Genomic Analysis of the Correlation between CRISPR Self-Targeting Spacers and Prophage-related Pathogenicity in *Vibrio cholerae* and *Staphylococcus aureus*

By

Protik Roy
ID: 22376002

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
October 2024

Declaration

It is hereby declared that,

1. The thesis submitted is my original work while completing m y 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 that 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 issues in an international journal.

Protik Roy

Protik Roy
ID: 22376002

Approval

The thesis " Genomic Analysis of the Correlation Between CRISPR Self-Targeting Spacers and Prophage-related Pathogenicity in *Vibrio cholerae* and *Staphylococcus aureus*" Submitted by Protik Roy (22376002) of Summer 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Biotechnology on 24th October 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Iftekhar Bin Naser
Associate Professor, MNS Department
BRAC University

Co-Supervisor:
(Member)

Tushar Ahmed Shishir
Lecturer, MNS Department
BRAC University

Program Director:
(Member)

Dr. Munima Haque
Associate Professor, MNS Department
BRAC University

External Expert
Examiner: (Member)

Dr. Zahid Hayat Mahmud
Scientist, Environmental Micro
Health Systems and Population Studies Division,
International Centre for Diarrheal Disease Research, Bangladesh (icddr,b)

Departmental Head:
(Chair)

Dr. Md. Firoze H. Haque
Associate Professor, MNS Department
BRAC University

Ethics Statement

No human or animal model was used in this study.

Abstract

The interactions between bacteria and phages are intricate and dynamic, manifesting through two main life cycles: the lytic cycle, in which phages destroy their hosts via cell lysis, and the lysogenic cycle, where phage DNA integrates into the bacterial genome, allowing for stable coexistence without immediate harm to the host. This interaction is often described as a competitive struggle, where bacteria evolve various defense strategies to evade phage attacks. In response, phages have developed strategies to bypass these defenses, securing their survival and replication. This investigation employed an in-silico approach to examine the pathogenicity of bacteria in relation to the CRISPR-Cas system and self-targeting spacers. For the analysis, two bacterial species were chosen: *Vibrio cholerae*, comprising 1794 strains, and *Staphylococcus aureus*, consisting of 16286 strains. Both strains demonstrated significantly distinct results across almost all parameters. The average pathogenic gene count with CRISPR and pathogenic gene count with STS demonstrated a notable decrease in *V. cholerae* strains. In contrast, *S. aureus* displayed a minor reduction in average pathogenic gene count with CRISPR (1.55), while revealing a significantly elevated average count in the presence of STS (2.50). A significant decrease in the pathogenic gene count of *V. cholerae* was recorded with CRISPR (from 1.12 to 0.13) and with STS (to 0.024), suggesting that the active CRISPR-Cas system may effectively target and interfere with the expression of genes linked to pathogenicity. Self-targeting spacers (STS) indicate a potential disruption of essential or pathogenicity-related genes, leading to a notable decrease in the count of pathogenic genes, which may be linked to auto-immunity or self-destruction mechanisms. In *S. aureus*, the CRISPR-Cas system appears to demonstrate a less pronounced suppression of pathogenic gene expression when contrasted with *V. cholerae*, as evidenced by a reduced count with CRISPR, suggesting a minimal regulatory influence on virulence factors. The significant increase in pathogenic gene count in *S. aureus* with STS indicates that STS could promote horizontal gene transfer or activate pathogenicity islands, leading to heightened virulence. This may occur if STS disrupts regulatory elements that govern the integration of virulence genes or if the bacteria compensate for STS by acquiring additional virulence factors. In addition, the *S. aureus* bacteria might employ some unknown mechanism to co-exist with STS-containing systems, increasing the overall pathogenic gene count of this bacteria.

Keywords: Bacteria**,** Bacteriophage, Prophage, *Vibrio cholerae, Staphylococcus aureus,* Genome, CRISPR, Self-Targeting Spacers (STS), Pathogenic genes

Dedication

Dedicated to my family and friends for their unwavering support and encouragement.

Acknowledgment

All my acknowledgments and gratitude are for the Almighty Creator for his glorious creation, where he not only made me a human being but also a sensible one to choose between better and worse. I am thankful to the great almighty till eternity.

I would like to thank Dr. Md. Firoze H. Haque, Associate Professor and Chairperson of the Department of Mathematics and Natural Sciences, BRAC University, for allowing me to continue my work at the BRAC University Bioinformatics laboratory.

I would like to especially thank Dr. Munima Haque, Associate Professor and Program Director, Department of Mathematics and Natural Sciences, BRAC University for the inspiration and guidance she provided me as a student.

I feel proud to express my deepest sense of gratitude and profound appreciation to my respected supervisor, Dr. Iftekhar Bin Naser, Associate Professor, Department of Mathematics and Natural Sciences, BRAC University, for his constant supervision, scholastic guidance, valuable suggestions, constructive criticisms and kind help throughout this research work and in preparing the manuscript of this thesis and presentation slides.

