Biased Acquisition of Spacers by the CRISPR System of the *V. cholerae*

By

Erena Tuzneen Supan

ID:17136010

Nayeema Haque

ID: 17136028

Monira Momtaz

ID: 17136005

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

Department Mathematics and Natural Sciences

Brac University

[September] [2021]

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing a 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. We have acknowledged all main sources of help.

Erena Tuzneen Supan	Nayeema Haque	Monira Momtaz
ID: 17136010	ID: 17136028	ID: 17136005

Approval

The thesis/project titled "Biased Acquisition of Spacers by the CRISPR System of the *V. cholerae*" submitted by Erena Tuzneen Supan (ID: 17136010), Nayeema Haque (ID: 17136028), Monira Momtaz (ID: 17136005) of Spring 2017 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Biotechnology on 9th September 2021.

Examining Committee:

Supervisor:
(Member)

Dr. Iftekhar Bin Naser
Assistant Professor
Department of Mathematics and Natural Sciences
BRAC University

Program Coordinator:
(Member)

Dr. Iftekhar Bin Naser
Assistant Professor
Department of Mathematics and Natural Sciences
BRAC University

Departmental Head:
(Chair)

A F M Yusuf Haider
Professor and Chairperson
Department of Mathematics and Natural Sciences
BRAC University

Ethics Statement

This thesis has been composed entirely by us and it has not been submitted in whole or in part in any previous institution for a degree or diploma.

Abstract

Cholera is a deadly waterborne disease that is caused by the serogroup of O1 or O139 of *Vibrio cholerae*. Also, there are many non-pathogenic *Vibrio cholerae* strains known as non-O1 and non-O139. Bacteriophages are bacteria eating viruses and are highly specific towards the host cell. They require host cellular machinery to perform their metabolic activities. Many of the *Vibrio cholerae* strains contain CRISPR which is an immune defense system that targets foreign DNA like plasmids, phage DNA and even other bacterial DNA that are inserted into CRISPR arrays. When a phage or any invader attacks bacteria, the CRISPR system inserts the fragments of phage or the invader's DNA into the CRISPR array. It is known as spacers which later work as an immune memory. This study includes analysis of *V. cholerae* strains that contain CRISPR, specifically the in depth analysis of spacers using different bioinformatics tools and softwares. One of the major findings of this study was that the spacers come from some specific regions of the phage whole genome.

Keywords: *Vibrio cholerae*; bacteriophage; CRISPR; spacer; PLE

Dedication

Dedicated to our family, friends and all the well wishers for their love and support.

Acknowledgement

To begin with, I would like to express my highest gratitude to Almighty Allah for giving us blessings and the ability to complete our thesis with success. Moreover, we would like to express our appreciation and love to our parents who always give us courage and support.

We would like to express our deepest gratitude to our thesis supervisor **Dr. Iftekhar Bin Naser**, (Assistant Professor, Coordinator, Biotechnology Program, Brac University), who guided us throughout this thesis. Moreover, his constant effort and encouragement towards our research allowed us to think critically and grow as researchers. It was a great honor to complete this work under his supervision.

We also hereby express our sincere gratitude towards Professor **A F M Yusuf Haider**, (Chairperson of MNS Department, Brac University) for allowing us and encouraging us to complete our undergraduate thesis.

We would also like to thank our dearest friends and well wishers for their constant support.

List of Acronyms

- NCBI- National Center for Biotechnology Information
- BLAST- Basic Local Alignment Search Tool
- RAST- Rapid Annotation using Subsystem Technology
- PLE- PICI Like Element
- CRISPR- Clustered Regularly Interspaced Short Palindromic Repeats
- DNA- Deoxyribonucleotide Acid
- RNA- Ribonucleic Acid
- *V. cholerae- Vibrio cholerae*
- *E. coli- Escherichia coli*
- ds- Double Stranded
- ss- Single Stranded
- bp- Base Pair
- ORF- Open Reading Frame

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Bacteriophage is a type of virus that has bacteria as their specific host. As phages continue their infection against bacteria, bacteria have developed defense mechanisms against them. One of them is known as CRISPR. CRISPR is a defense mechanism that helps to keep a record of all the phages that have attacked the bacteria previously. When a bacteria gets infected by a phage, the bacteria integrates the phage DNA on its own genome as spacers. Later, when the same phage attacks the bacteria again, the bacteria transcribe RNA molecules from it and the guide RNA guides an enzyme to the target to cut the foreign DNA. Thus, the phage DNA becomes harmless. These spacer sequences can also be passed down to the bacteria progeny. Due to the role of spacers in bacteria defense mechanisms we have decided to analyze their sequences, distribution, diversity etc.

From our study, we were expecting to find an increment in the average spacer count over the years. However, from our collected data it shows no significant change. Moreover, from all the spacer sequences the virus sequence was taking the least space. Among the bacteriophage and virus other than bacteriophage sequences, bacteriophage sequences were more prominent. Furthermore, most of the strains contained type I CRISPR-Cas systems aside from a few which contained both type I and type II. Also, we expected no PLE presence in these strains. However, 10 out of 208 strains contained PLE along with a CRISPR-Cas system. Among all the phage spacers the top 5 phage spacers sequence came from a specific location of that phage. So, it can be said that the spacer comes from a biased location. It seems that the proteins from the bias region are mainly phage tail protein.

1.2 Objective & Brief Methodology

Objective

The main objective of our research was to analyze the spacer sequences of *V. cholerae* CRISPR through different bioinformatics software.

Brief Methodology

1. Retrieving the *V. cholerae* sequences that contain CRISPR.
2. Finding the spacer sequences of those CRISPR.
3. Software based analysis had been focused on
    - Analysis of the Spacer Sequence (BLASTn)
    - CRISPR Type Identification (CRISPR-Cas++)
    - Phage Genome Annotation (RAST)
    - Study of Proteins (RAST)
    - Detection of PLE (BLAST)

Chapter 2

Literature Review

2.1 *Vibrio cholerae*

Cholera is a waterborne disease that has caused millions of deaths. It is estimated that each year 1.3 to 4.0 million cases are found and 21000 to 143000 deaths occur worldwide due to cholera (WHO, 2021). It is still a major reason for mortality in Asia and Africa. The serogroup of *Vibrio cholerae* O1 or O139 is the causative agent of cholera. These strains carry the genes for cholera toxin and *Vibrio* pathogenicity island. According to Faruque et al., (2003), *Vibrio cholerae* O139 was first reported in 1992 in Madras and in Southern Bangladesh. Apart from the deadly disease causing strains there are non pathogenic *Vibrio cholerae* strains known as non-O1 and non-O139.

2.1.1 Cholera Outbreak History

The first cholera outbreak started in 1817 near the South Asia region (Claeson & Waldman, n.d.). It first occurred in India and spread to Bangladesh. Later, in 1820 it occurred in Thailand and Indonesia. In the year of 1821, Basra, Iraq was also attacked. It further spread to Europe.

Claeson & Waldman also mentioned that the second cholera outbreak started from Europe and America in 1829. In the next year Moscow and St. Petersburg, Finland and Poland were also attacked. The following year England and Germany also became victims of this deadly pandemic. The disease was spread to Canada in 1832 and later the United States was also infected. The second outbreak of the pandemic was stopped with Mexico and Cuba in 1833.

They further added that the third outbreak was the deadliest. It again started in India in 1852 and quickly spread from Iran to Europe and to the US. The whole world fell victim to it once again. Among all the continents Africa was mostly affected. In the history of cholera pandemic 1854 is considered to be the worst of all.

The fourth cholera pandemic began in 1863 in the Bengal region and spread to Mecca. This disease was spread throughout the Middle East and was carried to Russia, Europe, Africa and North America (Claeson & Waldman, n.d.). .

The fifth cholera pandemics started in 1881. Naples in 1884, Spain in 1885, and Russia in 1893–94. China and Japan were infected between 1877 and 1879 . South America was affected by the 1890s.

The sixth pandemic started from 1899 and lasted until 1923. This time India, Arabia, and the North African coast were mostly affected. Some spread in Egypt, Russia and western Europe.

In 1961 the seventh cholera pandemic occurred in Indonesia. From Indonesia it spread throughout Asia during the 1960s. In the following decade the Middle East, Africa were also affected. In 1994 Democratic Republic of the Congo also fell victim to it. In 1998 the disease spread to South and North America.

A deadly cholera outbreak happened in Zimbabwe in the year of 2008–09. It is considered the 7th cholera pandemic. Another devastating outbreak occurred in Haiti in 2010–11 (Claeson, 2019). In the present time *V. cholerae* O139 is present in Bangladesh and India and people from this region are being mainly affected. Here is a graphical representation of the Cholera pandemic periods over the years (Fig-1).

