A STUDY TO INVESTIGATE THE CORRELATION OF ABUNDANCE BETWEEN THE SPACERS AND THE PROPHAGES IN BACTERIAL GENOME

By
Taslimun Jannat
18376005

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of Master of Science in Biotechnology

Department of Mathematics and Natural Sciences
BRAC University
September 2019

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material, which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.
5. I would like to request the embargo of my thesis for 12M from the submission date due to publication issue in international journal.

Taslimun Jannat

Taslimun Jannat

ID: 18376005

Approval

The thesis "A study to investigate the correlation of abundance between the spacers and the prophages in bacterial genome" submitted by **Taslimun Jannat** (18376005) of Summer, 2018 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Biotechnology on 14th November 2019.

Examining Committee:

Supervisor:
(Member)

Dr. M. Mahboob Hossain
Professor, MNS Department
BRAC University

Program Coordinator:
(Member)

Iftekhar Bin Naser
Assistant Professor, MNS Department
BRAC University

External Expert Examiner:
(Member)

Departmental Head:
(Chair)

A F M Yusuf Haider
Professor, MNS Department
BRAC University

Ethics Statement

The idea of this thesis was unique and mine therefore there was no conflict of interests.

Abstract

Bacteriophages, the most abundant and diverse entities in the biosphere which are composed of a nucleic acid molecule that is surrounded by a protein structure. These viruses can infect or kill bacteria through lytic and lysogenic cycle, however, the bacteria have evolved their own defense system to protect themselves from the integration of foreign mobile genetic elements through CRISPR-Cas that utilizes short foreign DNA sequences, known as spacers. Since it is their defense mechanism, therefore the correlation of abundance between the spacers and the prophages are expected to be inversely proportional. The project was designed to find the relationship between the total number of spacers and the prophages through computational approach. The finding of this thesis conflict with expected outcome since regression line do not demonstrate significant changes when analyzed. Moreover, almost 33% bacterial genomes were found to carry prophages even in the presence of spacers those align with them and the number of spacers increases with the increase of prophage length. In spite of being present in bacterial genome along with spacers, they do not share any core genes indicate that not any specific types of prophages have these special capacity to infect bacteria in presence of spacer rather they are of different species.

Keywords: Bacteria, Bacteriophage, Prophage, Spacers, Genome,

Dedication

Dedicated to my family and friends for their love, motivations and continuous support…

Acknowledgement

First of all, I would like to express my deepest gratefulness to Almighty Allah for everything he has blessed me with in my whole life and for giving me patience and strength to make this project a success.

I would like to thank Professor A F M Yusuf Haider, Ph.D., Professor and Chairperson of Department of Mathematics and Natural Sciences, BRAC University for allowing me to continue my work at the BRAC University microbiology laboratory.

I am gratified to my supervisor Professor Dr. M. Mahboob Hossain, Department of Mathematics and Natural Sciences, BRAC University for believing in me and giving me support, constant encouragement and guidance throughout this project whenever I needed.

I would like to especially thank Assistant Professor Iftekhar Bin Naser, Department of Mathematics and Natural Sciences, BRAC University for the inspiration and guidance she provided me as a student.

I express my sincere gratitude to Tokee Md. Tarek, Lecturer, Department of Mathematics and Natural Sciences, BRAC University for his constant inspiration, encouragement, care and guidance in my bioinformatics work.

I would like to extend my deepest gratitude and regards to Tushar Ahmed Shishir, Teaching Assistant, Department of Mathematics and Natural Sciences, BRAC University for his constant guidance, support and wise words throughout the work during this project.

Table of Contents

Declaration ..... **ii**

Approval ..... **iii**

Ethics Statement ..... **iv**

Abstract ..... **v**

Dedication ..... **vi**

Acknowledgement ..... **vii**

Table of Contents ..... **viii-ix**

List of Tables ..... **x**

List of Figures ..... **xi**

CHAPTER 1 ..... **1**

INTRODUCTION **1**

CHAPTER 2 ..... **1**

BACKGROUND STUDY **4**

2. Bacteriophage 5

2.1.1. Classification of Bacteriophages 5

2.1.2. History of Bacteriophages 7

2.1.3. Life cycle of Bacteriophages: 8

2.1.3.1. Lytic cycle: 8

2.1.3.2. Lysogenic cycle 10

2.2.1. Prophage biology 11

2.2.2. Abundance of Prophage 12
2.3. CRISPR 13

2.3.1. Locus structure 15

2.3.2. Cas genes and CRISPR subtypes: 15

CHAPTER 3 17

MATERIALS AND METHODS 17

3.1 Bacterial genome data download from NCBI.....18

3.2 Determining Spacers number using MinCED 19

3.3 Genome Annotation with Prokka 19

3.4 Finding Prophages using PhiSpy 21

3.5 BLAST 22

CHAPTER 4 24

RESULTS 24

4.1. Number of spacers containing genome: 25

4.2. Identification of prophage: 25

4.3. Number of spacers vs. number of prophages: 26

4.4. Spacers number vs Prophage length: 27

4.5. Spacers number vs. Blast hit: 29

4.6. Aligned spacers vs. Prophage 30

CHAPTER 5 31

DISCUSSION 31

CHAPTER 6 34

REFERENCES 34

List of Tables

Table 2.1: ICTV taxonomic classification of bacteriophage infecting bacteria and archaea 7

Table 2.2: Signature genes and their putative functions for the major and minor CRISPR-Cas types ..…….15-16

Table 3.1: Tools used by Prokka …….20

Table 3.2: Description of Prokka output files……...20

List of Figures

Figure 2.1: Different families of bacteriophages 6

Figure 2.2: The stages of lytic and lysogenic life cycle of bacteriophage 10

Figure 4.1: A dot plot (in R), where x-axis is representing the total number of prophages and y-axis is representing the total number of spacers found in the 2793 genomes…….26

Figure 4.2: A dot plot (in R), where x-axis and y-axis is showing prophage length and the number of spacers found in the 2793 genomes respectively……...27

Figure 4.3: A dot plot (using R) between the number of spacers found in 2793 bacterial genome, plotted on y-axis and the number of blast hits, plotted on x-axis, showing a positive relation between the number of spacers and number of blast hits as the regression is moderately skewed left…….……...29

Figure 4.4: A dot plot between the number of aligned prophage (plotted on x-axis) and the number of aligned prophage (plotted on y-axis) ……30

CHAPTER 1

INTRODUCTION

1. Introduction

Bacteriophages, the most prevailing and diverse entities in the environment present in almost every bacterium which are the most abundant across the ecosystem, present approximately as a number of 10^{31} in this planet (Comeau et al., 2008). More than 6000 different bacteriophages have been discovered, classified according to their morphology, genetic content (DNA vs. RNA), specific host, living places and their life cycle under the International Committee on Taxonomy of Viruses (ICTV) (Fauquet and Pringle, 2000).

It was discovered by a group of scientists over a long period but not at the same time. In 1896, a British bacteriologist Ernest Hanbury Hankin demonstrated that the waters from the Indian rivers Ganga and Yamuna contained a biological principle that destroyed cultures of cholera-inducing bacteria and known to retain larger microorganisms such as bacteria (Hankin, 1896). Frederick Twort, a British microbiologist, noted that "pure" cultures of bacteria may be associated with a filter-passing transparent material which may entirely break down bacteria of a culture into granules and this transparent material was described by Towrt and after two years later Félix d'Herelle independently described a similar experimental finding, while studying patients suffering or recovering from bacillary dysentery (Twort, 1915). D'Herelle described his discovery as a microbe that was a "veritable" microbe of immunity and an obligate bacteriophage and also demonstrated the activity of this anti-Shiga microbe. Max Delbrück, Alfred Hershey, and Salvador Luria were awarded the Nobel Prize in Physiology or Medicine for their discoveries of the replication of viruses and their genetic structure.

Bacteriophage/phage attaches to a bacterium and inserts its genetic elements into it during the infection made by bacteriophage, then it undergoes into one of its two life cycle: lytic life cycle and lysogenic life cycle. Temperate phages such as; lambda phage can reproduce using both the lytic and the lysogenic cycle. In the lytic cycle, the infecting phage kill the host cell to produce many progeny who rule over the cellular machinery and host cells become gradually weakened by phage enzymes that lead to the release of enormous (100-200) new phage progeny into the surrounding environment whereas in the lysogenic cycle the bacteriophage's genome is not expressed rather it is integrated into the bacteria's genome to form prophages and with the replication of bacteriophage during the cellular division of bacteria, the daughter cells contain prophages, known as lysogen which can switch to the lytic cycle at any time by induction process where prophage DNA is excised from the bacterial genome and is transcribed and translated to make coat proteins for the virus and regulate lytic growth (Holmes, 1996).

Much of the diversity observed in closely related bacterial strains is a result of the incorporation of diverse prophages into the core bacterial genome. Prophages enhance bacterial fitness by encoding many proteins important in virulence and antibiotic resistance. The presence or absence of prophages play a great role on a large fraction of the variation among individuals within a bacterial species, and phages are likely to be important vehicles for horizontal transfer of genetic information between bacteria. In complex microbial ecosystems such as the gastrointestinal tract, marine environment even in mixed culture fermentations bacteriophages are frequently found to be part of the microbial community. Moreover, prophages are frequently found to be a part of the microbial community (Rohwer, 2002).

CRISPR is a family of DNA sequences which is found within the genomes of prokaryotic organisms such as bacteria and archaea (Barrangou, 2015). These sequences are derived from DNA fragments of bacteriophages that have previously infected the prokaryote and which are used to detect and destroy DNA from similar phages during subsequent infections. These are the specialized region of DNA with two specific characteristics: the presence of nucleotide repeats and spacers.

