail. Conserved structures are also unclear since sequence variation among phages is usually large. Moreover, it has been suggested that the current identified strain count is estimated to be just 0.01-0.1% of all strains in the biosphere (Angly *et al.*, 2005).

1.6 Primary research questions

The three-dimensional structure of SBP8a was characterized by cryo-electron tomography (cryo-ET) (Fu *et al.*, 2011). The genome sequence of SBP8a was sequenced by Roche 454 pyrosequencing and yet has not been published and characterized. In this study, I characterized the genomic DNA of 48 naturally occurring phages from the same soil source that infected *Bacillus anthracis/cereus* in Fu *et al.*'s study, as well as re-interrogating the genomic DNA of SBP8a by high throughput sequencing. The raw reads and assembled contigs of 48 phage sequencing data were used to explore tailed phage structure genetics, including phage genome configuration and phage genome characterization.

1.6.1 Characterization of phage genome configuration of 48 Bacillus phage genome sequences

Three characteristics of phage NGS data or shotgun sequencing data have been reported in the literature as described below. First, the circularity of phage genome assemblies was reported in a sequencing library of sheared SPO1 DNA (Stewart *et al.*, 2009), and three *Bacillus cereus* phages were revealed after genome assembly of Roche 454 pyrosequencing reads (Grose *et al.*, 2014; Stewart *et al.*, 2009). Second, the nucleotide position with the highest read edge frequency was potentially the physical end of phage genome (Jiang *et al.*, 2011; Li *et al.*, 2014). Third, a coverage build-up region on the sequence redundant region was demonstrated by the research work of Gill *et al.* for SP10, of Li *et al.* for phiCbK and of Yee *et al.* for T3 phage NGS data (Gill *et al.*, 2012; Li *et al.*, 2014; Yee *et al.*, 2011). However, none of the literature reports

development of a method that could describe these characteristics in NGS data. The online tool PAUSE has been used for terminus prediction, and yet it did not describe the variables behind the tool in the literature (http://cpt.tamu.edu/pause/). Furthermore, none of these studies investigated more than two types of tailed phage genome termini.

The specific aim of this study was to determine whether these characteristics of NGS data are significant for identifying the physical ends of 48 *Bacillus anthracis* phage genomes. In order to deal with large numbers of raw read data, a computer-based method was developed to identify the characteristics of phage NGS data listed above. Dideoxynucleotide sequencing reactions were used to validate the genome configuration of isolated phages. This study also extends the investigation to every type of phage genome terminus that has been identified so far. Moreover, the relationships among the three characteristics of phage NGS data were scrutinized to describe the causality of contig circularity, read edge frequency and coverage build-up of tailed phage genomes.

1.6.2 Characterization of the genomic structures and gene annotations of complete genome sequence of 48 Bacillus anthracis phages

This study focused on the genomic structures of novel *Bacillus* phages. The putative open reading frames and gene functions were predicted by genomic characterization tools with homology searches in the databases. The morphology and classification were implied based on the putative protein functions in the genome. The genome comparison among closely related phages was conducted to explore the relationships and potential genes involved in novel mechanisms of phage pathogenesis, interaction with host cells, and evolutionary history. This study gives a basic knowledge

of genomic structure and molecular taxonomic classification of novel *Bacillus anthracis* phages.

CHAPTER 2

Predicting Genome Terminus Sequences of *Bacillus cereus*-group Bacteriophage using Next Generation Sequencing data*

ABSTRACT

Linear dsDNA is the genome feature that most tailed bacteriophages share. The genomic DNA is packaged by DNA-packaging motors that are similar among diverse phages. However, DNA packaging strategies and resulting terminus of packaged DNA have evolved differences. Characterizing novel *Bacillus cereus*-group phages requires understanding the complete genome sequence, which is crucial for subsequent studies. Identifying physical ends of the phage genome by restriction enzyme digestion could not specify the terminal sequence without reference genome. In this study, we sequenced 48 isolates of phages infecting *Bacillus cereus* and analyzed Next Generation Sequencing (NGS) data to predict the terminus sequence of novel phages. Most of assembled contigs featured reads that mapped to both ends of contig, suggesting that phage genomic assemblies by NGS data form circular contigs. However, known assemblers were not capable of reporting the underlying phage genome configuration for genome assemblies. To identify the physical map of sequenced phage *in silico,* a terminus prediction method was developed by means of 'neighboring coverage ratios' and the read edge frequency from read alignment files. Termini were confirmed by primer walking and supported by phylogenetic inference of large terminase protein sequences. Three novel *Bacillus* phages (SBP8a, I48 and Q8) featured direct terminal repeats. The genomic terminus prediction method was validated on nine published phages with known packaging mechanisms.

Complete phage genome sequences allowed a proposed characterization of the potential packaging mechanism and precise characteristics of genome annotation.

* This chapter is based on a manuscript for publication. The authorship is listed as follow: Cheng-Han Chung, Michael H. Walter, Luobin Yang, Shu-Chuan (Grace) Chen, Vern Winston and Michael A. Thomas.

2.1 INTRODUCTION

Tailed, double-stranded DNA bacteriophages (phages) are the most abundant type of phage (Brussow and Hendrix, 2002; Wommack and Colwell, 2000). Although tailed phages share a similar mechanism for DNA packaging, the diversity of protein sequences and underlying mechanisms has been characterized as driving different forms of DNA recognition and cleavage reactions, resulting in different types of genome terminus sequences (Casjens and Gilcrease, 2009; Fujisawa and Morita, 1997).

The protein components of the DNA packaging process were revealed in 1970's (Kaiser *et al.*, 1975; Luftig *et al.*, 1971). A capsid-formed prohead and a DNA translocating machinery were involved in dsDNA packaging of tailed-bacteriophage. The prohead is a preformed shell consisting of capsid protein. Terminase contains two subunits: large terminase subunit (TerL) carries out prohead binding (Sun *et al.*, 2008), DNA translocating activity at N-terminal domain powered by ATPase (Duffy and Feiss, 2002; Kondabagil *et al.*, 2006), and DNA cleavage activity at C-terminal domain (Hwang *et al.*, 2000; Smits *et al.*, 2009); and small terminase subunit (TerS) is responsible for DNA recognition and DNA binding activity (Catalano *et al.*, 1995; de Beer *et al.*, 2002). The substrate of DNA packaging motor in tailed phages is usually a linear DNA that has multiple copies of unit-length genome, known as concatemer, that was synthesized with a circular template multiple rounds without termination during phage DNA replication process. To generate mature DNA genome, the TerS binds to concatermeric DNA and forms a complex with TerL that functions to load genomic DNA into prohead through portal proteins and cleave one copy of genome sequence by endonuclease activity on TerL (Casjens, 2011; Rao and Feiss, 2008).

Several packaging strategies and corresponding nature of packaged DNA are classified based on different genomic termini. Phages lambda and P2 have 5' cohesive ends (Catalano *et al.*, 1995; Murray and Murray, 1973). Mycobacteriophage L5 and D29 (Donnelly-Wu *et al.*, 1993; Ford *et al.*, 1998), *Bacillus subtilis* phage phi105 (Ellis and Dean, 1985) and *E. coli* phage HK97 have 3' extensions (Juhala *et al.*, 2000). Circularly permuted headful packaging systems are used by phages such as P22, SPP1 and T4 (Ratcliff *et al.*, 1979; Rhoades *et al.*, 1968; Streisinger *et al.*, 1964). Phages T3 and T7 have non-permuted and relatively short direct terminal repeats with exact length in every virion genome (Dunn and Studier, 1983; Pajunen *et al.*, 2002), comparing to phages with long terminal repeats that are 10139 bp and 13185 bp on T5 and SPO1, respectively (Stewart *et al.*, 2009; Wang *et al.*, 2005). N4-like phages feature dynamic length of terminal repeats (Ohmori *et al.*, 1988). Phi29-like phages have covalent bound terminal proteins (Ito, 1978; Salas *et al.*, 1978). Headful packaging is evident in Mu-like phages with host DNA sequence flanking the integrating position of host chromosome (Bukhari and Taylor, 1975; George and Bukhari, 1981; Groenen and van de Putte, 1985). A collection of diverse types of genome termini has been described elsewhere (Born *et al.*, 2011; Casjens *et al.*, 2005).

Replication strategies and packaging mechanisms determine the different forms of genomic termini among phages. With the exception of P2-like, phi29-like and Mu-like phages that have monomeric DNA as packaging substrates, most tailed bacteriophages synthesize concatemeric DNA that contains up to 10 or more copies of genome before DNA packaging begins (Casjens, 2011). Phi29 only replicates monomeric linear DNA, whereas Mu-like phages duplicate and integrate their genome into the host DNA

chromosome. As for the packaging mechanism, a packaging cleavage site called *cos* was first identified in lambda phage (Feiss *et al.*, 1983). The terminase gpNu1 recognizes the *cosN* site and generates cohesive termini of DNA to initiate the DNA packaging until it identifies another *cos* site on the concatemer. The packaging machinery of phages T3 and T7 (direct terminal repeats) binds to a packaging recognition site and cut at the cleavage site with a unit length of virion genome from concatemeric DNA (Chung and Hinkle, 1990; Hashimoto and Fujisawa, 1992). It is believed that the terminal redundancy is duplicated at rear end of packaged genome after DNA replication since there is only one copy of the terminal repeat between the unit-length genome in head-to-tail concatemer (Dunn and Studier, 1983; Pajunen *et al.*, 2002). The terminal duplication models were proposed for unraveling the mechanism of DNA processing and maturation during DNA packaging in T7 (short direct terminal repeat phage) (Chung and Hinkle, 1990; Fujisawa and Morita, 1997; Zhang and Studier, 2004), though the mechanism of which long terminal redundancies are synthesized remains unknown. Terminase of circularly permuted phages P22 also recognizes the *pac* site to initiate DNA packaging (Wu *et al.*, 2002). It does not cleave at the second *pac* site identified on the concatemer; rather, it packages generally 102% to 110% of genome sequence until the prohead is full (Casjens and Hayden, 1988). The cleavage site is the start site of the next genome that to be packaged into the next prohead. Thus, the genome sequences have up to 10% genome redundancy and yet have different genome ends among virions.

Next generation sequencing (NGS) has greatly enhanced various genetic research including studies of bacteriophage. The high throughput volume of NGS allows one to acquire high sequencing coverage of phage genomes and to reveal genome characteristics

such as terminal redundancy and cleavage sites at genome ends. Yee and colleagues identified the build-up of coverage in the middle of coverage map in the whole genome sequence of SPO1-similar phage SP10. They defined the region of higher read depth as genome terminal redundancy of approximately 12 kb (Yee *et al.*, 2011). However, they did not describe the characteristics of the terminal sequence.

NGS read frequencies (from the five prime end) revealed that T4-like phage IME08, which has circular permutation, has a sequence preference at the genome terminus rather than a random cleavage site during headful packaging (Jiang *et al.*, 2011). In 2012, Gill and colleagues characterized five *Caulobacter crescentus* phages that are closely related to phiCbK. These phages showed 10 to 17 kb terminal redundancy based on a striking build-up of coverage over the assembled contig. The genomic terminal redundancies were further confirmed by tagging the genome terminus before high throughput sequencing with short nucleotide fragment as markers (Gill *et al.*, 2012). Li and others also tagged genome termini of T3 phage with ligated adaptors for high throughput sequencing to locate the terminus. Compared to the untagged T3 genome, they demonstrated that the high coverage region in the contig represented the genomic redundancy of T3 phage. A set of lytic phages isolated from sewage were NGS-sequenced and characterized to determine termini based on the nucleotide that had the high read frequency (Li *et al.*, 2014). However, the high frequency reads may vary in genome ends among phages and NGS sequencing methods.

In the present study, we calculated the read edge frequency and neighboring coverage ratios (NCR) to determine the potential region of termini. Raw sequencing data from nine published phages were used to calibrate the accuracy of the terminus-

determining method in this study. Phylogenetic analysis of terminase large subunits and 'primer walking' using Sanger sequencing were conducted to validate the putative termini from NGS data for the isolates we sequenced in this study.

A study of *Bacillus anthracis* phage assemblage characterized three novel phages from urban Iowa topsoil (Walter and Baker, 2003). One spore-binding phage named SBP8a has also been scrutinized in detail (Fu *et al.*, 2011). Here, we characterized genomic DNA of 48 naturally occurring phages which can infect *Bacillus anthracis/cereus* from the same soil source in Fu *et al.*'s study as well as interrogating genomic DNA of SBP8a by high throughput sequencing.

Partial phage genomes are often published. Incomplete phage genomes lead to insufficient information for physical map reconstruction, phage classification, gene annotation, characterization, and subsequent functional and host interaction studies. Therefore, documenting a complete genome, with terminus identification, is crucial to complete phage characterization. However, there is no common algorithm in genome contig assembler software to characterize physical termini of phage genomes to date. In this study, we demonstrated that the contig circularity is an important feature to acquire complete genome sequences from most of the tailed phage NGS data. The read edge frequency and 'neighboring coverage ratio' are proposed as two important criteria to predict the genome termini for reconstructing complete physical maps for novel phages.

2.2 MATERIALS AND METHODS

2.2.1 Culture media and bacteria strains

Host *Bacillus anthracis* Sterne and *Bacillus cereus* 569 UM20 were provided by Dr. J. Jackman's lab (Johns Hopkins University Applied Physics Laboratory, Laurel, MD) and Dr. Terri Koehler's lab (Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, Houston, TX), respectively. *B. anthracis* Sterne is a vaccine strain without virulent plasmids pXO1 and pXO2. Tryptic Soy Broth (TSB, Difco Bacto BBL 211824) was used for growth and solid media plates (1.5% Bacto-Agar) in this study.

2.2.2 Isolation, propagation/increase and DNA extraction of phages

Natural *B. anthracis* phages were isolated and purified as previously described (Walter and Baker, 2003). Culture lysis, lysate clarification, triple-serially transfer and isolation were conducted by standard phage methods (Thorne, 1968). *B. anthracis* phages were isolated and purified with the same procedure as SBP8a isolate (Fu *et al.*, 2011). Phage DNA was extracted followed by the method described in molecular biology laboratory manual (Sambrook and Russell, 2001). A 100ng aliquot of genomic DNA from each isolate was used to perform high throughput sequencing.

2.2.3 Genomic DNA sequencing

Phage genomic DNA was sequenced by *Ion PGM* 316 chip v2 (*Life Technologies*, CA, USA) and by *MiSeq* Reagent Kit 2x300 v3 (*Illumina*, San Diego, CA, USA) at the

Molecular Research Core Facility (Idaho State University, ID, USA). For the *Ion PGM* sequencer, a DNA library was prepared by Ion Xpress™ Plus Fragment Library Kit with Barcode Adapters 1-96 Kit (*Life Technologies*, CA, USA). For the *MiSeq* sequencer, a library was prepared by Nextera XT DNA Sample Preparation Kit and barcoded with the *MiSeq* Index Kit (*Illumina*, San Diego, CA, USA). A sequencing data of SBP8a phage through Roche/454 shared by Dr. Ian Molineux was also included as a replicate of SBP8a genome.

Nine published genome sequences of characterized phages were obtained from the lab of Dr. Graham Hatfull's lab (University of Pittsburg, Dept. Biological Sciences), including three *Bacillus* phages Adelynn (*Illumina*), Nigalana (*Roche/454*) and Troll (KF208639.2) (*Ion Torrent*); three Cluster C mycobacteriophages Zeenon (*Illumina*), Teardrop (*Roche/454*) and Breeniome (KF006817) (*Ion Torrent*) (Hatfull *et al.*, 2013); three Cluster A mycobacteriophages Equemioh13 (KJ959632) (*Illumina*), Zetzy (*Roche/454*) and Lilith (*Ion Torrent*).

2.2.4 Bioinformatics Analysis

Twenty-three and 26 phage isolates were DNA sequenced using *Ion PGM* and *MiSeq* sequencers, respectively. Genomic contigs were assembled with Newbler 2.9 (*454 Life Sciences*) and Velvet 1.2.10 (Zerbino and Birney, 2008) for sequence reads from *Ion PGM* and *MiSeq*, respectively. Among the contig sequences of I48-like isolates, over-call (insertion) and under-call (deletion) homopolymers from *Ion PGM* and/or *MiSeq* were corrected based on the majority call on the aligned position when homopolymers of more than two bases from *MiSeq* sequences varied in *Ion PGM* sequences.

‘Neighboring coverage ratio’ (NCR) and read edge frequency were used for investigating phage genomic termini, which are described in detail below. Predicting termini from mapped reads must pass twin criteria: high read edge frequency and increased coverage over neighboring coverages. NCRs were calculated by the average coverage within 100 nucleotide right window divided by coverage within an adjacent 100 nucleotide left window. A NCR fold-change of 1.8 and the enriched coverage region of 1.8 fold over average coverage were used as thresholds for high coverage change. The regional crests of NCR that exceeded two thresholds were recorded as a potential genome end regions. Another criterion that has to be satisfied to be a potential terminus is described next. The nucleotide positions of a read that mapped onto corresponding contig could be acquired from read mapping files. In the present study, the 5' edge of the read was defined as the first basepair that mapped leftward to the reference contig, and the 3' edge of reads was the last basepair rightward against contig position. The frequencies of read edge positions were calculated after all alignable reads were mapped onto the phage contig. The top three frequencies of 5’ or 3’ end position were listed as the possible first or the last nucleotide of genome, respectively. Theoretically, the position with highest read edge frequency is the physical end of genome if the viral genome is linear. More discussion and results in detail were described in *Section 3.2*. A nucleotide position that passes the thresholds in NCR screening and is high frequency position is necessary to be a ‘significant hit’ for terminal position. The NCRs, read edge frequency, judgment of circular/linear contig and flanking sequence were analyzed with *perl* script with integration of *samtools* package (Li *et al.*, 2009). Open reading frames (ORFs) were predicted with both Glimmer 3.02 and GeneMark. The search for homologous proteins

and protein function of each predicted ORF was conducted by BLASTP (comparing against non-redundant database in NCBI) and HHblits ('HMM-HMM-based lightning-fast iterative sequence search': HHblits; http://toolkit.genzentrum.lmu.de/hhblits/, comparing against uniprot20 database in HHsuite 2.0). Amino acid sequences of the large terminase subunits from selected phages were used for phylogenetic analysis with MEGA 6.0.

2.3 RESULTS AND DISCUSSION

2.3.1 Genome characterization by next generation sequencing

Q11 isolate was re-sequenced by *MiSeq* due to potential contamination and low coverage from *Ion PGM* sequencing. Spore-binding phage 8a (SBP8a) was sequenced by *MiSeq*, previously characterized as *Myoviridae* and characterized to have *B. anthracis* spore-binding activity (Fu *et al.*, 2011). The average lengths of reads generated were 292.29bp by the *Ion PGM* and 228.12bp by the *MiSeq* sequencer (Supplementary Table 2.1 and 2.2). Total yields in base pairs were 1.06 gigabase (Gb) by *Ion PGM* and 12.59 Gb by *MiSeq* in a single flow cell . Expected per-run yields based on manufacturers' specifications were 0.3-1.0 Gb for *Ion Torrent* and 13.2-15 Gb for *MiSeq* Reagent Kit v3.

Genome sequencing reads of 39 isolates were successfully generated single-contig assemblies. The sequence lengths of these 39 isolates fell into two major size categories: 35 isolates had approximately 158 kb genomes and four isolates had up to 26 kb genomes. Assembled sequence of isolates Q2, Q8 and Q10 had fewer than three nucleotide differences in pairwise comparisons (Supplementary Table 2.3). The contig

lengths of I3, I17, I46 and Q11 varied from 21,717 to 26,005 bp. These all contained identical sequence within aligned regions except I17 after multiple sequence alignment by ClustalOmega (Sievers *et al.*, 2011) (Supplementary Table 2.4). For the remaining 31 'I48-like' isolates (contig lengths of about 158 kb), pairwise alignments showed that the number of nucleotide differences between any two were less than eight after correction of indels (insertion/deletion) in homopolymeric regions between *Ion PGM* and *MiSeq* sequencer as described in the methods section (Supplementary Table 2.5).

Six of 23 *Ion PGM* –sequenced genomes (Isolate I2B, I6, I15, I37, Q11 and Q1) failed to generate single contigs by Newbler assembler (*454 LifeSciences*). These six genomes had more than 280 contigs after assembly and total contig lengths were too large for potential phage genomes. The second largest assembled contigs from each of the six isolates were subjected to BLAST searches against non-redundant nucleotide database in NCBI. The significant hits included 23S rRNA, plasmid and intergenic region of various bacteria strains, suggesting that those samples were contaminated by host or environmental bacteria DNA. Three assemblies (I33, I35 and Q5) had only two major contigs that have over 1,000 bp in length. They featured gaps of 12 to 167 bp between the two contigs after mapping I33 and I35 contigs to I48 genomes, and Q5 against Q8. I50 yielded two major contigs which appeared to represent partial genomes of large-genome phages and small-genome phages. To avoid the ambiguity and incorrect assemblies, these ten isolates that were sequenced by *PGM* (I2B, I6, I15, I37, Q11, Q1, I33, I35, Q5 and I50) were excluded from subsequent analyses. Among the remaining 39 single-contig genome assemblies, the average coverage of the 25 *MiSeq* contigs was 1927.21, which was 7.02 times higher than that of the 14 *Ion PGM* contigs (274.57).

Five out of six incomplete genome assemblies were sequenced by *Ion PGM*. This suggests that the high coverage of assemblies from the *MiSeq* system could increase the chance to resolve complete viral genome.

The isolates I48 (157,912 bp; large-genome strain), Q8 (158,180 bp; QCM-crystal:spore-binding strain (unpublished data)), Q11 (26,005bp; small-genome, QCM-crystal:spore-binding strain) and SBP8a (158,819 bp) were chosen as representative strains for genome similarity comparison (Table 2.1). Isolate Q8 and SBP8a were both selected by adherence to spores (SBP8a) and by simultaneous adherence to quartz-crystal micro balance electrodes and spores (QCM-unpublished data) but were propagated by the same methods. However, the overall identity between them was only 13%. Intriguingly, I48 and SBP8a shared approximately 90% identity over genomic sequence. Q11 possessed a one-sixth genome size in comparison to the other three strains. The sequence identity that Q11 shared to other selected isolates is less than 1.327% over 26,005 bp.

2.3.2 Prediction of genomic termini by NGS data

The random start sites and orientation of contig sequences among 31 I48-like isolates was revealed after genome alignment (Supplementary Figure 2.1). The read mapper *bowtie2* (Langmead and Salzberg, 2012) and sequence alignment file generated by *samtools* (Li *et al.*, 2009) also showed that some reads could map to both ends of contig sequence. It indicated that the assemblers tended to produce random cleavage on circularized contigs. There was no algorithm in *Newbler* or *Velvet* to judge the genomic terminus if the assembled contig has a circular map. Furthermore, assemblers did not detect circularity, but cleaved contigs arbitrarily in order to report a contig as linear. This

gave rise to a ‘random’ (and probably inaccurate) genomic terminus following assembly. A manual technique was used to define the circularity of contigs in the following analyses (data not shown). Since 31 I48-like isolates showed nearly identical genome sequences in different orders and orientations, we had a sufficient sample size to address the potential termini of I48-like phage strains.