I sincerely appreciate my co-supervisor, Tushar Ahmed Shishir, Lecturer, Department of Mathematics and Natural Sciences, BRAC University, for his unwavering inspiration, encouragement, support, and guidance in my endeavors.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acr	Anti-CRISPR
AMR	Antimicrobial Resistance
Blast	Basic Local Alignment Search Tool
bp	Base pair
Cas gene	CRISPR associated gene
CRISPR	Clustered regularly interspaced short palindromic repeats
DNA	Deoxyribonucleic acid
ds	double-stranded
FASTA	Fast Adaptive Shrinkage Threshold Algorithm
LPS	Lipopolysaccharide
lpt	phage-encoded lipoprotein
Pf	Pathogenic factor
RefSeq	Reference Sequence
RNA	Ribonucleic acid
STS	Self-targeting spacer

Chapter 1
INTRODUCTION

1.1 Bacteria and their Immune System

Nathan Wolfe said, "If aliens came to Earth, they would most likely study humans and try to find out more about the dominant form of life on our planet." The American virologist is pointing to the massive microorganisms that have evolved to live in a variety of habitats, including harsh ones (such as extremophiles). In order to be effectively adjusted, they need to use a variety of tactics. Occasionally, bacteria may experience modifications in their cell wall development, mutations or deletions in their genetic architecture, mutualistic relationships with their hosts, and other physiological changes. In addition to this, they have used a variety of immune response tactics to defend themselves against external threats. These tactics may resemble processes that involve trial and error. Natural selection will be obliged to either use a new approach or reassign the older one if one strategy fails.

This thesis will focus on bacteria, their phages, CRISPR-Cas, self-targeting spacers and pathogenicity. Estimates suggest that there are more than 10^{31} phages on Earth, surpassing the combined numbers of bacteria and other organisms, thereby establishing them as the most abundant biological entities in the biosphere(Chevallereau et al., 2022). Bacteriophages are viruses that infect bacteria and replicate within them. Viruses may possess single-stranded or double-stranded nucleic acids and can replicate through either lytic or lysogenic pathways. The host cell wall is frequently lysed, resulting in the release of new phage particles. Infecting bacteria and evading the bacterial immune system are prerequisites for the successful multiplication of phages. Phages employ a variety of strategies to achieve this. Methylation is commonly utilized by gut phages to circumvent bacterial immunity, as exemplified by Dazbok phages of *Bacillus mycoides* (Wei & Zhou, 2024) .

Furthermore, phages have developed strategies to circumvent bacterial immunity. Phages possess the capability to encode various anti-defense proteins (Gad1, Tad1, Tad2, and Had1) that can obstruct bacterial defense mechanisms, such as the Thoeris, Gabija, and Hachiman systems(Gao & Feng, 2023). Phages can effectively infect host cells while evading bacterial defenses. Phages can neutralize bacterial defense mechanisms by encoding proteins that inhibit the cleavage of phage-derived DNA or that sequester immune signaling molecules generated by bacterial defense systems in response to phage infection. Additionally, in response to bacterial adaptive immunity, there are reports of the production of Acr proteins.