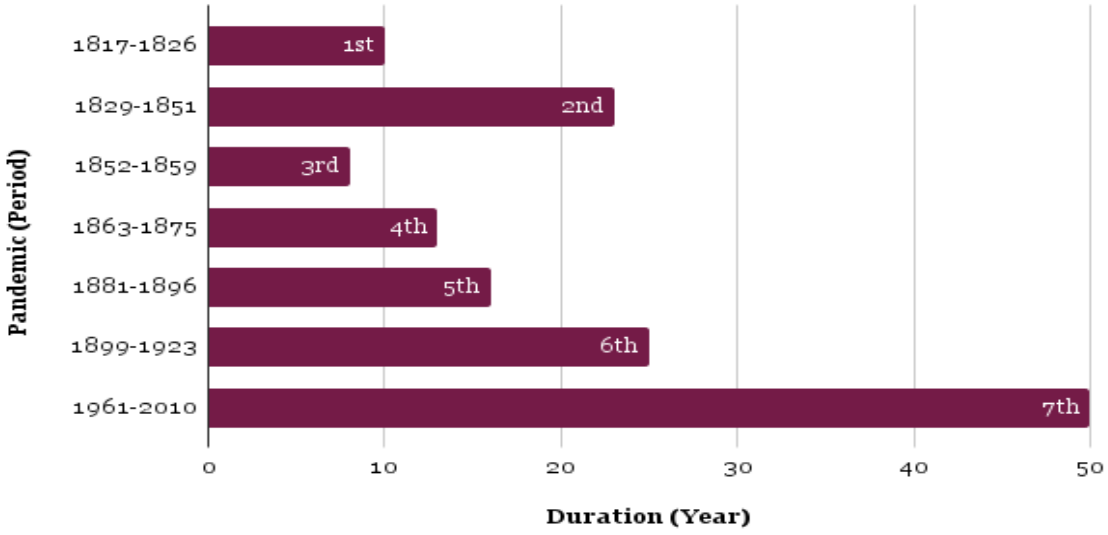


Figure 1: Duration of the Cholera Pandemic. The Y-axis indicates the pandemic period. The X-axis indicates the duration of the pandemic in years.

2.2 Bacteriophage

As humans have been fighting against bacteria similarly bacteria have been fighting against viruses. Virus is a non living infectious agent that can only multiply upon contact with its host. The host range determines the class of virus. Viruses that infect bacteria only are known as bacteriophages. Phages do not contain any plasma membrane, internal organelles, or metabolic machineries. They lack the ability to reproduce also. Hence, they infect host cells and insert their genome for producing virus particles by using the host's replication process. Phage genome sizes range vastly. It starts from ~3,300 nucleotides. Hatfull & Hendrix, (2012) mentioned that the smallest one is the ssRNA phage of *E. coli.* There are some bacteriophages whose genome size consists of 540-735 kb, which are even larger than the genome size of bacteria. These phages are known as Mahaphage. The largest genome size for bacteriophages known till date is the one with 735kb (Shayeb et al., 2020).

2.2.1 History of Phage

Among all the organisms bacteriophages are the most abundant in nature. While studying the bacteriophage researchers discovered that it originates to the early Precambrian Era, approximately 4.5 billion years ago (BACTERIOPHAGES – AN INTRODUCTION TO PHAGES, n.d.). An English bacteriologist Ernest Hanbury Hankin reported in 1896, Ganges water in India seemed to have some antibacterial properties. According to Clokie et al. (2011), it was first discovered in 1915 by William Twort. Later, in 1917 Felix d'Herelle found out that bacteriophage can kill bacteria. He gave it the name bacteriophage or bacteria-eater. The word came from the Greek word *phagein*, which means to devour. Researchers believe bacteriophages control the levels of bacteria in nature. In the following years the studies of bacteriophage focused on some model phages which primarily infected *E. coli*. These research provided the foundation of modern molecular biology.

2.2.2 Structure and Classification of Bacteriophage

All bacteriophages share some similar characteristics. Phages are highly specific towards the host cell. They attack a specific species or strain of bacteria. They are obligate intracellular parasites and require host cellular machinery to perform their metabolic activities. All bacteriophages have the same basic structure. It mainly has a polyhedral head, a short collar and a helical tail (Fig-2). The head consists of 2000 capsomeres with nucleic acid inside. The tail is made of an inner hollow tube. Also, the tail is surrounded by a sheath with annular rings. The distal end included a basal plate with tail fibres at each corner. Tail fibres help the bacteriophage to attach itself to the bacteria.

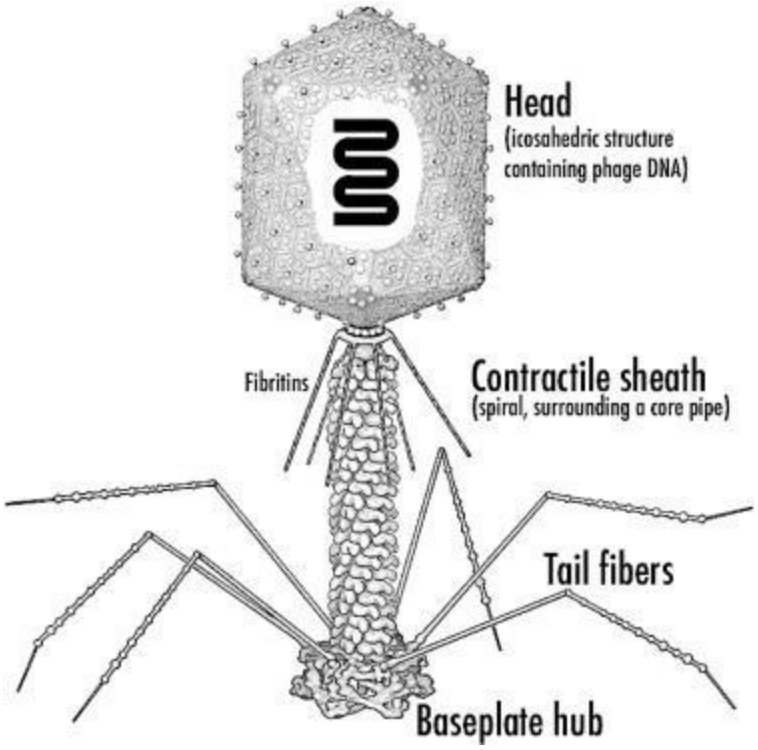
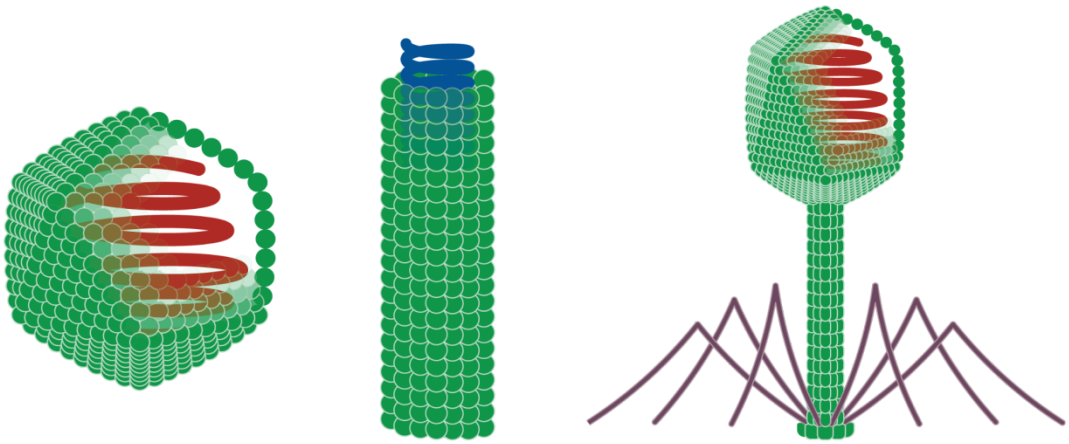


Figure 2: Basic Structure of a Bacteriophage (Harada et al., 2018). The basic structure of a bacteriophage contains polyhedral head, contractile sheath, tail fibers and baseplate hub.

Phages have three basic structural forms (Fig-3). They are: i) an icosahedral head without a tail, ii) an icosahedral head with a tail, and iii) a filamentous form.



(i) (ii) (iii)

Figure 3: The basic structure forms the bacteriophages. The forms are i) icosahedral head without a tail ii) filamentous and iii) icosahedral head with a tail.

Either DNA or RNA can be the genetic material of phages . According to Harada et al. (2018) based on genetic materials it can be divided into four groups. i) single stranded DNA phages (ssDNA), ii) double stranded DNA phages (dsDNA), iii) single stranded RNA phages (ssRNA), and iv) and double stranded RNA phages (dsRNA) (Fig-4).