An evident coverage crest was observed on every coverage map of I48-like isolates (Supplementary Figure 2.2 and 2.3). It is known that the category of phage genome with long direct repeat features terminal redundancy. Sequencing coverage will be higher than non-terminal repeat regions since there are two copies of the repeat sequence in one complete packaged genome, resulting in twice the probability to acquire the reads located within repeat regions from the sequencer. Based on the coverage distribution of I13, an I48-like phage (Figure 2.1b), we hypothesize that the increasing or decreasing edge of high coverage regions is the overlapped mapping reads within the repeat region of the viral genome (Figure 2.1a). Assuming that the cumulative local coverage and a high frequency of read edge position is the consequence of randomly fragmented genome, the positions of genomic termini can be predicted by coverage distribution and read mapping.

A 1.8 fold-change of coverage between left and right ‘window’ was used as a threshold to define higher coverage regions. Given this window size, neighboring coverage ratios (NCR) were calculated using the average coverage within a right 100-nucleotide window divided by coverage within an adjacent 100-nucleotide left window. When scanning over an assembled phage contig, the ratio was expected to increase abruptly when the left edge of a 3' window panel was oriented on a potential terminus

region and expected to drop sharply as the right edge of 5' window approached the end of another potential terminus (Figure 2.1c). The NCR map of I13, which belongs to I48-like phage strain, suggested a 2,750 bp high coverage region located within the internal region of I13 contig (Table 2.2). Instead of assuming a circular genome (rare for Myoviridae), we suggested a linear genome with direct terminal repeats based on the NCR results. The NCR analysis for every single-contig phage was summarized in *Supplementary Table 2.6- 2.10.*

We next analyzed read edge frequency. Library preparations for *Ion Torrent* and *MiSeq* sequencing have genome fragmentation/shotgun steps. Considering every nucleotide position that is fragmented through genome shotgun process during sequencing library preparation is equally likely, the fragments that contain the first base pair or the last base pair of a linear genome should have highest frequency in genome fragmentation pool. Assuming a non-biased amplification and sampling for sequencing, the expected high frequency of read edge position that underlies genome terminus should be presented in sequence data. As a result, fragments of DNA that contain genome termini will be sequenced once for every copy of phage genome DNA that was input to NGS. The expected outcome of this feature is a higher occurrence of read edge positions at the potential terminus based on the read mapping file. The mapping result indicated that there are 110 reads that had their 5' end aligned at nucleotide position 111,610 bp on I13 contig, while 3' end position occurred 274 times at position 114359 bp (Figure 2.1d). Furthermore, the positions with the highest read edge frequency in I13 isolate (sequenced by *PGM*) matched the predicted terminal positions by NCR (Table 2.2), in which case its position was located right on the window edge (3' end position) or within the window

panel (5’ end position). However, the phenomenon might not always be true because of considerable variations in every step over sequencing. It is known that there is a PCR clean-up step during high throughput sequencing so that too short fragments were eliminated for sequencing. Indeed, the phage sequencing data indicated that the predicted terminal sequence by NCR was not the position with highest read edge frequency in the NGS data. For example, left terminus of I12 was determined based on the second highest frequency of read edge position where the position matched the predicted terminus from NCR method. Nevertheless, most of terminus predicting result for I48-like phages featured 5’ genomic terminal sequence: 5'- AGGTTTTTCT while the 3’ terminus is CATACGGTTT-3'.

The I48-like isolates appeared to have a linear genome about 158 kb in length with an additional 2,750 bp direct terminal repeat (DTR). We were not able to locate potential termini at the relative positions based on the NCR and read edge frequencies in *MiSeq* sequencing though 20 *Miseq*-sequenced I48-like isolates possessed almost identical contig sequences to 11 *PGM*-sequenced I48-like phages. This might be the consequence of biased fragmentations during the library preparation. According to previous studies regarding NGS DNA fragmentation methods, the sonicated DNA fragments (Grokhovsky *et al.*, 2011) as well as the neubulization and Covais method (Poptsova *et al.*, 2014) show sequence preferences and therefore non-randomly distributed cleavage sites. The negative result from *MiSeq* may also be due to the saturation of coverage, in which case the signal of terminus prediction in both methods might be perturbed and disagree with each other among the sequenced samples in one batch.

The prediction of SBP8a termini showed one hit of 'significantly increasing' NCR at position 111,794 bp and decreasing ratio at 114,616 bp, which resulted in a potential terminal repeat genome with two specific terminal cleavages from SBP8a *Roche/454* sequencing data. Simultaneously, the highest frequency of either 5' or 3' read edge occurred at the same position as NCR's prediction (Table 2.2). However, NCR method and read edge position were failed to conclude a possible position for SBP8a by using *MiSeq* sequencing data (Supplementary Table 2.9, data sheet 'SBP8a'), although a relative high coverage region was observed (Supplementary Figure 2.4). Note that the sequencing data of SBP8a by MiSeq was conducted with the same batch of sequencing of I48-like isolates that were unable to find the expected characteristic of terminal position, in which case a possible justification was described before.

For NGS data analyses of Q8-like isolates, one consistent terminus was found with flanking sequence 5'-AGGTTTTTGTG among Q2, Q8 and Q10 (Supplementary Table 2.7, data sheet 'Q8-like'), close to the apparent boost of coverage seen in coverage distribution (Supplementary Figure 2.5). There was no sudden coverage change with fold-change larger than 1.8 at 3' downstream of the consistent terminus in NCR analysis. All Q11-like isolates had no hit in NCR analysis (Supplementary Table 2.8, data sheet 'Q11-like'). The coverage distributions showed fluctuating oscillation but not one abrupt build-up across the contig (Supplementary Figure 2.6). I3, I17, I46 and Q11 have nearly identical sequence over the whole contig and Q11 has largest contig from assembly. Furthermore, there is no read that was aligned across both ends of corresponding contig in isolates of Q11-like group, in which case these contigs appeared to be linear. This indicates that the contig sequences of Q11-like isolates were incomplete from assemblies.

2.3.3 Terminal sequence validation from primer walking method

Taq polymerase generates an artificial 'A' nucleotide call after the last base of template, which is expected to be seen from chromatograms in Sanger sequencing. The 'primer walking' takes advantage of this feature for defining the last base pair of a genome where termini were predicted *in silico* such as NCR and read edge frequencies. When the primer was designed within genomic redundant region and elongated toward terminus using a template genome with direct terminal repeat, the output of primer walking are expected to have a mixture signal from two templates. While one copy of direct repeat would terminate the sequencing reactions at the physical end of template DNA and add the untemplated 'A' peak after the last base pair, another copy of direct repeat continues the elongation into non-redundant region. As a result, the chromatography that reaches the physical ends would engage signal intensity decreased by about half afterward, along with an artificial 'A' call generated by *Taq* polymerase. On the other hand, conducting primer walking with a non-redundant genome template would expect to see a complete termination of sequencing after the untemplated 'A' call is observed. However, the primer walking could not identify either 5' or 3' sequence without a terminal ligation step before direct sequencing.

Our primers were designed around 150-250 bp upstream of predicted termini. Primer walking was used to confirm the genome termini for isolates and to validate the termini predictions from NGS data (Table 2.3). By comparing predictions from NGS data to primer walking methods, the genome terminus location for SBP8a phage matched most closely, having only a two base-pair difference on 5' terminus and an exact match

on 3’ terminus. The 5' termini of I13 and I22 were 73 bp upstream of the ‘coverage-predicted’ terminal sequences, while 3’ termini had less than three base pair difference from predicted positions. It is worth mentioning that the predicted terminus at 3’ end that used for primer walking was based on one intersected position from the top 10 rank of 3’ read edge frequencies that three Q8-like shared (TATTTTTCGA-3'), rather than the highest frequency among read edge positions (Supplementary Table 2.7, data sheet ‘Q8-like’). However, the primer designed a couple of hundred base pairs upstream of the intersected position still helped us locate the physical end of Q8-like strains by primer walking. The exact terminal position was revealed at position 5,099 against Q8 contig, which is 74 bp downstream of the predicted position 5,025 from NGS data. The genome configuration of Q8 was determined to have a 6,731 bp direct terminal repeat. Interestingly, this terminal sequence starts with 5'-AGGTTTT... was shared among SBP8a (AGAAAAACCT-3’), I13 (AGAAAAACCT-3’), I22 (AGAAAAACCT-3’), Q8 (5’- AGGTTTTGTG) and Q10 (5’- AGGTTTTGTG). The terminal sequence of SBP8a and I48-related phages was identical (5'- AGGTTTTTCT), which suggests that the sequence-specific terminus is probably the packaging recognition site on concatermeric DNA where the terminase complex initiates DNA packaging. Based on the result of primer walking, it is confirmed that SBP8a, I48-like and Q8-like phages had direct terminal repeat with exact length and a position-specific cleavage site. This suggested that these packaging enzymes possessed similar DNA pattern recognition activity. Although the terminus was thoroughly defined in this study, the relationship between the specific terminus and large terminase protein is unknown.

For phage Q11, we designed primers at around 150 bp upstream of contig ends and performed the Sanger sequencing to verify that the Q11 genome is complete from NGS data. Surprisingly, primer-walking revealed that Q11-like phage genomes had two position-specific cleavage sites outside of the contig. The artificial 'A' call occurred at 9 bp upstream of the first nucleotide and 43 bp downstream of last nucleotide of Q11 contig. The possibility of insufficient sequencing data should be excluded since the average coverage among Q11-like phages is between 2077.66 and 12345.86, which is relatively higher than other sequenced phages in this study. Smaller, phi-29 like phages have terminal proteins at the genome terminus that might interrupt genome sequencing (Inciarte *et al.*, 1976; Ito *et al.*, 1976). This could be tested for our smaller genomes by using an additional step of protease K treatment during the DNA extraction in the future experiments. It is also possible that the terminal sequence have tendency to form secondary structure that are difficult to address during library preparation or sequencing.

2.3.4 Calibration of terminus prediction with published NGS data from nine phages

Our terminus predicting tool performed fairly accurate prediction for novel phage isolates from topsoil sample. We additionally calibrated the prediction tool against published NGS data from nine phages with known packaging mechanisms and genome configuration. Three known types of phage genomes termini were included in this calibration: direct terminal repeats, circularly permuted and 3' overhang termini. Three different sequencing platforms for each type of genome terminus allowed us to investigate whether the sequencer has an effect on terminus prediction. Table 2.2 summarized terminus prediction of nine phage NGS data in this calibration. The

predicted terminus of Adelynn and Nigalana that were reported to have direct terminal repeats were successfully identified. It suggested that *Bacillus* phage Adelynn has a size of 2,693 bp direct terminal repeat, while Nigalana has 2,867 bp terminal redundancy. Intriguingly, the 5' terminus of Adelynn (5'- GGGTTTTTAT) and Nigalana (5'- AGGTTTTTCT) are mostly conserved against the suggested initiating terminus on SBP8a, I48-like and Q8-like phages. Our prediction tool failed to define the terminus though there was a coverage build-up indicated between 85kb and 88kb of Troll phage (Supplementary Figure 2.7). The phages with known circularly permuted genomes tended to have fluctuating coverage distribution without a consistent terminus since the phage genome has a diversity of physical terminus, though the *pac* site that the packaging motor recognizes on the concatemer is consistent. The terminus prediction of Breeniome, Teardrop and Zeenon (with known circular permutation) failed to identify a distinguishable terminus except for the 3' end of phage Teardrop. This terminus may be the first unit-length genome that was cleaved from concatemer, which was supposed to have a consistent cleavage by the *pac* site. However, the relatively low coverage of Teardrop NGS data might weaken this conclusion. Among the known 3' overhang phage genomes, our prediction tool identified the terminal position immediately adjacent to the reported positions of terminal sequence in the Mycobacteriophage database (http://phagesdb.org/) in the analysis of Equemioh13(Supplementary Figure 2.8). Zetzy was defined as a linear contig with a predicted terminus at position 34,529 bp by NCR and read edge frequency, which is inconsistent with the 3' overhang terminus (48455-48463, 5'-CGGGTGGTAA) reported on database. However, a linear contig is not assumed to have a terminus from internal sequence when the linear contig is complete,

which is inadequate to perform the terminus prediction *in silico*. Lilith phage genome had no clear coverage build-up region sufficient to surpass our criteria for terminus prediction (Supplementary Figure 2.7), so a possible terminal sequence was not offered.

2.3.5 Phylogenetic clustering by terminase large subunit and implications for genome end and packaging mechanisms types.

The annotated amino acid sequences of TerL from genomes of I48, Q8, SBP8a, and nine phages for predicting validation were defined based on ORF prediction and homologous sequence search with BLASTP as well as HHblits. These twelve TerL were then aligned with 69 known large terminase protein sequences of phages, covered as many packaging strategies as available from the current literature (Figure 2.2). Neither significant E-value ($< 10^{-10}$ in BLASTP) nor probability (> 80% in HHblits) for terminase large subunit was not found on Q11 genome, which indicated that Q11 genome might not have TerL gene. I48, Q8 and SBP8a clustered with SPO1-related phages, inferring that those genomes have long direct terminal repeats, which matched the prediction from NGS data for I48, Q8 and SBP8a. Adelynn, Nigalana and Troll were also clustered in the SPO1-like clade known to have direct terminal repeats. Equimioh13, Zetzy and Lilith were closely related to L5 and D29 that are known to have 3' overhangs at genome terminus. Breeniome, Zeenon and Teardrop were clustered with P22-like phages which have circularly permuted genomes. Overall, the reconstructed phylogeny based on amino acid sequence of phage terminase large subunit produced clusters corresponding to types of genome terminus. This result supported the prediction by NGS data and primer walking methods that the three novel phages have linear genomes with

direct terminal repeats. While the amino acid sequences of TerL were widely used to correlate with the type of DNA packaging and genome configuration, one should keep in mind that genetic recombination or horizontal gene transfer that causes the change of TerL gene across phages over evolutionary history could perturb the inference of 'phage clustering' by phylogenetic analysis (Casjens and Thuman-Commike, 2011). Therefore, the genome structure should be determined with caution by further experimental analysis rather than TerL phylogeny only.

2.3.6 A complete phage genome assembly should form a circular contig

The circularity of contig was found in a majority of NGS sequenced phages analyzed in this study (35 out of 39 single-contig novel isolates; eight out of nine published phages) (Supplementary Table 2.10). According to current understanding of tailed phage genome configuration, at least a portion of genome sequence has redundancy in one virion genome, except the packaging strategies found in Mu-like phage and phi29-like phage. To be more specific, a NGS data of tailed phage with 5' or 3' protruding end should have a region of overlapped mapping at cohesive sequences. This results in a circular map in assemblers and reports an arbitrary linear genome. A completely assembled contig of direct terminal repeat genome, as we demonstrated in the study, should be circular. In these cases, the high coverage and high occurrence of terminal nucleotide should exist in NGS data; and a terminus prediction method as the work in this study should be able to identify the terminal sequences. The sequencing data of a circularly permuted phage would of course form a circular-like map during assembly since redundant sequences among different headful packaged virions are diverse.

However, a study that analyzed highly sequenced reads of T4-like phage IME08 in NGS data revealed that the T4-like phage might have a sequence-specific cleavage on one terminus of genome (Jiang *et al.*, 2011). We observed that Q11-like contigs are linear because, again, there is no read that could be mapped to both end of contig. It also appeared incomplete since a missing portion of genome sequence was obtained using the primer-walking method. Therefore, the circularity of phage contig is necessary to retrieve a complete tailed phage genome, as well as to address terminus prediction using the methods in this study or similar programs. It is noteworthy that a phage with 5' cohesive end was not examined by the developed method in the study. There is also a limitation that using terminus predicting method *in silico* and direct sequencing without terminal ligation that the study performed are insufficient to identify the precise overhang sequence for either 5' cohesive end or 3' cohesive end. Deciding whether a tailed phage genome is completely sequenced and assembled through NGS requires further work.

2.4 CONCLUSIONS

Phage genome terminus prediction by NGS data is an efficient, but not absolutely precise, method to identify terminal sequences. The identification of phage genome terminus also allowed insight into potential DNA packaging mechanisms. Upstream sequence of predicted termini can be used to design primers for genome walking. We demonstrated that the Sanger sequencing yields promising data for determining physical ends of the genome. Primer walking confirmed that I13 and I22 (I48-like strain) had 2,822 bp direct terminal repeats. Our prediction tool identified an accurate position for one site-specific terminus, yet another site-specific terminus was 74 bp mispredicted from

NGS data. With aid from primer walking, the terminus prediction method also confirmed that the Q8 and Q10 (Q8-like strain) had about 6,730 bp direct terminal repeats. A conserved terminal sequence was found in SBP8a, I48-like strain and Q8-like strain. Q11 and I46 were characterized to have a linear genome without genome redundancy. In conclusion, this work suggests that NGS data could be useful to identify the terminal sequences of *de novo* linear phage genomes when single-contig and circularized assemblies were generated.

ACKNOWLEDGEMENT

This publication was made possible by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant #P20GM103408.

Figures:

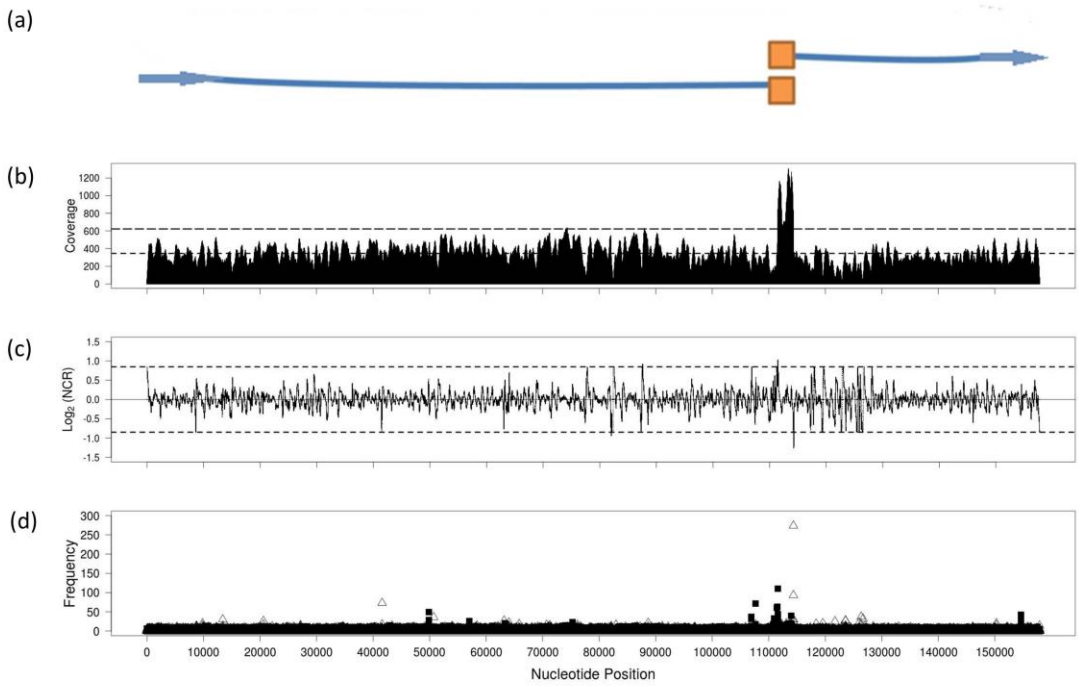


Figure 2.1. The map of coverage distribution, neighboring coverage ratio, and read edge frequency of I13 isolate. The visualized figure (a) illustrates the hypothetical type of genome configuration based on the coverage distribution (b), base-two logarithm of Neighboring Coverage Ratio (c) and 5' or 3' read edge frequency (d), resulting in a potential characteristic of direct terminal repeat. In figure (b), the lower dash line represents the average coverage of I13 sequencing reads. The upper dash line represents the 1.8 times of average coverage, which was defined as the criteria of genome redundant region. The dash lines on figure (c) indicate the cut-off of 1.8 fold change of NCR at level of (-0.848, 0.848). In figure (d), filled black squares point the frequency of 5' read edge position; triangles indicate the frequency of 3' read edge position.

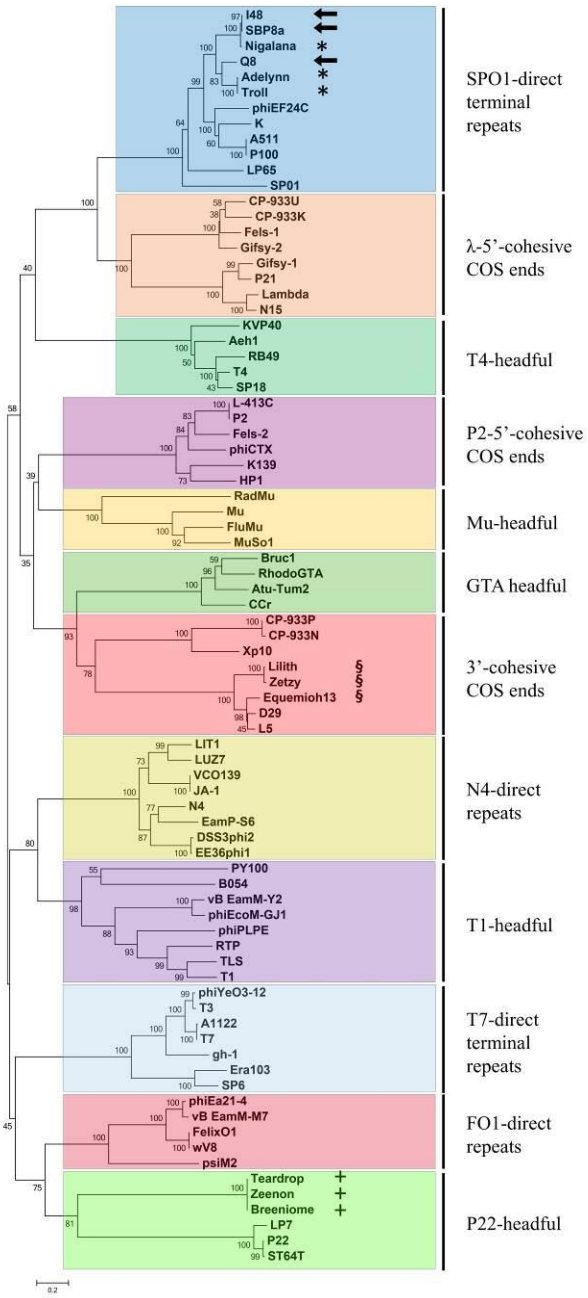


Figure 2.2. Neighbor-joining phylogeny of large terminase amino acid sequences. The alignment of protein sequence was generated by ClustalW2 (Larkin *et al.*, 2007). The phylogeny was reconstructed using Maximum Likelihood method based on the Poisson correction model. The number next to the internal nodes indicates the bootstrap value divided by trials size of 1,000. The names of phages were illustrated at the tip of the phylogeny. The root of the phylogeny was arbitrarily chosen for visualization purpose. Arrows: three novel *Bacillus* phages including SBP8a, I48 and Q8. Asterisks: nine phages with known types of genome terminus.

Tables:

Table 2.1. Genome similarity comparison among representative strains.

		I48	Q8	SBP8a	Q11
I48	Query Coverage;		1700/158180;	151518/158819;	70/26005;
	Identities		72%	94%	81%
	Overall Identity		0.774%	89.679%	0.002%
Q8	Query Coverage;			27614/158819;	57/26005;
	Identities			75%	89%
	Overall Identity			13.040%	0.196%
SBP8a	Query Coverage;				367/26005;
	Identities				94%
	Overall Identity				1.327%

Table 2.2. Summary of terminus prediction on selected isolated in this study and nine published phages.