Spacer DNA or intergenic spacers (IGS) is a region of non-coding DNA between genes (Lackie, 2007). The size of the spacer DNA sequences is only few nucleotides long in bacteria and in eukaryotes, their size can be great and includes repetitive DNA which comprises the majority of the DNA of the genome. The spacers are taken from viruses that previously attacked the organism (in case of bacteria). They serve as a bank of memories, which enables bacteria to recognize the viruses and protect the cell from future attacks.

So, theoretically, when the number of spacers in a bacterial genome is large, the number of prophages will be small. In this project, attempts were made to know whether the number of spacers has any effect on the number/length of prophages of bacterial genome or not and if they have, how they are correlated: directly proportional or inversely proportional.

This project has been conducted through computational approach by using 2793 whole bacterial genome, determining the number of spacers and the prophages, annotating the bacterial genome and finding their relationship.

CHAPTER 2

BACKGROUND STUDY

2.1. Bacteriophage:

A bacteriophage is a type of virus that infects bacteria, composed of a nucleic acid molecule that is surrounded by a protein structure. It attaches itself to a susceptible bacterium and infects the host cell through forcing the cell to produce viral components instead of producing bacterial component. They are the most common and diverse entities in the biosphere and found wherever bacteria exist (Grath and Sinderen, 2007). Over the last 30 years or so, it has become clear that phages are the most abundant organism on Earth. There are estimated 10^{31} phage particles on the planet, a great number that translates into approximately a trillion phages for every grain of sand in the world (Comeau et al., 2008). Bacteriophages play significant roles in a large number of biological and environmental processes, it is estimated that phages can kill and lyse between 15% and 40% of the ocean's bacteria every day which influences the ratio of particulate to dissolve carbon, rates of phytoplankton productivity and oxygen production, perhaps even global climate and weather patterns (Danovaro et al., 2011) and they also play as significant drivers in the evolution of bacteria, especially temperate phages which are prominent agents of horizontal gene transfer. It is believed worldwide that phages mediate gene transfer events between bacteria through transduction process up to 20 million times per second (Chibani – Chennoufi, 2004).

2.1.1. Classification of Bacteriophages:

More than 6000 different bacteriophages have been discovered and described morphologically, including 6196 bacterial and 88 archaeal viruses and vast majority of these viruses are tailed while a small proportion are polyhedral, filamentous or pleomorphic (Ackermann and Prangishyili, 2012). They may be classified according to their morphology, their genetic content (DNA vs. RNA), their specific host (for instance the staphylococcal phage family, the *Pseudomonas* phage family, and so on), the place where they live (marine virus vs. other habitats), and their life cycle (Deghorain and Van, 2012). Evolving classification formats have been proposed over time and abbreviations for these viruses were proposed by Fauquet and Pringle in 2000 (Fauquet and Pringle, 2000).

In 1971, the International Committee on Taxonomy of Viruses (ICTV) classified phages into six genera corresponding to five of Bradley's basic type, namely the T4, λ, ΦX174, MS2, and fd phage groups. After that new phages were added over the time. Currently, nineteen families are recognized by the ICTV that infect archaea and bacteria among of these only two families have RNA genomes, and only five families are surrounded by an envelope and only two from

the viral family with DNA genomes have single-stranded genomes. Eight of the viral families with DNA genomes have circular genomes while nine have linear genomes and nine families infect bacteria only, nine infect archaea only, and one (*Tectiviridae*) infects both bacteria and archaea.

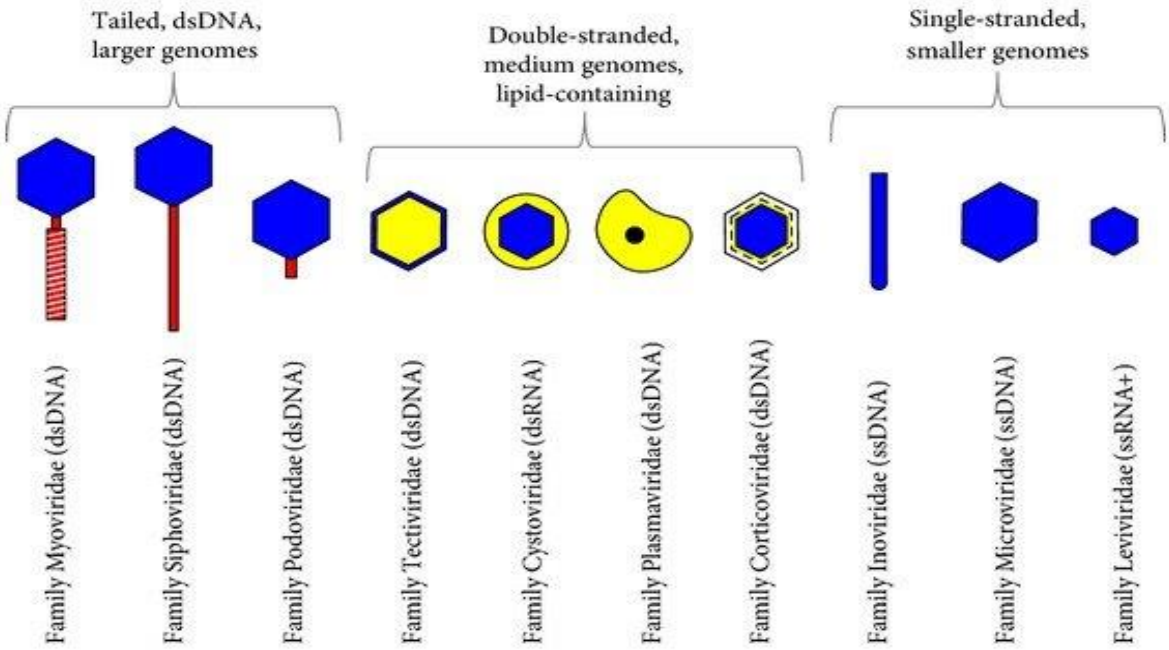

Figure 2.1: Different families of bacteriophages

Table 2.1: ICTV taxonomic classification of bacteriophage infecting bacteria and archaea

Order	Family	Morphology	Nucleic acid	Examples	Sub families	Genera
Caudovirales	Ackermannviridae		dsDNA		2	4
	Myoviridae	Nonenveloped, contractile tail	Linear dsDNA	T4 phage, Mu, PBSX, P1Puna-like, P2, I3, Bcep 1, Bcep 43, Bcep 78	6	41
	Siphoviridae	Nonenveloped, noncontractile tail (long)	Linear dsDNA	λ phage, T5 phage, phi, C2, L5, HK97, N15	11	100
	Podoviridae	Nonenveloped, noncontractile tail (short)	Linear dsDNA	T7 phage, T3 phage, Φ29, P22, P37	3	23
Ligamenvirales	Lipothrixviridae	Enveloped, rod-shaped	Linear dsDNA	Acidianus filamentous virus 1		3
	Rudiviridae	Nonenveloped, rod-shaped	Linear dsDNA	Sulfolobus islandicus rod-shaped virus 1		1
	Siphoviridae	Nonenveloped, noncontractile tail (long)	Linear dsDNA	λ phage, T5 phage, phi, C2, L5, HK97, N15	11	100
	Podoviridae	Nonenveloped, noncontractile tail (short)	Linear dsDNA	T7 phage, T3 phage, Φ29, P22, P37	3	23
Unassigned	Ampullaviridae	Enveloped, bottle-shaped	Linear dsDNA			1
	Bicaudaviridae	Nonenveloped, lemon-shaped	Circular dsDNA			1
	Clavaviridae	Nonenveloped, rod-shaped	Circular dsDNA			1
	Corticoviridae	Nonenveloped, isometric	Circular dsDNA			1
	Cystoviridae	Enveloped, spherical	Segmented dsRNA			1
	Fuselloviridae	Nonenveloped, lemon-shaped	Circular dsDNA			2

Continuation of table 2.1….

	Globuloviridae	Enveloped, isometric	Linear dsDNA			1
	Guttaviridae	Nonenveloped, ovoid	Circular dsDNA			2
	Inoviridae	Nonenveloped, filamentous	Circular ssDNA	M13		7
	Leviviridae	Nonenveloped, isometric	Linear ssRNA	MS2, Qβ		2
	Microviridae	Nonenveloped, isometric	Circular ssDNA	ΦX174	2	6
	Plasmaviridae	Enveloped, pleomorphic	Circular dsDNA			1
Tectiviridae	Nonenveloped, isometric	Linear dsDNA			2	

2.1.2 History of Bacteriophages:

In 1896, a British bacteriologist Ernest Hanbury Hankin, working as the Chemical Examiner and Bacteriologist to the Government of the United Provinces and of the Central Provinces of India, demonstrated that the waters from the Indian rivers Ganga and Yamuna contained a biological principle that destroyed cultures of cholera-inducing bacteria. This substance could pass through milli - pore filters, known to be able to retain larger microorganisms such as bacteria. He published his work in the Annals of the Pasteur Institute (Hankin, 1896). In 1915, while he was studying the growth of vaccinia virus on cell-free agar media, Frederick Twort, a British microbiologist, noted that "pure" cultures of bacteria may be associated with a filter- passing transparent material which may entirely break down bacteria of a culture into granules (Twort, 1915). This transparent material, which was found to be unable to grow in the absence of bacteria, was described by Twort as a ferment secreted by the microorganism for some purpose not clear at that time.

Two years after this report, Félix d'Herelle independently described a similar experimental finding, while studying patients suffering or recovering from bacillary dysentery. He isolated from stools of recovering shigellosis patients a so-called "anti-Shiga microbe" by filtering stools that were incubated for 18 h. This active filtrate, when added either to a culture or an emulsion of the Shiga bacilli, was able to cause arrest of the culture, death and finally lysis of the bacilli (Ttwort, 1917). D'Herelle described his discovery as a microbe that was a "veritable" microbe of immunity and an obligate bacteriophage. He also demonstrated the activity of this anti-Shiga microbe by inoculating laboratory animals as a treatment for shigellosis, seeming to

confirm the clinical significance of his finding by satisfying at least some of Koch's postulates. He even introduced treatment with intravenous phage for invasive infections, and he summarized all these findings and observations in 1931 (Twort, 1931). More than a half a century later, in 1969, Max Delbrück, Alfred Hershey, and Salvador Luria were awarded the Nobel Prize in Physiology or Medicine for their discoveries of the replication of viruses and their genetic structure.