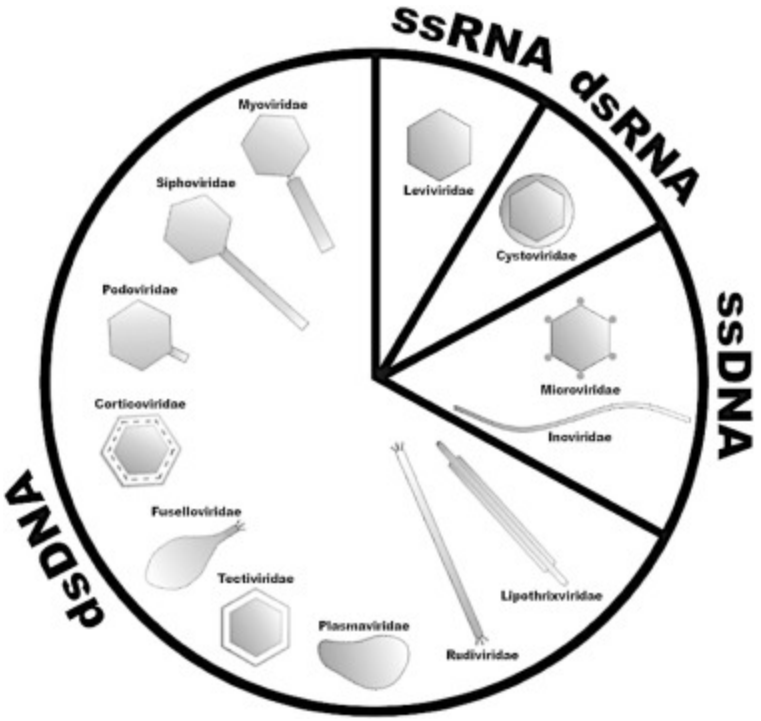


Figure 4: Classification of Bacteriophage (Harada et al., 2018). The main four groups i) single stranded DNA phages (ssDNA), ii) double stranded DNA phages (dsDNA), iii) single stranded RNA phages (ssRNA), iv) and double stranded RNA phages (dsRNA) groups and their sub classification is shown.

2.2.3 Life Cycle of Bacteriophages

Bacteriophages use either lytic cycle or lysogenic cycle to infect their hosts.

2.2.3.a Lytic Cycle

In the lytic cycle, the bacteriophage attaches itself to the host cell and enters inside. In order to produce their own progeny they use the cellular machinery of the host cell. After inserting itself into the host cell, early proteins are synthesized by the phage genome so that it can break down the host DNA. As a result, the phage takes control of the host cellular machinery. After that, the phage uses the host cell to synthesize the other proteins needed to build new phage particles. While the phage is synthesizing its protein the bacteria cell becomes weak by phage enzymes and bursts (Fig-5). In each lytic cycle an average of 100-200 new phage progeny are produced.

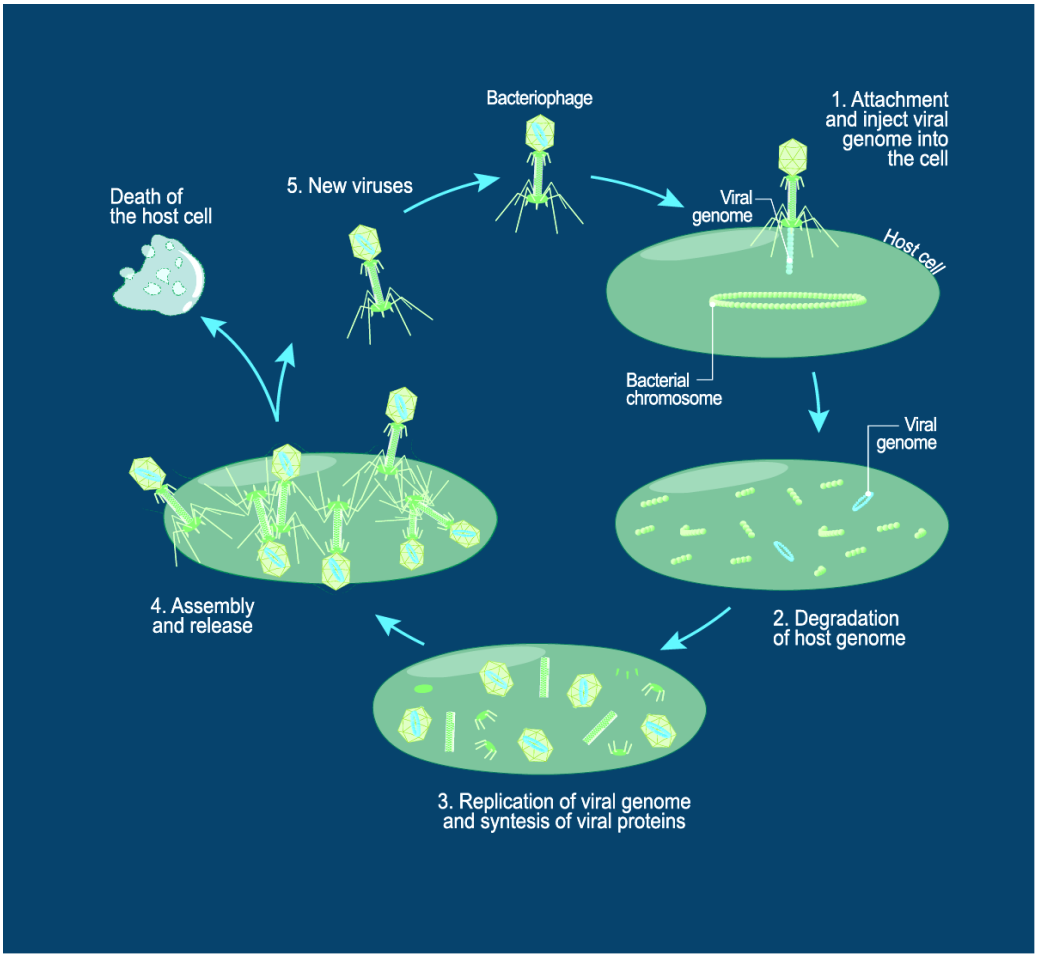


Figure 5: Lytic Cycle of Bacteriophage (Steward, 2018). 1) The phage inserts its viral genome into the host cell after attaching itself to it. 2) The host genome slowly degrades. 3) Using the host machinery virus replicates its genome and synthesizes proteins. 4) All the particles assemble and get released. 5) The host cell dies and a new phage is born.

2.2.3.b Lysogenic Cycle

Upon infecting a bacterium the bacteriophage inserts its nucleic acid into the bacterial chromosome. It allows the phage DNA to be replicated and passed on with the host's own DNA. After the condition of the host cell worsens, the virus becomes active. The lysis of the host cell

occurs during the reproductive cycle. As long as the host cell's offspring reproduce the virus reproduces as well.

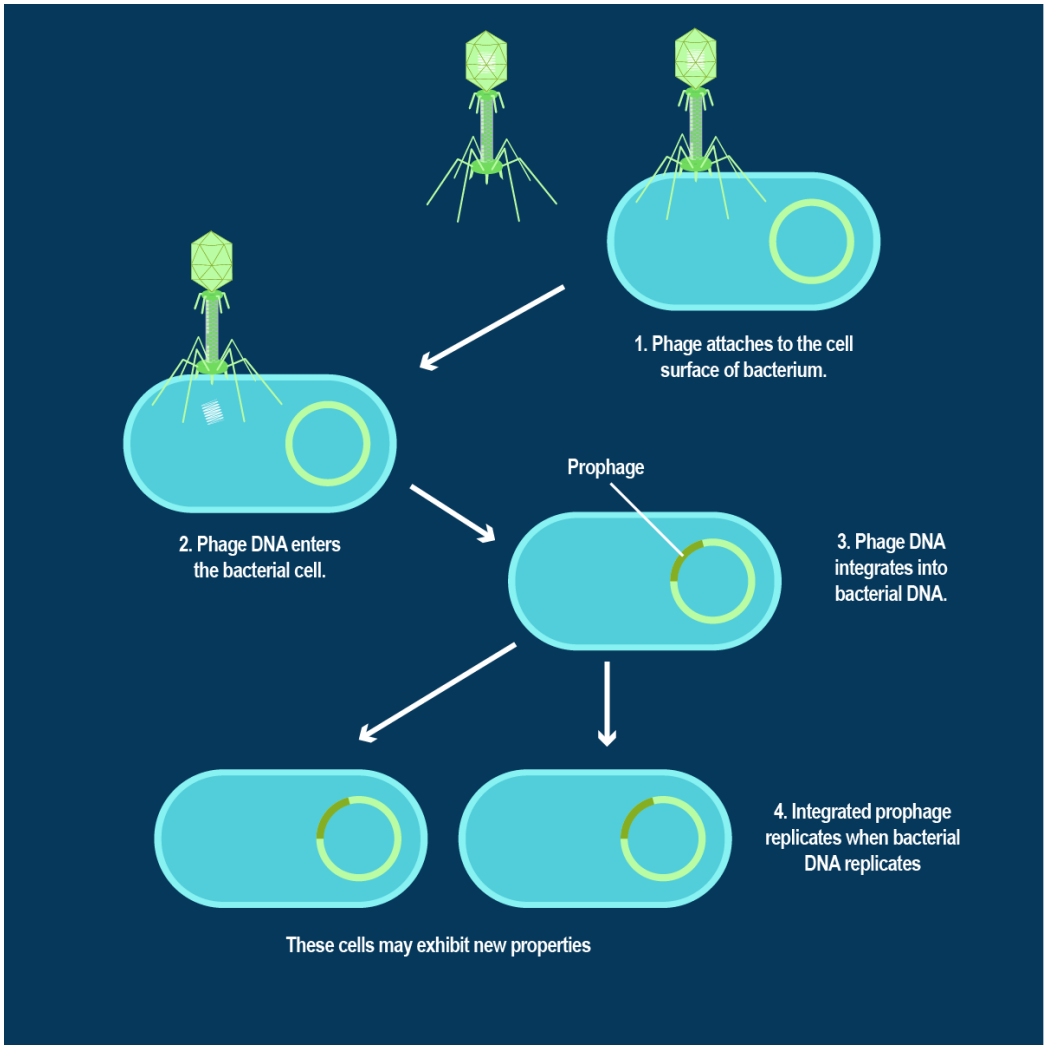


Figure 6: Lysogenic Cycle (Steward, 2018). 1) The phage attaches itself to the bacteria. 2) After attaching the phage DNA enters the bacterial cell. 3) The phage DNA integrates into the bacterial chromosome. 4) The integrated DNA replicates alongside the bacterial DNA.