Phage	Sequence Platform	# reads	Coverage	Contig size	Contig form	Ave. Freq.	5' edge position (FC>1.8)	NCR	Position with highest 5' edge frequency	Freq.	3' edge position (FC>1.8)	NCR	Position with highest 3' edge frequency	Freq.	5' terminus flanking sequence	3' terminus flanking sequence	Predicted termini
Novel Phages																	
SBP8a	*Roche/454*	38455	87.27	158794	Circular	0.24	**111794**	1.874	**111794**	117	**114616**	0.159	**114616**	238	5'-TCAGGTAGAA	AGAAAAACCT-3'	Site-specific termini
SBP8a	*MiSeq*	4662176	4670.87	158822	Circular	29.35	29634	1.889	30078	2434	32372	0.526	29983	2455	undef	undef	Can't be determined
I13 (I48-like)	*PGM*	186893	345.92	157905	Circular	1.18	**111600**	2.036	**111610**	110	**114359**	0.42	**114359**	274	5'-AAACCGTATG	AGAAAAACCT-3'	Site-specific termini
I48 (I48-like)	*MiSeq*	2975178	4042.41	157912	Circular	18.84	72862	1.819	73296	1332	74149	0.524	73201	1261	undef	undef	Can't be determined
Q8 (Q8-like)	*PGM*	72799	115.53	158180	Circular	0.46	**156549**	68.497	**156549**	965	no hit	N/A	156836	114	5'-AGGTTTTGTG	undef	5' site-specific terminus
Q11	*MiSeq*	2411528	15285.66	26005	Linear	92.73	no hit	N/A	15118	6717	no hit	N/A	15126	6090	undef	undef	Can't be determined
Published phage with known packaging mechanism																	
Direct Terminal Repeat																	
Adelynn	*MiSeq*	250419	215.43	162356	Circular	1.54	**18032**	3.346	**18032**	261	**20724**	0.499	**20724**	171	5'-GGGTTTTTAT	CCGCCTACCC-3'	Site-specific termini
Nigalana	*Roche/454*	35451	98.06	160174	Circular	0.22	**6458**	2.296	**6458**	121	**9324**	0.542	**9324**	95	5'-AGGTTTTTCT	CGTTCTACCT-3'	Site-specific termini
Troll	*PGM*	32949	29.17	163019	Circular	0.20	no hit	N/A	62795	7	no hit	N/A	43962	9	undef	undef	Can't be determined
Circular permutation																	
Breeniome	*MiSeq*	138039	60.10	154434	Circular	0.89	no hit	N/A	26061	17	no hit	N/A	48207	19	undef	undef	Can't be determined
Teardrop	*Roche/454*	9807	27.05	155389	Circular	0.06	no hit	N/A	10666	26	**11109**	0.497	**11109**	28	undef	CCGCTCCGTT-3'	3' Site-specific terminus
Zeenon	*PGM*	203150	179.03	155292	Circular	1.31	no hit	N/A	139104	23	no hit	N/A	10008	22	undef	undef	Can't be determined
3' overhangs																	
Equemioh13	*MiSeq*	150088	394.34	53042	Circular	2.83	**40880**	4.803	**40880**	276	no hit	N/A	41030	191	5'-TGCGGCCGCC	undef	5' Site-specific terminus
Zetzy	*Roche/454*	9157	80.92	48463	Linear	0.19	**34529**	3.321	**34529**	116	no hit	N/A	34586	70	5'-CCTGTGCGCC	undef	5' Site-specific terminus
Lilith	*PGM*	215716	668.86	50827	Circular	4.24	no hit	N/A	5180	89	no hit	N/A	3846	113	undef	undef	Can't be determined

NCR: neighboring coverage ratio; Freq.: frequency of read edge position.

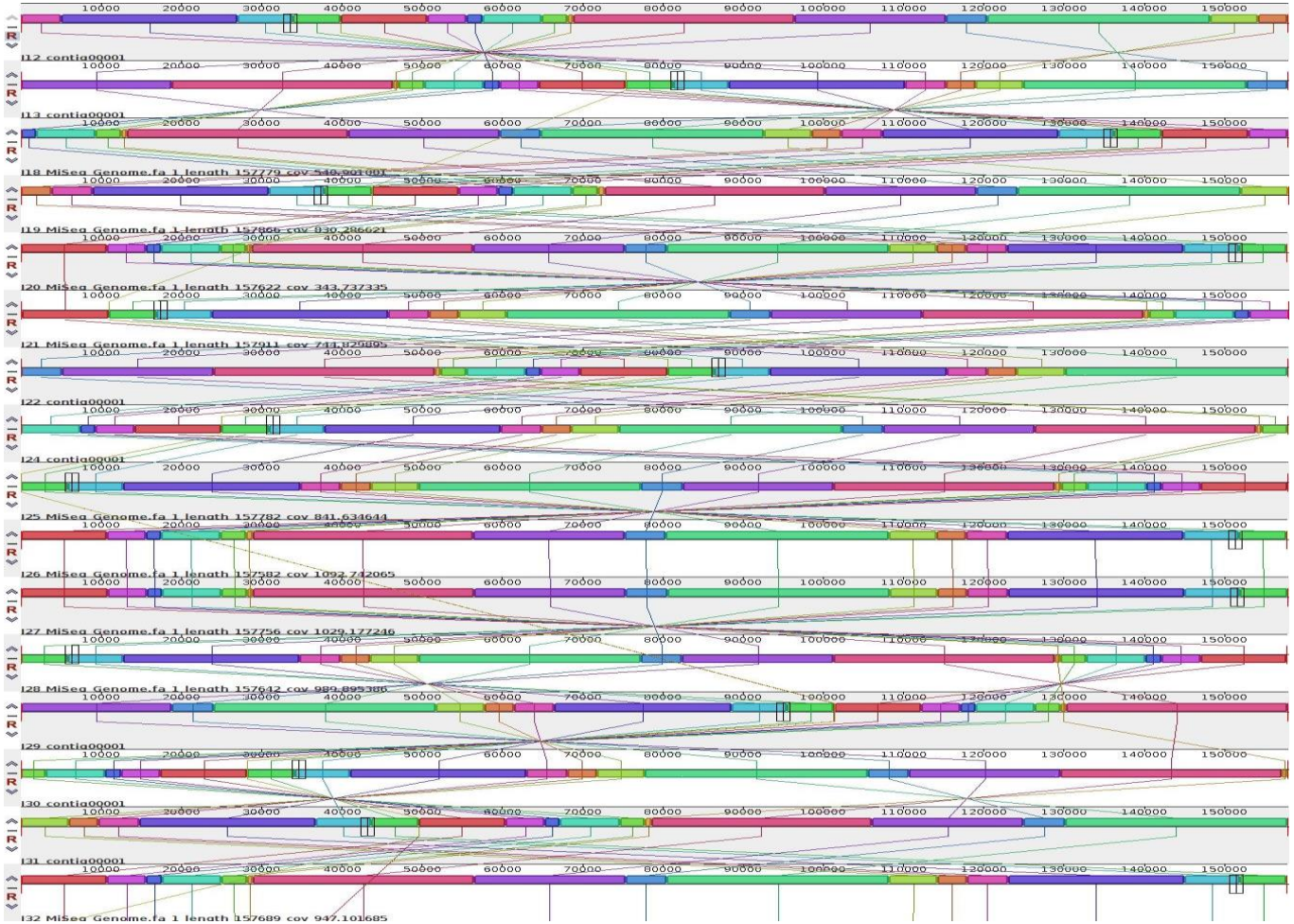
Table 2.3. Comparison of terminal position between the prediction from NGS data and identification from primer walking method.

Name	Terminus prediction from NGS data							Terminus identification from primer walking method						
	Sequencer	Contig size	5' terminus position	5' terminus flanking sequence	3' terminus position	3' terminus flanking sequence	size of direct terminal repeat	5' terminus position	position difference to prediction	5' terminus flanking sequence	3' terminus position	position difference to prediction	3' terminus flanking sequence	size of direct terminal repeat
SBP8a	*Roche/454*	158794	111794	5'-TCAGGTAGAA	114616	AGAAAAACCT-3'	2822	111796	+2	5'-AGGTAGAACG	114616	0	AGAAAAACCT-3’	2820
I13	*PGM*	157905	111610	5'-AAACCGTATG	114359	AGAAAAACCT-3'	2749	111537	-73	5’- AGGTAGAACG	114359	0	AGAAAAACCT-3’	2822
I22	*PGM*	157889	106421	5'-AAACCGTATG	109166	CCAGAGAAAA-3'	2745	106348	-73	5’- AGGTAGAACG	109169	+3	AGAAAAACCT-3’	2821
Q8	*PGM*	158180	156549	5'-AGGTTTTGTG	5025†	TATTTTTCGA-3'	6656	156549	0	5’- AGGTTTTGTG	5099	+74	GGGTCTACCC-3’	6730
Q10	*PGM*	158174	86402	5-'AGGTTTTGTG	93056†	TTATTTTTCG-3'	6654	86402	0	5’- AGGTTTTGTG	93130	+74	GGGTCTACCC-3’	6728
Q11	*MiSeq*	26005	n/a		n/a		n/a	-9*	n/a	5’ AAAATGTAAACC	+43*	n/a	GTACAATATACATTT-3’	n/a
I46	*MiSeq*	24896	n/a		n/a		n/a	-740*	n/a	5’ AAAATGTAAACC	+421*	n/a	GTACCATATACATTT-3’	n/a

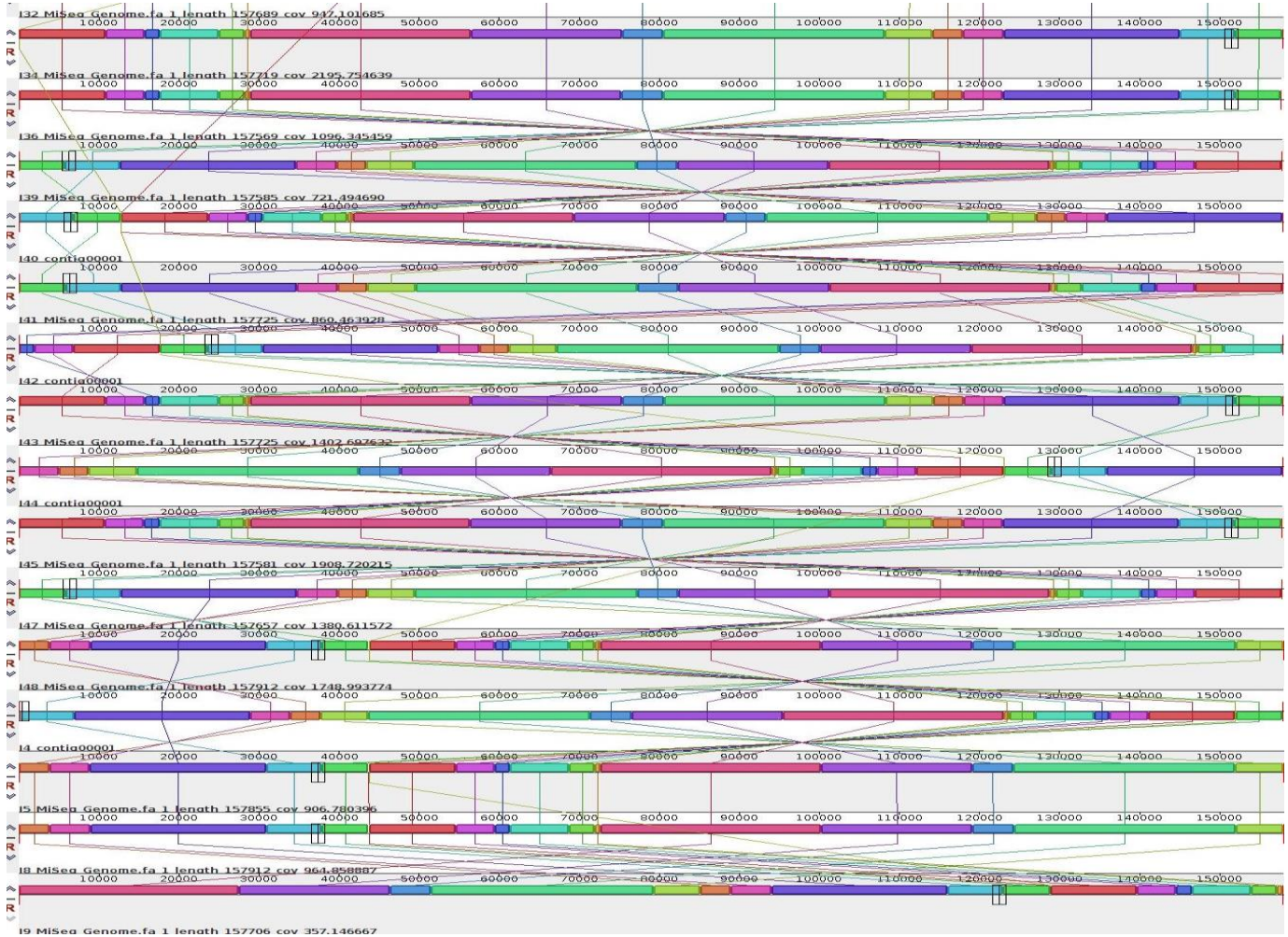
†: The predicted terminus for Q8 and Q10 was the intersection of terminal prediction among Q2, Q8 and Q10 instead of most promising prediction individually.

* : The position outside of contig sequence was calculated by the relative position of sequence from primer walking and assembled contig sequence from NGS data.

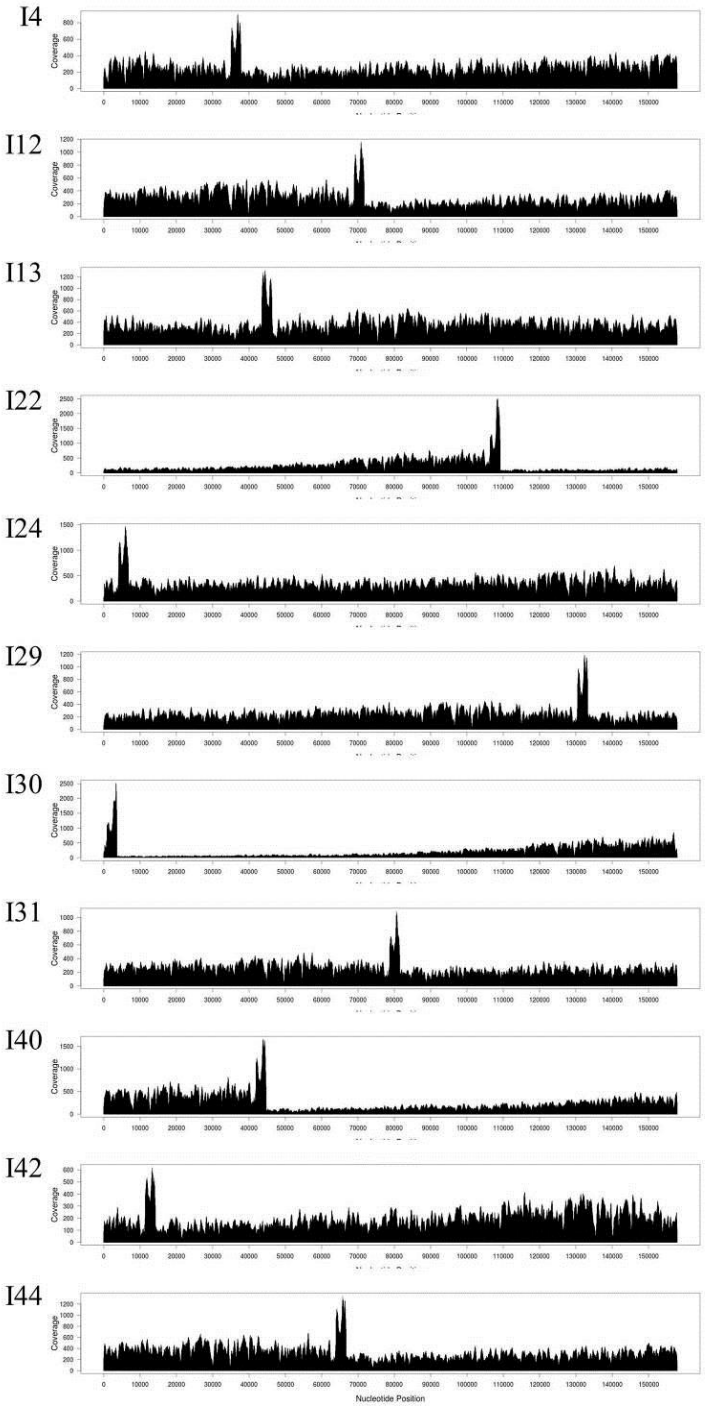
Supplementary Figures:



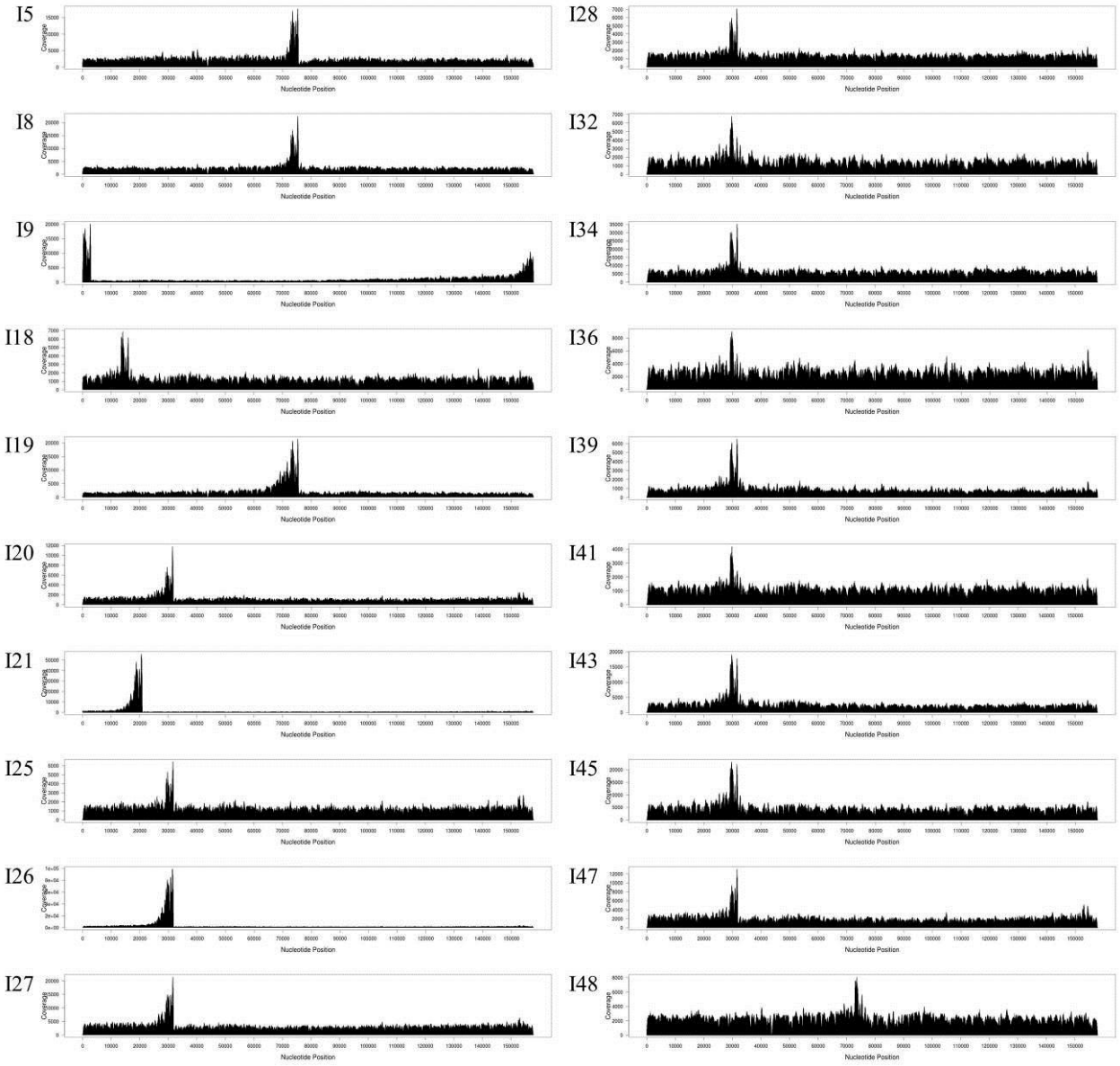
Supplementary Figure 2.1 (Continue on next page)



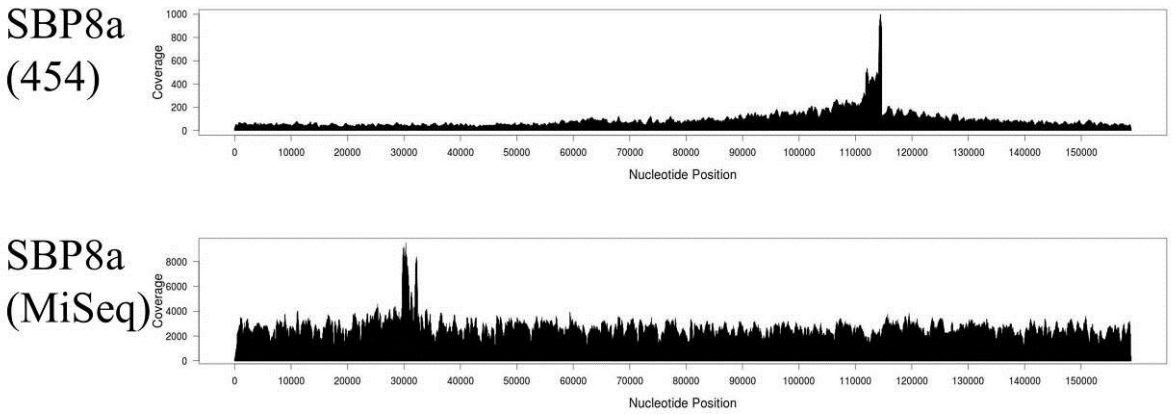
Supplementary Figure 2.1. A genome alignment of 31 I48-like isolates. Whole genome alignment using Mauve (Darling *et al.*, 2004) generated 16 Locally Collinear Blocks (LCB) conserved among isolates. A homologous block showed the same color on each horizontal genome.