2.1.2. Life cycle of Bacteriophages:

Bacteriophages use two type of cycles to infect their bacterial hosts: the lytic cycle and lysogenic cycle.

2.1.2.1. Lytic cycle:

The lytic life cycle is where phages infect and rapidly kill their infected host cells, thereby shaping bacterial population dynamics and occasionally assisting in their long term evolution via generalized transduction (Abedon, 2008; Weinbauer et al., 2004; Wilhelm et al., 1999).

Stages of lytic cycle:

1. Attachment: Proteins which present in the tail of the phage bind to a certain receptor on the surface of the bacterial cell.
2. Entry: The phage injects its double-stranded DNA genome into the cytoplasm of the bacterium.
3. DNA copying and protein synthesis: Then the DNA of phage is copied and phage genes are expressed to make proteins, such as capsid proteins.
4. Assembly of new phage: From the capsid proteins, capsids are assembled and are stuffed with DNA to make lots of new phage particles.
5. Lysis: At the end of the lytic cycle, the phage expresses genes for proteins that poke holes in the plasma membrane and cell wall. The holes let water flow in and make the cell expand and burst like an overfilled water balloon. This step releases hundreds of new phages, which can find and infect other host cells nearby.

By this way lytic infection can let the phages spread like wildfire through a bacterial population.

The lysogenic life cycle in contrast, is where phages instead of directly killing their hosts, integrate into their host genome, or exist as plasmids within their host cell (Waldor et al., 2005). This cycle can be stable for thousands of generations and the bacteriophage may alter the

phenotype of the bacterium by expressing genes that are generally not expressed in the general course of infection in a process which is known as lysogenic conversion and a popular example of this is the gene associated with *Vibrio cholerae* that encodes the toxin which cause cholera symptoms (Los et al., 2010).

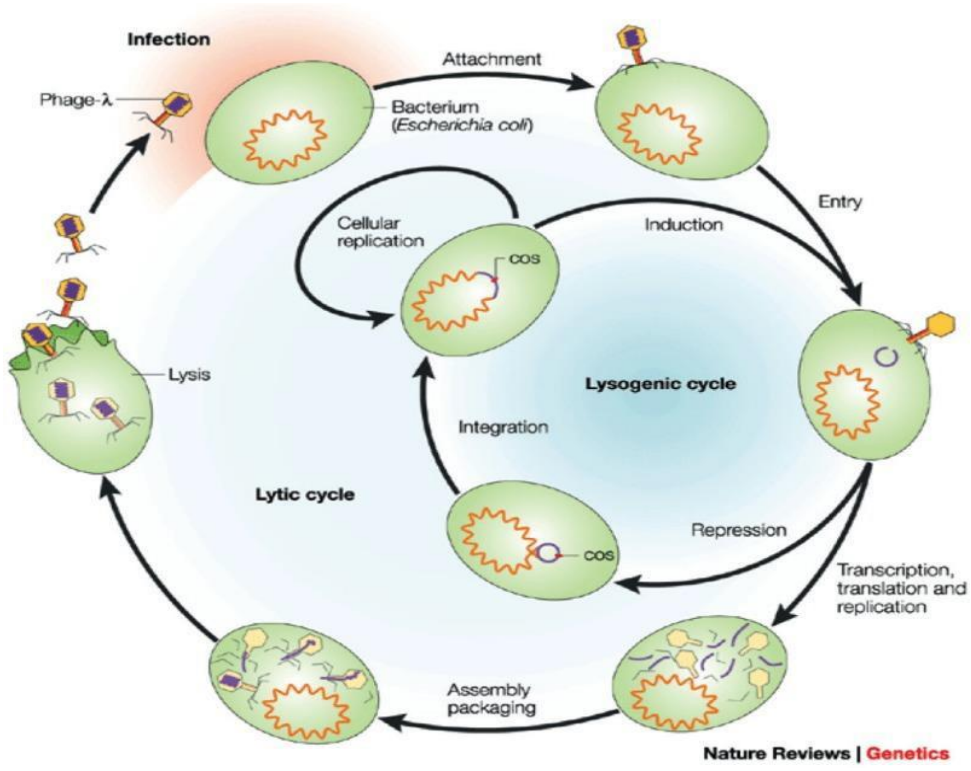

Figure 2.2: The stages of lytic and lysogenic life cycle of bacteriophage

2.1.2.2. Lysogenic cycle: Some phages can only use the lytic cycle but some other phages such as lambda phage can switch between the two cycles. Most phages are either lytic (sometimes called virulent phages) or capable of switching between these lytic and lysogenic cycles which are sometimes called temperate phages. However, there is an exception to almost most of the rules, Filamentous phages are secreted from the cell in a process that does not lyse or kill the cell even after the cell is properly activated and producing new phages particles (Holmes et al., 1996; Rakonjac, 2012)

In the lysogenic cycle, the first two steps (attachment and DNA injection) occur just as the same what happens in the lytic cycle. However, once the phage DNA is inside the cell, it is not immediately copied or expressed to make proteins. Instead, it recombines with a particular region of the bacterial chromosome. This causes the phage DNA to be integrated into the

chromosome. Not all phages integrate their DNA into the genome of their host during the lysogenic cycle. Some instead keep their genome in the cell as a separate, circular piece of DNA (Holmes, 1996). This is still considered a lysogenic cycle because the phage does not drive production of new virus particles or kill the cell. Instead, it remains "silent" and is passively copied along with the cell's DNA.

However, the integrated phage DNA which is called a prophage is not active and its genes are not expressed and it does not drive production of new phages. The prophage genome is then replicated passively along with the host genome as the host cell divides for as long as it remains there and does not form the proteins required to produce progeny. As the phage genome is generally comparatively small, the bacterial hosts are normally relatively unharmed by this process.

2.2.1. Prophage biology:

Phages are ubiquitous and can be found in any environment where their bacterial hosts are present. It has been suggested that there exist ~100 million phage species (Rohwer, 2003) and only a small fraction of phages have so far been characterized. As such, phages likely play a major role in defining the dynamics of microbial community structure and function. In fact, much of the diversity observed in closely related bacterial strains is a result of the incorporation of diverse prophages into the core bacterial genome (Rohwer et al., 2002). Prophages enhance bacterial fitness by encoding many proteins important in virulence and antibiotic resistance. The presence or absence of prophages play a great role on a large fraction of the variation among individuals within a bacterial species, and phages are likely to be important vehicles for horizontal transfer of genetic information between bacteria (Ojaimi et al., 2003; Banks et al., 2002; Casjens et al., 2003). In order to understand the information in bacterial whole- genome nucleotide sequences, it is essential to recognize and understand prophages when they are present. Properly functional prophages can induce a round of lytic growth to start; however not all prophage-like existence in bacterial genomes encode functional bacteriophages. There are four additional types of prophage related existence have been characterized:

A) Defective prophages: these are also called cryptic prophages which are in a state of mutational decay. Defective prophages are unable to program the full phage replication cycle (Campbell, 1994). Several defective prophages in *E. coli* K- 12, Rac e14, DLP12 and QIN and in *Bacillus subtilis*, 186 and SKIN were discovered before genomic sequencing became

possible and have been studied in some detail. Each of these harboers some functional genes. For example, Rac encodes the RecE homologous recombination system, QIN harbours intact cell lysis genes and PBSX encodes the synthesis of a virion- like particle (Kaiser and Murray, 1979; Greener and Hill, 1980; Lindsey *et al*., 1989; Espio *et al*., 1983; Krogh *et al*., 1996; Takemaru *et al*., 1995; Mizuno *et al*., 1996). For example, Rac encodes the RecE homologous recombination system, QIN harbours intact cell lysis genes and PBSX encodes the synthesis of a virion- like particle (Kaiser and Murray, 1979; Espion *et al.*, 1983; Okamato *et al.*, 1968).

B) Satellite phages are those that do not carry their own virion structural protein genes, and have chromosomes that have been evolutionarily designed to be encapsulated by the virion proteins of other specific phages. The best example of such a parasitic relationship is that between satellite phage P4 and fully functional phage P2 (Ruzin *et al.*, 2001).

C) Some bacteria produce bacteriocins (proteins that kill other bacteria) that resemble phage tails (e.g. Gratia, 1989; Thaler *et al*., 1995; Zink *et al*., 1995; Nguyen *et al*., 1999; Nakayama *et al*., 2000). Two of these that have been characterized, the type F and R bacteriocins of *Pseudomonas aeruginosa* PAO1, are similar to phage λ tails and phage P2 tails respectively (Nakayama *et al*., 2000). The gene clusters encoding them have nearly complete sets of λ and P2 tail gene homologues in nearly the same order as they are found in those phages.

D) Gene transfer agents (GTAs) are encoded by some bacterial genomes (Yen *et al*., 1979; Starich *et al*., 1985; Rapp and Wall, 1987; Humphrey *et al*., 1997). GTAs are tailed phage-like particles which encloses random fragments of the bacterial genome. These particles cannot propagate as viruses because most of them do not carry the genes that encode the GTA and, those that do contain a DNA fragment that is too short to include the full set of GTA genes. These particles can deliver their DNA into the same species but in different bacterium of that species where the DNA replace the resident cognate chromosomal region through homologous recombination. The best characterized GTA is encoded by a cluster of genes on the *Rhodobacter capsulatus* chromosome (Lang *et al*., 2000; Lang and Beatty, 2001).