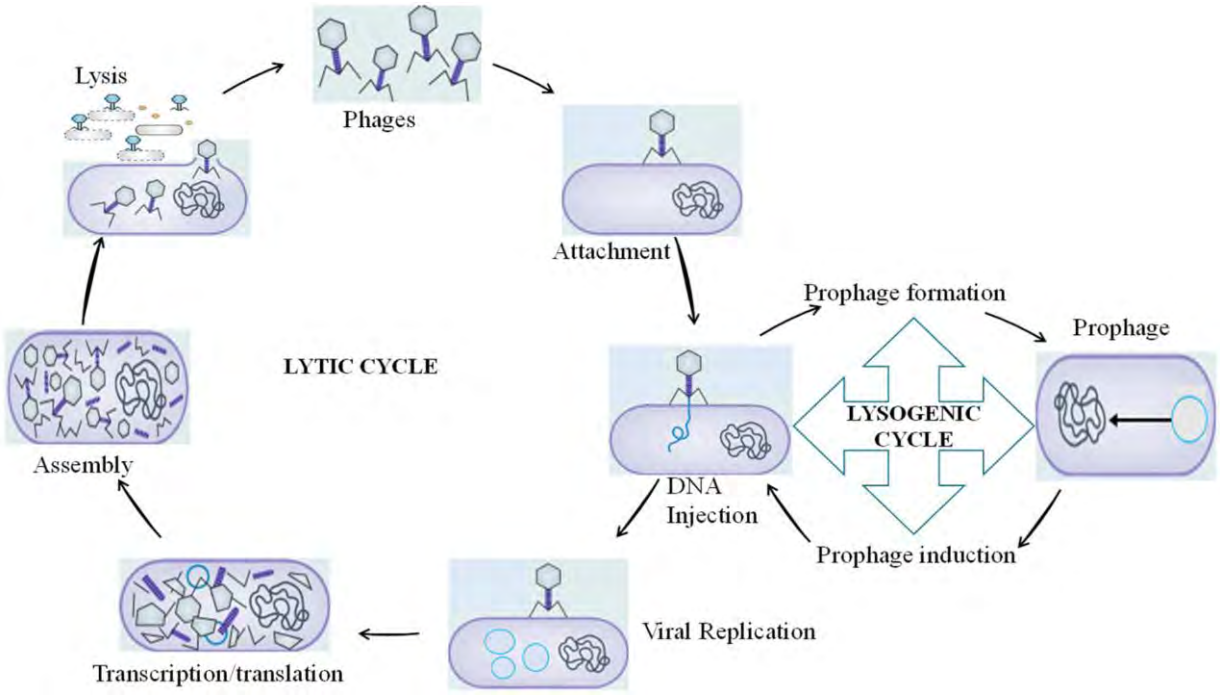


***Figure 1**: Lytic and lysogenic pathway of bacteriophages*

(Source: https://pubmed.ncbi.nlm.nih.gov/37896176/)

Bacteria possess a variety of mechanisms to defend against foreign invaders. Similar to eukaryotes, the bacterial immune system can be categorized into two types: innate and adaptive immunity. Figure 2 provides a concise overview of the immune responses exhibited by the bacteria.