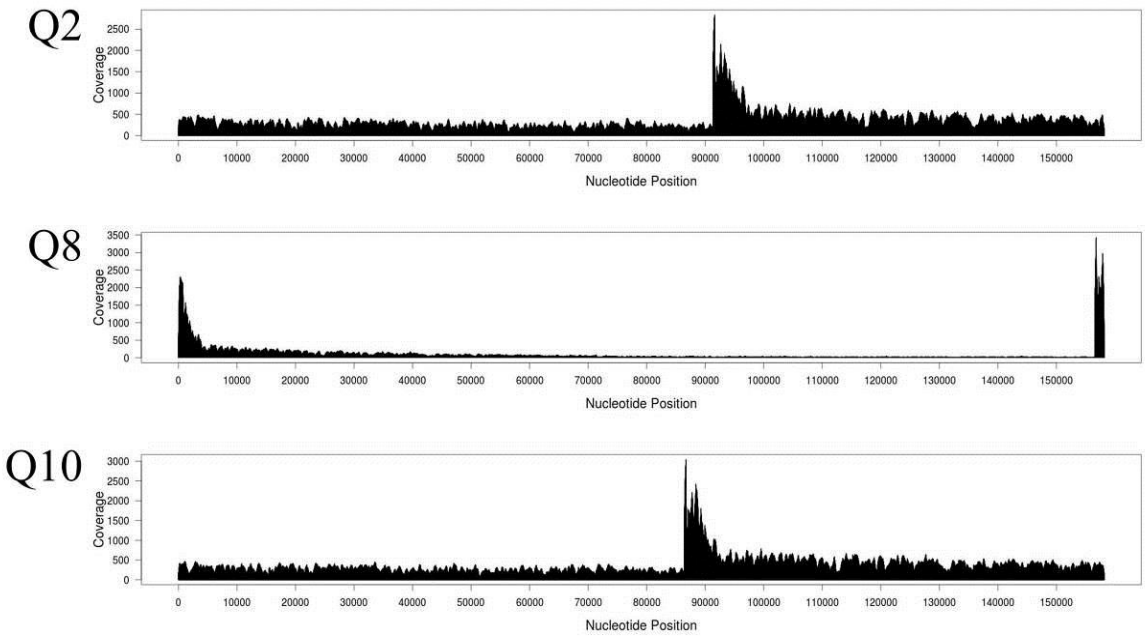
Supplementary Figure 2.2. Coverage distribution of 11 I48-like isolates sequenced by *Ion Torrent PGM*. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



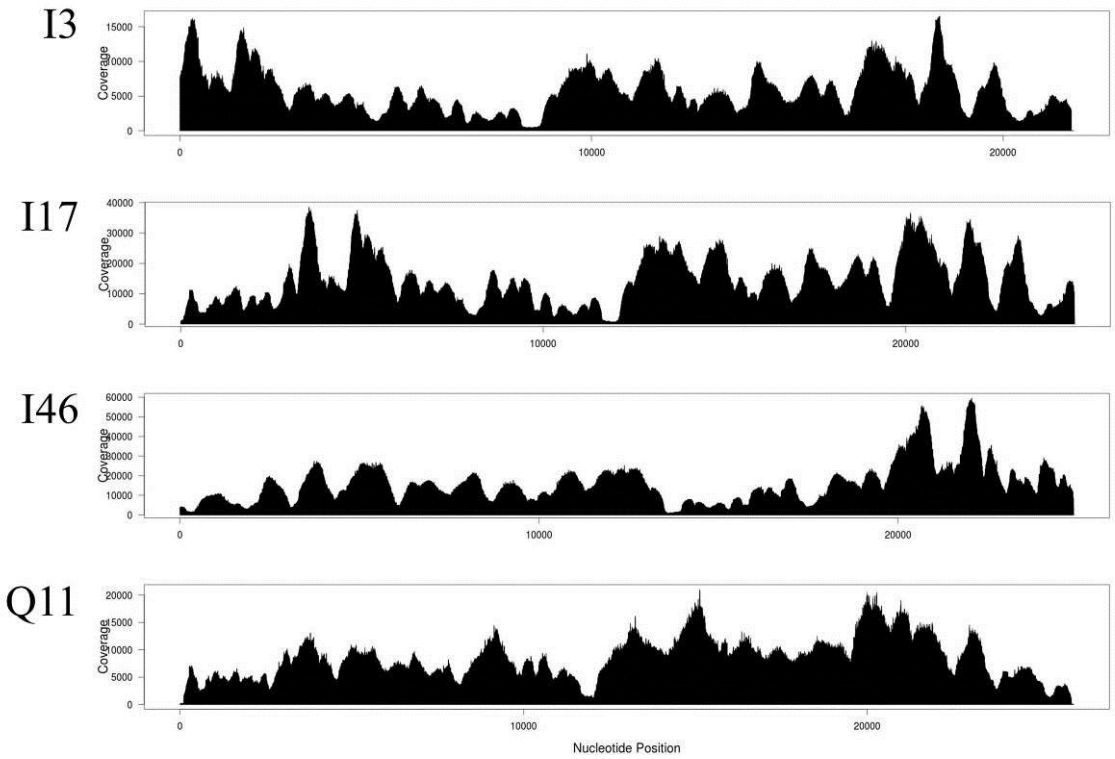
Supplementary Figure 2.3. Coverage distribution of 20 I48-like isolates sequenced by *MiSeq* genome sequencer. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



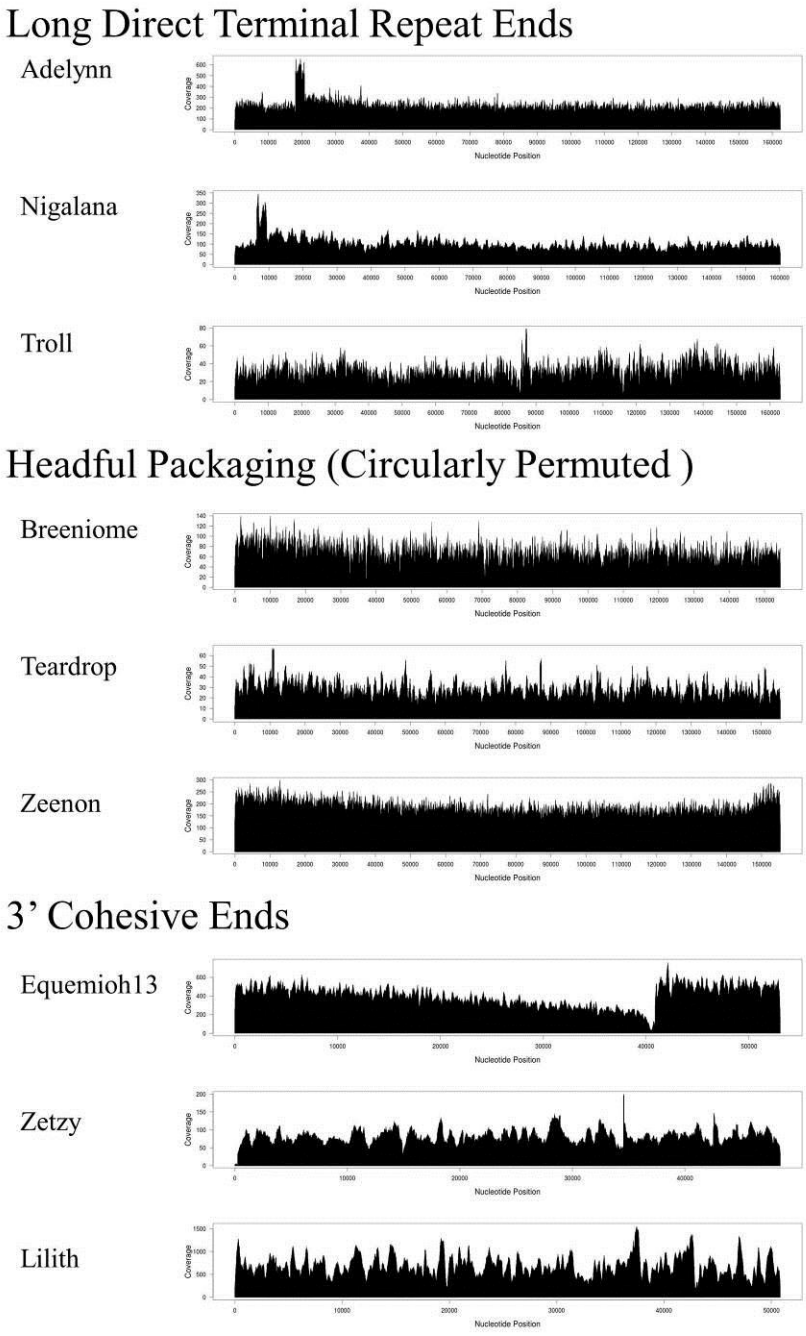
Supplementary Figure 2.4. Coverage distribution of SBP8a isolates sequenced by *Roche/454* or *MiSeq* genome sequencer. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



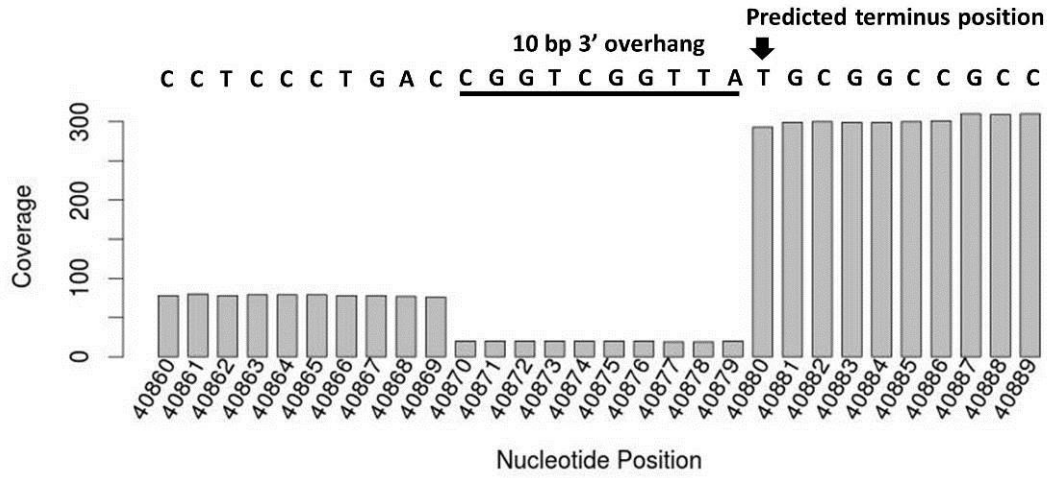
Supplementary Figure 2.5. Coverage distribution of three Q8-like isolates sequenced by *PGM* genome sequencer. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



Supplementary Figure 2.6. Coverage distribution of four Q11-like isolates sequenced by *MiSeq* genome sequencer. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



Supplementary Figure 2.7. Coverage distribution of nine previously sequenced phages by *MiSeq, Roche/454 or PGM* genome sequencer. X-axis represents the nucleotide position of assembled contig; y-axis represents the coverage on the corresponding position.



Supplementary Figure 2.8. Coverage distribution of Equemioh13 from position 40,860 to 40,889. The underlined sequences indicate the 10 bp 3' overhang sequence that was reported on database (http://phagesdb.org/). The significant hit of predicted terminus position on 40,880 is adjoined to the 3' overhang region.

Supplementary Tables:

Supplementary Table 2.1. Summary of 23 phage genome sequencing by *Ion PGM* and genome assembly.

Isolates	Read yield	Base yield	Average Length of read	number of contig (>500 bp)	Length of largest contig	Coverage
I2B	130478	36448132	279.34	408	157773	54.27
I12	149573	45096131	301.50	1	157910	276.26
I13	186893	56984029	304.90	1	157905	345.87
I15	166588	45725098	274.48	281	157854	81.34
I22	127336	37591312	295.21	1	157889	235.49
I24	192759	57331193	297.42	1	157885	325.63
I29	133229	39199737	294.23	1	157878	244.76
I30	122668	34307906	279.68	1	157864	214.81
I31	151854	40604960	267.39	1	157866	238.08
I33	139978	40400219	288.62	2	150224	250.21
I35	168585	49682300	294.70	2	102317	310.81
I37	161427	48105963	298.00	6	156872	286.96
I40	136464	41551515	304.49	1	157892	258.58
I42	96238	30157209	313.36	1	157871	188.73
I44	171708	52082227	303.32	1	157845	325.45
I4	131860	39400251	298.80	1	157919	241.91
I6	144911	42897782	296.03	517	157836	55.78
Q10	206095	61662555	299.19	1	158174	383.87
Q11	226973	67165785	295.92	2005	41047	8.09
Q1	227943	66445270	291.50	3198	131247	8.53
Q2	202252	58905221	291.25	1	158178	367.94
Q5	180990	52939743	292.50	2	126876	328.17
Q8	72799	18992728	260.89	1	158180	118.82

Supplementary Table 2.2. Summary of 26 phage genome sequencing by *MiSeq* paired-end sequencing and genome assembly.

Isolate	R1 yield	Average Length of R1	R2 yield	Average Length of R2	number of contig (>500 bp)	Length of largest contig	Coverage
I17	850507	236.85	850507	237.19	1	22845	6160.02
I18	471409	245.58	471409	246.06	1	157911	540.90
I19	726027	247.80	726027	248.07	1	158002	830.29
I20	512028	190.63	512028	191.11	1	157756	343.74
I21	953536	217.41	953536	217.48	1	158063	744.83
I25	972340	208.21	972340	208.50	1	157912	841.63
I26	1296142	208.05	1296142	208.64	1	157714	1092.74
I27	1348561	202.70	1348561	204.31	1	157886	1029.18
I28	946695	229.80	946695	230.48	1	157772	989.90
I3	330818	218.72	330818	217.84	1	21717	2077.66
I32	750439	271.95	750439	271.74	1	157823	947.10
I34	2339966	220.97	2339966	221.96	1	157851	2195.75
I36	862625	256.57	862625	256.43	1	157701	1096.35
I39	548444	255.90	548444	255.28	1	157715	721.49
I41	802973	233.30	802973	233.39	1	157855	860.46
I43	1050199	267.06	1050199	266.77	1	157857	1402.70
I45	1686621	252.02	1686621	253.10	1	157713	1908.72
I46	1693118	237.74	1693118	237.83	1	24896	12345.86
I47	1435984	219.28	1435984	219.92	1	157787	1380.61
I48	1487589	249.58	1487589	249.85	1	158046	1748.00
I50	962336	229.31	962336	229.52	2	158044	789.02
I5	971776	219.53	971776	220.07	1	157985	905.78
I8	977737	221.90	977737	222.40	1	158042	963.86
I9	455199	227.39	455199	228.58	1	157856	356.15
Q11	1205764	189.69	1205764	189.83	1	26005	5342.89
SPB8a	2331088	173.27	2331088	173.78	1	158951	1353.72

Supplementary Table 2.3. Number of nucleotide differences among Q strain.

	Q2	Q8	Q10
Q2			
Q8	3		
Q10	1	2	

Supplementary Table 2.4. Number of nucleotide differences among 25kb small-genome strain.

	I3	I17	I46	Q11
I3				
I17	68			
I46	0	68		
Q11	0	68	0	

Supplementary Table 2.5. Number of nucleotide differences among 158kb large-genome strain.

	I19	I18	I20	I21	I25	I26	I27	I28	I32	I34	I36	I39	I41	I43	I45	I47	I48	I5	I8	I9	I40	I42	I13	I44	I12	I29	I30	I4	I31	I22	I24
I19																															
I18	2																														
I20	2	0																													
I21	2	0	0																												
I25	3	1	1	1																											
I26	3	1	1	1	2																										
I27	3	1	1	1	2	2																									
I28	2	0	0	0	1	1	1																								
I32	2	0	0	0	1	1	1	0																							
I34	1	1	1	1	2	2	2	1	1																						
I36	3	1	1	1	2	2	2	1	1	2																					
I39	2	0	0	0	1	1	1	0	0	1	1																				
I41	4	2	2	2	3	3	3	2	2	3	3	2																			
I43	3	1	1	1	2	2	2	1	1	2	2	1	3																		
I45	3	1	1	1	2	2	2	1	1	2	0	1	3	2																	
I47	1	1	1	1	2	2	2	1	1	0	2	1	3	2	2																
I48	2	0	0	0	1	1	1	0	0	1	1	0	2	1	1	1															
I5	1	1	1	1	2	2	2	1	1	0	2	1	3	2	2	0	1														
I8	4	2	2	2	3	3	3	2	2	3	3	2	4	3	3	3	2	3													
I9	2	2	2	2	3	3	3	2	2	1	3	2	4	3	3	1	2	1	4												
I40	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5											
I42	7	5	5	5	6	6	6	5	5	6	6	5	5	6	6	6	5	6	7	7	2										
I13	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5	0	2									
I44	8	6	6	6	7	5	7	6	6	7	7	6	8	7	7	7	6	7	8	8	3	3	3								
I12	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5	0	2	0	3							
I29	6	4	4	4	5	5	5	4	4	5	5	4	4	5	5	5	4	5	6	6	1	1	1	4	1						
I30	6	4	4	4	5	5	5	4	4	5	3	4	6	5	3	5	4	5	6	6	1	3	1	4	1	2					
I4	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5	0	2	0	3	0	1	1				
I31	6	4	4	4	5	5	5	4	4	5	5	4	6	5	5	5	4	5	6	6	1	3	1	4	1	2	2	1			
I22	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5	0	2	0	3	0	1	1	0	1		
I24	5	3	3	3	4	4	4	3	3	4	4	3	5	4	4	4	3	4	5	5	0	2	0	3	0	1	1	0	1	0	

Supplementary Table 2.6. Genome terminus prediction of 31 I48-like isolates using NGS data.

						NCR							Read edge frequency						
Sequencer	Isolate	Genome Size	Contig form	Coverage	1.8 FC	L coverage	L start	L end	R coverage	R start	R end	Coverage ratio	5' edge position	Frequency	flanking sequence	3' edge position	Frequency	flanking sequence	Size of direct repeat
PGM	I12	157910	Circular	275.18	495.32	285.73	68887	68986	655.91	**68987**	69086	**2.296**	65061	116		**71748**	**216**	AGAAAAACCT-3'	2750
						528.02	71649	**71748**	158.98	71749	71848	**0.301**	**68998**	**114**	5'-AAACCGTATG	156889	64		
													64312	53		71747	50		
PGM	I13	157905	Circular	345.92	622.66	400.11	111500	111599	814.75	**111600**	111699	**2.036**	**111610**	**110**	5'-AAACCGTATG	**114359**	**274**	AGAAAAACCT-3'	2749
						840.61	114260	**114359**	352.68	114360	114459	**0.42**	107673	71		114358	93		
													111537	63		41587	73		
PGM	I22	157889	Circular	229.43	412.98	245.47	101642	101741	548.76	101742	101841	2.236	**106421**	**221**	5'-AAACCGTATG	**109166**	**864**	CCAGAGAAAA-3'	2745
						545.48	106313	106412	1091.87	**106413**	106512	**2.002**	102485	162		109169	47	AGAAAAACCT-3'	
						1199.59	109063	**109162**	96.25	109163	109262	**0.08**	106222	97		83473	37		
PGM	I24	157885	Circular	343.21	617.78	370.93	3962	4061	810.86	**4062**	4161	**2.186**	**4072**	**114**	5'-AAACCGTATG	**6821**	**193**	AGAAAAACCT-3'	2749
						687.41	6722	**6821**	354.31	6822	6921	**0.515**	137	89		91943	55		
													3999	58		6820	49		
PGM	I29	157878	Circular	240.61	433.10	270.82	130387	130486	677.08	**130487**	130586	**2.5**	**130501**	**114**	5'-AAACCGTATG	**133251**	**380**	AGAAAAACCT-3'	2750
						795.58	133152	**133251**	244.41	133252	133351	**0.307**	130428	102		60471	63		
													126563	78		133250	49		
PGM	I30	157864	Circular	211.04	379.87	209.41	173	272	397.13	273	372	1.896	**883**	**197**	5'-AAACCGTATG	**3632**	**852**	AGAAAAACCT-3'	2749
						524.39	773	872	1009.19	**873**	972	**1.925**	154809	143		3631	182		
						1375.04	3533	**3632**	46.21	3633	3732	**0.034**	683	88		3630	50		
						233.8	153966	154065	532.79	154066	154165	2.279							
PGM	I31	157866	Circular	246.09	442.97	243.08	78611	78710	532.76	**78711**	78810	**2.192**	**78720**	**120**	5'-AAACCGTATG	**81469**	**161**	AGAAAAACCT-3'	2749
						483.95	81370	**81469**	205.9	81470	81569	**0.425**	74782	81		8689	41		
													78521	48		88762	32		

PGM	I40	157892	Circular	252.48	454.46	202.37	37157	37256	465.31	37257	37356	2.299	**41935**	**211**	5'-AAACCGTATG	**44680**	**479**	AGAAAAACCT-3'	2745
						485.58	41825	41924	1027.15	**41925**	42024	**2.115**	38000	119		129813	62		
						725.15	44577	**44676**	78.36	44677	44776	**0.108**	41933	81		151399	36		
PGM	I42	157871	Circular	184.19	331.54	177.48	11349	11448	373.87	**11449**	11548	**2.107**	**11459**	**57**	5'-AAACCGTATG	**14207**	**116**	AGAAAAACCT-3'	2748
						304.98	14108	**14207**	122.52	14208	14307	**0.402**	7522	27		99298	39		
													54441	26		120889	30		
PGM	I44	157845	Circular	317.89	572.21	384.71	63829	63928	853.12	**63929**	64028	**2.218**	**63943**	**142**	5'-AAACCGTATG	**66692**	**317**	AGAAAAACCT-3'	2749
						772.23	66593	**66692**	258.02	66693	66792	**0.334**	60007	107		66691	80		
													63870	74		151773	71		
PGM	I4	157919	Circular	240.17	432.30	224.56	34940	35039	524.3	**35040**	35139	**2.335**	**35050**	**83**	5'-AAACCGTATG	**37799**	**139**	AGAAAAACCT-3'	2749
						478.61	37700	**37799**	211.47	37800	37899	**0.442**	31112	65		122939	60		
													35048	37		37798	52		
MiSeq	I18	157911	Circular	1277.56	2299.61	2033.25	13288	13387	4109.29	13388	13487	2.021	13438	631	5'-GTATAAAGGA	14079	574	ACTTCATGAT-3'	
						2978.71	14576	14675	1376.35	14676	14775	0.462	15881	611		13727	560		
						3560.73	16019	16118	1432.9	16119	16218	0.402	13822	604		14230	484		
MiSeq	I19	157866	Circular	2003.08	3605.54	9713.33	74063	74162	4456.9	74163	74262	0.459	73308	1915	ATATAATAGT	73565	1787	ACTTCATGAT	
						9779.37	75512	75611	1666.96	75612	75711	0.17	72924	1762		73213	1533		
													75367	1761		75375	1427		
MiSeq	I20	157756	Circular	1172.60	2110.68	1307.44	25751	25850	2363.28	25851	25950	1.808	31466	1508	GTTCTAGGGA	31474	1433	TGTTCTAGGG	
						4048.63	30160	30259	1876.14	30260	30359	0.463	31358	855		29664	845		
						4515.82	31192	31291	8291.69	31292	31391	1.836	31264	842		31423	735		
						4756.88	31616	31715	782.34	31716	31815	0.164							
MiSeq	I21	157912	Circular	1330.90	2395.62	26947.26	19313	19412	11786.11	19413	19512	0.437	20617	10789	GTTCTAGGGA	20625	9804	TGTTCTAGGG	
						21668.23	20768	20867	421.24	20868	20967	0.019	20566	7824		18815	8507		
													18558	7736		20574	7195		
MiSeq	I25	157912	Circular	1291.70	2325.06	1703.38	29009	29108	3176.44	29109	29208	1.865	31602	1563	GTTCTAGGGA	31610	1312	TGTTCTAGGG	
						2867.16	30298	30397	1513.71	30398	30497	0.528	29159	894		29800	1098		