2.2.2. Abundance of Prophage: The first approaches that helped to realize the abundance of phages were based on epifluorescent microscopy which follows DNA staining that suggested that in sea water there are around 10 phages in existence for each bacterial/archaeal cell (Suttle, 2005; Fuhrman, 1999; Fuhrman, 1995). Similar figures have been shown for freshwater environments, but for other more complex environments the situation is less clear and virus numbers may either be higher or lower than that of their bacterial/archaeal hosts. The

abundance and distribution of phages is generally based on their host organism.

Therefor to make sense of viral abundance, it must be established where the host exists. Most of the Earth's Bacteria and Archaea are found in the open ocean, the soil and in ocean sediments, and terrestrial sub-surfaces where there are an estimated 1.2×10^{29}, 2.6×10^{29}, 3.5×10^{30} and $0.25– 2.5 \times 10^{30}$ cells respectively (Whitman, 1998) in humans the majority of prokaryotes are found in the colon, so multiplying the total human population of 6.8×10^{9} by the number of prokaryotes per gram of human colonic matter (3.2×10^{11}), by the average amount of colonic material per human of 220 grams, gives an estimated total of 4.8×10^{23} prokaryotes (Fuhrman, 1995).

Two researchers examined 173 *Salmonella enterica* (serovar Typhimurium) isolates and found that 136 released functional phages (Schicklmaier *et al*., 1998; Schmieger and Schicklmaier, 1999) and the LT2 isolate of *S. enterica* that is commonly used in laboratory studies carries four intact, fully functional prophages (Yamamoto, 1967; 1969; Figueroa- Bossi and Bossi, 1999; McClelland *et al*., 2001). Other studies have ran out about the presence of particular prophage features in multiple isolates of the same bacterial species. In the *E. coli* chromosome, the attachment site of the λ- like (lambdoid) phage 21 is occupied by phage- like sequences in 28 of 77 strains examined (Wang *et al*., 1997), the lambdoid phage Atlas attachment site is occupied in 23 of 72 strains examined (Milkman and Bridges, 1990; Sandt and Hill, 2000), and four of 33 strains examined have something (probably λ- like in two cases) inserted at the phage λ attachment site (Kuhn and Campbell, 2001).

Hybridization of DNA from various bacterial strains with authentic phage or prophage DNA probes has shown that related prophages are often present in a substantial fraction of other isolates of the same species, for example: Gram- negative enterobacteria (Anilions *et al*., 1980; Lindsey *et al*., 1989; Faubladier and Bouche, 1994), *Wolbachia* (Masui *et al*., 2000) and *Haemophilus* (Chang *et al*., 2000), *Spirochaete Borrelias* (Casjens *et al*., 1997) etc. Finally, a substantial fraction of searches for strain- specific bacterial sequences for use in the typing of related bacterial isolates have found prophage sequences, such as: *enterobacteria, campylobacter, Neisseria* and *Lactobacillus*. Certainly, prophages are common in diverse bacterial species.

2.3. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and Spacers:

CRISPR is a family of DNA sequences which is found within the genomes of prokaryotic organisms such as bacteria and archaea (Barrangou, 2015). These sequences are derived

from DNA fragments of bacteriophages that have previously infected the prokaryote and which are used to detect and destroy DNA from similar phages during subsequent infections. These are the specialized region of DNA with two specific characteristics: the presence of nucleotide repeats and spacers. Repeated sequences of nucleotides which are called the building blocks of DNA, distributed throughout a CRISPR region.

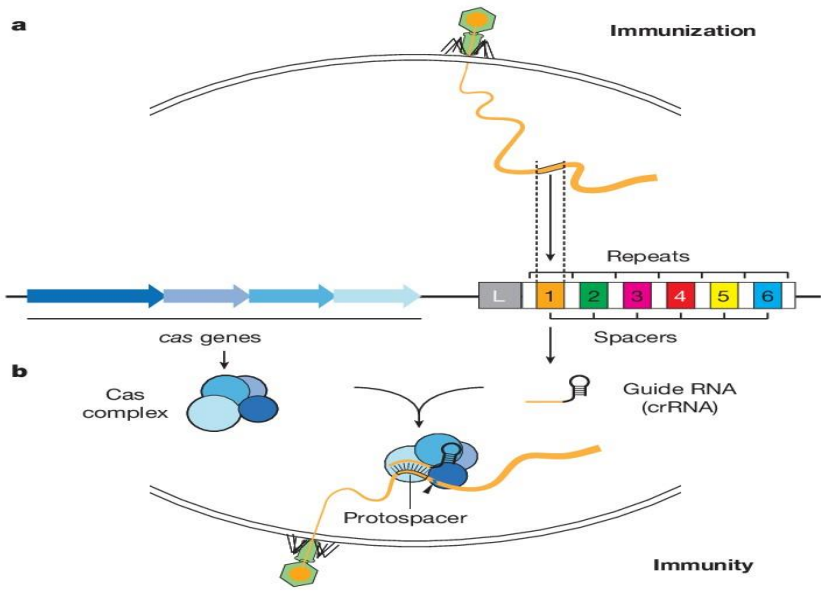

Figure 3.1: Stages of CRISPR-Cas immunity

Spacer DNA or intergenic spacers (IGS) is a region of non-coding DNA between genes. The size of the spacer DNA sequences is only few nucleotides long in bacteria and in eukaryotes, their size can be large and includes repetitive DNA which comprises the majority of the DNA of the genome (Lackie, 2007).

In ribosomal DNA, the spacers remain within and between gene clusters which are called internal transcribed spacer (ITS) and external transcribed spacers (ETS) respectively whereas in animals, the mitochondrial DNA genes have very short spacers.

The spacers are taken from viruses that previously attacked the organism (in case of bacteria). They serve as a bank of memories, which enables bacteria to recognize the viruses and protect the cell from future attacks.

This was first demonstrated experimentally by Rodolphe Barrangou and a team of researchers at Danisco which was a food ingredients company. In 2007 a research paper published in the journal Science, where the researchers used *Streptococcus thermophilus* bacteria, which are commonly found in yogurt and other dairy cultures, what they have chosen for test model. They observed that after a virus attack, new spacers were incorporated into the CRISPR region.

Moreover, the DNA sequence of these spacers was identical to parts of the virus genome. They also manipulated the spacers by taking them out or putting in new viral DNA sequences (Barrangou, 2007). Through this way, they were able to alter the bacteria's resistance to an attack by a specific virus. Thus, the researchers confirmed that CRISPRs play a role in regulating bacterial immunity.

2.3.1. Locus structure:

The CRISPR array is made up of an AT-rich leader sequence followed by short repeats that are separated by unique spacers. CRISPR repeats typically range in size from 28 to 37 base pairs (bps), though there can be as few as 23 bp and as many as 55 bp (Barrangou, 2015). Some show dyad symmetry, implying the formation of a secondary structure such as a stem-loop ('hairpin') in the RNA, while others are designed to be unstructured. The size of spacers in different CRISPR arrays is typically 32 to 38 bp (range 21 to 72 bp) (Lackie, 2007).

2.3.2. Cas genes and CRISPR subtypes:

Small clusters of *cas* genes are often located next to CRISPR repeat-spacer arrays. Collectively the 93 *cas* genes are grouped into 35 families based on sequence similarity of the encoded proteins. 11 of the 35 families form the *cas* core, which includes the protein families Cas1 through Cas9. A complete CRISPR-Cas locus has at least one gene belonging to the *cas* core.

Table 2.2: Signature genes and their putative functions for the major and minor CRISPR-Cas types.

Class	Cas type	Signature protein	Function
1	IA	Cas3	Single-stranded DNA nuclease (HD domain) and ATP-dependent helicase
	IB	Cas8a, Cas5	Subunit of the interference module. Important in Targeting of invading DNA by recognizing the PAM sequence
	IC	Cas8b	
	ID	Cas8c	
	IE	Cas10d	contains a domain homologous to the palm domain of nucleic acid polymerases and nucleotide cyclases
	IF	Cse1, Cse2	
	IU	Csy1,Csy2, Csy3	Not determined
	III	GSU0054	
	IIIA	Cas10	Homolog of Cas 10d and Cse1
	IIIB	Csm2	Not determined

Continuation of table 2.2….

	IIIC	Cmr5	Not determined
	IIID	Cas10 or Csx11	
	IV	Csx10	
	IVA		
	IVB		
2	II	Cas9	Nucleases RuvC and HNH together produce DSBs, and separately can produce single-strand breaks. Ensures the acquisition of functional spacers during adaptation
	IIA	Csn2	Ring-shaped DNA-binding protein. Involved in primed adaptation in Type II CRISPR system
	IIB	Cas4	Not determined
	IIC		Characterized by the absence of either Csn2 or Cas4
	V	Cpf1, C2c1, C2c3	Nuclease RuvC. Lacks HNH
	VI	Cas13a, Cas13b, Cas13c, Cas13d	RNA-guided RNase

CHAPTER 3

MATERIALS AND METHODS

Genomics is a multifaceted field of biology focusing on the function, structure, evolution, mapping and editing of genomes. A genome is an organism's complete set of DNA, which includes all of its genes. It targets at the collective characterization and qualification of all of an organism's genes, their interrelations and influences on the organism. Genomics also involves the sequencing and analysis of genomes by using of high throughput DNA sequencing and bioinformatics to assemble and analyze the function and structure of entire genomes (Culver, 2002).

In this project, correlation between bacterial genome and spacers have been investigated through some computational approaches such as genome data downloading, finding the spacers, annotating the whole genome bacterial sequences, finding the prophages etc. by using NCBI database, Minced, Prokka, Phispy tools respectively.