2.2.4 Application of Phage in Biotechnology

Bacteriophages are a group of distinct viruses. They have promising future uses in biotechnology and research. Recently, researchers have discovered that bacteriophages can be an important tool in modern biotechnology. Some of the future applications can be substitutes for antibiotics, biocontrol agents in agriculture. It can be vehicles for both DNA and protein vaccines. It can detect pathogenic bacterial strain and can work as a display system for proteins and antibodies.

2.2.4.a Phage Therapy

Phage therapy is a therapy that uses viruses to treat bacterial infections or diseases. Lin, Koskella & Lin (2017), said that the principle of this therapy is to use the phage to infect the bacteria and lyse it. Aside from the naturally occurring phages, scientists are also trying to use engineered phages in this therapy. In recent years the increase in superbugs have become a main issue. In this time of crisis, phage therapy can work as an alternative for antibiotics. One of the potential advantages of phage therapy can be that as phage only infect bacteria it should not have any side effect on the human patient. Keeping the potential advantages of phage therapy in mind more research should be done on this topic so that the alternative for antibiotics can be found in the near future.

2.2.5 CRISPR

The acronym of CRISPR is “Clustered Regularly Interspaced Short Palindromic Repeats". According to Karimi et al (2018), CRISPR is a DNA sequence that is found in 48% of bacteria and 90% of archaea. Scientist Ishino along with his coworkers first noticed a short repeat in the *E. coli* genome in the 1980s (Ishino, Krupovic & Forterre, 2018). Remarkably, this was the first step in discovering CRISPR. CRISPR is an array of repeated short nucleotides DNA sequences interspersed at uniform intervals between distinctive nucleotide sequences derived from the DNA of phages that had attacked the bacteria and works as their immune defense system.

CRISPR defense systems are dependent on small RNAs for sequence specific detection. These RNAs silence the foreign nucleic acids.

2.2.5.a Anatomy of CRISPR Locus

CRISPR locus is an array of short direct repeat DNA sequences interspersed with spacer sequences. CRISPR is an immune defense system that targets foreign DNA like plasmids, phage DNA and even other bacterial DNA that are inserted into CRISPR arrays. These inserted sequences are known as spacers (Fig-7). These spacer sequences transcribe when they are attacked by the same previous invaders. It works as a memory and guide to identify and inactivate the invaders. In a given locus the repeats are identical in both length and sequence. Even though the length of the spacers are similar, the sequence is different. Moreover, the repeat length can differ in various species. It can be from 21 bp to 47 bp, and spacers can be of 20–72 bp (Grissa, Vergnaud & Pourcel, 2007). Most CRISPR repeats can form hairpin structures because they are partly palindromic even though their sequences are diverse. In most cases, a single CRISPR loci is found. However, there have been some cases where multiple CRISPR is also seen. Grissa, Vergnaud & Pourcel (2007), also said that the leader sequences are about 80% identical in a genome. CRISPR loci are surrounded by a similar group of conserved protein coding genes. Karginov & Hannon (2010), said that Jansen and colleagues presented cas 1-4 core CRISPR associated gene families. Most CRISPR do not carry all four cas genes. They also said that, in most cases, cas 1 is common along with one or more cas genes.

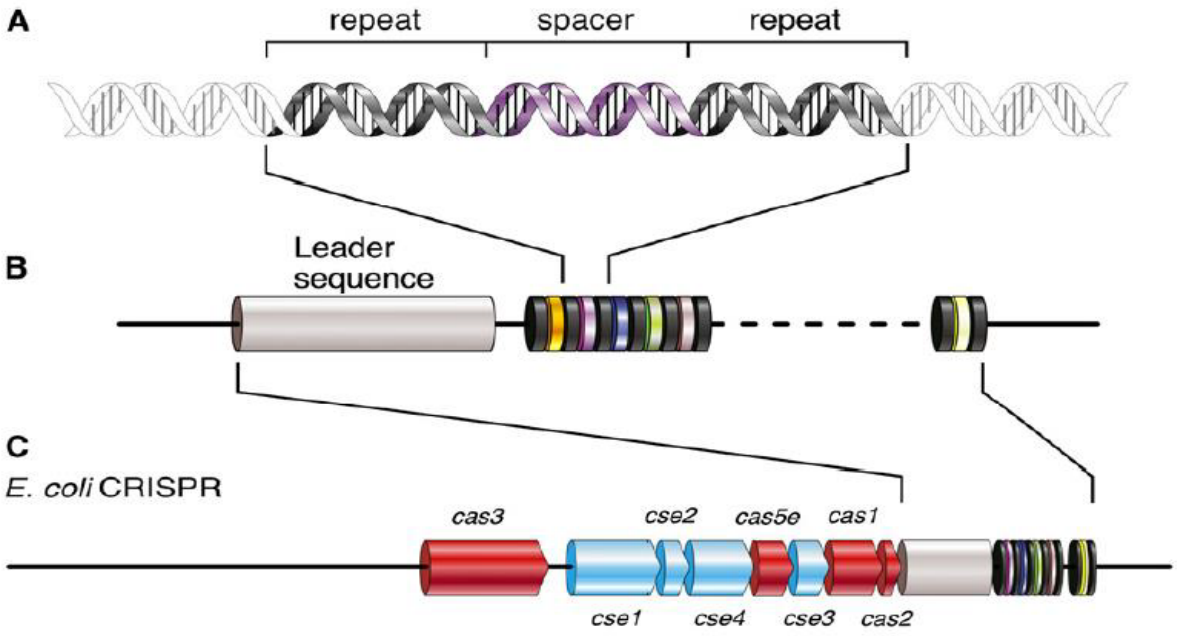


Figure 7: Anatomy of CRISPR (Karginov & Hannon, 2010). A) Repeat sequences are interspersed by varying spacer sequences. B) A conserved leader sequence which is highlighted in gray color is located on one side of the cluster. C) Cas genes surround the CRISPR locus. An example of *E. coli* CRISPR is shown.

2.2.5.b CRISPR as a Defense System of Bacteria

It is known that CRISPR arrays help the bacteria to remember the viruses that have attacked it previously. When a bacteria gets infected by a phage, it inserts phage DNA into the bacterial genome as spacers. The spacers of the CRISPR sequence is a record of all the phages that have infected the bacteria. Then the bacteria transcribe RNA molecules from it. The guide RNA guides an enzyme to the target to cut the foreign DNA when it attacks again. In this way, the viral DNA becomes harmless. Moreover, this record can also be passed down to the bacteria progeny. Aside from being part of bacteria's immune system, CRISPR is also involved in controlling endogenous transcription and regulating bacterial pathogenicity (Shabbir et al., 2019).

2.2.5.c CRISPR Classification

The CRISPR-cas system can be classified into three subtypes. They are: i) CRISPR-cas system type I, ii) CRISPR-cas system type II and iii) CRISPR-cas system type III (Fig-8). It is based on the presence of the cas gene. According to Makarova & Koonin (2015), in case of type I cas3 gene is present. In type II cas9 gene and in case of type III cas10 gene is found. All CRISPR systems have cas1 and cas2 proteins in common as they are important for Spacer Acquisition.

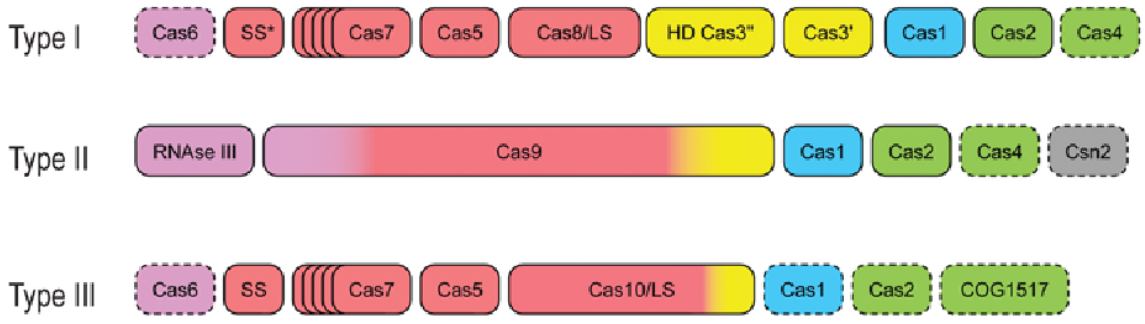


Figure 8: CRISPR Cas System Classification (Makarova et al, 2015). Type I, II and III with their respective Cas genes are shown in this figure.