						3497.77	31741	31840	1028.73	31841	31940	0.294	31428	824		29951	925	
MiSeq	I26	157714	Circular	3154.20	5677.55	51563.21	30210	30309	25446.17	30310	30409	0.493	31515	12188	GTTCTAGGGA	31523	12023	TGTTCTAGGG
						36293.41	31668	31767	1136.86	31768	31867	0.031	31351	8486		29713	8329	
													31313	8466		31472	7975	
MiSeq	I27	157886	Circular	3200.16	5760.29	9270.49	30299	30398	4683.81	30399	30498	0.505	31592	2432	GTTCTAGGGA	31600	2531	TGTTCTAGGG
						9887.3	31731	31830	2184.55	31831	31930	0.221	31428	1689		29790	1607	
													31390	1524		31549	1450	
MiSeq	I28	157772	Circular	1388.84	2499.92	3183.38	30197	30296	1468.49	30297	30396	0.461	31502	1484	GTTCTAGGGA	31510	1150	TGTTCTAGGG
						4235.59	31640	31739	1449.86	31740	31839	0.342	29059	1053		29700	1017	
													31451	1014		29348	913	
MiSeq	I32	157823	Circular	1482.64	2668.76	2898.95	30252	30351	1574.96	30352	30451	0.543	29498	654	ATATAATAGT	36918	593	GGTCTACAAT
													29114	520		157669	465	
													10821	472		29755	458	
MiSeq	I34	157851	Circular	6301.01	11341.82	9217.01	29000	29099	19214.61	29100	29199	2.085	31602	3775	GTTCTAGGGA	31610	2812	TGTTCTAGGG
						15297.62	30297	30396	6869.42	30397	30496	0.449	29159	3135		29448	2789	
						22189.93	31740	31839	8324.02	31840	31939	0.375	31551	2794		29800	2612	
MiSeq	I36	157701	Circular	2687.83	4838.09	2861.11	28995	29094	5387.68	29095	29194	1.883	29145	697	GTATAAAGGA	29434	661	AGTCTTATAC
													29529	624		29786	648	
													29426	529		36949	642	
MiSeq	I39	157715	Circular	901.02	1621.84	1871.55	28986	29085	3676.8	29086	29185	1.965	29146	930	GTATAAAGGA	29787	985	ACTTCATGAT
						3836.39	31728	31827	1185.79	31828	31927	0.309	31589	913		31597	791	
													29530	882		31827	788	
MiSeq	I41	157855	Circular	1123.39	2022.10								29573	699	AGTCTTATAC	29478	654	ACTATTATAT
													29189	635		29830	646	
													29127	519		36993	585	
MiSeq	I43	157857	Circular	2567.46	4621.43	9753.79	31738	31837	3495.08	31838	31937	0.358	29540	1708	ATATAATAGT	29445	1394	AGTCTTATAC
													29156	1382		29948	1346	

													31599	1158		29797	1324		
MiSeq	I45	157713	Circular	4586.84	8256.32	12179.72	30239	30338	4869.41	30339	30438	0.4	29484	2188	ATATAATAGT	29741	1951	ACTTCATGAT	
						13229.9	31680	31779	6104.7	31780	31879	0.461	29100	2035		29389	1890		
													31543	1847		29492	1605		
MiSeq	I47	157787	Circular	2030.77	3655.39	5602.18	30288	30387	2594.35	30388	30487	0.463	31592	2770	GTTCTAGGGA	31600	2447	TGTTCTAGGG	
						6235	31731	31830	1292.95	31831	31930	0.207	31390	1727		29790	1901		
													31428	1658		29941	1597		
MiSeq	I48	157912	Circular	4042.41	7276.34	5387.15	72762	72861	9797.35	72862	72961	1.819	73296	1332	ATATAATAGT	73201	1261	AGTCTTATAC	
						7137.53	74050	74149	3742.53	74150	74249	0.524	72912	1258		80716	1212		
													72850	1052		73553	1206		
MiSeq	I5	157855	Circular	2442.58	4396.64	9282.02	74044	74143	4534.12	74144	74243	0.488	75350	1680	GTTCTAGGGA	75358	1662	TGTTCTAGGG	
						6883.42	75495	75594	1151.3	75595	75694	0.167	73291	1413		73548	1580		
													75148	1238		73699	1266		
MiSeq	I8	157912	Circular	2442.91	4397.25	8994.41	74052	74151	4096.48	74152	74251	0.455	75357	2357	GTTCTAGGGA	75365	1945	TGTTCTAGGG	
						12038.09	75499	75598	2782.52	75599	75698	0.231	75306	1654		75595	1732		
													72914	1614		73555	1634		
MiSeq	I9	157856	Circular	1165.90	2098.62	9268.69	1323	1422	4068.06	1423	1522	0.439	2627	2145	GTTCTAGGGA	825	1759	ACTTCATGAT	
						8988.98	2778	2877	592.24	2878	2977	0.066	184	1911		473	1668		
													568	1879		2635	1588		

L= Left window; R=Right window

Supplementary Table 2.7. Genome terminus prediction of 3 Q8-like isolates using NGS data.

						NCR							Read edge frequency						
Sequencer	Isolate	Genome Size	Contig form	Coverage	1.8 FC	L coverage	L start	L end	R coverage	R start	R end	Coverage ratio	5' edge position	Frequency	flanking sequence	3' edge position	Frequency	flanking sequence	Size of direct repeat
PGM	Q8	158180	Circular	115.5256	207.9461	130.95	5026	5125	236.63	5126	5225	1.807	**156549**	**965**	5'-AGGTTTTGTG	156836	114	GCTTCAGAAG-3'	287
						27.73	156449	156548	1899.41	**156549**	156648	**68.497**	156551	475		156843	103	AAGAATAATA-3'	294
													157214	331		156848	100	TAATACATAG-3'	299
																156835	90		286
																156838	71		289
																158138	69		1589
																156846	62		297
																156842	60		293
																5025	56	TATTTTTCGA-3'	6656
PGM	Q10	158174	Circular	377.1076	678.7937	216.19	86302	86401	1778.86	**86402**	86501	**8.228**	**86402**	**730**	5-'AGGTTTTGTG	86748	194	GTTATTTGAG-3'	346
													86404	400		86737	116	TTCTTTTTTG-3'	335
													87067	370		**93056**	**99**	TTATTTTTCG-3'	6654
PGM	Q2	158178	Circular	360.5248	648.9447	207.55	91198	91297	1838.6	**91298**	91397	**8.859**	**91298**	**597**	5-'AGGTTTTGTG	91633	146	TTCTTTTTTG-3'	335
						746.64	93881	93980	1349.59	93981	94080	1.808	91300	558	5'-GTTTTGTGTT	91644	141	GTTATTTGAG-3'	346
						497.43	97856	**97955**	272.14	97956	98055	**0.547**	91963	286	5'-AAGATATTGA	**97953**	**93**	TTATTTTTCG-3'	6655

L= Left window; R=Right window

Supplementary Table 2.8. Genome terminus prediction of 4 Q11-like isolates using NGS data.

						NCR							Read edge frequency						
Sequencer	Isolate	Genome Size	Contig form	Coverage	1.8 FC	L coverage	L start	L end	R coverage	R start	R end	Coverage ratio	5' edge position	Frequency	flanking sequence	3' edge position	Frequency	flanking sequence	Size of direct repeat
MiSeq	I17	22845	Linear	14997.09	26994.77	No hit							1	8761	5'-GGTGTACATA	22721	6338		N?A
													11931	3337		13814	4909		
													13806	3281		2487	3763		
MiSeq	I3	21717	Linear	5746.133	10343.04	No hit							1	7787	5-'GCCAACGGAT	21656	2971	TTTGCATTAC-3'	N?A
													11760	1350		441	2175		
													128	1332		11768	1850		
MiSeq	I46	24896	Linear	15957.76	28723.96	No hit							21829	9126		24896	15570	TAAAAAAGAT-3'	N?A
													1	7978		22142	8871		
													21736	7579		22755	6861		
MiSeq	Q11	26005	Linear	8476.481	15257.67	No hit							15118	6717	5'-TCCTTACACTG	15126	6090		N?A
													19654	4268		3892	3549		
													20270	4032		20278	3460		

L= Left window; R=Right window

Supplementary Table 2.9. Genome terminus prediction of SPB8a isolate using NGS data.

						NCR							Read edge frequency						
Sequencer	Isolate	Genome Size	Contig form	Coverage	1.8 FC	L coverage	L start	L end	R coverage	R start	R end	Coverage ratio	5' edge position	Frequency	flanking sequence	3' edge position	Frequency	flanking sequence	Size of direct repeat
MiSeq	SBP8a	158822	Circular	2522.96	4541.32	2676	29533	29632	4989.64	29633	29732	1.865	30078	2434	5'-ATATAATAGT	29983	2455	AGTCTTATAC-3'	
						6572.65	32272	32371	3413.58	32372	32471	0.519	59407	2021		30335	1958		
													29975	1937		30086	1934		
Roche 454	SBP8a	158794	Circular	87.27	157.08	239.567	111764	111793	448.9	**111794**	111823	**1.874**	**111794**	**117**	5'-TCAGGTAGAA	114618	371	AAAAACCTGA-3'	2822
						779	114587	**114616**	124.033	114617	114646	**0.159**	111796	61	5'-AGGTAGAACG	**114616**	**238**	AGAAAAACCT-3'	
													114119	36		114612	39		

L= Left window; R=Right window

Supplementary Table 2.10. Genome terminus prediction of 9 published isolates using NGS data.

						NCR							Read edge frequency						
Sequencer	Isolate	Genome Size	Contig form	Coverage	1.8 FC	L coverage	L start	L end	R coverage	R start	R end	Coverage ratio	5' edge position	Frequency	flanking sequence	3' edge position	Frequency	flanking sequence	Size of direct repeat
MiSeq	Adelynn	162356	Circular	215.4328	387.779	148.81	17932	18031	497.85	**18032**	18131	**3.346**	**18032**	**261**	5'-GGGTTTTTAT	**20724**	**171**	CCGCCTACCC-3'	2693
						508.3	20625	**20724**	253.41	20725	20824	**0.499**	20574	111		18182	148		
													19366	42		19489	43		
Roche454	Nigalana	160174	Circular	98.05883	176.5059	118.84	6358	6457	272.83	**6458**	6557	**2.296**	**6458**	**121**	5'-AGGTTTTTCT	**9324**	**95**	CGTTCTACCT-3'	2867
						212.07	9225	**9324**	114.93	9325	9424	**0.542**	8834	63		6966	34		
													97125	14		31230	18		
PGM	Troll	163019	Circular	29.16864	52.50355	no hit							62795	7		43962	9		
													32333	6		88320	7		
													146300	6		87302	7		
MiSeq	Breeniome	154434	Circular	60.10079	108.1814	no hit							26061	17		48207	19		
													1949	17		94014	17		
													31738	15		22652	14		
Roche454	Teardrop	155389	Circular	27.04679	48.68423	63.26	11010	**11109**	31.44	11110	11209	**0.497**	10666	26		**11109**	**28**	CCGCTCCGTT-3'	
						52.73	87219	87318	26.45	87319	87418	0.502	86912	19		87318	21		
													77060	17		149602	17		
PGM	Zeenon	155292	Circular	179.0281	322.2506	no hit							139104	23		10008	22		
													153938	22		14389	21		
													26084	21		139787	21		
MiSeq	Equemioh13	53042	Circular	394.3426	709.8167	86.24	40780	40879	414.22	**40880**	40979	**4.803**	**40880**	**276**	5'-TGCGGCCGCC	41030	191		
													45395	46		40869	56		
													1073	45		41851	43		

Roche454	Zetzy	48463	Linear	80.92188	145.6594	47.61	34429	34528	158.09	**34529**	34628	**3.321**	**34529**	**116**	5'-CCTGTGCGCC	34586	70
													42513	29		48463	37
													245	28		28907	25
PGM	Lilith	50827	Circular	668.8557	1203.94	412.35	5083	5182	796.31	5183	5282	1.931	5180	89		3846	113
													32979	82		778	82
													13025	81		37658	65

CHAPTER 3

Complete Genome Sequence and genetic comparison of Three Novel Phages That Infect *Bacillus cereus*-group Bacteria*

ABSTRACT

In this study, we reported the genome sequence of *Bacillus anthracis* phage BJC, QCM8 and QCM11. BJC, SBP8a and QCM8 were predicted as *Myoviridae* using DNA packaging strategy similar to SPO1. The non-permuted genome sequences of *Myoviridae* phages BJC and QCM8 have long direct terminal repeats with lengths of 2,821 bp and 6,731 bp, respectively. There were 296 putative ORFs identified in unit length of BJC, and 276 ORFs were identified in QCM8. The QCM11 genome was predicted to have 41 ORFs in the genome. The nucleotide sequence of BJC and SBP8a were 89.68% similar and shared nearly identical characterized genes and corresponding order. Nine published phages were selected for genome comparisons based on nucleotide identity against BJC, SBP8a or QCM8. Four phi29-like phages were used to perform genome comparisons against QCM11. A total of 17 phage genomes were classified into three sequence-related clusters according to nucleotide identity as well as the number of homologous genes they shared to each other. Thirty-nine of 48 QCM8 characterized genes with corresponding functions of QCM8 showed conservation across BJC-related and QCM8-related phages. The genome characterization of QCM11 revealed that 12 putative proteins, including a terminal protein that phi29-like phages featured. The terminal proteins harbored on QCM11 and MG-B1 were conserved. QCM11 is the first *Podoviridae* phage to be sequenced that infects *Bacillus anthracis*.

* This chapter is based on a manuscript for publication. The authorship is listed as follow: Cheng-Han Chung, Michael H. Walter, Shu-Chuan (Grace) Chen, and Michael A. Thomas

3.1 INTRODUCTION

The bacteriophage (phage) is an entity of viruses that infect bacteria. Angly *et al*. suggested that the number of virus particles was estimated at 10^{31} in the biosphere (Angly *et al.*, 2005). They possess high specificity and sensitivity to target host bacteria, creating potential medical and scientific research applications (Adhya *et al.*, 2014; Hyman *et al.*, 2012). With the studies of virulence factors from human pathogens, bacteriophages are well-recognized as key agents in the evolution of bacterial pathogenesis (Boyd, 2012). The interchange between lytic cycle and lysogenic cycle allows phages undergo vertical and horizontal gene transfer within bacterial populations, which derives the acquisition of novel genes and functions in the co-evolution between phages and bacteria (Hendrix *et al.*, 2000).

Although numbers of *Bacillus* phages have been identified, there is scant knowledge about the genomic characteristics and diversity of these organisms. Grose *et al.* collected 101 Bacillus phage genomes available on either phagesDB (http://bacillus.phagesdb.org/) or Genbank by June 2014 (Grose *et al.*, 2014). Approximately 30 additional complete sequences have been published since then (updated in April 2015). Walter and Baker characterized three *Bacillus anthracis* phages from urban Iowa topsoil (Walter and Baker, 2003). One spore-binding phage named SBP8a has also been characterized morphologically by cryo-electron tomography (cryo-ET) (Fu *et al.*, 2011). Forty-eight plaque forming phages that infect *Bacillus anthracis/cereus* from the same soil source in Fu *et al.*'s study were isolated (unpublished data; Chung *et al.*). The genomic DNA of 48 isolates along with SBP8a were sequenced by high throughput sequencing. We identified three major clusters of *Bacillus* phages

among 48 novel isolates excluding SBP8a. Here we address the whole-genome analysis and genome annotation in detail with the complete genome sequence and genome configuration of BJC, QCM8, QCM11 as well as SBP8a.

3.2 MATERIALS AND METHODS

3.2.1 Culture media and bacteria strains

Host *Bacillus anthracis* Sterne and *Bacillus cereus* 569 UM20 were provided by Dr. J. Jackman's lab (Johns Hopkins University Applied Physics Laboratory, Laurel, MD) and Dr. Terri Koehler's lab (Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, Houston, TX), respectively. *Bacillus anthracis* Sterne is a vaccine strain without virulent plasmid pXO1 and pXO2. Tryptic Soy Broth (TSB, Difco) was used for growth and solid media plates (1.5% Bacto-Agar) in this study.

3.2.2 Isolation of Bacillus anthracis phages

Natural *B. anthracis* bacteriophage BJC was enriched from 'prairie-planting' soil (Cedar Falls, IA) by previously described methods (Fu *et al.*, 2011). Culture lysis, lysate clarification, triple-serially transfer and isolation were conducted by standard phage methods (Thorne, 1968). *B. anthracis* phage BJC was isolated and purified with the same procedure as SBP8a isolate (Fu *et al.*, 2011). Well-separated phage plaques were picked and processed through triple-serially plaque-dilution/transfers and isolates preserved in

200μL TSB at 4°C. Phage DNA was extracted followed by the method described in molecular biology laboratory manual (Sambrook and Russell, 2001). A 100ng aliquot of genomic DNA from each isolate was used to perform high throughput sequencing. QCM8 and QCM11 that simultaneously bind to quartz crystal microbalance (QCM) and *Bacillus anthracis* Sterne spore were collected (unpublished data, Walter *et al.*), following by the same triple-serially isolating and propagating procedure as BJC. Genome sequences of BJC, SBP8a, QCM8 and QCM11 were acquired by high throughput sequencing of *PGM* and *MiSeq* (unpublished data, Chung *et al.*).

3.2.3 Bioinformatics Analysis and genomic comparison

Genomic contigs were assembled with Newbler 2.9 and Velvet 1.2.10 for sequence reads from *Ion PGM* and *MiSeq* (Zerbino and Birney, 2008). Complete genome sequences and configuration of BJC, SBP8a, QCM8 and QCM11 were determined by 'neighboring coverage ratio' (NCR) and read edge frequency (unpublished data; Chung *et al.*). The terminal ends of four novel phages were verified by direct sequencing. Nucleotide identities between genomic sequences were calculated by percentage of aligned regions multiplying identities within aligned regions in pairwise alignment using blastn. Open reading frames (ORFs) were predicted with both Glimmer 3.02 (Delcher *et al.*, 1999; Salzberg *et al.*, 1998) and GeneMark (Besemer and Borodovsky, 2005). The potential tRNAs in the genome were predicted by Aragorn (Laslett and Canback, 2004). The search for homologous proteins and protein functions were defined as at least one best hit with described protein function among BLASTP (e-value < 10e-4, non-redundant database on NCBI), rpsBLAST (e-value < 10e-4, NCBI Conserved Domain Database

constructed by position-specific score matrices (PSSMs)) and hhblits (probability >80% and e-value< 10e-4, uniprot20 Hidden-Markov Model (HMM) database updated on March 2012). Dot plots were generated by Gepard (Krumsiek *et al.*, 2007) after the terminal repeats of BJC, SBP8a and QCM8 were removed and input sequences were rearranged by 5' terminal sequences of BJC or QCM8. The summary of genome annotation and Genbank file of novel phages and representative *Bacillus* phages were prepared using DNAMaster (http://cobamide2.bio.pitt.edu/). Genomic maps of phages were prepared by Phamerator (Cresawn *et al.*, 2011). A pham was defined as the homologous proteins when at least two putative proteins had pairwise identity greater than 32.5% and the e-value of the pairwise BLASTP less than 1e-50 among the pool of predicted ORFs in phages that input to Phamerator.

3.3 RESULTS

3.3.1 Basic characteristics of novel Bacillus phage isolates

Phage BJC and SBP8a have cross-infectivity between *B. anthracis* and *B. cereus*. QCM8 and QCM11 were collected by QCM: spore-binding technique. QCM11 formed relatively small and turbid plaques when increasing on TSA plates and log-phase *B. cereus* 569 UM20 (data not shown). It has been known that temperate phages usually produce turbid plaques due to lower efficiency to lyse their host cells since 1950's (Lwoff, 1953). This may suggest that QCM11 is in lysogenic cycle under the experimental condition.

BJC and QCM8 were found to have terminal repeat of 2,821 bp and 6,731 bp, respectively. QCM11 presented a linear sequence without genome redundancy. The terminal positions were predicted by methods that were developed in previous study (unpublished data; Chung *et al.*). A 5' terminal sequence AGGTTTTKYK was found to be conserved in phage BJC, SBP8a and QCM8. The conserved 5' terminus was used for rearranging genome sequences for phage BJC, SBP8a and QCM8. The unit-length genome sequence was used to search related phage isolates against non-redundant nucleotide collection on NCBI using blastn. The BLASTN search found five BJC-related phages (Hakuna, Megatron, BPS13, BPS10C and W.Ph.) with greater than 35% overall nucleotide identity against BJC and SBP8a (e-value < 1e-150). Four documented QCM8-related phages (BCU4, BCP78, Bcp1 and vB_BceM_Bc431v3) passed 35% nucleotide identity criteria against QCM8 (e-value < 1e-169). There was only phage MG-B1 with over 35% nucleotide identity against QCM11 (e-value <1e-96) in the nucleotide database on NCBI. The next hit had less than 3% nucleotide identity in blast result of QCM11. However, genome annotation of MG-B1 suggested that it possessed a putative gene encoding terminal protein and presumably classified into phi29-like phages (Redondo *et al.*, 2013). Therefore, phi29 and two other phi29-like phage GA-1 and B103 were included for genome comparison. Genome redundancies of selected phages were examined and removed by self-BLAST, and the sequences of conserved 5' terminus were used as standards to complement and rearrange the genome of related phages. Genome sequences of 17 related isolates were concatenated by order for dot plot (Figure 3.1). The diagonals among BJC-related and QCM8-related isolates indicated that not only

nucleotide identities but also the sequence arrangement were similar within groups (Figure 3.1A).

3.3.2 Genome annotation and prediction of gene functions

Predicted ORFs were summarized using GeneMark and Glimmer open reading frame predicting method embedded in DNA Master phage sequence annotator. Two hundred and seventy-six and 296 putative ORFs were identified in unit length of QCM8 and BJC, respectively. Three putative tRNAs were identified in phage BJC, whereas phage QCM8 anchored 13 tRNAs in the genome. Phage QCM11 was predicted to have 41 ORFs without computationally predictable tRNA in the genome (Table 3.1). Phage SBP8a had nearly identical organization and homology with BJC. The gene functions of 47 putative ORFs were assigned on phage BJC and SBP8a (Figure 3.2), resulting in 249 uncharacterized genes. A large portion of putative ORFs in QCM8 (232 out of 291) also remained uncharacterized. Three tRNAs in phage BJC and SBP8a encode two negative charged amino acids (aspartate and glutamate) and asparagine. Thirteen predicted tRNAs that carry thirteen different amino acids were identified on phage QCM8 including methionine. A phage that carries its own tRNAs might benefit in efficient codon usage and would be predicted to compensate for lack of sufficient codons in late lytic phase in host (Pope *et al.*, 2014).

3.3.3 Predicted function of proteins on BJC, SBP8a, QCM8 and homologous genes among selected Bacillus phages

The phage encoded proteins with annotated functions were classified by mechanisms including virion structure protein and assembly, host cell lysis, DNA replication, transcriptional regulation, biosynthetic process and host proteins (Table 3.2). A genome map illustrates high nucleotide similarity and coordinative gene locations between phage BJC and SBP8a (Figure 3.2). Phage QCM8 shared conserved proteins with phage BJC and SBP8a although the nucleotide similarity was relatively low against both of them. Among 48 annotated proteins, 39 of them were conserved and assigned as 'phams' across genomes of BJC-related phages and QCM8-related phages (Table 3.2).