3.1 Bacterial genome data download from NCBI:

Bacterial genome sequencing which was started by an approach made on genome analysis through sequencing and assembly of unselected pieces of DNA to get the complete nucleotide sequence of the genome from the whole chromosome in the year of 1995 which led to a promising breakthrough in the area of microbiology and infectious disease research.

NCBI, The National Centre for Biotechnology Information advances science and health by providing access to biomedical and genomic information. NCBI has a multi-disciplinary research group that consists of computer scientists, molecular biologists, mathematicians, biochemists, research physicians concentrating on basic and applied research in computational molecular biology. NCBI assumed responsibility for the GenBank DNA sequence database in October 1992, the stuff of NCBI who have the advanced training in molecular biology build the database from various sequences submitted by different individual laboratories and by exchanging data with European Molecular Biology Laboratory (EMBL), the DNA Database of Japan (DDBJ) and the international nucleotide sequence databases. NCBI has the arrangements with the U.S. Patent and Trademark Office that enable the integration of patented sequence data.

It also supports and distributes a multifarious of databases for the medical and scientific communities, these databases include the Molecular Modeling Database (MMDB) of 3D protein structures, the Online Mendelian Inheritance in Man (OMIM), a Gene Map of the Human Genome, the Taxonomy Browser, and the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute.

From NCBI, a number of 12000 bacterial genome sequences have been downloaded.

3.2 Determining Spacers number using MinCED:

Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are a novel type of direct repeat found in a wide range of bacteria and archaea. CRISPRs work by defending their hosts against invading extrachromomal elements such as viruses. The CRISPR arrays are identified using minCED (mining CRISPRs in environmental data sets), a derivative of CRISPR Recognition Tool that is more conservative in repeat calling and allows more flexible user outputs (Bland, 2007). It is a program to find Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) in full genomes or environmental datasets such as assembled contigs from metagenome. Custom code determines the orientation of the repeats, generates the consensus repeat sequences, and returns the number of repeats, indicating the size of the array. After the identification of CRISPR loci, the types and subtypes are assigned by using the presence or absence of genes, detect multiple systems in a genome, and by identifying the missing repeats and cas proteins it determines the completeness of the system.

3.3 Genome Annotation with Prokka:

Genome annotation is the process of identifying and labelling all the relevant features on a genome sequence (Richardson and Watson, 2013). Genome annotation can be divided into three basic categories. At first, the annotation is a nucleotide level annotation which seeks to identify the physical location of DNA sequences so that it can determine where components, for example: genes, RNAs and repetitive elements are present. The second category of annotation is a protein level annotation that aims to determine the possible functions of genes and identifying that which one a given organism does or does not have and the third one is process level annotation that targets to identify the pathways and processes in which different genes interact. In the last two levels, sequencing errors may compromise the interference of the true gene function because of reduced similarity (Miller et al., 2010; Reeves et al., 2009; Stein, 2001).

Prokka is a command line software tool which can be installed on any Unix system to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files through the coordination of some existing software tools.

Table 3.1: Tools used by Prokka

Tool	Features predicted
Prodigal	Coding sequences (CDS)
RNAmmer	Transfer RNA genes
SignalP	Signal leader peptides
Aragorn	Ttransfer RNA genes
Infernal	Non-coding RNA

This tool finds and annotates features both protein coding regions and RNA genes, such as tRNA, rRNA present on a sequence. It generally uses a two-step process for the annotation of protein coding regions: first, protein coding regions on the genome are identified using Prodigal; second, the function of the encoded protein is predicted by similarity to proteins in one of many protein or protein domain databases.

The mandatory parameter of Prokka for the input file is preassembled genomic DNA sequences in FASTA format. Prokka creates 10 files in the specified output directory, these are given below:

Table 3.2: Description of Prokka output files

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FAST file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

3.4 Finding Prophages using PhiSpy:

Phages when infect their host and after that remain inside the microbial cell replicating with genome during their lysogenic growth, are called prophages. These mobile elements can have major impact on their bacterial hosts' genomes and phenotypes, which may lead to diversification, increased virulence or antibiotic resistance. A prophage generally integrates into a genome by site-specific recombination which is catalyzed by integrases (a family of proteins) (Cambell et al., 1963). These proteins recognize sequences on both the phage (attP which denotes the attachment site in the phage genome) and bacterial (attB, denotes the attachment site in the bacterial genome) genomes, and homologous recombination between these sites results in the duplication of a short stretch of DNA in the continuity of the chromosome which results in the duplicated sites, attL and attR, flanking the inserted prophage and the they get ready for the reverse reaction, excision of the phage from the chromosome. The att regions vary widely in total length and its specific integration site is within a bacterial genome (Landy, 1977; Shimada et al.,1972; Rausch et al., 1991) Phages often integrate into tRNA/tmRNA genes but do not exclusively use those loci as the target site for integration (Fouts, 2006).

PhiSpy, a bioinformatic tool, written in C++, Python and R, developed for identifying prophages in microbial genomes which focuses on the similarity-based and composition-based strategies. This tool is based on seven distinctive similarity-agnostic characteristics. These characteristics are protein length, transcription strand directionality, customized AT and GC skew, the abundance of unique phage DNA sequence words, phage insertion points and the similarity of phage proteins. With the first five characteristics, identification of prophages without any sequence similarity with known phage is possible. Optimized metrics were designed to quantify each of these characteristics and the random forest classification algorithm was used to predict prophages by ranking genomic regions based on those characteristics. PhiSpy also uses similarity-based approaches which enables a complete identification of prophages in a genome. Finally, each predicted prophage region was evaluated by the identification of duplicate *att* sites and by phage protein similarity. PhiSpy found 94% of prophages in 50 bacterial genomes with a 6% false-negative rate and a 0.66% false-positive rate (Sajia et al., 2012).

Before running this tool some softwares had to be installed, such as: Python (version 3.4 or later), Biopython (version 1.58 or later), gcc – GNU project C and C++ compiler (version 4.4.1 or later), The R project for Statistical Computing (version 2.9.2 or later).

As PhiSpy requires seed annotation directory for the input, GenBank files were separated which contain the sequences of the genome from the annotated files which was done by Prokka and converted the "genbank to seed" annotation directory and then ran the PhiSpy.

There are three output files generally located in the output directory. They are:

A. Prophage.tbl:

It contains two columns which are separated by tabs (id, location). The id is in the format: pp_number, where number is a sequential number of the prophage (starting at 1). Location is be in the format: contig_start_stop that encompasses the prophage.

B. Prophage_tbl.tsv:

This is a tab seperated file. The file contains all the genes of the genome. The tenth colum represents the status of a gene. If this column is 1 then the gene is a phage like gene; otherwise it is a bacterial gene.

This file has 16 columns:(i) fig_no: the id of each gene; (ii) function: function of the gene; (iii) contig; (iv) start: start location of the gene; (v) stop: end location of the gene; (vi) position: a sequential number of the gene (starting at 1); (vii) rank: rank of each gene provided by random forest; (viii) my_status: status of each gene based on random forest; (ix) pp: classification of each gene based on their function; (x) Final_status: the status of each gene. For prophages, this column has the number of the prophage as listed in prophage.tbl above; If the column contains a 0 we believe that it is a bacterial gene. If we can detect the att sites, the additional columns will be: (xi) start of attL; (xii) end of attL; (xiii) start of attR; (xiv) end of attR; (xv) sequence of attL; (xvi) sequence of attR.

C. Prophage_coordinates.tsv:

This file has the prophage ID, contig, start, stop, and potential att sites identified for the phages.

3.5 BLAST:

Basic Local Alignment Search Tool, an algorithm for finding regions of similarity between biological sequences through comparing nucleotide or protein sequence databases and calculates the statistical significance, one of the most widely used bioinformatics programs for sequence searching. BLAST uses the heuristic algorithm which is much faster than other approaches for example: calculating an optimal alignment. It is more time-efficient than

FASTA by searching only for the more significant patterns in the sequences, yet with comparative sensitivity. BLAST determines which bacterial species have a protein that is lineage to a certain protein with known amino-acid sequence. It also helps to find out genes which encode proteins that exhibit structures or motifs such as ones that have never been identified. BLAST takes FASTA or Genbank format sequences and weight matrix as input.

The output of BLAST comes in a variety of formats such as, HTML, plain text, XML formatting etc. For NCBI's web-page, the default format for output is HTML. When performing a BLAST on NCBI, the results are given in a graphical format showing the hits found, a table showing sequence identifiers for the hits with scoring related data, as well as alignments for the sequence of interest and the hits received with corresponding BLAST scores for these

In Unix, the BLAST result contains various columns such as: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, e-value, bit score, sequence.

CHAPTER 4

RESULTS

4.1. Number of spacers containing genome:

From NCBI, a total number of 12,000 bacterial genomes was downloaded and the intergenic spacers (approximately 26 to 72 nucleotide variable sequences) which are the region of non-coding DNA between genes, one of the major specific characteristics in the specialized region of the DNA, known as CRISPR was found in 5680 genomes from which a number of 2793 genomes were used in this project by using MinCED tool.

Average spacer has been found around 49 per genome. The highest number of spacers per genome detected was 812 which is carried by a myxobacteria named *Haliangium ochraceum*, and 44 spacers per genome was the smallest number of spacers, found in a polyunsaturated fatty acid-rich and steroid producing soil myxobacterium named *Minicystis rosea.*

4.2. Identification of prophage:

Total number of prophages identified was 1,36,294 from the number of 2793 whole genome bacterial sequences by using prophage identifying tool PhiSpy which works based on some distinctive characteristics such as: protein length, transcription strand directionality, customized AT and GC skew, the abundance of unique phage DNA sequence words, phage insertion points and the similarity of phage proteins.

Average number of prophages per genome is 11.5 and while they do not align with spacers is 9.4.

4.3. Number of spacers vs. number of prophages:

It was hypothesized that with the increase of the number of spacers, the number of prophages will decrease.