Type I System

According to Makarova et al (2015), the Type I system is the most common in bacteria and archaea. This system is again classified into six subtypes (A–F). Cas3 gene is the unique gene for type I system.

Type II System

Type II is only present in bacteria. This type is simpler than the other CRISPR Cas system. This system has four genes: cas1, cas2, cas9, and cas4 in type II-B and an additional csn2 gene is

found in the case of type II-A (Makarova et al, 2015). Among these Cas9 genes is the distinctive gene for type II systems.

Type III System

Type III is mainly present in archaea. This type is divided into two groups. i) Type III-A and ii) Type III-B. The type III-B system is only present along with other CRISPR types. Makarova et al (2015), said that the type III system has both cas6 and cas10 genes. The Cas10 gene is unique for this system.

Chapter 3

Materials & Methods

3.1 Bacterial Genome Data Collection from NCBI

From NCBI 12000 *V. cholerae* sequences with more than 20,000 bp length were downloaded and analyzed with CRISPRFinder. The sequences that contained confirmed CRISPR were chosen for further analysis of our study. A total of 208 *V. cholerae* strains were selected. FASTA format of both whole genome sequences and the contigs/nodes/scaffolds containing CRISPR were downloaded. Specific information like year of isolation, location of isolation and publication were retrieved of these 208 strains.

3.2 CRISPR Type Identification with CRISPR-Cas++

After downloading the whole genome sequences of the strains, CRISPR-Cas++ was used to determine the CRISPR type. Both whole genome sequences and contigs/nodes/scaffolds of the CRISPR containing strains were analyzed through CRISPR-Cas++.

3.3 Downloading Spacers with CRISPRFinder

The strains containing confirmed CRISPR were analyzed and spacers were downloaded using CRISPRFinder for further research.

3.4 Analyzing Spacer Sequences Using BLAST

The downloaded spacer's sequences were analyzed through BLAST to know the origin of the spacer sequences.

3.5 Confirming the Presences of PLE Using BLAST

The FASTA format of 3 PICI Like Element (PLE) sequences were retrieved from NCBI. Then the whole genome sequences of the bacterial strains and 3 PLE sequences were BLAST to confirm the presence of PLE.

3.6 Phage Sequence Analysis

The similarities between the phage sequences were checked using BLAST. The FASTA format of the whole genome sequences of the top 5 phages were downloaded from NCBI. The spacers of specific phages were selected and BLAST with their respective whole genome sequence.

.

3.7 Phage Annotation using RAST

The whole genome sequences of top 5 phages were uploaded on RAST for annotation. The protein names from the specific bias regions were retrieved using RAST.

Chapter 4

Result

4.1 Average Spacer Count

As bacteria are invaded by foreign DNA the CRISPR in it tries to acquire spacer sequences from it. Therefore we speculate that *V. cholerae* that was isolated long before may contain less spacers than the *V. cholerae* isolated recently. From all the 208 bacterial strains, we could retrieve the isolation date of 152 strains from NCBI. Using this information we counted the average spacers in years (Fig-9). From this, we could not get any significant information about the increment of spacer number in recent isolates than previous isolates.

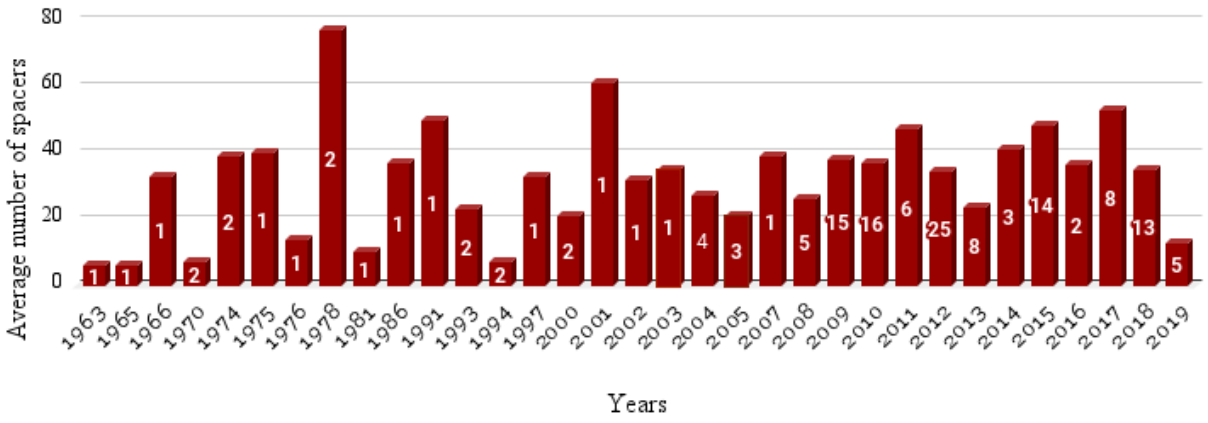


Figure 9: Average Number of Spacers in Years. The Y- axis indicates the average number of spacers and the X- axis indicates the years. The numbers on the columns represent the strain's number in each year that was used.

4.2 Percentage Of Unknown, Bacteria and Virus Sequence

A huge number of bacteria is degrading now and then, releasing their DNA into the environment. So, it might be easier for competent bacterial cells to uptake those free DNA as spacers. For this reason we expected that bacteria spacer sequences would cover the largest part of the total spacers. However, from the BLAST result of the spacers we calculated, the percentage of unknown sequences were highest. Then came bacteria and lastly virus sequences. Among all the

spacers 60.55% were unknown, 28.84% were bacteria and 10.62% were virus sequences (Fig-10).

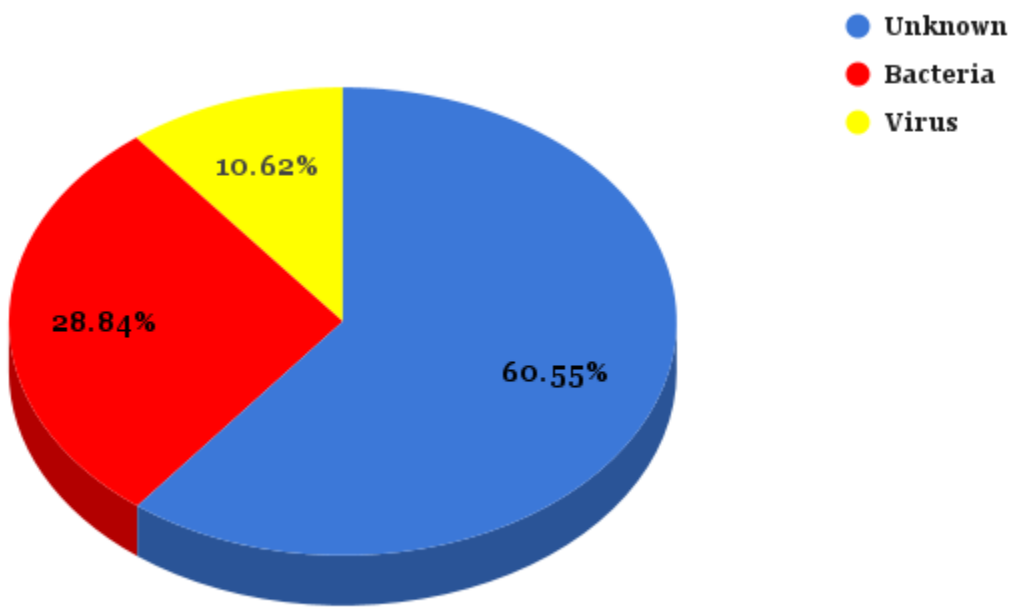


Figure 10: Percentage Of Unknown, Bacteria and Virus Sequence. Here, the blue pie slice indicates unknown sequences, red pie slice indicates bacteria sequences and the yellow pie slice indicates the virus sequences.

4.3 Bacteriophage vs Virus Other than Bacteriophage Count

From all the bacteriophage and virus other than bacteriophage spacer sequences, we expected that bacteriophage number would be higher than the virus other than bacteriophage because bacteriophages are supposed to attack bacteria more frequently. From the BLAST result, we compared the number of virus and bacteriophage spacer sequences. Among 2451 sequences, 1934 of them were of bacteriophage and 517 of them were of virus sequences (Fig-11).