Sixteen putative proteins known to be structural proteins were conserved in coordinative orders among 12 compared *Bacillus* phages, except one putative structural protein (QCM8_gp61) were annotated in QCM8-related cluster. Terminase large subunit in BJC-related phages and QCM8-related phages were clustered as a significant pham. This result recalled that these phages possess similar DNA packaging mechanism as SPO1-like phages with long terminal repeats according to a phylogeny reconstructed by protein sequence alignment of terminase large subunit (unpublished data, Chung *et al.*). The tail lysin 1 and 2 were closely related to tape measure protein chaperone on phage SPO1 and known to include a +1 programmed translational frameshift between two adjacent ORFs (Xu *et al.*, 2004). The tail fiber protein has been demonstrated to be required for host receptor recognition on γ phage (Davison *et al.*, 2005). The similar tail fiber protein sequences among selected phages inferred a similar host recognition mechanism that BJC, SBP8a and QCM8 share to host *Bacillus anthracis* and *Bacillus cereus*. However, a non-homologous ORF was located within the tail fiber protein of QCM8 (Figure 3.2, QCM8 gp86). Further experimental analysis is needed to examine

whether the inserted ORF containing uncharacterized domain in tail fiber protein affects the mechanism of host recognition and host range of QCM8.

An endolysin and a holin that are essential for phage to complete lytic cycle and release of viral particles were identified on BJC phage cluster (Hanlon, 2007). The holin was found to be conserved among 12 compared phages. No putative ORF in the genome of QCM8 was predicted to encode a conserved endolysin gene in the database. However, gp42 of QCM8 possesses N-acetylmuramoyl-L-alanine amidase domain that was homologous of Bc431v3 gp011 (El-Arabi *et al.*, 2013). This N-acetylmuramoyl-L-alanine amidase domain containing protein might be the possible endolysin for QCM8.

Two DNA polymerase proteins were found to be conserved among three novel phages. Several predicted proteins involved in DNA recombination were also identified, including exonuclease, SbcCD recombination nuclease and resolvase. The recombination systems of phage were known to play critical roles in the evolutionary history between tailed phages and hosts (Casjens and Thuman-Commike, 2011; Nafissi and Slavcev, 2014). BJC, SBP8a and QCM8 harbors two distinct transcription factors and a cII phage related protein. Gene expression control is crucial for phages to respond different conditions in host cells. For example, *cII* gene in lambda phages controls the dynamic of lifestyle between lytic cycle and lysogenic cycle of phages (Kobiler *et al.*, 2007).

3.3.4 Predicted function of proteins of QCM11 and homologies in other Bacillus phages

QCM11 has over 35% overall nucleotide identity and a similar gene arrangement to MG-B1. MG-B1 was claimed to be the first phage found that infects *Bacillus weihenstephanensis* (Redondo *et al.*, 2013). The annotation of MG-B1 discovered 14 homologous genes harbored on phi29-like phages. The phi29 phages family were classified into three groups based on DNA sequence, protein functions as well as morphological and serological characteristics (Pecenkova and Paces, 1999). Three phi29-like phages including phi29 (group I), B103 (group II) and GA-1 (group III) were selected to represent the nucleotide sequence diversity among phi29-like phages. The dot plot suggests that these phages were distinctly related except distinguishable diagonals between phage QCM11 and MG-B1 and between phage phi29 and B103 (Figure 3.1B).

A feature of phi29-like phage genome possesses is the inverted terminal repeats at both ends of linear dsDNA (Meijer *et al.*, 2001). The physical ends of phi29 and B103 present six-nucleotide inverted terminal repeat (3'- TTTCAT-5'), whereas phage GA-1 has eight reverse complementary nucleotide (3'-TTTATCTT-5') (Blanco *et al.*, 1992). The inverted repeat is involved in a sliding-back mechanism of the replication initiation (Bravo and Salas, 1997). This feature is conserved in protein-primed phages and adenovirus (Illana *et al.*, 1996; King and Vandervliet, 1994; Martin *et al.*, 1996). Indeed, phage QCM11 possesses a short inverted terminal repeat of 7 nucleotides (5'-AAATGTA-3').

The protein functions and corresponding mechanisms on phi29-like phages have been studied over half of a century (Meijer *et al.*, 2001). This allowed us to identify 12 functionally annotated proteins. Six of these are conserved genes across five QCM11-related phages with validated functions (Table 3.1). The morphological feature of phi29-

like phage is the polyhedral head and very short non-contractile tail, which has been classified as *Podoviridae* (Ackermann, 2009). The core structural proteins were conserved among QCM11-related phages, including major head protein (QCM11 gp25), tail protein (QCM11 gp26), and neck connector proteins that consist of upper collar connector (QCM11 gp27) and low collar protein (QCM11 gp28) (Figure 3.3). One of the two DNA polymerase presents in phage phi29 genome is conserved in phage QCM11 (gp13). The last homologous gene encodes DNA encapsidation protein (gp33). This protein involves in the binding capability to prohead then to DNA (Guo *et al.*, 1987).

One of the most important features of phi29-like phages is the protein-primed DNA sequence. The terminal protein (TP) covalently binds to the 5' DNA ends and serves as the primer for DNA polymerase to initiate DNA replication (Hermoso *et al.*, 1985). TP can attach to DNA ends as the origins of replication (parental TP). TP also forms a heterodimer with DNA polymerase (primer TP) to interact with parental TP for subsequent replication steps (Gutierrez *et al.*, 1986; Gutierrez *et al.*, 1986). However, the homologous TP protein (gp14) can only be found in MG-B1 but not in QCM11. TPs in phi29 (gp3), B103 (gp2) and GA-1 (gp11) formed a pham in protein clustering.

The predicted genes that are only conserved between QCM11 and MG-B1 include two DNA binding proteins (gp17 and gp23), two transcription regulators (gp15 and gp40) and dUTPase (gp3). dUTPase is a protein in tailed phages that functions as G-protein-like regulator and found to be crucial for controlling the transfer of virulence factors on *staphylococcal* phage (Tormo-Mas *et al.*, 2013).

3.4 DISCUSSION

Although the availability of sequence data and computational methods have been improved, insufficient evidence exists to apply protein function and corresponding mechanisms to sequences in phage genomics. In this study, over 80% of putative ORFs were uncharacterized after alignment against protein databases. With approximately 130 *Bacillus* phages sequenced, we report the genomic characterization of four *Bacillus anthracis* /*cereus* phages. Phage BJC, SBP8a and QCM8 were suggested to be new members of *Myoviridae* phages. Phage QCM11 was characterized as *Podoviridae* phage.

Besides the phage-related proteins, two putative genes in QCM8 encode bacterial proteins were identified. Phage QCM8 encoded sporulation protein YhbH. The homologous *yhbH* in *Escherichia coli* is involved in ribosome complex maturation (Ueta *et al.*, 2005). Ribosome stabilization in the stationary stage in bacteria might be influential for cell survival (Maki *et al.*, 2000). PhoH-like protein, which is also encoded by phage QCM8, is an ATP-binding protein that belongs to a phosphatase regulon. Though this gene participates in lipid and RNA metabolism, the function of phosphatase regulon is still not clear (Kazakov *et al.*, 2003).

This study characterized phage QCM11 as a phi29-like phage. The QCM11, to our knowledge, is the first *Podoviridae* phage found to infect *Bacillus anthracis*. The overall nucleotide identity between QCM11 and phage other than MG-B1 was less than 3%. However, six conserved phams in QCM11 were identified across different subgroups from phi29-like phage. It is worth mentioning that the isolated locations of phi29-like phages from different subgroups had geographical coincidence. Phi29 and phages that

belongs to subgroup I were all isolated in the U.S., subgroup II including B103 were discovered in Japan, and subgroup III phage such as GA-1 were all identified in Europe (Meijer *et al.*, 2001).

Figures:

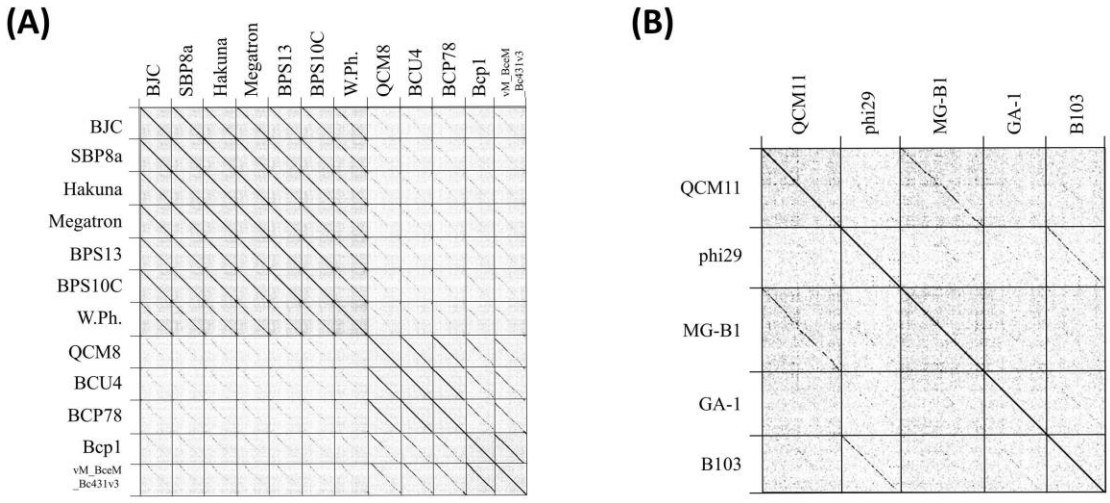


Figure 3.1. Nucleotide dot plot of BJC, SBP8a, QCM8 and QCM11 genome sequence and relative *Bacillus* phages. Dot plots were generated by Gepard (Krumsiek *et al.*, 2007) with a word length of 10.

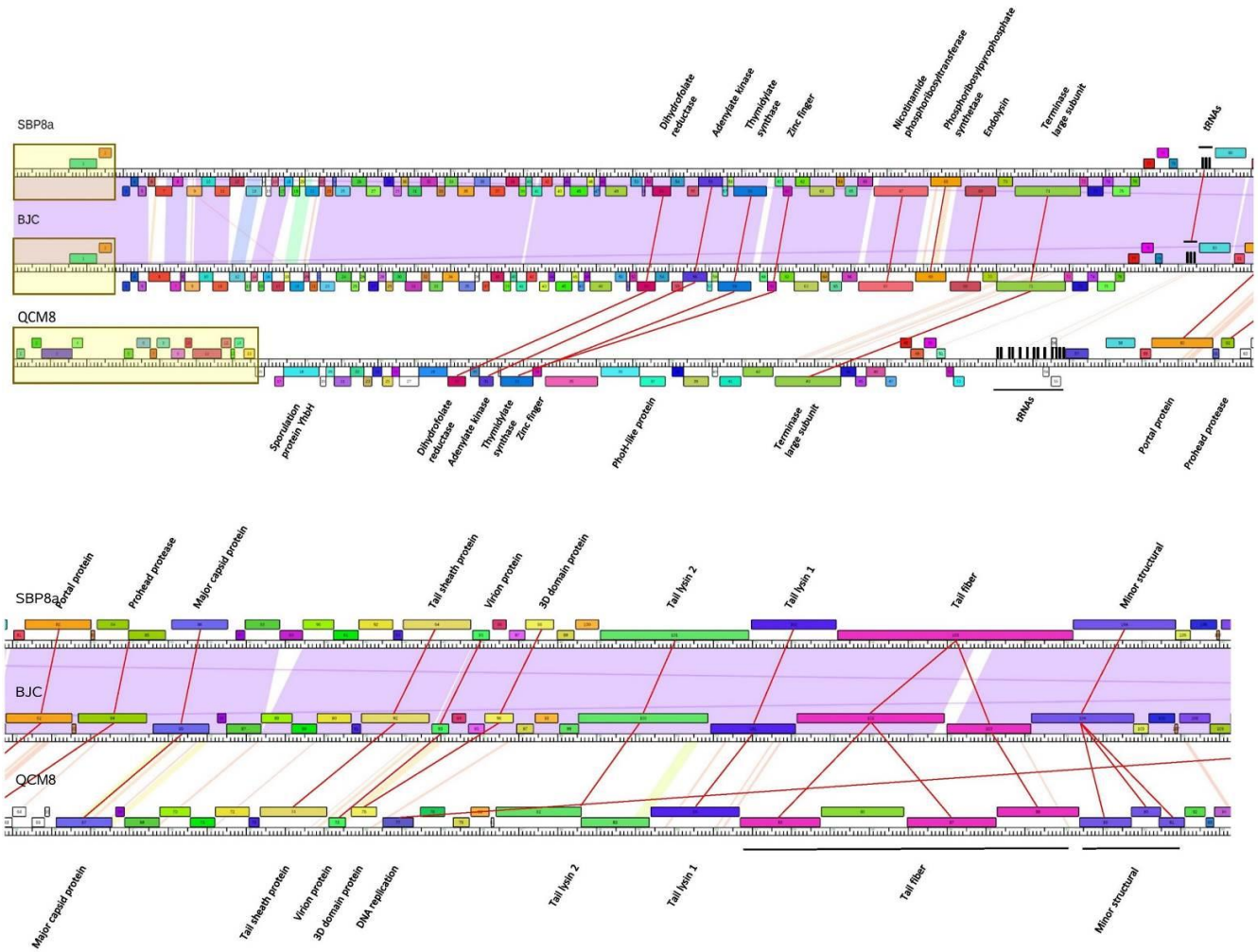


Figure 3.2 (Continue on next page)

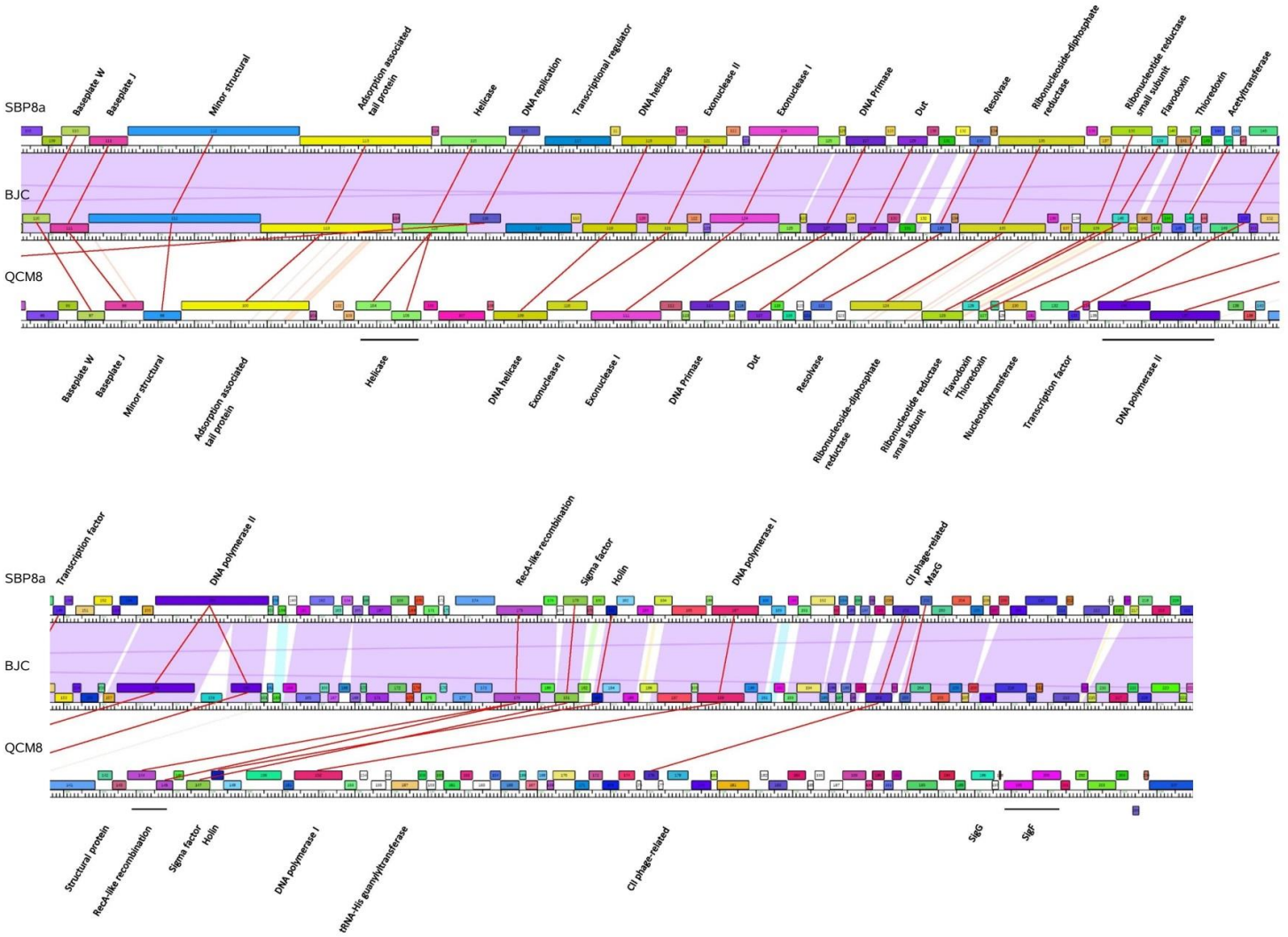


Figure 3.2 (Continue on next page)

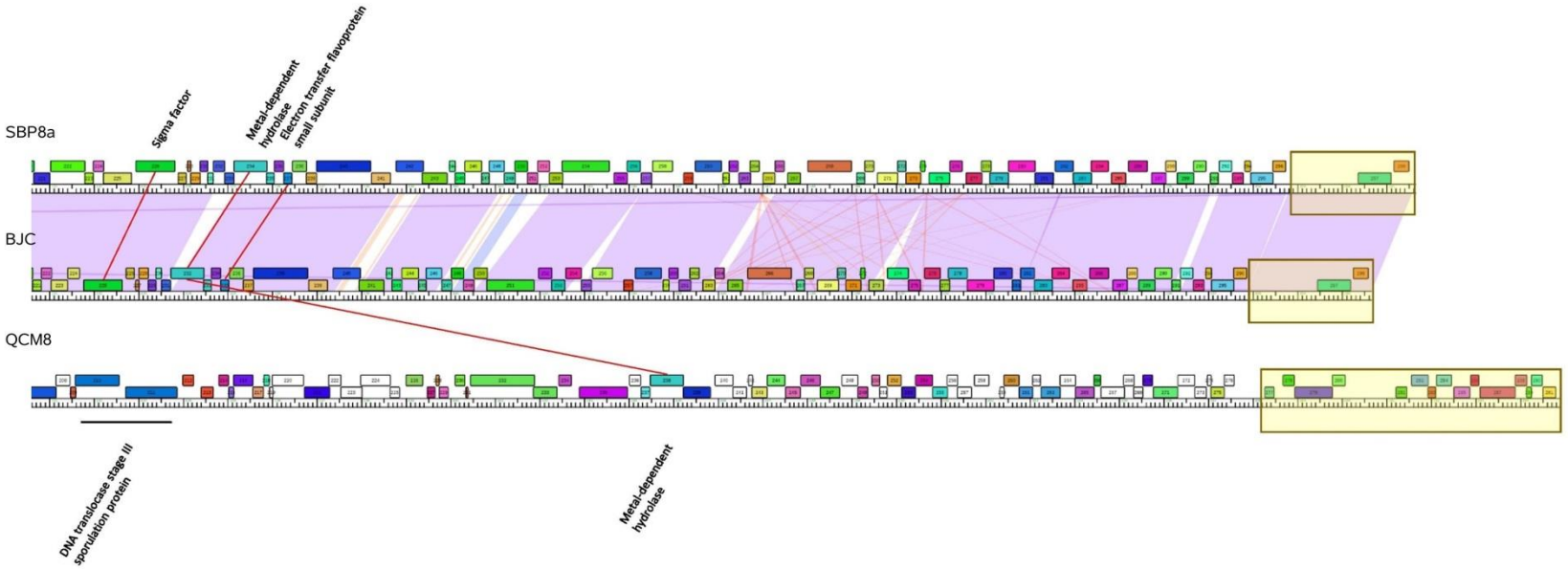


Figure 3.2. Genome maps of BJC, SBP8a and QCM8. Light and shaded regions between phages indicate the sequence of high nucleotide similarity. Boxes for putative ORFs are labeled when a protein function was predicted by gene annotation (see Method 2.4 for assigning protein functions in detail). Black boxes represent the putative ORFs of tRNAs. Dark red lines connecting boxes among phages represent homologous pham at corresponding locations. Yellow shaded boxes denote the terminal repeat regions at both ends of genome. SigG, RNA polymerase sigma 28 subunit G; SigF, RNA polymerase sigma 28 subunit F. Phage map was generated using Phamerator.

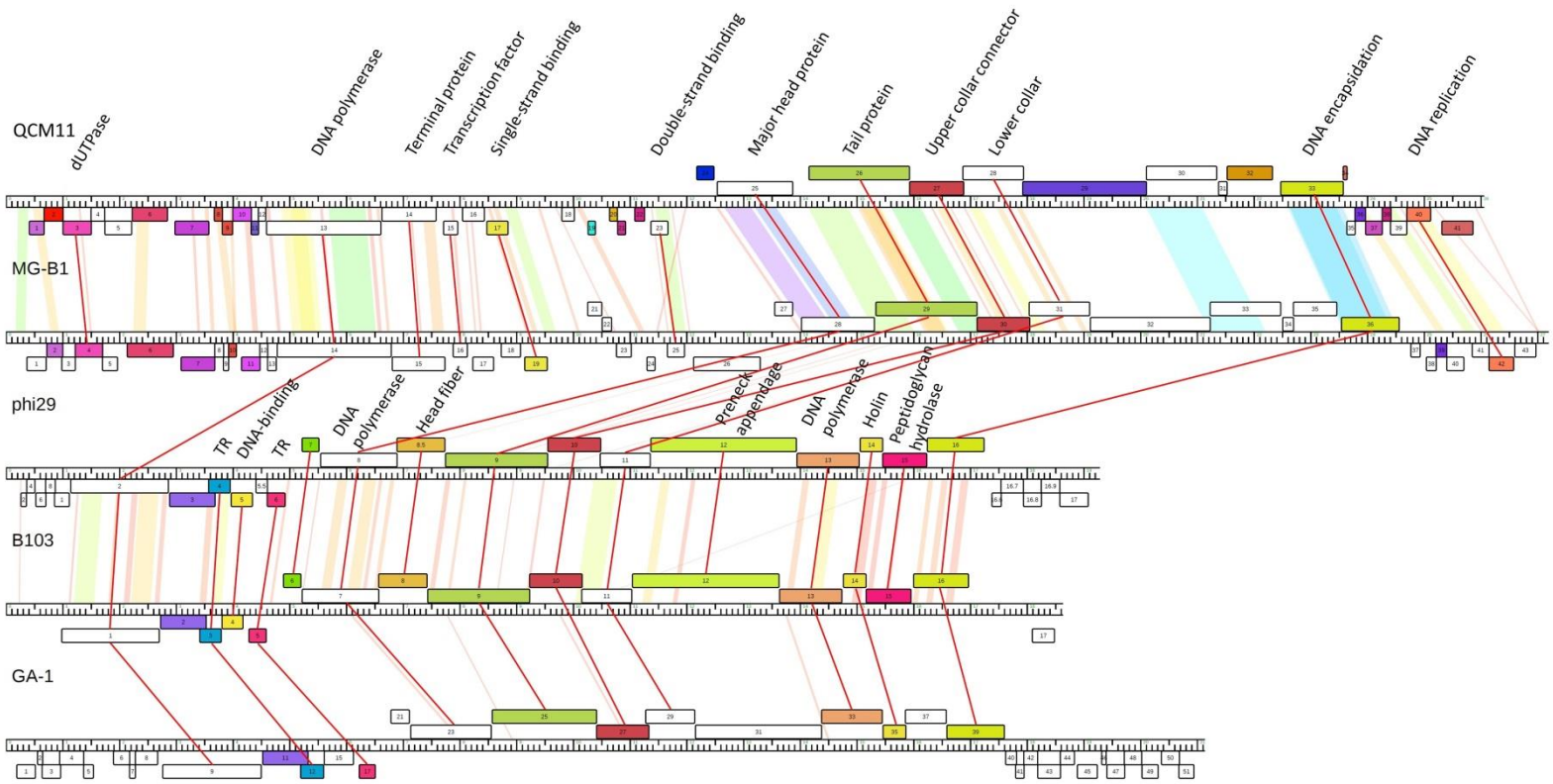


Figure 3.3. Genome maps of QCM11 and selected phages. Light and shaded regions between phages indicate the sequence of high nucleotide similarity. Boxes for putative ORFs are labeled when a protein function was predicted by gene annotation. Dark red lines connecting boxes among phages represent homologous pham at corresponding locations. Yellow shaded boxes denote the terminal repeat regions at both ends of genome. Phage map was generated using Phamerator. TR, Transcription regulator.