A graph is plotted by using R, between the total number of spacers from the 2793 genomes and the total number of 2793 prophages.

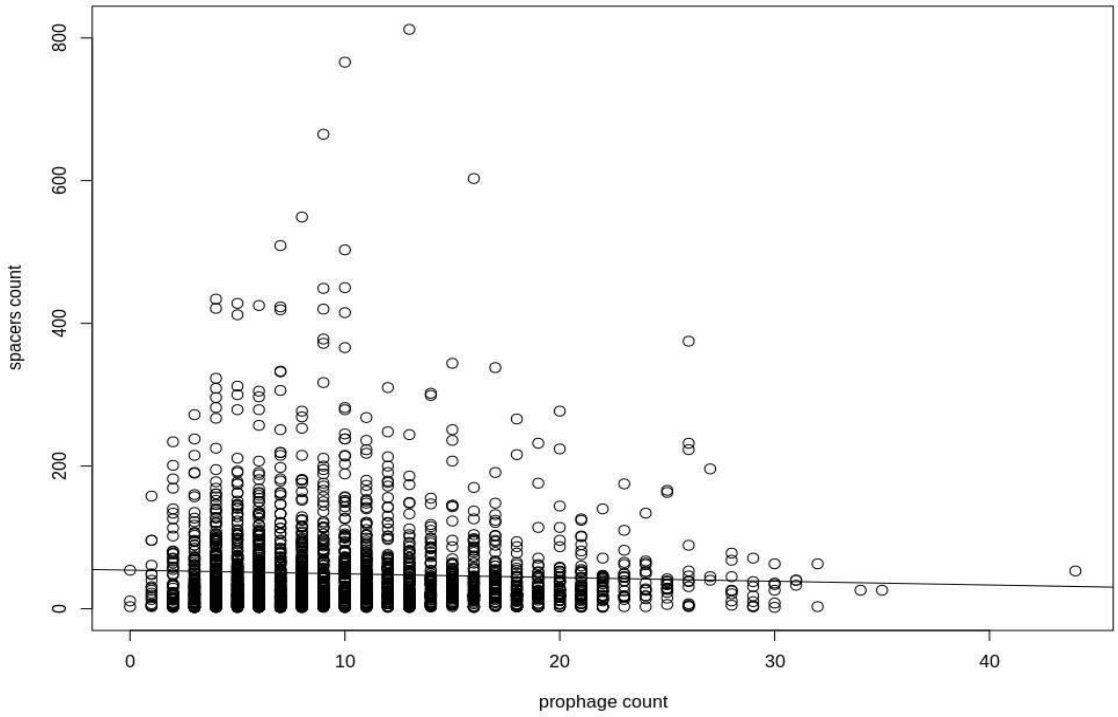

Figure 4.1: A dot plot (in R), where x-axis is representing the total number of prophages and y-axis is representing the total number of spacers found in the 2793 genomes.

According to the above illustration, the regression line is not sloped at any angle which indicates that there is no positive or negative relationship between the variables, instead this regression line is flat or horizontal, showing that there is no strong relationship between the number of spacers on the number of the prophages as the graph indicates that y has no strong effect on the dependent variable x or vice versa.

4.4. Spacers number vs Prophage length:

A graph is plotted by using R, between the total number of spacers from the 2793 genomes and the total length of 2793 prophages.

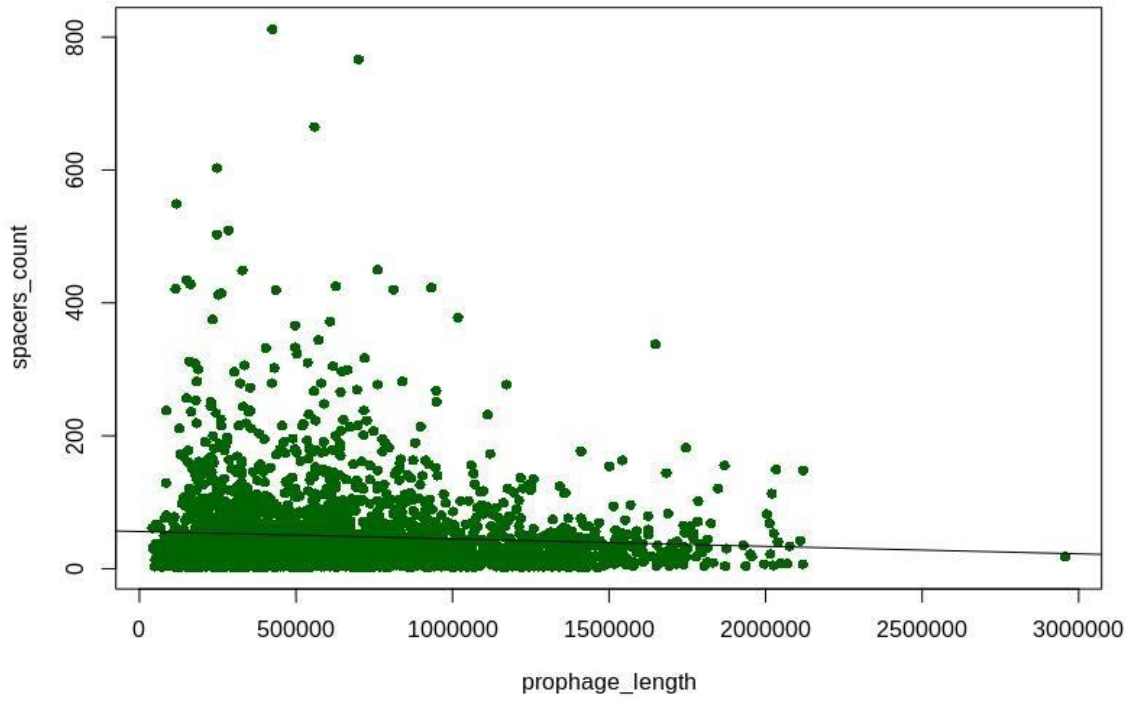

Figure 4.2: A dot plot (in R), where x-axis and y-axis is showing prophage length and the number of spacers found in the 2793 genomes respectively.

In this graph, the regression line is not sloped at any angle which indicates that there is no positive or negative relationship between the variables, instead this regression line is flat or horizontal, showing that there is no strong relationship between the number of spacers on the length of the prophages as the graph indicates that y has no strong effect on the dependent variable x or vice versa.

From the data, it has been observed that, there are some complete genome of bacteria which have few spacer number but the prophage length is large as well as there are some where the number of spacers increased with the length of the prophages; for example: the complete genome of *Caldilinea aerophile* (DSM 14535 strain) has the number of spacer of 665 and the prophage length is 55,9453 whereas *Helicobactor pylori* has only 2 spacer but the prophage length of this is 10,04,575. Few other examples: A genome of *Escherichia coli* (0111:H

serotype and 11128 strain) has only 10 spacers and 11,22,896 of prophage length, *Saccharopolyspora erythrea* (NRRL 2338 strain) contains 10 spacers and the length of prophage is 20,39,123.

4.5. Spacers number vs. Blast hit:

According to the hypothesis of this research, it was predicted that with the increase of the number of spacers, the length of the prophages will decrease which also led to the decrease of the number of blast hits, performed by BLAST, an algorithm for finding regions of similarity between biological sequences through comparing nucleotide or protein sequence databases and calculates the statistical significance.

A graph (dot plot) between the number of spacers found in the 2793 bacterial genome against the number of blast hits was plotted which made a positive slope, indicating strong positive association: with the increase of spacers number, the number of blast hits also increased.

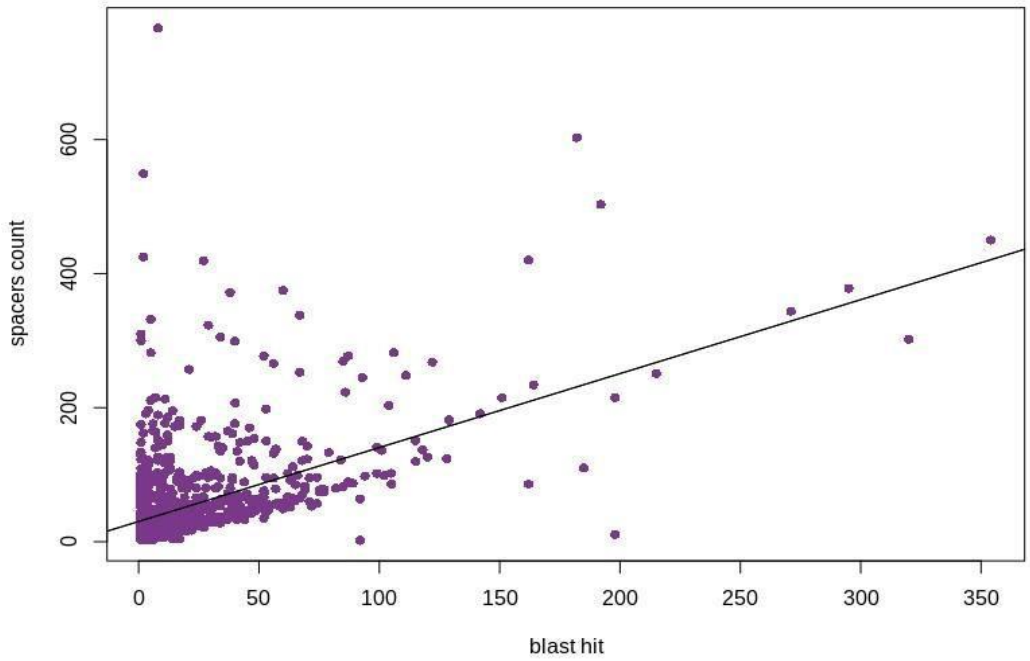

Figure 4.3: A dot plot (using R) between the number of spacers found in 2793 bacterial genome, plotted on y-axis and the number of blast hits, plotted on x-axis, showing a positive relation between the number of spacers and number of blast hits as the regression is moderately skewed left.