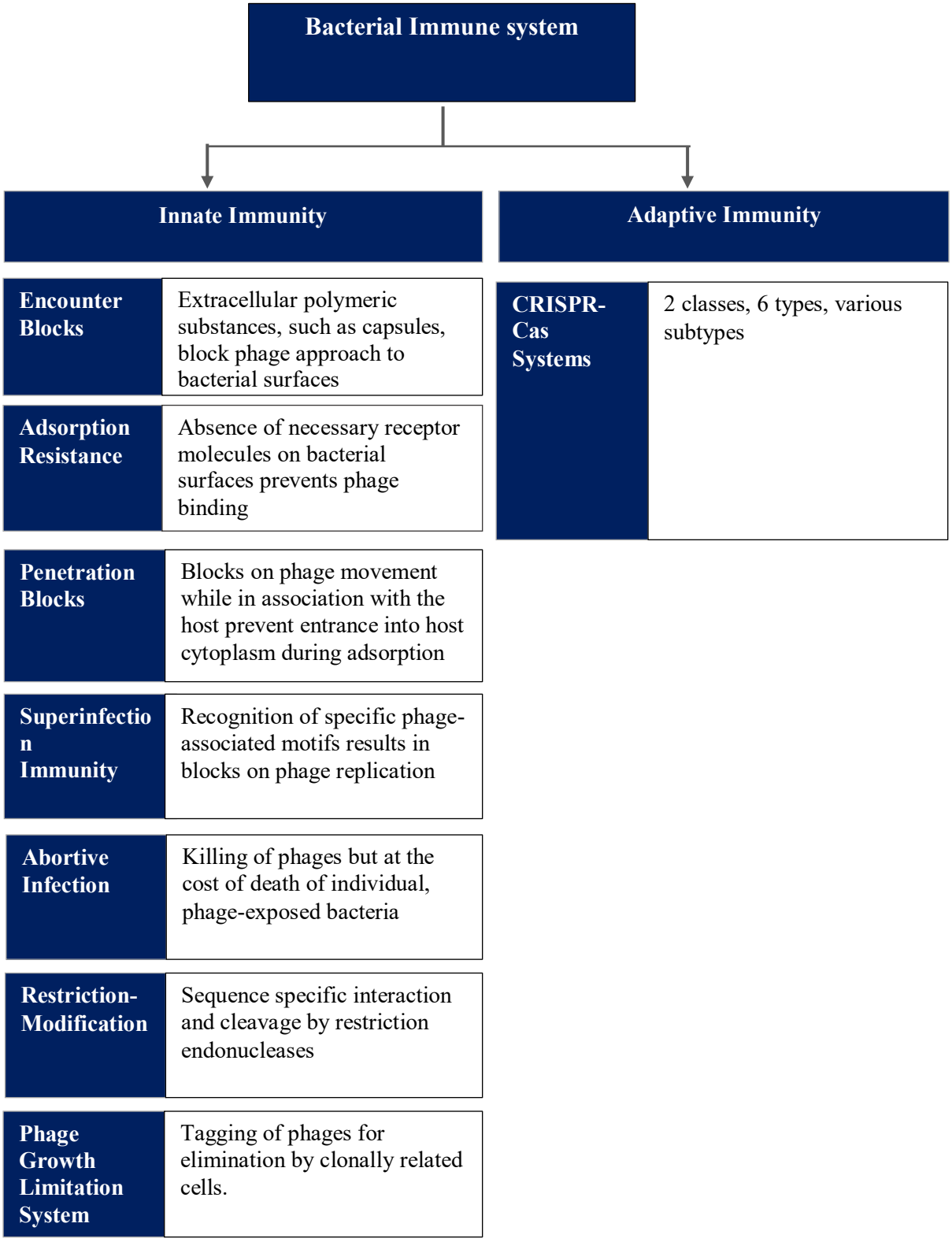


***Figure 2**: Immune responses by bacteria*

1.2 Bacterial Adaptive Immune System: CRISPR-Cas

"CRISPR" denotes Clustered Regularly Interspaced Short Palindromic Repeats, an adaptive defensive mechanism bacteria and archaea utilize. In 2000, it was revealed that the CRISPR array was associated with many actively transcribed genes. Researchers discovered that CRISPR sequences constitute a component of the adaptive Immune systems in bacteria and archaea. These organisms identify and defend against further infections by the same virus by employing CRISPR sequences to acquire and retain viral DNA pieces.

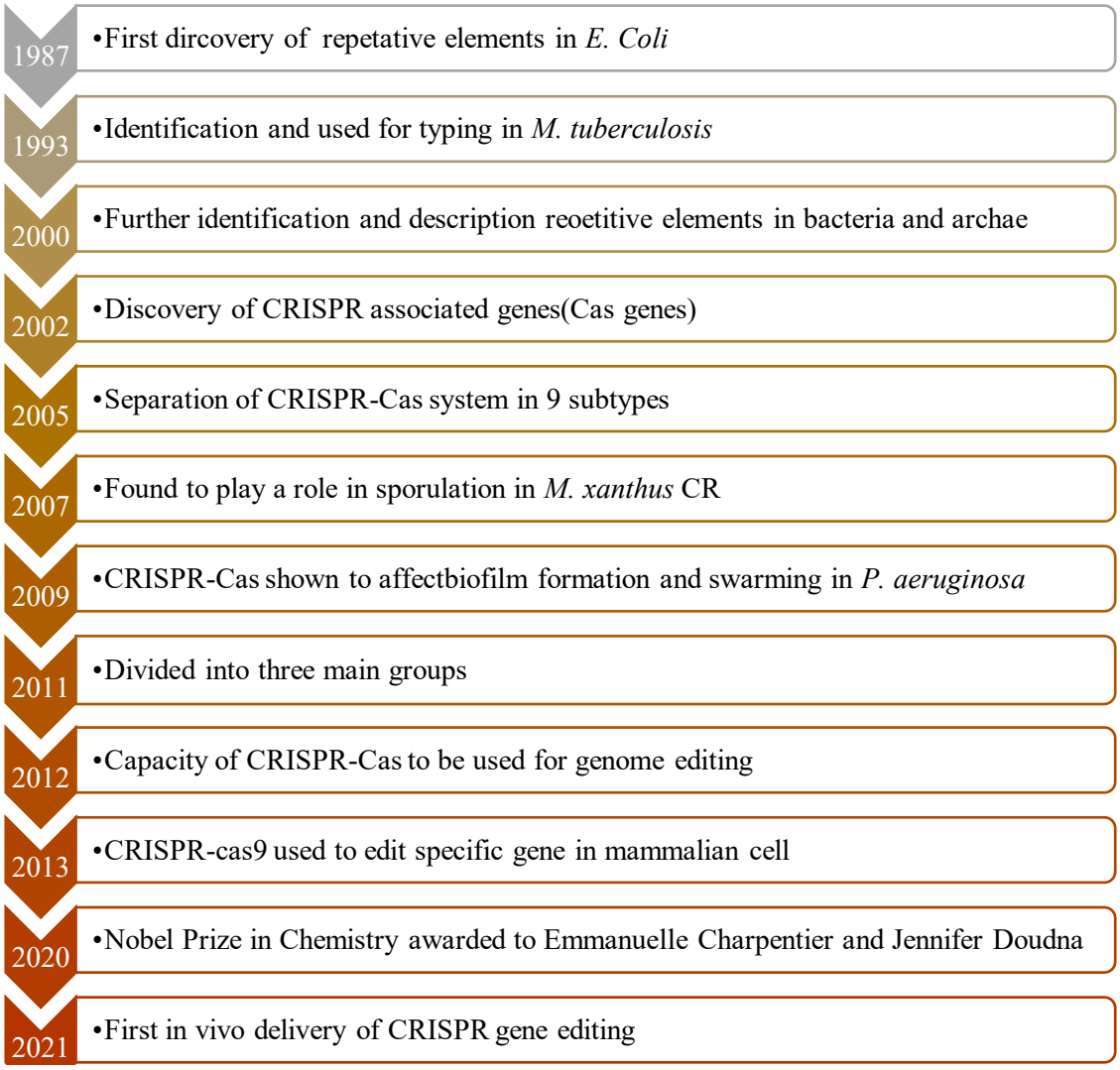


***Figure 3**: The landmarks in CRISPR research*

Subsequent research has revealed the involvement of CRISPR in the bacterial immune system. The landmarks in CRISPR research are summarized in figure 3. Researchers Ruud Jansen, Philippe Horvath, and Francisco Mojica identified that CRISPR-associated (Cas) proteins play