Tables:

Table 3.1. Characteristics of BJC, SBP8a, QCM8 and QCM11.

Isolate	Genome size (bp)		Gene counts				Gene density[c]	G+C contents (%)
	Unit length	DTR[a] length	ORF[b]	tRNA	in DTR (ORF, tRNA)	Function assigned ORF		
BJC	157909	2823	298	3	(2, 0)	48	92.70%	38.743
SBP8a	158822	2821	298	3	(2, 0)	47	93.07%	38.667
QCM8	158180	6731	291	13	(15, 0)	59	90.23%	39.898
QCM11	26054	N/A	41	0	N/A	17	84.51%	30.44

[a] DTR: Direct terminal repeat

[b] ORF: Open reading frame

[c] The gene density was calculated by the summation of nucleotide within ORFs and tRNA divided by the length of genome with DTR.

Table 3.2. Homologous genes with predicted functions among BJC, SBP8a, QCM8 and selected *Bacillus* phages

Protein function	# of members	BJC	SBP8a	Hakuna	Megatron	BPS13	BPS10C	W.Ph.	QCM8	BCP78	BCU4	Bcp1	vB_BceM
Structural proteins													
Terminase large subunit	12	gp71	gp71	gpgp069	gp070	gp6	gp006	gp003	gp43	gp0012	gp0005	gp057	gp010
Portal protein	12	gp82	gp82	gpgp079	gp081	gp264	gp266	gp012	gp60	gp0225	gp0214	gp070	gp235
Prohead protease	12	gp84	gp84, gp85	gpgp081, gpgp082	gp083, gp084	gp262, gp263	gp264, gp265	gp013, gp014	gp62	gp0222, gp0223	gp0211, gp0212	gp071, gp072	gp233, gp234
Major capsid protein	12	gp85	gp86	gpgp083	gp085	gp261	gp263	gp015	gp67	gp0221	gp0210	gp074	gp232
Tail sheath protein	12	gp92	gp94	gpgp090	gp092	gp254	gp255	gp022	gp74	gp0214	gp0203	gp081	gp225
Virion protein	12	gp93	gp95	gpgp091	gp093	gp253	gp254	gp023	gp75	gp0213	gp0202	gp082	gp224
Tail lysin 2	12	gp100, gp99	gp101	gpgp097	gp099	gp247	gp248	gp029	gp82, gp83	gp0206	gp0195	gp087	gp218
Tail lysin 1	12	gp101	gp102	gpgp098	gp100	gp246	gp247	gp030	gp84	gp0205	gp0194	gp088	gp217
Tail fiber protein	12	gp102, gp103	gp103	gpgp099	gp101	gp245	gp246	gp031	gp85, gp87, gp88	gp0201, gp0204	gp0190, gp0193	gp089, gp091, gp092	gp214, gp216
Minor structural protein	12	gp104	gp104	gpgp100	gp102	gp244	gp245	gp032	gp89, gp90, gp91	gp0200	gp0189	gp094	gp213
Baseplate assembly protein W	12	gp110	gp110	gpgp107	gp108	gp239	gp240	gp037	gp97	gp0194	gp0183	gp100	gp207
Baseplate j protein	12	gp111	gp111	gpgp108	gp109	gp238	gp239	gp038	gp98	gp0193	gp0182	gp101	gp206
Phage minor structural protein	12	gp112	gp112	gpgp109	gp110	gp237	gp238	gp039	gp99	gp0192	gp0181	gp102	gp205
Adsorption associated tail protein	12	gp113	gp113	gpgp110	gp111	gp236	gp237	gp040	gp100	gp0191	gp0180	gp103, gp104	gp204
Deoxyuridine 5-triphosphate, Dut	12	gp129	gp129	gpgp126	gp127	gp221	gp222	gp057	gp117	gp0178	gp0168	gp117	gp191
Holliday junction resolvase	12	gp133	gp133	gpgp130	gp131	gp216	gp217	gp061	gp122	gp0174	gp0164	gp121	gp187
Structural protein	5								gp142	gp0157	gp0147	gp140	gp167
Cell lysis													
Endolysin	7	gp69	gp69	gpgp067	gp068	gp8	gp008	gp001					
Holin	12	gp183	gp181	gpgp178	gp179	gp174	gp172	gp109	gp148	gp0152	gp0142	gp145	gp163

Gene regulation													
Transcriptional regulator	7	gp117	gp117	gpgp114	gp115	gp232	gp233	gp044					
Transcription factor 1	12	gp150	gp149	gpgp146	gp147	gp201	gp200	gp078	gp133	gp0164	gp0154	gp131	gp175
Transcriptional regulator	5								gp106, gp107	gp0187	gp0176	gp108	gp200
CII phage-related protein	11	gp201	gp201	gpgp198	gp197		gp157	gp127	gp176	gp0126	gp0116	gp178	gp129
RecA-like recombination repair protein	12	gp179	gp175	gpgp173	gp174	gp179	gp178	gp105	gp144, gp145	gp0155	gp0145	gp141	gp166
DNA replication													
Dihydrofolate reductase	12	gp53	gp53	gpgp051	gp052	gp23	gp023	gp256	gp29	gp0025	gp0017	gp043	gp024
Zinc finger protein 92 isoform	12	gp61	gp61	gpgp059	gp060	gp16	gp016	gp265	gp34	gp0021	gp0013	gp047	gp020
Helicase 2	12	gp115	gp115	gpgp112	gp113	gp234	gp235	gp042	gp104, gp105	gp0188	gp0177	gp107	gp201
DNA replication protein	10	gp116	gp116	gpgp113	gp114	gp233	gp234	gp043	gp77	gp0211	gp0199		
Replicative DNA helicase	12	gp119	gp119	gpgp116	gp117	gp230	gp231	gp046	gp109	gp0185	gp0174	gp109	gp199
Exonuclease II; Nuclease SbcCD D subunit	11	gp121	gp121	gpgp118	gp119		gp229	gp048	gp110	gp0184	gp0173	gp111	gp197
Exonuclease I	12	gp124	gp124	gpgp121	gp122	gp226	gp226	gp051	gp111	gp0183	gp0172	gp112	gp196
DNA Primase	12	gp127	gp127	gpgp124	gp125	gp223	gp224	gp054	gp114	gp0181	gp0170	gp115	gp194
Ribonucleoside-diphosphate reductase	12	gp135	gp135	gpgp132	gp133	gp214	gp215	gp064	gp124	gp0173	gp0163	gp122	gp186
Ribonucleotide reductase small subunit	1	gp138											
Ribonucleotide reductase small subunit 2	10	gp139	gp138	gpgp135	gp136	gp211	gp212	gp067	gp125	gp0172	gp0162		
DNA polymerase II	12	gp158, gp160	gp156	gpgp154, gpgp155	gp155, gp157	gp194	gp193	gp086, gp088	gp136, gp137	gp0162	gp0152	gp136	gp171
RNA polymerase sigma factor sigma	12	gp181	gp178	gpgp175	gp177	gp176	gp175	gp107	gp147	gp0153	gp0143	gp143	gp164
DNA polymerase I	12	gp189	gp187	gpgp184	gp185	gp169	gp167	gp115	gp152	gp0148	gp0138	gp149	gp158
RNA polymerase sigma factor	7	gp225	gp226	gpgp223	gp220	gp137	gp136	gp149					
RNA polymerase sigma 28 subunit SigG	5								gp196	gp0109	gp0099	gp199	gp110
RNA polymerase sigma 28 subunit SigF	3								gp199, gp200	gp0108	gp0098		

Biosynthetic process													
Adenylate kinase	12	gp56	gp56	gpgp054	gp055	gp20	gp020	gp259	gp31	gp0023	gp0015	gp045	gp022
Thymidylate synthase	12	gp59	gp59	gpgp057	gp058	gp18	gp018	gp262	gp33	gp0022	gp0014	gp046	gp021
Phosphoribosylpyrophosphate synthetase	7	gp68	gp68	gpgp066	gp067	gp9	gp009	gp274					
Thioredoxin	9	gp143	gp142		gp140	gp206	gp207	gp071	gp127	gp0170	gp0160		
Acetyltransferase	6	gp146	gp145	gpgp142	gp143		gp204	gp074					
Nucleoside Triphosphate Pyrophosphohydrolase; MazG	7	gp203	gp202	gpgp200	gp199	gp157	gp156	gp129					
Electron transfer flavoprotein small subunit	4	gp235	gp237	gpgp234	gp231								
Nucleotidyltransferase	5								gp130	gp0167	gp0157	gp127	gp178
tRNA-His guanylyltransferase	5								gp157	gp0146	gp0136	gp164	gp144
Metal-dependent hydrolase	12	gp232	gp234	gpgp231	gp228	gp131	gp130	gp156	gp238	gp0081	gp0071	gp226	gp080
Host functions/ pathogenesis													
Sporulation protein YhbH	5								gp18	gp0034	gp0026	gp027	gp041
PhoH-like protein	3								gp37	gp0018	gp0010		
Nicotinamide phosphoribosyl transferase	7	gp67	gp67	gpgp065	gp066	gp10	gp010	gp272					
3D domain protein	12	gp96	gp98	gpgp094	gp096	gp250	gp251	gp026	gp76	gp0212	gp0201	gp083	gp222
Flavodoxin	10	gp140	gp139	gpgp136	gp137	gp210	gp211	gp068	gp126	gp0171	gp0161		
DNA translocase stage III sporulation protein	5								gp210, gp211	gp0100	gp0090	gp210	gp099

Table 3.3. Homologous genes with predicted functions between QCM11 and selected *Bacillus* phages

protein function	number of members	QCM11	MG-B1	GA-1	phi29	B103
Structural protein						
Terminal protein	2	QCM11_14	MG-B1_015			
Major head protein	5	QCM11_25	MG-B1_028	GA-1p23	phi29_gp8	B103_7
Tail protein	5	QCM11_26	MG-B1_029	GA-1p25	phi29_gp9	B103_9
Upper collar connector	5	QCM11_27	MG-B1_030	GA-1p27	phi29_gp10	B103_10
Lower collar	5	QCM11_28	MG-B1_031	GA-1p29	phi29_gp11	B103_11
DNA encapsidation protein	5	QCM11_33	MG-B1_036	GA-1p39	phi29_gp16	B103_16
Morphogenesis protein	3			GA-1p33	phi29_gp13	B103_13
Head morphogenesis protein	2				phi29_gp7	B103_6
Head fiber protein	2				phi29_gp8.5	B103_8
Preneck appendage protein	2				phi29_gp12	B103_12
Cell lysis						
Holin	3			GA-1p35	phi29_gp14	B103_14
Peptidoglycan hydrolase	2				phi29_gp15	B103_15
Gene regulation						
Transcription factor	2	QCM11_15	MG-B1_016			
Single-strand binding protein	2	QCM11_17	MG-B1_019			
Double-strand binding protein	2	QCM11_23	MG-B1_025			
DNA replication organizer	2	QCM11_40	MG-B1_042			
Transcription regulator	3			GA-1p12	phi29_gp4	B103_3
Transcription regulator	3			GA-1p17	phi29_gp6	B103_5
Single stranded DNA-binding protein	2				phi29_gp5	B103_4
DNA replication/ metabolism						
dUTPase	2	QCM11_3	MG-B1_004			
DNA polymerase	5	QCM11_13	MG-B1_014	GA-1p09	phi29_gp2	B103_1

CHAPTER 4: CONCLUSIONS

This thesis interrogates the phage genome configuration (Chapter 2) and genomic structures (Chapter 3) of novel *Bacillus anthracis* phages by use of NGS data *in silico,* published genome analysis programs, custom written *perl* programs, and known databases.

4.1 Summary of empirical findings

In Chapter 2, the in-depth investigation of phage genome terminus demonstrates that the NGS data contains the characteristics that are significant for determining the physical ends of phage genome. The circularity of the phage genome assembly is the first trait that is required for identification of complete phage genomes. Secondly, the nucleotide positions with highest read edge frequency represented genome termini. Moreover, coverage build-up is an evidence of genome redundancy and physical ends of phage sequences.

In Chapter 3, BJC, SBP8a, and QCM8 were predicted as *Myoviridae* with DNA packaging strategy similar to SPO1, possessing the exact length of direct terminal repeats. The genome characterization of QCM11 discovers a close relationship to MG-B1, which reveals that QCM11 is a phi29-like phage belonging to *Podoviridae*. This suggests that QCM11 is the first *Podoviridae* phage to be sequenced that infects *Bacillus anthracis*.

4.2 Significance, comments and limitations of this study

The results of phage NGS data suggest that a complete phage genome assembly from high throughput data forms a circular contig if the tailed phage generates a concatemeric genome during replication. According to the DNA packaging strategies of concatemer-generating phages currently found in literature, the type of genome configuration can be 5' cohesive ends, 3' cohesive ends, long/short direct terminal repeats or circularly permuted genome. All the packaged dsDNA genomes have certain terminal redundancies that make the assembled contig circular. Furthermore, this characteristic is required for the terminus prediction to be accessible. Thirty-five out of 39 analyzed assemblies were circular based on the observation that more than one reads mapped to both ends of corresponding contig simultaneously. The other four linear-form contigs, however, that belong to QCM11-like phages appeared to be incomplete assemblies when misassembled terminus sequences were recovered by direct sequencing, suggesting that they have different DNA packaging strategy than that of a concatemer-generating phage. The BLAST search against non-redundant nucleotide collection in NCBI found that QCM11-like phage is closely related to phi29-like *Bacillus* phage MG-B1 (Redondo *et al.*, 2013), which suspected of utilizing the protein-primed mechanism to replicate the unit-length of DNA for progenies (Salas, 1991). Eight of nine published contigs also had circularity in NGS data, save for Mycobacteriophage Zetzy. It was reported to have 3' cohesive ends of 10 bp. Low coverage Zetzy sequencing data might have limited the assembly to recover the circularity. Nevertheless, the linear contig is inadequate for the terminus prediction method developed in this study so that the potential terminus of Zetzy cannot be adopted to validate the terminus prediction method.

This thesis describes a terminus prediction method developed using two criteria by dissecting raw reads generated by high throughput sequencing: 1) the highest or lowest Neighboring Coverage Ratio and, 2) the highest read edge frequency. Both criteria were set to look for the template ends of a linear genome. It is known that most of tailed phages have linear dsDNA genomes. It is also known that the circularity of assembled contig is the outcome of a concatemer-generating phage because of the portion of genome redundancy that was generated during DNA packaging. This gives rise to an arbitrary ends of assembly generated by assembler to in order to initiate and terminate the sequence assembly with a given algorithm. The arbitrarily generated terminus of linear output then loses information regarding the physical ends of linear genomes. However, the information is retained in raw reads. Every copy of a linear genome template that enters into the sequencing contains one fragment that possesses the 5' terminal end and another that contains the 3' terminal end. This gives rise to a higher proportion of reads that initiate or terminate the sequencing call at the physical ends. The characteristic was confirmed by tagging the physical ends of the T3 phage genome with a sequenced oligo-nucleotide before sequencing library preparation (Li *et al.*, 2014). It indicated that the nucleotide positions where the edge of reads located on the most represented genome termini. The nucleotide position with highest read edge frequency was found to be the exact position or flanking position (varied from 2- 74 bp upstream or downstream) of terminal ends in this study except the batch of phages that were sequenced by *MiSeq*. The potential reasons that the MiSeq NGS data did not inherit this characteristic might be due to the saturation of coverage or the fragmentation filtering step during library preparation. The average coverage of *MiSeq* sequencing in our study (1927.21 times sequenced /bp)

was over seven times higher than that of *PGM* sequencing (274.57 times sequenced /bp). However, evident coverage build-up could still be found in the coverage distribution (Supplementary Figure 2.3). The high coverage region could have over 20,000 coverages at the peak. Another possibility is that a selected size-range band of DNA fragments was excised from electrophoresis gel for adaptor ligated fragments in order to remove unligated adaptors or over-size fragments. The PCR clean-up step of Nextera XT DNA Library Preparation selected tagged fragments that were greater than 500 bp for 2x300 paired-end reads in this study. This step might remove a higher proportion of reads that contain terminal sequences. Nevertheless, the published *MiSeq* sequence cluster of phage Adelynn and Equemioh13 preserved this characteristic and potential termini were identified using this developed method.

The third characteristic is regional coverage build-up on coverage maps. It exists when the phage has either long or short direct terminal repeats. The developed program was designed to identify the edge of coverage build-up region by Neighboring Coverage Ratio (NCR). Therefore, instead of searching for high coverage peaks, the program tried to identify the peak of coverage change in terms of a coverage ratio. However, the ratio fluctuated a lot while comparing the coverage of one base pair against the next one. The fluctuation of ratio was leveraged by taking an average of coverage within a range of nucleotide positions. It indeed improved the situation and better identified the potential terminus. In the case of phages with direct terminal repeats, the coverage ratio increased gradually while the adjacent position of two windows was approaching the start of high coverage position of the genome redundant region, whereas the ratio decreased when the border of two windows got close to the end of the high coverage region. Windows size of

100 bp and the cutoff of NCR fold-change were empirically selected by repetitive examinations on our NGS data. They were designed as variables in the developed program for users to define different scale to identify the hits of potential terminus. For example, 3' cohesive end Equemioh13 had only 10 bp genome redundancy. While a 100 bp window size was able to predict the position of potential terminus, a smaller window size for this type of phages might better locate the predicted terminus on the physical ends of phage genomes.

BJC, SBP8a and QCM8 were suggested to be new members of *Myoviridae* phages based on nucleotide identities against known *Myoviridae* phages. Morphological evidence is needed to validate the structure of these phages such as TEM and cryo-electron tomography. QCM11 harbors 12 putative proteins including terminal protein conserved between QCM11 and MG-B1. Six of them are further conservative in terms of protein sequence similarities in phi29 phage. This is a unique cluster of phages that uses a protein-primed DNA replicating mechanism. This explained the outcome of linear assemblies in QCM11-like phage NGS data because none of a portion of terminal redundancy was observed to date in phi29-like phages. However, it has been reported that there is a six to eight bp inverted terminal repeat harbored at the physical ends of phi29-like genome. This feature was essential for a sliding-back mechanism of the replication initiation (Bravo and Salas, 1997) and being conserved among protein-primed DNA phages (Illana *et al.*, 1996; King and Vandervliet, 1994; Martin *et al.*, 1996). QCM11 also featured a 7-bp inverted repeat and was thought to be essential for replication and life cycle. However, the amino acid sequences of terminal protein in QCM11 and MG-B1 showed a relatively distinct relationship against other phi29-like phages that were

documented. Furthermore, QCM11 is the first *Podoviridae* phage to be sequenced that infects *Bacillus anthracis*. These indicate that a broad range of phage species is still missing in the picture of phage evolution, which urges the community to put more efforts on projects for phage hunting.

4.3 Future work

The significance of computational findings is limited without proof of direct evidence by using experimental analyses. Several directions that could help this project to be further conclusive and better improve the similar projects to identify novel phages are listed as follows:

4.3.1 The diversity of isolated phage from top soil sample

Though the replicates of similar or identical isolates helped the terminus prediction method be valid and conclusive, the diversity of naturally occurred phages that infect *Bacillus anthracis* has not been observed. Isolating more distinct *Bacillus anthracis* phages is warranted to better understand the phages genomes as well as their relationships to hosts in the close evolution process between phages and bacteria. The enrichment along with the triple-serially isolation was suspected to be a critical factor of sampling bias during the isolating step. Metagenomic sequencing is recommended to eliminate artificial selection. This would be a challenging and yet efficient method to discover a diverse range of phages from soil samples.

4.3.2 The identification of phage morphology

It is essential to characterize the phage morphologically for classifying novel phages against well-studied ones in literature. Although the genetic evidence suggested that BJC, QCM8 belong to *Myoviridae* and QCM11 is novel *Podoviridae* that infects *Bacillus anthracis*, the morphological evidence is inevitable to assert the class of phages. Transmission electron microscopy (TEM) is a widely used technique to identify the morphology of phages and cryo-electron tomography is useful for the composition of structural proteins and corresponding conformational changes of phages.

4.3.3 A validation of terminus prediction on 5' cohesive end phages

The study scrutinized types of phage genome configuration that covered 3' cohesive end, exact direct terminal repeat and circularly permuted phages. Although it is expected that the genome redundancy of 5' protruding end will give rise to the circularity of assembled contig, a set of phage NGS data is needed to conclude the characteristics of circularity of assembly and highly frequent position of read edges.

4.3.4 Simulation of NGS data

The procedure of high throughput sequencing involves variations and bias that are difficult to address. Although the highest read edge frequency was demonstrated to be a robust characteristic for identifying genome terminus of tailed phages, it remained

uncertain how this characteristic of phage NGS data was generated. I assumed that the possibility of every nucleotide position that was fragmented through genome fragmentation process was equally likely during sequencing library preparation. If the assumption was valid, the terminus-containing fragments should appear highest frequency in genome fragmentation pool. The enzymatic fragmentation step and subsequent PCR amplification and PCR clean-up may bring the variations to interrupt the distribution of read edge frequency. This might disrupt the result of terminus prediction, which may be the case that the batch of *MiSeq* sequencing experienced in this study.

A phage genome with a selected type of terminus discussed in the literature can be the input sequence of simulation. The process of sequencing could be simulated computationally starting from a batch of genome sequences, following by the simulated outcome of each step including genome fragmentation, read amplification and PCR clean up as similar as the protocol of genome sequencing does. This framework can demonstrate the distribution of read edge frequency, and emphasizes more on the importance of phage genome configuration in phage genome research.

4.3.5 Validating the frameshift of putative ORFs

The adjacent ORFs that represented the same putative proteins in the database were considered as frameshift of translation. Though tail fiber protein that experienced +1 programmed translational frameshift between two adjacent ORFs was well studied (Xu *et al.*, 2004), the frameshift happens more frequently when a insertion/deletion

presents within a ORF region of a novel genome. This could be solved by using direct sequencing that walks through the predicted ORFs that had implausible frameshift.