It has been found from the data that average number of spacers while it is aligned with prophage is 56 and the average number of spacers when they do not align with prophage is 45, overall the average two of these is 44 spacers per bacterial genome. And average number of prophages per genome while align with spacers is 11.5 and while they do not align with spacers is 9.4.

4.6 Aligned spacers vs. Prophage:

Spacers which are taken from viruses that previously attacked the bacteria, enables a bacterium to recognize the viruses and protect the cell from future attacks, thus they are supposed to immunize bacteria from foreign DNA. But in this research work, it has been found that though spacers are present, prophage still integrated.

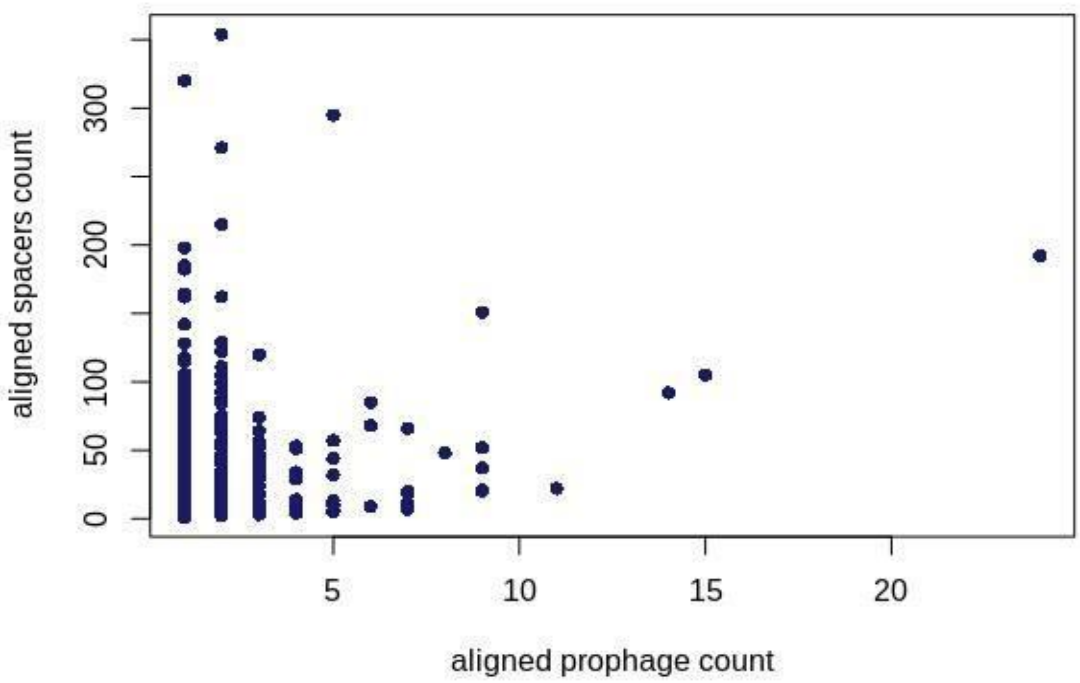

Figure 4.4: A dot plot between the number of aligned prophage (plotted on x-axis) and the number of aligned prophage (plotted on y-axis).

This graph illustrates that numerous spacers aligned with many integrated prophages in bacterial genome that they were not supposed to do.

CHAPTER 5

DISCUSSION

Bacteriophages are the viruses that infect bacteria, the most diverse and prevailing entities in the environment (Comeau et al., 2008). They attach themselves to a susceptible bacterium and infect the host cell through forcing the cell to produce viral components instead of producing bacterial components. During making the infection they undergo either lytic or lysogenic life cycle. In lytic phages, they make progeny of them by destroying the host cell and in the lysogeny they remain dormant in the cell and with the replication of bacteriophage during the cellular division, the daughter cells contain the prophages (Holmes, 1996). On the other hand, bacteria have their own defensive mechanism to protect the attack of bacteriophages, known as CRISPR which functions through nucleotide repeats and spacers. Spacers are the non-coding region in the DNA sequences and these can remember, recognize the viruses and prevent further attack (Lackie, 2007). So, theoretically, when the bacterial immune system, provided by CRISPR through spacers works properly then the bacteria will be less infected or prophages will be less integrated in the bacteria. On the basis of this theory this project work was conducted to observe the correlation between the spacers and the prophages in bacterial genome.

This project has been carried out by 12,000 bacterial whole genome sequences where 5680 genomes were found that contained spacers. Finally, a total number of 2973 genomes were used in this project to investigate the correlation between the abundance of the number of spacers and the number of prophages. From the 2973 genomes, the total number of spacers was 1,36,294 where average spacers per genome was around 49. The highest number of spacers per genome detected was 812 which is carried by a myxobacteria named *Haliangium ochraceum*, and 44 spacers per genome was the smallest number of spacers, found in a polyunsaturated fatty acid- rich and steroid producing soil myxobacterium named *Minicystis rosea*. Average number of prophages per genome is 11.5 and while they do not align with spacers is 9.4.

It has been found that with the increase number of spacers, the number of prophages were not decreasing rather the regression line found after making the statistical graph between the number of spacers and the number of prophages, showing a horizontal line indicating no significant relationship between these two variables. When the comparison of total number of spacers and the total length of prophages was made it was observed that the number of spacers has almost no effect on the length of prophages, several statistics from random data were also made to observe the changes but all of the graph represented a flat regression line which indicates no relationship between these two variables. Moreover, it was expected that when the number of spacers increases, the number of similarities between the sequences of spacers and

prophages will decrease but from some dot plot (using R) made between the number of spacers and the number of blast hit and it was observed that with the increase of the spacers count, the number of blast hit also increases; this showed a strong positive relation of the number of spacers on the number of blast hits.

It has been also observed that numerous spacers aligned with many integrated prophages which they were not supposed to do because, spacers which are taken from virus that previously attacked the bacteria, enables the bacteria to recognize the viruses and prevent future attacks: a natural defense mechanism of bacterial by CRISPR (Barrangou, 2015).

It was hypothesized before this project work that when the spacer number in a genome is large then the number of the prophage will be short. The finding of this thesis conflict with expected outcome since regression line did not demonstrate any significant changes when analyzed. Moreover, almost 33% bacterial genomes were found to carry prophages even in the presence of spacers those align with them and the number of spacers increases with the increase of prophage length. In spite of being present in bacterial genome along with spacers, they do not share any core genes indicate that not any specific types of prophages have this special capacity to infect bacteria in presence of spacer rather they are of different species.

CHAPTER 6

REFERENCES

Abedon ST. (2008). Phages, Ecology, Evolution. In: Abedon ST, editor. Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses. Cambridge University Press; 2008. pp. 1–28

Ackermann HW, Prangishvili D. (2012). Prokaryote viruses studied by electron microscopy. Arch Virol. 2012 Oct;157(10):1843-9.

Ackermann HW. (1987). Bacteriophage taxonomy in 1987. PMID: 3153614.

Ackermann HW. (1996). Frequency of morphological phage descriptions in 1995. Arch Virol. 1996;141(2):209-18. DOI: 10.1007/bf01718394.

Anilionis, A., Ostapchuk, P., and Riley, M. (1980) Identification of a second cryptic lambdoid prophage locus in the E. coli K12 chromosome. Mol Gen Genet 180: 479–481

Bacteriophages: characteristics and 1st stages of a classification. Pathol Biol (Paris). 1969 Nov;17(21):1003-24.

Banks, D.J., Beres, S.B., and Musser, J.M. (2002) The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends Microbiol 10: 515–521.

Barrangou R (2015). The roles of CRISPR-Cas systems in adaptive immunity and beyond. Curr Opin Immunol. 2015 Feb; 32:36-41. doi: 10.1016/j.coi.2014.12.008 PMID: 25574773.

Barrangou R1, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007 Mar 23;315(5819):1709-12. DOI: 10.1126/science.1138140 PMID: 17379808.

Bertani, E., and Six, E. (1988) The P2- like phages and their parasite P4. In The Bacteriophages, Vol. 2. Calendar, R. (ed.). New York: Plenum Press, pp. 73–143.

Bland C, Ramsey TL, Sabree F, et al. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007 Jun 18;8:209. DOI: 10.1186/1471-2105-8-209 PMCID: PMC1924867.

Campbell A, Schneider S, Song B. (1992). Lambdoid phages as elements of bacterial genomes (integrase/phage21/Escherichia coli K-12/icd gene) Genetica. 1992;86:259–267

Campbell A. (1963). Episomes. Adv. Genet. 1962;11:101–145

Campbell, A. (1994) Comparative molecular biology of lambdoid phages. Annu Rev Microbiol 48: 193–222

Casjens, S., and Hendrix, R. (2003) Bacteriophage roles in bacterial chromosome evolution. In The Bacterial Chromosome. Higgins, P. (ed.). Washington, DC: American Society for Microbiology Press, (in press).

Casjens, S., Van Vugt, R., Tilly, K., Rosa, P.A., and Stevenson, B. (1997). Homology throughout the multiple 32- kilobase circular plasmids present in Lyme disease spirochetes. J Bacteriol 179: 217–227.

Chang, C.C., Gilsdorf, J.R., DiRita, V.J., and Marrs, C.F. (2000) Identification and genetic characterization of Haemophilus influenzae genetic island 1. Infect Immun 68: 2630–2637

Culver KW, Labow MA (8 November 2002). Genomics. In Robinson R (ed.). Genetics. Macmillan Science Library. Macmillan Reference USA. ISBN 978-0-02-865606-9.

d'Herelle F. (1917). Sur un microbe invisible antagoniste des bacilles dysentériques. C R Acad Sci Paris. 1917;165:173–5.

d'Herelle F. (1931). Bacteriophage as a treatment in acute medical and surgical infections. Bull N Y Acad Med. 1931;7:329–48

Deghorain M. and Melderen L. (2012). The Staphylococci Phages Family: An Overview. Viruses. 2012 Dec; 4(12): 3316–3335.