a role in cleaving viral DNA (Ishino et al., 2018). The Cas9 protein's capacity to split DNA rendered it particularly significant. In 2012, Jennifer Doudna and Emmanuelle Charpentier designed the CRISPR-Cas system to facilitate gene editing. By amalgamating tracrRNA and crRNA into a singular guide RNA (sgRNA), they optimized the system and demonstrated that Cas9 from Streptococcus pyogenes could be directed to a specific location to enable double-strand DNA cleavage. This discovery rapidly elevated CRISPR-Cas9's status as a powerful genome editing instrument. In 2013, Feng Zhang and George Church independently demonstrated CRISPR-Cas9's capability to alter genes in mammalian cells (Cong et al., 2013). This discovery illuminated the potential of CRISPR for various applications in genetics and biotechnology. Jennifer Doudna and Emmanuelle Charpentier were awarded the 2020 Nobel Prize in Chemistry for their pioneering development of CRISPR-Cas9 as a genome editing tool.

1.3 CRISPR Components and Types

The four fundamental components of the CRISPR locus are the Cas genes, the AT-rich leader sequence, spacers, and repeats. Cas genes exhibit diversity and vary across different types. However, the Cas1 and Cas2 genes are nearly conserved across all types. CRISPR arrays undergo processing into functional RNA molecules facilitated by Cas proteins, which subsequently guide these RNAs to target complementary sequences in foreign DNA or RNA (Nuñez et al., 2014). Cas gene cassettes are preceded by a leader sequence.

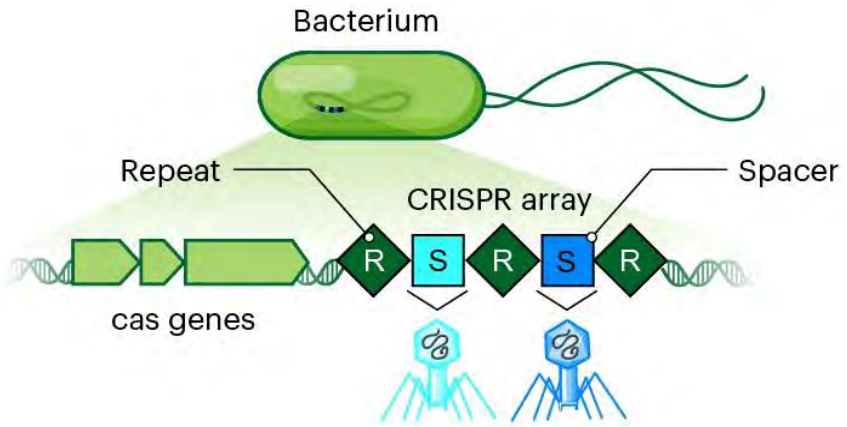


***Figure 4**: Components of CRISPR*

(Source: https://innovativegenomics.org/crisprpedia/crispr-in-nature/)

The promoter elements within the leader sequence facilitate the transcription of CRISPR arrays, ensuring the expression of defense-related CRISPR RNAs (crRNAs) (Bernal-Bernal et al., 2018).The conserved sequence motifs of the leader sequence facilitate adaptation by

directing the Cas1 and Cas2 proteins to the correct integration site on the CRISPR array(Kieper et al., 2019). The leader sequence is essential for the proper functioning of the CRISPR-Cas system, as it coevolves with Cas proteins and CRISPR repeats. The subsequent component consists of spacers, which are distinct sequences ranging from 18 to 72 nucleotides (M. Li et al., 2017), interspersed with palindromic repeats. They serve as immunological memory, enabling the host to identify and target previously encountered pathogens. The subsequent component repeats are crucial for the biogenesis of CRISPR RNA (crRNA) molecules, derived from the spacer sequence of the CRISPR transcript.

CRISPR systems are categorized into two primary classes and six types according to the properties of the associated CRISPR (Cas) proteins (Makarova et al., 2015). Some of these kinds target DNA, whereas others target RNA.