REFERENCES

Ackermann, H.W. Phage classification and characterization. *Methods in molecular biology* 2009;501:127-140.

Ackermann, H.W., Azizbekyan, R.R., Emadi Konjin, H.P., Lecadet, M.M., Seldin, L. and Yu, M.X. New Bacillus bacteriophage species. *Archives of virology* 1994;135(3-4):333-344.

Adhya, S., Merril, C.R. and Biswas, B. Therapeutic and prophylactic applications of bacteriophage components in modern medicine. *Cold Spring Harbor perspectives in medicine* 2014;4(1):a012518.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J. and Rohwer, F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC bioinformatics* 2005;6:41.

Besemer, J. and Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research* 2005;33(Web Server issue):W451-454.

Blanco, L., Bernad, A., Esteban, J.A. and Salas, M. DNA-independent deoxynucleotidylation of the phi 29 terminal protein by the phi 29 DNA polymerase. *The Journal of biological chemistry* 1992;267(2):1225-1230.

Blevins, S.M. and Bronze, M.S. Robert Koch and the 'golden age' of bacteriology. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2010;14(9):e744-751.

Born, Y., Fieseler, L., Marazzi, J., Lurz, R., Duffy, B. and Loessner, M.J. Novel virulent and broad-host-range Erwinia amylovora bacteriophages reveal a high degree of mosaicism and a relationship to Enterobacteriaceae phages. *Applied and environmental microbiology* 2011;77(17):5945-5954.

Boyd, E.F. Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. *Advances in virus research* 2012;82:91-118.

Bravo, A. and Salas, M. Initiation of bacteriophage phi29 DNA replication in vivo: assembly of a membrane-associated multiprotein complex. *Journal of molecular biology* 1997;269(1):102-112.

Brown, E.R. and Cherry, W.B. Specific identification of Bacillus anthracis by means of a variant bacteriophage. *The Journal of infectious diseases* 1955;96(1):34-39.

Brussow, H. and Hendrix, R.W. Phage genomics: small is beautiful. *Cell* 2002;108(1):13-16.

Bukhari, A.I. and Taylor, A.L. Influence of insertions on packaging of host sequences covalently linked to bacteriophage Mu DNA. *Proceedings of the National Academy of Sciences of the United States of America* 1975;72(11):4399-4403.

Bukhari, A.I. and Zipser, D. Random Insertion of Mu-1 DNA within a Single Gene. *Nature* 1972;236:240-243.

Casjens, S. and Hayden, M. Analysis in vivo of the bacteriophage P22 headful nuclease. *Journal of molecular biology* 1988;199(3):467-474.

Casjens, S., Winn-Stapley, D.A., Gilcrease, E.B., Morona, R., Kuhlewein, C., Chua, J.E., Manning, P.A., Inwood, W. and Clark, A.J. The chromosome of Shigella flexneri

bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *Journal of molecular biology* 2004;339(2):379-394.
Casjens, S.R. The DNA-packaging nanomotor of tailed bacteriophages. *Nature reviews. Microbiology* 2011;9(9):647-657.
Casjens, S.R. and Gilcrease, E.B. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods in molecular biology* 2009;502:91-111.
Casjens, S.R., Gilcrease, E.B., Winn-Stapley, D.A., Schicklmaier, P., Schmieger, H., Pedulla, M.L., Ford, M.E., Houtz, J.M., Hatfull, G.F. and Hendrix, R.W. The generalized transducing Salmonella bacteriophage ES18: complete genome sequence and DNA packaging strategy. *Journal of bacteriology* 2005;187(3):1091-1104.
Casjens, S.R. and Thuman-Commike, P.A. Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology* 2011;411(2):393-415.
Catalano, C.E., Cue, D. and Feiss, M. Virus DNA packaging: the strategy used by phage lambda. *Molecular microbiology* 1995;16(6):1075-1086.
Chung, Y.B. and Hinkle, D.C. Bacteriophage T7 DNA packaging. II. Analysis of the DNA sequences required for packaging using a plasmid transduction assay. *Journal of molecular biology* 1990;216(4):927-938.
Cohen, A., Ben-Ze'ev, H. and Yashouv, J. Outgrowth of Bacillus cereus spores harboring bacteriophage CP-51 DNA. I. Initiation of bacteriophage development. *Journal of virology* 1973;11(5):648-654.
Cregg, J.M. and Stewart, C.R. Terminal redundancy of "high frequency of recombination" markers of Bacillus subtilis phage SPO1. *Virology* 1978;86(2):530-541.
Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R.W. and Hatfull, G.F. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC bioinformatics* 2011;12:395.
Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 2004;14(7):1394-1403.
Davison, S., Couture-Tosi, E., Candela, T., Mock, M. and Fouet, A. Identification of the Bacillus anthracis (gamma) phage receptor. *Journal of bacteriology* 2005;187(19):6742-6749.
de Beer, T., Fang, J., Ortega, M., Yang, Q., Maes, L., Duffy, C., Berton, N., Sippy, J., Overduin, M., Feiss, M. and Catalano, C.E. Insights into specific DNA recognition during the assembly of a viral genome packaging machine. *Mol Cell* 2002;9(5):981-991.
Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic acids research* 1999;27(23):4636-4641.
Donnelly-Wu, M.K., Jacobs, W.R., Jr. and Hatfull, G.F. Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria. *Molecular microbiology* 1993;7(3):407-417.
Duffy, C. and Feiss, M. The large subunit of bacteriophage lambda's terminase plays a role in DNA translocation and packaging termination. *Journal of molecular biology* 2002;316(3):547-561.

Dunn, J.J. and Studier, F.W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *Journal of molecular biology* 1983;166(4):477-535.
El-Arabi, T.F., Griffiths, M.W., She, Y.M., Villegas, A., Lingohr, E.J. and Kropinski, A.M. Genome sequence and analysis of a broad-host range lytic bacteriophage that infects the Bacillus cereus group. *Virology journal* 2013;10:48.
Ellis, D.M. and Dean, D.H. Nucleotide sequence of the cohesive single-stranded ends of Bacillus subtilis temperate bacteriophage phi 105. *Journal of virology* 1985;55(2):513-515.
Feiss, M., Widner, W., Miller, G., Johnson, G. and Christiansen, S. Structure of the bacteriophage lambda cohesive end site: location of the sites of terminase binding (cosB) and nicking (cosN). *Gene* 1983;24(2-3):207-218.
Ford, M.E., Sarkis, G.J., Belanger, A.E., Hendrix, R.W. and Hatfull, G.F. Genome structure of mycobacteriophage D29: implications for phage evolution. *Journal of molecular biology* 1998;279(1):143-164.
Fouet, A. and Mock, M. Differential influence of the two Bacillus anthracis plasmids on regulation of virulence gene expression. *Infection and immunity* 1996;64(12):4928-4932.
Fu, X.F., Walter, M.H., Paredes, A., Morais, M.C. and Liu, J. The mechanism of DNA ejection in the Bacillus anthracis spore-binding phage 8a revealed by cryo-electron tomography. *Virology* 2011;421(2):141-148.
Fujisawa, H. and Morita, M. Phage DNA packaging. *Genes to cells : devoted to molecular & cellular mechanisms* 1997;2(9):537-545.
George, M. and Bukhari, A.I. Heterogeneous host DNA attached to the left end of mature bacteriophage Mu DNA. *Nature* 1981;292(5819):175-176.
Gill, J.J., Berry, J.D., Russell, W.K., Lessor, L., Escobar-Garcia, D.A., Hernandez, D., Kane, A., Keene, J., Maddox, M., Martin, R., Mohan, S., Thorn, A.M., Russell, D.H. and Young, R. The Caulobacter crescentus phage phiCbK: genomics of a canonical phage. *BMC genomics* 2012;13:542.
Gillis, A. and Mahillon, J. Phages Preying on Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis: Past, Present and Future. *Viruses-Basel* 2014;6(7):2623-2672.
Granum, P.E. Bacillus cereus and its toxins. *Society for Applied Bacteriology symposium series* 1994;23:61S-66S.
Groenen, M.A. and van de Putte, P. Mapping of a site for packaging of bacteriophage Mu DNA. *Virology* 1985;144(2):520-522.
Grokhovsky, S.L., Il'icheva, I.A., Nechipurenko, D.Y., Golovkin, M.V., Panchenko, L.A., Polozov, R.V. and Nechipurenko, Y.D. Sequence-specific ultrasonic cleavage of DNA. *Biophysical journal* 2011;100(1):117-125.
Grose, J.H., Belnap, D.M., Jensen, J.D., Mathis, A.D., Prince, J.T., Merrill, B.D., Burnett, S.H. and Breakwell, D.P. The genomes, proteomes, and structures of three novel phages that infect the Bacillus cereus group and carry putative virulence factors. *Journal of virology* 2014;88(20):11846-11860.
Grose, J.H., Jensen, J.D., Merrill, B.D., Fisher, J.N., Burnett, S.H. and Breakwell, D.P. Genome Sequences of Three Novel Bacillus cereus Bacteriophages. *Genome announcements* 2014;2(1).

Guo, P., Peterson, C. and Anderson, D. Prohead and DNA-gp3-dependent ATPase activity of the DNA packaging protein gp16 of bacteriophage phi 29. *Journal of molecular biology* 1987;197(2):229-236.
Gutierrez, J., Garcia, J.A., Blanco, L. and Salas, M. Cloning and template activity of the origins of replication of phage phi 29 DNA. *Gene* 1986;43(1-2):1-11.
Gutierrez, J., Vinos, J., Prieto, I., Mendez, E., Hermoso, J.M. and Salas, M. Signals in the phi 29 DNA-terminal protein template for the initiation of phage phi 29 DNA replication. *Virology* 1986;155(2):474-483.
Hanlon, G.W. Bacteriophages: An appraisal of their role in the treatment of bacterial infections. *Int J Antimicrob Ag* 2007;30(2):118-128.
Hashimoto, C. and Fujisawa, H. DNA sequences necessary for packaging bacteriophage T3 DNA. *Virology* 1992;187(2):788-795.
Hatfull, G.F., Science Education Alliance Phage Hunters Advancing, G., Evolutionary Science, P., KwaZulu-Natal Research Institute for, T., Course, H.I.V.M.G., University of California-Los Angeles Research Immersion Laboratory in, V., Phage Hunters Integrating, R. and Education, P. Complete genome sequences of 63 mycobacteriophages. *Genome announcements* 2013;1(6).
Hendrix, R.W., Lawrence, J.G., Hatfull, G.F. and Casjens, S. The origins and ongoing evolution of viruses. *Trends in microbiology* 2000;8(11):504-508.
Hermoso, J.M., Mendez, E., Soriano, F. and Salas, M. Location of the serine residue involved in the linkage between the terminal protein and the DNA of phage phi 29. *Nucleic acids research* 1985;13(21):7715-7728.
Hwang, Y., Hang, J.Q., Neagle, J., Duffy, C. and Feiss, M. Endonuclease and helicase activities of bacteriophage lambda terminase: changing nearby residue 515 restores activity to the gpA K497D mutant enzyme. *Virology* 2000;277(1):204-214.
Hyman, P., Abedon, S.T. and C.A.B. International. Bacteriophages in health and disease. Wallingford, Oxfordshire: CABI; 2012.
Ibrahim, M.A., Griko, N., Junker, M. and Bulla, L.A. Bacillus thuringiensis: a genomics and proteomics perspective. *Bioengineered bugs* 2010;1(1):31-50.
Illana, B., Blanco, L. and Salas, M. Functional characterization of the genes coding for the terminal protein and DNA polymerase from bacteriophage GA-1. Evidence for a sliding-back mechanism during protein-primed GA-1 DNA replication. *Journal of molecular biology* 1996;264(3):453-464.
Inciarte, M.R., Lazaro, J.M., Salas, M. and Vinuela, E. Physical map of bacteriophage phi29 DNA. *Virology* 1976;74(2):314-323.
International Committee on Taxonomy of Viruses and King, A.M.Q. Virus taxonomy : classification and nomenclature of viruses : ninth report of the International Committee on Taxonomy of Viruses. London ; Waltham, MA: Academic Press; 2012.
Ito, J. Bacteriophage phi29 terminal protein: its association with the 5' termini of the phi29 genome. *Journal of virology* 1978;28(3):895-904.
Ito, J., Kawamura, F. and Yanofsky, S. Analysis of phi 29 and phi 15 genomes by bacterial restriction endonucleases, EcoR1 and Hpal. *Virology* 1976;70(1):37-51.
Jiang, X., Jiang, H., Li, C., Wang, S., Mi, Z., An, X., Chen, J. and Tong, Y. Sequence characteristics of T4-like bacteriophage IME08 benome termini revealed by high throughput sequencing. *Virology journal* 2011;8:194.

Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F. and Hendrix, R.W. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *Journal of molecular biology* 2000;299(1):27-51.
Kaiser, D., Syvanen, M. and Masuda, T. DNA packaging steps in bacteriophage lambda head assembly. *Journal of molecular biology* 1975;91(2):175-186.
Kazakov, A.E., Vassieva, O., Gelfand, M.S., Osterman, A. and Overbeek, R. Bioinformatics classification and functional analysis of PhoH homologs. *In silico biology* 2003;3(1-2):3-15.
King, A.J. and Vandervliet, P.C. A Precursor Terminal Protein Trinucleotide Intermediate during Initiation of Adenovirus DNA-Replication - Regeneration of Molecular Ends in-Vitro by a Jumping Back Mechanism. *Embo Journal* 1994;13(23):5786-5792.
Kobiler, O., Rokney, A. and Oppenheim, A.B. Phage lambda CIII: a protease inhibitor regulating the lysis-lysogeny decision. *PloS one* 2007;2(4):e363.
Koehler, T.M. Bacillus anthracis physiology and genetics. *Molecular aspects of medicine* 2009;30(6):386-396.
Kondabagil, K.R., Zhang, Z. and Rao, V.B. The DNA translocating ATPase of bacteriophage T4 packaging motor. *Journal of molecular biology* 2006;363(4):786-799.
Krumsiek, J., Arnold, R. and Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 2007;23(8):1026-1028.
Langmead, B. and Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* 2012;9(4):357-359.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23(21):2947-2948.
Laslett, D. and Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research* 2004;32(1):11-16.
Lee, W.J., Billington, C., Hudson, J.A. and Heinemann, J.A. Isolation and characterization of phages infecting Bacillus cereus. *Letters in applied microbiology* 2011;52(5):456-464.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.
Li, S., Fan, H., An, X., Fan, H., Jiang, H., Chen, Y. and Tong, Y. Scrutinizing virus genome termini by high-throughput sequencing. *PloS one* 2014;9(1):e85806.
Ljungquist, E. and Bukhari, A.I. State of prophage Mu DNA upon induction. *Proceedings of the National Academy of Sciences of the United States of America* 1977;74(8):3143-3147.
Luftig, R.B., Wood, W.B. and Okinaka, R. Bacteriophage T4 head morphogenesis. On the nature of gene 49-defective heads and their role as intermediates. *Journal of molecular biology* 1971;57(3):555-573.
Lwoff, A. Lysogeny. *Bacteriological reviews* 1953;17(4):269-337.
Maki, Y., Yoshida, H. and Wada, A. Two proteins, YfiA and YhbH, associated with resting ribosomes in stationary phase Escherichia coli. *Genes to Cells* 2000;5(12):965-974.

Martin, A.C., Blanco, L., Garcia, P., Salas, M. and Mendez, J. In vitro protein-primed initiation of pneumococcal phage Cp-1 DNA replication occurs at the third 3' nucleotide of the linear template: a stepwise sliding-back mechanism. *Journal of molecular biology* 1996;260(3):369-377.
Meijer, W.J., Horcajadas, J.A. and Salas, M. Phi29 family of phages. *Microbiology and molecular biology reviews : MMBR* 2001;65(2):261-287 ; second page, table of contents.
Murray, K. and Murray, N.E. Terminal nucleotide sequences of DNA from temperate coliphages. *Nature: New biology* 1973;243(126):134-139.
Nafissi, N. and Slavcev, R. Bacteriophage recombination systems and biotechnical applications. *Applied microbiology and biotechnology* 2014;98(7):2841-2851.
Ohmori, H., Haynes, L.L. and Rothman-Denes, L.B. Structure of the ends of the coliphage N4 genome. *Journal of molecular biology* 1988;202(1):1-10.
Pajunen, M.I., Elizondo, M.R., Skurnik, M., Kieleczawa, J. and Molineux, I.J. Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3. *Journal of molecular biology* 2002;319(5):1115-1132.
Pecenkova, T. and Paces, V. Molecular phylogeny of phi29-like phages and their evolutionary relatedness to other protein-primed replicating phages and other phages hosted by gram-positive bacteria. *Journal of molecular evolution* 1999;48(2):197-208.
Pope, W.H., Anders, K.R., Baird, M., Bowman, C.A., Boyle, M.M., Broussard, G.W., Chow, T., Clase, K.L., Cooper, S., Cornely, K.A., DeJong, R.J., Delesalle, V.A., Deng, L., Dunbar, D., Edgington, N.P., Ferreira, C.M., Weston Hafer, K., Hartzog, G.A., Hatherill, J.R., Hughes, L.E., Ipapo, K., Krukonis, G.P., Meier, C.G., Monti, D.L., Olm, M.R., Page, S.T., Peebles, C.L., Rinehart, C.A., Rubin, M.R., Russell, D.A., Sanders, E.R., Schoer, M., Shaffer, C.D., Wherley, J., Vazquez, E., Yuan, H., Zhang, D., Cresawn, S.G., Jacobs-Sera, D., Hendrix, R.W. and Hatfull, G.F. Cluster M mycobacteriophages Bongo, PegLeg, and Rey with unusually large repertoires of tRNA isotypes. *Journal of virology* 2014;88(5):2461-2480.
Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.V., Oparina, N.Y., Polozov, R.V., Nechipurenko, Y.D. and Grokhovsky, S.L. Non-random DNA fragmentation in next-generation sequencing. *Scientific reports* 2014;4:4532.
Rao, V.B. and Feiss, M. The bacteriophage DNA packaging motor. *Annual review of genetics* 2008;42:647-681.
Ratcliff, S.W., Luh, J., Ganesan, A.T., Behrens, B., Thompson, R., Montenegro, M.A., Morelli, G. and Trautner, T.A. The genome of Bacillus subtilis phage SPP1: the arrangement of restriction endonuclease generated fragments. *Molecular & general genetics : MGG* 1979;168(2):165-172.
Redondo, R.A., Kupczok, A., Stift, G. and Bollback, J.P. Complete Genome Sequence of the Novel Phage MG-B1 Infecting Bacillus weihenstephanensis. *Genome announcements* 2013;1(3).
Rhoades, M., MacHattie, L.A. and Thomas, C.A., Jr. The P22 bacteriophage DNA molecule. I. The mature form. *Journal of molecular biology* 1968;37(1):21-40.
Salas, M. Protein-priming of DNA replication. *Annual review of biochemistry* 1991;60:39-71.
Salas, M., Mellado, R.P. and Vinuela, E. Characterization of a protein covalently linked to the 5' termini of the DNA of Bacillus subtilis phage phi29. *Journal of molecular biology* 1978;119(2):269-291.

Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. Microbial gene identification using interpolated Markov models. *Nucleic acids research* 1998;26(2):544-548.
Sambrook, J. and Russell, D.W. Molecular cloning - A laboratory manual, 3rd edition. *Science* 2001;292(5516):446-446.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W.Z., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D. and Higgins, D.G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;7.
Skalka, A.M. DNA replication--bacteriophage lambda. *Current topics in microbiology and immunology* 1977;78:201-237.
Smits, C., Chechik, M., Kovalevskiy, O.V., Shevtsov, M.B., Foster, A.W., Alonso, J.C. and Antson, A.A. Structural basis for the nuclease activity of a bacteriophage large terminase. *EMBO reports* 2009;10(6):592-598.
Stewart, C.R., Casjens, S.R., Cresawn, S.G., Houtz, J.M., Smith, A.L., Ford, M.E., Peebles, C.L., Hatfull, G.F., Hendrix, R.W., Huang, W.M. and Pedulla, M.L. The genome of Bacillus subtilis bacteriophage SPO1. *Journal of molecular biology* 2009;388(1):48-70.
Streisinger, G., Edgar, R.S. and Denhardt, G.H. Chromosome Structure in Phage T4. I. Circularity of the Linkage Map. *Proceedings of the National Academy of Sciences of the United States of America* 1964;51:775-779.
Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell* 2008;135(7):1251-1262.
Takahashi, S. The starting point and direction of rolling-circle replicative intermediates of coliphage lambda DNA. *Molecular & general genetics : MGG* 1975;142(2):137-153.
Thorne, C.B. Transducing bacteriophage for Bacillus cereus. *Journal of virology* 1968;2(7):657-662.
Tormo-Mas, M.A., Donderis, J., Garcia-Caballer, M., Alt, A., Mir-Sanchis, I., Marina, A. and Penades, J.R. Phage dUTPases control transfer of virulence genes by a proto-oncogenic G protein-like mechanism. *Mol Cell* 2013;49(5):947-958.
Tourasse, N.J., Helgason, E., Okstad, O.A., Hegna, I.K. and Kolsto, A.B. The Bacillus cereus group: novel aspects of population structure and genome dynamics. *Journal of applied microbiology* 2006;101(3):579-593.
Ueta, M., Yoshida, H., Wada, C., Baba, T., Mori, H. and Wada, A. Ribosome binding proteins YhbH and YfiA have opposite functions during 100S formation in the stationary phase of Escherichia coli. *Genes to cells : devoted to molecular & cellular mechanisms* 2005;10(12):1103-1112.
Vilain, S., Luo, Y., Hildreth, M.B. and Brozel, V.S. Analysis of the life cycle of the soil saprophyte Bacillus cereus in liquid soil extract and in soil. *Applied and environmental microbiology* 2006;72(7):4970-4977.
Walter, M.H. and Baker, D.D. Three Bacillus anthracis bacteriophages from topsoil. *Current microbiology* 2003;47(1):55-58.
Wang, J., Jiang, Y., Vincent, M., Sun, Y., Yu, H., Wang, J., Bao, Q., Kong, H. and Hu, S. Complete genome sequence of bacteriophage T5. *Virology* 2005;332(1):45-65.
Wommack, K.E. and Colwell, R.R. Virioplankton: viruses in aquatic ecosystems. *Microbiology and molecular biology reviews : MMBR* 2000;64(1):69-114.

Wu, H., Sampson, L., Parr, R. and Casjens, S. The DNA site utilized by bacteriophage P22 for initiation of DNA packaging. *Molecular microbiology* 2002;45(6):1631-1646.
Xu, J., Hendrix, R.W. and Duda, R.L. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol Cell* 2004;16(1):11-21.
Yee, L.M., Matsumoto, T., Yano, K., Matsuoka, S., Sadaie, Y., Yoshikawa, H. and Asai, K. The genome of Bacillus subtilis phage SP10: a comparative analysis with phage SPO1. *Bioscience, biotechnology, and biochemistry* 2011;75(5):944-952.
Zerbino, D.R. and Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 2008;18(5):821-829.
Zhang, X. and Studier, F.W. Multiple roles of T7 RNA polymerase and T7 lysozyme during bacteriophage T7 infection. *Journal of molecular biology* 2004;340(4):707-730.