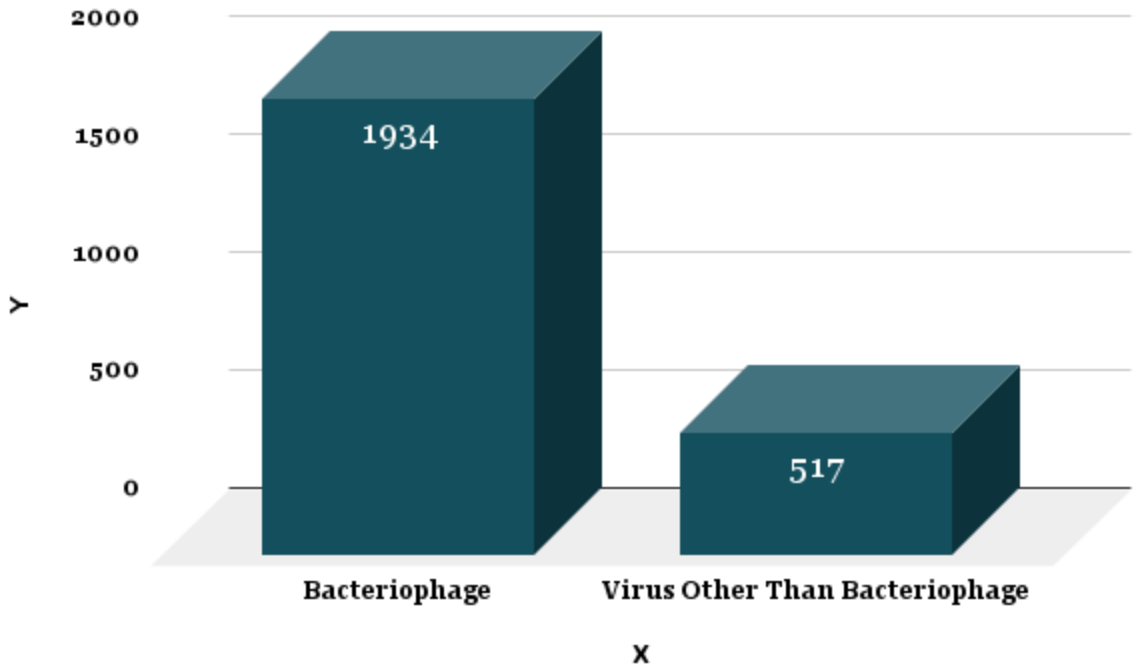


Figure 11: Bacteriophage vs Virus Other than Bacteriophage Count. The Y- axis indicates the number of spacers and the X- axis indicates the bacteriophage and virus other than bacteriophage respectively. The numbers on the columns represent the spacer's number of each type.

4.4 *Vibrio* vs Other Bacteria Count

Within the same species the transfer of mobile genetic elements is a lot easier and happens frequently. For this reason we speculated that we would get a higher number of *Vibrio cholerae* than other bacterial spacers. From the BLAST result, we compared the number of *Vibrio* and other bacterial spacer sequences. Among 2324 bacterial spacer sequences, 2245 of them were of *Vibrio* and 79 of them were of other bacteria sequences (Fig-12).

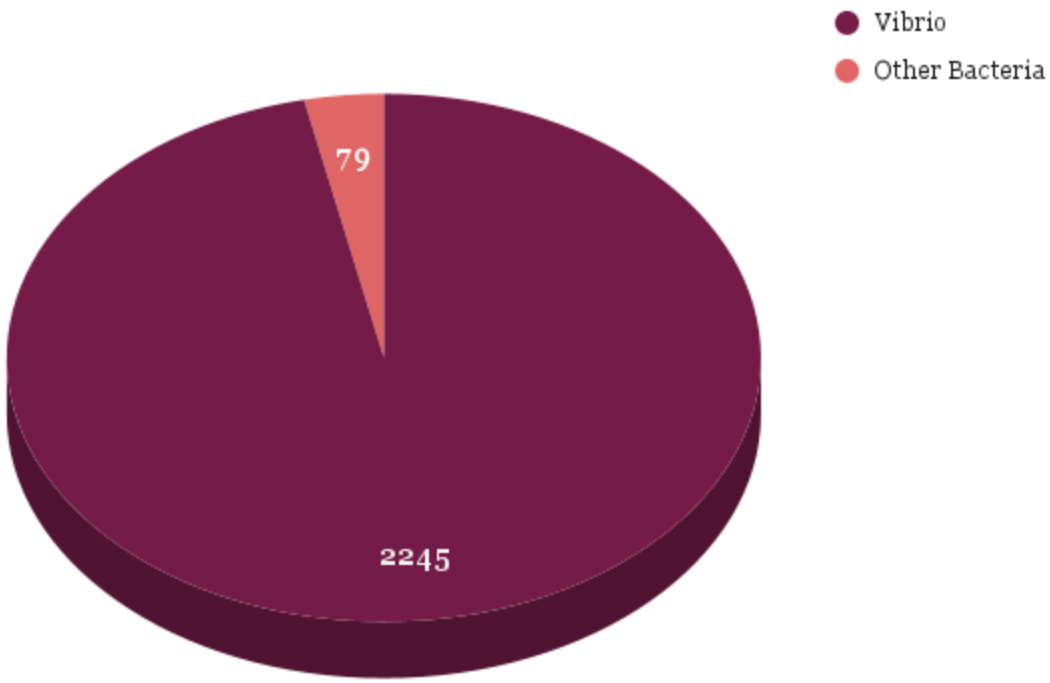


Figure 12: *Vibrio* vs Other Bacteria Count. Here, the purple pie slice represents the number of *Vibrio* bacteria spacers and the pink pie slice represents the number of other bacteria spacers.

4.5 CRISPR Type Identification

In general, bacteria contain one type of CRISPR within itself. So, we expected to get either type I or type II CRISPR for all the strains. After analyzing the contigs/nodes/scaffolds we got CRISPR type I for all the strains. However, in a few strains questionable CRISPR type II were found along with CRISPR type I when the whole genome sequences were analyzed using CRISPR-Cas++.

4.6 Presence of PICI Like Element (PLE)

As a result of evolution most of the bacteria has lost PLE and gained the CRISPR-Cas system as their defense system. So, we expected that the bacterial strains will not contain any PLE as they already have CRISPR as their defense system. However, after analyzing the 208 *Vibrio cholerae* strains along with PLE using BLAST we found that 10 of the strains contain both the CRISPR-Cas system and all the 3 PLEs.

4.7 Spacer Acquisition

We assumed that the spacer acquisition would be random as it is the natural process. However, after analyzing the spacers using BLAST we came to the conclusion that the acquisition of these spacers was biased. The spacers of top 5 phages Vibrio phage VcP032, Vibrio phage VPUSM 8, Vibrio phage X29, Vibrio phage Rostov 7 and Vibrio phage phi 2 were selected and BLAST with their respective whole genome sequence to find if most of the spacers are acquired from any specific region of the phage. We have found that 38.36% of the spacers that are acquired from Vibrio phage VcP032 comes from the 17000-23000 region (Fig-13).

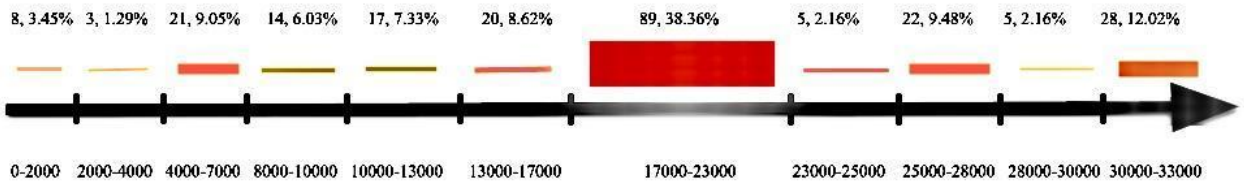


Figure 13: Spacer Acquisition Location of Vibrio phage VcP032. Here, the X axis indicates the length of the phage. The intensity of the color and thickness of the lines indicates the spacer density. The darkest and thickest red box indicates the most dense region.

In the case of Vibrio phage VPUSM 8, 31.86% of the spaces acquired from this phage come from the location 19000-24000 (Fig-14).

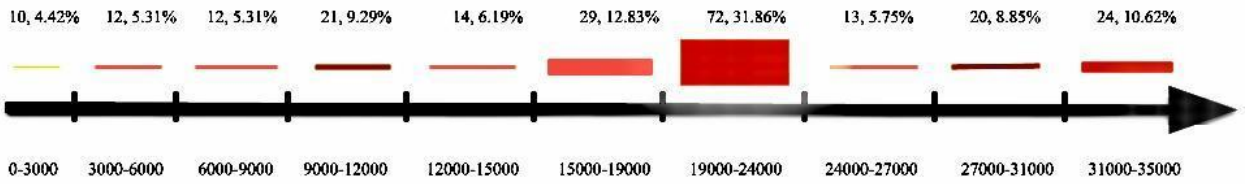


Figure 14: Spacer Acquisition Location of Vibrio phage VPUSM 8. Here, the X axis indicates the length of the phage. The intensity of the color and thickness of the lines indicates the spacer density. The darkest and thickest red box indicates the most dense region.

For Vibrio phage X29, 18.09% of the spaces acquired from this specific phage come from the 10000-14000 region (Fig-15).

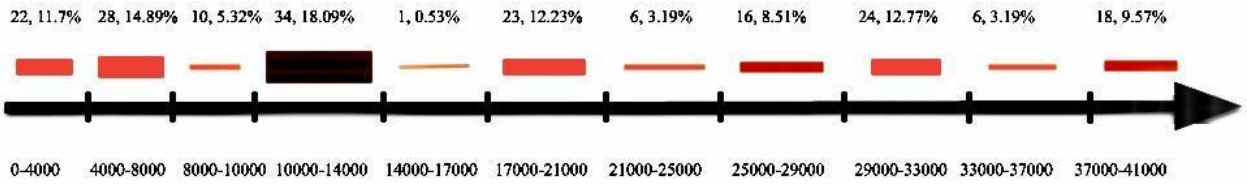


Figure 15: Spacer Acquisition Location of Vibrio phage X29. Here, the X axis indicates the length of the phage. The intensity of the color and thickness of the lines indicates the spacer density. The darkest and thickest red box indicates the most dense region.