Table 1. *Different classes of CRISPR and their types*

CRISPR class	**Types**	**Signature protein**	**Mechanism of action**
Class 1	**Type I**	Cascade protein complex	Uses a multi-subunit complex (Cascade) to bind and unwind the target DNA, followed by cleavage by the Cas3 protein.
	Type III	Csm/Cmr complexes	Target both DNA and RNA, depending on the subtype. Involves multi-protein complexes with different Cas proteins
	Type IV	Csf proteins	Not well-characterized but involves interference with mobile genetic elements
Class 2	**Type II**	Cas9	Uses a single guide RNA (sgRNA) to direct Cas9 to the target DNA, where it creates a double-strand break
	Type V	Cpf1 (also known as Cas12)	Cas12a and related proteins create staggered cuts in the target DNA, which can be useful for certain types of genetic modifications
	Type VI	C2c2 (also known as Cas13)	Cas13a and related proteins bind and cleave target RNA, offering potential for RNA manipulation and antiviral applications

CRISPR arrays offer a unique perspective on the genetic material that prokaryotic evolution encountered and chose to avoid. As a molecular fossil record of previous invasions, we can deduce the sequence in which a particular organism encountered different genomes. Consequently, there has been a lot of interest in studying the spacer sequences' origin. Eighty to ninety percent of spacers that can be mapped to sequenced genomes do so to phage genomes. Genes linked to mobile genetic elements are matched by the remaining mapped spacers. Unexpectedly, the great majority of spacer sequences (>90%) come from unidentified sources, including the CRISPR "dark matter" (McGinn & Marraffini, 2019).

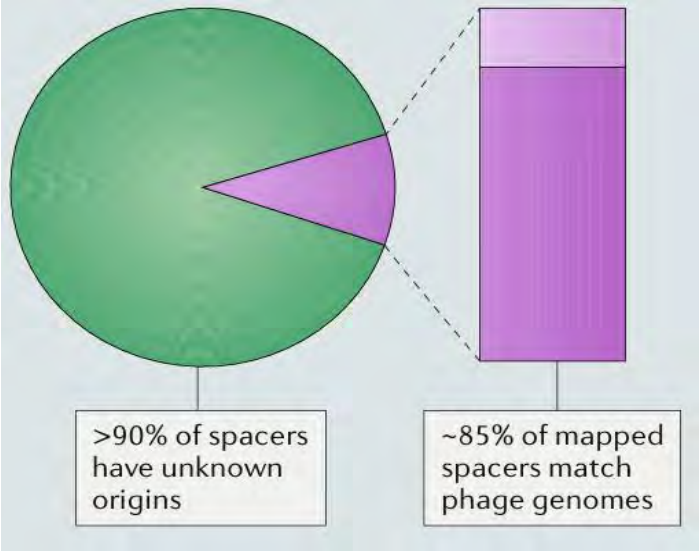


***Figure 5**: Dark matter of CRISPR spacers*

(Source: https://www.nature.com/articles/s41579-018-0071-7)

Probably the most reasonable explanation for these enigmatic spacer sequences is that a wide variety of phage and mobile genetic element (MGE) sequences are absent from sequence databases. Numerous strategies exist for the acquisition of spacer sequences, such as naïve acquisition and primed adaptation. They are mostly obtained from phages, plasmids, and other mobile genetic elements (MGEs). During the acquisition of spacers and their integration into the genome for memory retention, bacteria occasionally struggle to distinguish between self and non-self-entities. Consequently, they occasionally obtain their own DNA fragment and incorporate it into their CRISPR locus. This sort of spacer is referred to as self-targeting spacers (STSs) (Stern et al., 2010).

The role of the spacer is to direct the CRISPR complex during the CRISPR-Cas dependent destruction of foreign entities. Conversely, the self-targeting spacers target the host's DNA,

resulting in cell death or disintegration. Consequently, establishing the autoimmunity. Besides autoimmunity, numerous studies have demonstrated the significance of STS in endogenous gene regulation (Wimmer & Beisel, 2019). This generated discussion regarding the precise function of self-targeting spacers. Moreover, STSs have been identified as playing an evolutionary role in the bacterial genome via affecting genome remodeling (F et al., 2018). To evade autoimmune responses, the host has developed various mechanisms, such as the complete removal of the CRISPR array, rendering it non-functional, mutating the CRISPR array, eliminating self-targeting spacers, removing Cas gene operons, and utilizing anti-CRISPR (Acr) proteins. Here we have utilized two group of bacteria e.g., *V. cholerae* and *S. aureus*. *Vibrio cholerae* is a gram-negative, comma-shaped bacterium primarily responsible for cholera, an acute diarrheal disease. It is predominantly transmitted via the fecal-oral route, often through contaminated water or food. The bacterium has over 200 identified serogroups, but only serogroups O1 and O139 are associated with epidemic cholera outbreaks (Ramamurthy et al., 2022). On the other hand, *Staphylococcus aureus* is a gram-positive coccal bacterium commonly found on the skin and mucous membranes of humans. While many strains are harmless, some can cause a wide range of infections, from minor skin infections to severe conditions like pneumonia and sepsis (Tong et al., 2015). Here, we tried to investigate the CRISPR system, self-targeting spacers (STS) and prophage count in both bacteria to see the relation with pathogenic factor genes.