Espion, D., Kaiser, K., and Dambly- Chaudiere, C. (1983) A third defective lambdoid prophage of Escherichia coli K12 defined by the lambda derivative, lambdaqin111. J Mol Biol 170: 611–633.

Espion, D., Kaiser, K., and Dambly- Chaudiere, C. (1983) A third defective lambdoid prophage of Escherichia coli K12 defined by the lambda derivative, lambdaqin111. J Mol Biol 170: 611–633.

F.W. Towrt (1915). AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES. Volume 186, Issue 4814, P1241-1243, December 04, 1915

Faubladier, M., and Bouche, J.P. (1994) Division inhibition gene dicF of Escherichia coli reveals a widespread group of prophage sequences in bacterial genomes. J Bacterial 176: 1150–1156

Fauquet CM1, Pringle CR. (2000). Abbreviations for bacterial and fungal virus species names. Arch Virol. 2000;145(1):197-203. DOI: 10.1007/s007050050017.

Fouts D. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 2006;34:5839–5851

Fuhrman JA, Noble RT. (1998). Viruses and protists cause similar bacterial mortality seawater. Limnol Oceanogr. 1995;40:1236–1242

Gratia, J.P. (1989) Products of defective lysogeny in Serratia marcescens SMG 38 and their activity against Escherichia coli and other Enterobacteria. J Gen Microbiol 135: 25–35

Greener, A., and Hill, C.W. (1980) Identification of a novel genetic element in Escherichia coli K- 12. J Bacteriol 144: 312–321

Hankin EH. (1896). L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera. Ann Inst Pasteur (Paris) 1896;10:511–23.

Holmes, R. K. and Jobling, M. G. (1996). Genome organization. In S. Barron (Ed.), Medical microbiology (4th ed.). Galveston, TX: University of Texas Medical Branch at Galveston.

Humphrey, S.B., Stanton, T.B., Jensen, N.S., and Zuerner, R.L. (1997) Purification and characterization of VSH- 1, a generalized transducing bacteriophage of Serpulina hyodysenteriae. J Bacteriol 179: 323–329.

Jed A. Fuhrman (1999). Marine viruses and their biogeochemical and ecological effects. Nature

399, 541-548(1999).

Kaiser, K., and Murray, N.E. (1979) Physical characterisation of the ‘Rac prophage’ in E. coli K12. Mol Gen Genet 175: 159–174.

Kaiser, K., and Murray, N.E. (1979) Physical characterisation of the ‘Rac prophage’ in E. coli K12. Mol Gen Genet 175: 159–174.

Krogh, S., O'Reilly, M., Nolan, N., and Devine, K.M. (1996) The phage- like element PBSX and part of the skin element, which are resident at different locations on the Bacillus subtilis chromosome, are highly homologous. Microbiology 142: 2031–2040.

Lackie, J. M., ed. (2007), The Dictionary of Cell & Molecular Biology (4th ed.), Burlington, MA: Academic Press, p. 394

Landy A, Ross W. (2007). Viral integration and excision: structure of the lambda att sites. Science. 1977;197:1147–1160.

Lang, A.S., and Beatty, J.T. (2001). The gene transfer agent of Rhodobacter capsulatus and ‘constitutive transduction’ in prokaryotes. Arch Microbiol 175: 241–249

Lang, A.S., Beatty, J.T., LeBlanc, H., Towers, G., Harris, J., Lang, G., et al. (2000) Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus. Proc Natl Acad Sci USA 97: 859–864

Lindsey, D.F., Mullin, D.A., and Walker, J.R. (1989) Characterization of the cryptic lambdoid prophage DLP12 of Escherichia coli and overlap of the DLP12 integrase gene with the tRNA gene argU. J Bacteriol 171: 6197–6205.

Little JW. (2005). Loysogeny, Prophage Induction, and Lysogenic Conversion. In: Waldor MK, Friedman DI, Adhya S, editors. Phages Their Role in Bacterial Pathogenesis and Biotechnology. Washington DC: ASM Press; 2005. pp. 37–54

Los M, Kuzio J, McConnell MR, Kropinski AM, Wegrzyn G, Christie GE. Lysogenic conversion in bacteria. In: Sabour PM, Griffiths MW, editors. Bacteriophages in the Control of Food- and Waterborne Pathogens. Washington, DC: ASM Press; 2010

Masui, S., Kamoda, S., Sasaki, T., and Ishikawa, H. (2000) Distribution and evolution of bacteriophage WO in Wolbachia, the endosymbiont causing sexual alterations in arthropods. J Mol Evol 51: 491–497

Mizuno, M., Masuda, S., Takemaru, K., Hosono, S., Sato, T., Takeuchi, M., et al. (1996) Systematic sequencing of the 283 kb 210 degrees- 232 degrees region of the Bacillus subtilis genome containing the skin element and many sporulation genes. Microbiology 142: 3103–3111.

Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., et al. (2000) The R- type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol 38: 213–231.

National Human Genome Research Institute (8 November 2010). "A Brief Guide to Genomics". Genome.gov. Retrieved 2011-12-03.

Nguyen, A.H., Tomita, T., Hirota, M., Sato, T., and Kamio, Y. (1999) A simple purification method and morphology and component analyses for carotovoricin Er, a phage- tail- like bacteriocin from the plant pathogen Erwinia carotovora Er. Biosci Biotechnol Biochem 63: 1360–1369.

Ojaimi, C., Brooks, C., Casjens, S., Rosa, P., Elias, A., Barbour, A., et al. (2003) Profiling temperature- induced changes in Borrelia burgdorferi gene expression using whole genome arrays. Infect Immun 71: 1689–1705.

Okamato, K., Mudd, J., Mangon, J., Huang, W.M., and Marmur, J. (1968) Properties of the defective phage of Bacillus subtilis. J Mol Biol 34: 413–428.

Rakonjac, J. (2012). Filamentous bacteriophages: Biology and applications. In eLS. Chichester, UK: John Wiley and Sons.

Rapp, B., and Wall, J. (1987) Genetic transfer in Desulfovibrio desulfuricans. Proc Natl Acad Sci USA 84: 9128–9130.

Rausch H, Lehmann M. (1991). Structural analysis of the actinophage phi C31 attachment site. Nucleic Acids Res. 1991;19:5187–5189

Richardson EJ, Watson M. (2013). The automatic annotation of bacterial genomes. Brief Bioinform. 2013 Jan;14(1):1-12. doi: 10.1093/bib/bbs007.

Rohwer F, Edwards R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol. 2002 Aug;184(16):4529-35. DOI: 10.1128/jb.184.16.4529-4535.2002.

Rohwer F., Edwards R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol. 2002 Aug;184(16):4529-35.

Rohwer, F. (2003). Global phage diversity. Cell 113, 141. doi: 10.1016/S0092-8674(03)00276-9

Rohwer, F., and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. J. Bacteriol. 184, 4529–4535. doi: 10.1128/JB.184.16.4529-4535.2002

Ruzin, A., Lindsay, J., and Novick, R.P. (2001) Molecular genetics of SaPI1 – a mobile pathogenicity island in Staphylococcus aureus. Mol Microbiol 41: 365–377

Sajia Akhter,1,* Ramy K. Aziz,2,3 and Robert A. Edwards1,2,4,* (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 2012 Sep; 40(16): e126. doi: 10.1093/nar/gks406.

Seemann T1 (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153.

Shimada K, Weisberg R, Gottesman M. (1972). Prophage lambda at unusual chromosomal locations. I. Location of the secondary attachment sites and the properties of the lysogens. J. Mol. Biol. 1972;63:483–503.

Starich, T., Cordes, P., and Zissler, J. (1985) Transposon tagging to detect a latent virus in Myxococcus xanthus. Science 230: 541–543.

Suttle CA. (2005). Viruses in the sea. Nature. 2005;437:356–361. doi: 10.1038/nature04160.

Takemaru, K., Mizuno, M., Sato, T., Takeuchi, M., and Kobayashi, Y. (1995) Complete nucleotide sequence of a skin element excised by DNA rearrangement during sporulation in Bacillus subtilis. Microbiology 141: 323–327.

Thaler, J.O., Baghdiguian, S., and Boemare, N. (1995) Purification and characterization of xenorhabdicin, a phage tail- like bacteriocin, from the lysogenic strain F1 of Xenorhabdus nematophilus. Appl Environ Microbiol 61: 2049–2052

Weinbauer MG, Rassoulzadegan F. (2004). Are viruses driving microbial diversification and diversity Environ Microbiol. 2004;6:1–11. doi: 10.1046/j.1462-2920.2003.00539.x.

Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: The unseen majority. Proc Natl Acad Sci USA. 1998;95:6578–6583

Wilhelm SW, Suttle CA. (1999). Viruses and nutrient cycles in the sea: Viruses play critical roles in the structure and function of aquatic food webs. Bioscience. 1999;49:781–788

William S. Klug, Michael R. Cummings, Charlotte A. Spencer, Michael A. Palladino (2012). Concepts of genetics (10th ed.). San Francisco: Pearson Education. ISBN 978-0-321-72412-0

Yen, H.C., Hu, N.T., and Marrs, B.L. (1979) Characterization of the gene transfer agent made by an overproducer mutant of Rhodopseudomonas capsulata. J Mol Biol 131: 157–168.

Zink, R., Loessner, M.J., and Scherer, S. (1995) Characterization of cryptic prophages (monocins) in Listeria and sequence analysis of a holin/endolysin gene. Microbiology 141: 2577–2584