25.56% of the spacers that are acquired from Vibrio phage Rostov 7 comes from the 30000-35000 region (Fig-16).

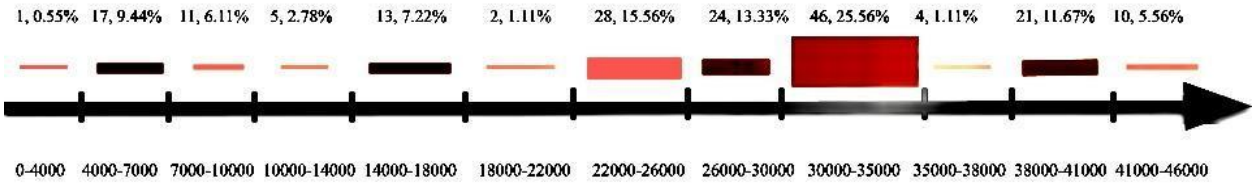


Figure 16: Spacer Acquisition Location of Vibrio phage Rostov 7. Here, the X axis indicates the length of the phage. The intensity of the color and thickness of the lines indicates the spacer density. The darkest and thickest red box indicates the most dense region.

In the case of Vibrio phage phi 2, 18.89% of the spacers that are acquired from this phage comes from the location 10000-14000 (Fig-17).

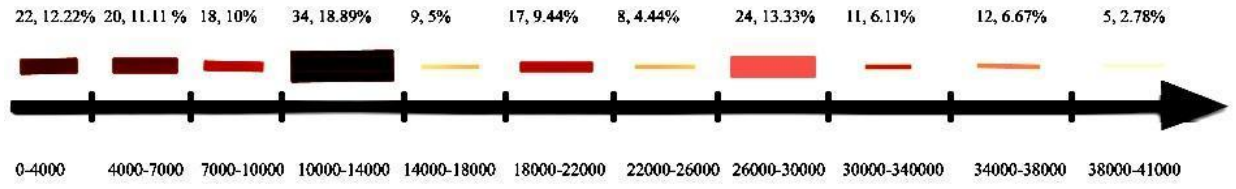


Figure 17: Spacer Acquisition Location of Vibrio phage phi 2. Here, the X axis indicates the length of the phage. The intensity of the color and thickness of the lines indicates the spacer density. The darkest and thickest red box indicates the most dense region.

4.8 RAST Annotation

We assumed that the proteins from the biased regions would be of the same type. We annotated Vibrio phage VcP032, Vibrio phage VPUSM 8, Vibrio phage X29, Vibrio phage Rostov 7 and

Vibrio phage phi 2 these top 5 phages through RAST. Most of the proteins were phage tail protein.

Table 1: Protein Analysis of Vibrio phage VcP032. The start and stop indicates the sequence length.

Start	Stop	Protein Name
16540	17550	Phage major capsid protein GpN
17566	18282	Phage terminase, endonuclease subunit GpM
18389	18850	Phage head completion-stabilization protein
18847	19335	Phage protein
19322	19981	Hypothetical Protein
19983	21092	Phage tail fibers
21092	21550	Phage tail assembly protein
21771	21998	orf27
21985	22572	Phage lysin (EC 3.2.1.17) # Phage lysozyme or muramidase (EC 3.2.1.17)
22547	22888	Hypothetical Protein
22854	23099	orf30

Table 2: Protein Analysis of Vibrio phage VPUSM 8. The start and stop indicates the sequence length.

Start	Stop	Protein Name
18621	19337	Phage terminase, endonuclease subunit GpM
19444	19905	Phage head completion-stabilization protein
19902	20390	Phage protein
20377	21036	Hypothetical Protein
21038	22147	Phage tail fibers
22147	22605	Phage tail assembly protein
22620	22829	Hypothetical Protein
22826	23053	orf27
22973	23275	repeat region
23040	23627	Phage lysin (EC 3.2.1.17) # Phage lysozyme or muramidase (EC 3.2.1.17)
23602	23943	Hypothetical Protein
23993	24154	orf30

Table 3: Protein Analysis Vibrio phage X29. The start and stop indicates the sequence length.

Start	Stop	Protein Name
9731	10234	Hypothetical Protein
10212	10502	Hypothetical Protein
10502	11971	Hypothetical Protein
11986	12507	Hypothetical Protein
12555	12767	Hypothetical Protein
12760	13038	Hypothetical Protein
13041	13202	Hypothetical Protein
13217	15853	Phage tail, tail length tape-measure protein H

Table 4: Protein Analysis of Vibrio phage Rostov 7. The start and stop indicates the sequence length.

Start	Stop	Protein Name
29896	30210	Hypothetical Protein
30210	30539	Hypothetical Protein
30554	31120	Hypothetical Protein
31117	31620	Hypothetical Protein
31598	31888	Hypothetical Protein

31888	33357	Hypothetical Protein
33372	33893	Hypothetical Protein
33944	34150	Hypothetical Protein
34143	34421	Hypothetical Protein
34424	34585	Hypothetical Protein
34600	37068	Hypothetical Protein

Table 5: Protein Analysis of Vibrio phage phi 2. The start and stop indicates the sequence length.

Start	Stop	Protein Name
9731	10234	Hypothetical Protein
10212	10502	Hypothetical Protein
10502	11971	Hypothetical Protein
11986	12507	Hypothetical Protein
12525	12767	Hypothetical Protein
12760	13038	Hypothetical Protein
13041	13202	Hypothetical Protein
13217	15853	Phage tail, tail length tape-measure protein H

Chapter 5

Discussion

5.1 Average Spacer Count

We counted the average spacers in years for 152 strains of *V. cholerae.* We expected the spacer count to increase over the years as a result of evolution. However, as the number of bacterial strains were not even in each year, the nonuniform data could not give us the expected result. Also, it could be that the time frame was not enough to show a significant change in the spacer numbers those bacterial strains have acquired in these 50+ years.

5.2 Percentage of Unknown, Bacteria and Virus Sequence

From all the *V. cholerae* strains contained a total of 7845 spacers. Among all the spacers, 4750 spacer sequences were unknown, 2262 were bacterial sequences and 833 spacers were virus sequences. After calculating, unknown 60.55%, bacteria 28.84% and virus covered 10.62% of the total spacers. More than half of the spacers were of unknown sequences. It could be that only a small fraction of living entities sequencing have been done till date. There are many bacteria, many phages, many living organisms that are still to be sequenced. If the data of all the living organisms sequencing was available then we might get all the origins of the spacers as all of those supposed to come from any living organisms. The second highest proportion was of bacterial sequences because the major fraction of them might be plasmid DNA. When a plasmid DNA inserts as a foreign DNA, if the bacteria already contains a CRISPR then it takes a huge number of spacers from the plasmid DNA. Another reason could be that the virus is in lysogenised condition in the bacteria and those DNA are also showing as bacterial DNA.

5.3 Bacteriophage vs Virus Other than Bacteriophage Count

A total of 2451 viral sequences were found. Among them, 1934 of them were bacteriophage sequences and 517 of them were viruses other than bacteriophage sequences. Maybe the reason for getting more bacteriophage sequences is that bacteriophages are more likely to attack a bacteria than any other viruses. Another reason could be that the genomic sequences of some

phages are closely related. So, one sequence of a spacer is shown to be matched with several bacteriophages. Thus, it increases the total bacteriophage count. However, the reason for getting 517 viruses other than bacteriophage sequences might be that there are also some other lysogenic or filamentous phages which do not kill bacteria but attack bacteria. Upon the attack bacteria keep those as spacers.

5.4 *Vibrio* vs Other Bacteria Count

A total of 2324 of bacterial spacer sequences were found. Among them, 2245 of them were *Vibrio* sequences and 79 of them were other bacterial sequences. We can assume some of the reasons for *V. cholerae* to acquire spacers from *Vibrio* and other bacteria. It is known that DNA or genetic elements are always entering bacterial cells. Transformation is one of the ways for a foreign DNA molecule to enter a competent bacterial cell. As we know, CRISPR is a DNA mediated immune system. So, the CRISPR system takes the foreign molecules as spacers. The reasons why some other bacterial spacer sequences were found is that in nature, bacterial cells are continuously getting degraded. When a bacterial cell gets degraded, numerous DNA fragments get released in the environment. With transformation these DNA fragments can enter into a competent cell. For example, if an *E. coli* phage attacks an *E. coli* bacterial cell, the cell bursts open and many phages get released into the environment. At the same time, the fragments of the *E. coli* DNA also get released into the environment. Those *E. coli* DNA fragments can enter into a *V. cholerae* cell. Here, we got different strains of *E. coli, E. albertii, Proteus mirabilis* etc as *V. cholerae* spacers. This is how other bacterial DNA spacer sequences were found in *V. cholerae.* Moreover, the most prominent spacers among *Vibrio* were of *V. cholerae*. Within the same species the transfer of mobile genetic elements is a lot easier and happens frequently. When these *V. cholerae* mobile genetic elements enter another *V. cholerae* cell, it takes those MGEs as spacers. This might be how *Vibrio* and other bacterial spacer sequences were acquired by *V. cholerae*.