1.4 Understanding the CRISPR- Cas Defense System

In a manner akin to eukaryotes, prokaryotes possess both innate and adaptive immunity (Dimitriu et al., 2020). Innate immunity operates primarily through the action of endonuclease cleavage targeting foreign invaders, a process known as Restriction-Modification (RM). The adaptive immune defense operates via the CRISPR system, acquiring memory in the form of spacer sequences. Approximately RM has been identified in 90 percent of sequenced prokaryote genomes. Conversely, CRISPR-Cas systems have been identified in 30-40 percent of bacterial genomes and 90 percent of archaeal genomes, respectively (Marraffini & Sontheimer, 2010). This section will concentrate on the second type and the function of spacer sequences. Initially, we will examine the fundamental mechanism of CRISPR and the process by which spacers are acquired. The CRISPR-Cas mediated adaptive immune response is categorized into three distinct stages: adaptation, expression, and interference.

Adaptation: During the adaptation phase, protospacers from foreign invaders are integrated into the host genome. The inclusion is facilitated by Cas1 and Cas2 proteins (Nuñez et al., 2014). Exceptions have also been documented. In the type V-C system, the overall process is mediated by a putative effector molecule C2C3 and a Cas homologue. Additionally, the Cas4 protein is essential throughout the adaption phase. In a type II-A system, the acquisition of spacers necessitates Csn2, Cas9, and tracr RNA (Amitai & Sorek, 2016). In summary, different types of CRISPR systems necessitate varied combinations of proteins. In most instances, Cas1 and Cas2 facilitate spacer acquisition. Diverse strategies for the selection of proto spacers in various prokaryotes have been documented. The target sequences incorporated into the CRISPR locus are not selected arbitrarily. The protospacer adjacent motif (PAM), a brief sequence located adjacent to the protospacer (shown in figure 6) in type I, type II, and type V CRISPR-Cas systems, is critical for both acquisition and interference (Gleditzsch et al., 2018). The PAM-recognizing domain of Cas9 is responsible for protospacer selection in type II-A CRISPR-Cas systems.

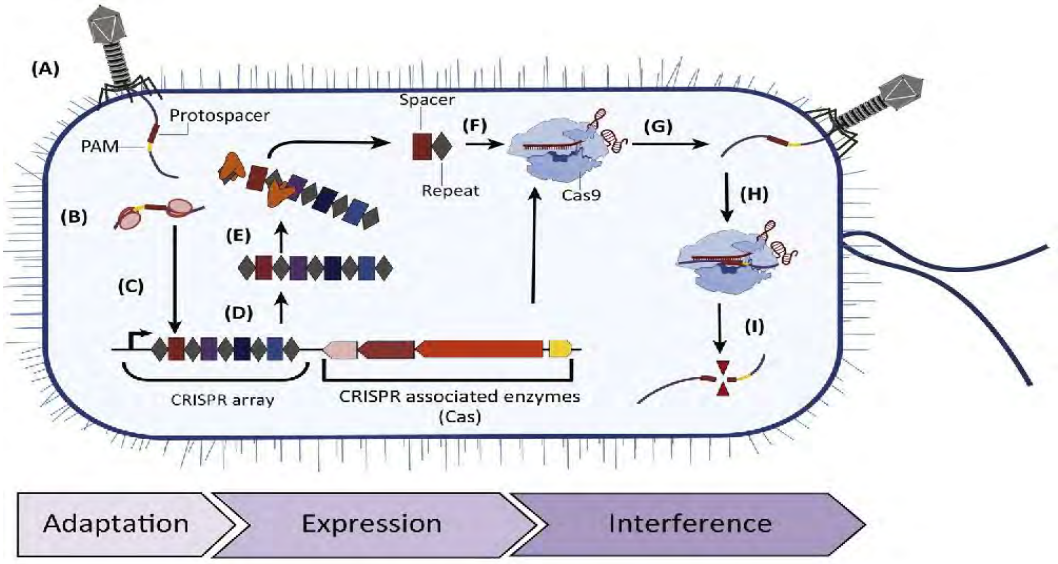


***Figure 6**: Basic mechanism of CRISPR-Cas mediated immunity*

(Source: https://www.semanticscholar.org/paper/Advances-in-Industrial-Biotechnology-Using-Systems.-Donohoue-Barrangou/5a03dd665dc085b07976158b6bddececd12ccf24)

Expression: The subsequent phase is expression, specifically the expression of the CRISPR array to generate crRNA, which undergoes multiple processing stages to produce mature guide RNA (gRNA). crRNA is bordered by a 5' tag, and the processing stages are typically executed by Cas6 proteins (Hille & Charpentier, 2016). However, an exception is observed in the type

I-C system, which utilizes the Cas5d protein in place of the Cas6 protein (Tsui & Li, 2015). Furthermore, tracrRNA is an essential component in the processing of crRNA.