5.5 CRISPR Type Identification

In case of identifying the strain's CRISPR type, both the contigs/nodes/scaffolds and whole genome sequence were used. We expected that we would get either type I or type II in each strain as it is rare to contain both types together. After analyzing the sequences using CRISPR-Cas++, we found that all of the contigs/nodes/scaffolds contained CRISPR type I. The similar result was given when we analyzed the whole genome sequences of the bacterial strains. However, in the case of whole genome sequence analysis we found a few questionable CRISPR type II along with type I. Further analysis is needed to confirm the presence of type II CRISPR in these strains. It is known that the presence of Cas9 determines the type II CRISPR. However, in the software CRISPR-Cas++ we found out that it was showing type II without having Cas9.

5.6 Presence of PICI Like Element (PLE)

Some of the bacteria contain PICI Like Elements (PLE). Among them, *V. cholerae* El Tor biotype strains contain PLE instead of a CRISPR-Cas system to defend themselves against phage (Naser, I. B., et al, 2017, November 1). When phage infects bacteria PLE acts as their 1st line of defense. PLE is a mobile genomic island. When the phage DNA enters the bacteria, the PLE comes out of the chromosome as plasmid. This plasmid interacts with the phage genome and inactivates it. We expected that *V. cholerae* non-O1 will not contain any PLE. As a result of evolution it has lost PLE and gained the CRISPR-Cas system as their defense line. However, after analyzing all the *V. cholerae* strains, we found that 10 of the strains among the 208 contained all 3 of the PLE along with the CRISPR-Cas system. It might be that, due to the presence of the CRISPR-Cas system the PLE sequences are randomly dispersed among them and so only few of them could carry both PLE and CRISPR sequences.

5.7 Spacer Acquisition

When a bacteria is attacked by a phage or other invader it's sequence is inserted into the CRISPR array and later works as a memory. These spacer sequences transcribe when they are attacked by the same previous invaders. It is natural for the acquisition of spacers to be random. In our case, most of the spacers come from genetic sequences. Also, they were from a specific region of that phage. So, it can be said that when the *V. cholerae* CRISPR system acquires spacers it shows a clear bias. We have also found that, among the top 5 phage spacer sequences, phage tail proteins were prominent. Naser et al., (2017), stated that upon invading the bacteria, bacteriophages interact with the PLE sequences of bacteria first. Thus, it will take PLE sequences as spacers. It can be assumed that when attacking a bacteria, initially the phage attaches to the bacteria surface with its tail. As bacteria encounter the phage tail first so they take those sequences as spacers. As these sequences were important, they were taken as spacers first.

Chapter 6

Conclusion

6.1 Limitation

- As the number of bacterial strains were not the same in each year, the data did not give any significant result. Besides, not every strain had the information on the year of isolation. This has also affected the result.
- Due to the high number of unknown spacer sequences, it did not give a clear representation of the percentage of bacteria and virus spacer sequences.
- In general, type I CRISPR consists of the Cas3 gene and type II CRISPR consists of the Cas9 gene. However, the CRISPR-Cas++ software gave results that did not confirm the presence of Cas3 or Cas9 to determine the types. So, this result might not be the most accurate.

6.2 Recommendation

- Further analysis should be conducted for the identification of the unknown spacer sequences.
- To confirm the CRISPR type, the presence of Cas3 and Cas9 should be detected.
- For in depth analysis, the source of all hypothetical proteins should be confirmed.

References

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., ... Banfield, J. F. (2020, February 12). Clades of huge phages from across Earth's ecosystems. Nature. https://www.nature.com/articles/s41586-020-2007-4

Bacteriophage. ScienceDirect (n.d.). https://www.sciencedirect.com/topics/engineering/bacteriophages

BACTERIOPHAGES – AN INTRODUCTION TO PHAGES (n.d.). Bacteriophages.news https://www.bacteriophage.news/bacteriophages-an-introduction-to- phages/

Bacteriophage. (n.d.). Encyclopaedia Britannica. https://www.britannica.com/science/bacteriophage.

Barrangou, R., & Marraffini, L. A. (2015, April 24). CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. PMC. https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC4025954/

Cholera. World Health Organization. (2021, February 5.). World Health Organization. https:// www.who.int/news-room/fact-sheets/detail/cholera

Claeson, M., & Waldman, R., (n.d.). Encyclopaedia Britannica. https://www.britannica.com/sciencecholera

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC bioinformatics*, *8*, 172. https://doi.org/10.1186/1471-2105-8-172

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., & Qadri, I. (2012, January 10). Bacteriophages and their implications on future biotechnology: A review. Virology journal. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398332/.

Harada, L. K., Silva, E. C., Campos, W. F., Fiol, F. S. D., Vila, M., Dąbrowska, K., Krylov, V. N., & Balcão, V. M. (2018, April 30). Biotechnological applications of bacteriophages: State of the art. Microbiological Research. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0944501318301332.

Hatfull, G. F., & Hendrix, R. W. (2012, October 1). Bacteriophages and their genomes. Current opinion in virology. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC3199584/.

Huge bacteria-eating viruses close gap between life and non-life. ScienceDaily. (2020, February 12). ScienceDaily. https://www.sciencedaily.com/releases/2020/02/ 200212131458.htm

Iftikhar, N. (2019, January 14). What Is Phage Therapy?. Healthline. https://www.healthline.com/health/phage-therapy

Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *Journal of bacteriology*, *200*(7), e00580-17. https://doi.org/10.1128/JB.00580-17

Karimi, Z., Ahmadi, A., Najafi, A., & Ranjbar, R. (2018). Bacterial CRISPR Regions: General Features and their Potential for Epidemiological Molecular Typing Studies. *The open microbiology journal*, *12*, 59–70. https://doi.org/10.2174/1874285801812010059

Karginov, F. V., & Hannon, G. J. (2010). The CRISPR system: small RNA-guided defense in bacteria and archaea. *Molecular cell*, *37*(1), 7–19. https://doi.org/10.1016/j.molcel.2009.12.033

Lin, D. M., Koskella, B., & Lin, H. C. (2017, August 6). Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. World journal of gastrointestinal

pharmacology and therapeutics. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547374/.

Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 6.3, Viruses: Structure, Function, and Uses. https://www.ncbi.nlm.nih.gov/books/NBK21523/

Makarova, K. S., & Koonin, E. V. (2015). Annotation and Classification of CRISPR-Cas Systems. *Methods in molecular biology (Clifton, N.J.)*, *1311*, 47–75. https://doi.org/10.1007/978-1-4939-2687-9_4

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nature reviews. Microbiology, 13(11), 722–736. https://doi.org/10.1038/nrmicro3569

Naser, I. B., Hoque, M. M., Nahid, M. A., Tareq, T. M., Rocky, M. K., & Faruque, S. M. (2017, November 1). Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of *Vibrio cholerae* in Bangladesh. Nature. https://www.nature.com/articles/s41598-017-14839-2.

Sapkota, A. (2020, December 28). Bacteriophage- definition, structure, life cycles, Applications, phage therapy. Microbe Notes. https://microbenotes.com/bacteriophage/

Shabbir, M. A. B., Shabbir, M. Z., Wu, Q., Mahmood, S., Sajid, A., Maan, M. K., Ahmed, S., Naveed, U., Hao, H., & Yuan, Z. (2019, July 5). CRISPR-cas system: Biological function in microbes and its use to treat antimicrobial resistant pathogens. Annals of Clinical Microbiology and Antimicrobials. BMC https://ann-clinmicrob.biomedcentral.com/articles/10.1186/s12941-019-0317-x

Steward, K. (2018, August 28). Lytic vs Lysogenic – Understanding Bacteriophage Life Cycles. Technology Networks. https://www.technologynetworks.com/immunology/articles/lytic-vs-lysogenic-understanding-bacteriophage-life-cycles-308094

Summers, W. C. (2012, April 1). The strange history of phage therapy. Bacteriophage. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3442826/.

Taylor, M. W. (2014, July). The Discovery of Bacteriophage and the d'Herelle Controversy. ResearchGate. https://www.researchgate.net/publication/312755787_The_Discovery_of_Bacteriophage_and_the_d%27Herelle_Controversy

Vibrio cholerae. (n.d.). ScienceDirect. https://www.sciencedirect.com/topics/medicine-and-dentistry/vibrio-cholerae

What are genome editing AND Crispr-cas9?. (2020, September 18). MedlinePlus. U.S. National Library of Medicine. NIH. https://medlineplus.gov/genetics/understanding/genomicresearch/genomeediting/.

Zhan, Y., & Chen, F. (2018, August 29). The smallest ssDNAphage infecting a marine bacterium. environmental microbiology. Society for Applied Microbiology. https://sfamjournals.onlinelibrary.wiley.com/ doi/abs/10.1111/1462-2920.14394