Interference: Mature crRNAs function as guides to accurately disrupt invasive nucleic acids during the last phase of immunization. In class 1 systems, Cascade-like complexes facilitate target degradation, but in class 2 systems, target interference is accomplished with a single effector protein (Chylinski et al., 2014). Type I, II, and V systems explicitly delineate the PAM sequence located upstream (types I and V) or downstream (type II) of the protospacer to avert self-targeting (Doudna & Charpentier, 2014). Alongside the PAM sequence, the adjacent region known as the "Seed" is essential for target degradation. The seed sequence is an 8-10 nucleotide long segment, close to the PAM in the invader's DNA, and complete complementarity is a critical need for effective target degradation (Jiang & Doudna, 2017).

In addition, there is another step called primed adaption. When the bacteria are subsequently affected by the same phages, it incorporates a new spacer into its CRISPR array. These phenomena are known as primed acquisition or primed adaption.

1.5 Objective of the Study

The main goal was to investigate the pathogenicity of bacteria in relation to the CRISPR-Cas system and self-targeting spacers in *V. cholerae* and *S. aureus.* Besides, we have several other aims

1. Determining the CRISPR types of both *Vibrio cholerae* and *Staphylococcus aureus*.
2. Sorting out the variation in both spacer and repeat number, length, and identity.
3. Finding out the number of prophages per strain
4. Looking for the pathogenic factor genes in the strains and perform a quantitative analysis.

Chapter 2
BACKGROUND INFORMATION

2.1 Phages and Bacteria Relationship

A persistent competition exists between bacteria and their phages for survival. Viruses are defined as entities that necessitate hosts for their replication. Bacteriophages, in a similar manner, necessitate a bacterial host for their subsequent generations. The Greek term phage translates to “eater,” which implies that bacteriophage specifically means “bacteria eater.” The bacteriophage was first discovered in 1886 by Ernest Hankin, and the term itself was coined by Félix d'Hérelle in 1917 (Abedon et al., 2011). To achieve a successful generation utilizing the bacterial host, it is essential for phages to circumvent the host's immune system. A variety of strategies have been utilized to achieve successful infection and replication, including the methylation of DNA and the production of anti-CRISPR proteins (Acr) (Choudhary et al., 2023). Within the host, phages can enter either the lysogenic stage or the lytic stage. During the lysogenic stage, the phage DNA integrates into the host genome and replicates alongside the host's chromosomal DNA. The host is subsequently referred to as the “Lysogen,” while the integrated phage is designated as a “prophage”(Garriss & Henriques-Normark, 2020). Conversely, in the lytic pathway, the host cell undergoes destruction while a multitude of virions is produced. The presence of the “cl” gene in the phage genome determines the choice between the lytic and lysogenic pathways (Hochschild & Lewis, 2009). It is important to note that successful phage multiplication can take place without causing any disruption to the bacterial cell wall (e.g., f1 coliphages), where the single-strand DNA is stripped from the bacterial cell and bound by coat proteins (Hay & Lithgow, 2019).

During the lysogenic stage, the expression of most viral genes is inhibited by a repressor protein encoded by the viral gene “cI,” which consequently prevents the production of assembly proteins and the lysis of bacterial cells. The extended suppression of viral genes occasionally results in the emergence of “cryptic phage,” during which they forfeit their capacity to detach from the bacterial genome (Wang & Wood, 2016). Reports indicate that phages may comprise approximately 20 percent of the bacterial genome (Sharma et al., 2023). Bacteriophages play a significant role in enhancing host fitness, particularly in the survival of human pathogenic bacteria. Therefore, it can be concluded that, rather than exhibiting a pathogenic relationship, bacteriophages sustain a symbiotic relationship with their host. The factors that enhance bacterial fitness are encoded by certain non-conserved prophage genes, referred to as “Morons,” which vary from phage to phage. As they incorporate additional information (more on) regarding the bacterial hosts, Roger Hendrix has referred to them as morons (Juhala et al.,

2000). The diverse phage entities play a crucial role in bacterial cell survival, as summarized in table 2.

2.2 Fitness Factors are enhanced by Prophages

Following integration into the bacterial chromosome, phage morons confer enhanced resistance to subsequent phage infections, notably against the identical phage strain. These non-conserved proteins offer limited protection against subsequent phage infection. In the primary defense mechanism, alter the cell surface receptors that serve as binding sites for various phages. In gram-positive bacteria, lipopolysaccharide (LPS) functions as a cell surface receptor for various phage infections. These phage-encoded proteins alter the structure of the outermost layer through two distinct mechanisms. Modification can occur through alterations in the intra-repeat linkages of O-antigen or through the modification of glucose or acetyl side groups. For example, the serotype converting unit present in *P. aeruginosa* phages D3, 297, and ϕKZ, along with *Salmonella anatum* phage ε15, includes an inhibitor of α-polymerase (iap) that obstructs the O-antigen polymerase (Wzyα) and replaces it with the phage-encoded wzyβ, which polymerizes the O-antigen repeats in a modified conformation (Krylov et al., 2013).

***Table 2**. Some examples of fitness factors enhanced by prophages*

Bacteria	Phage	Gene	Function
Pseudomonas aeruginosa	D3	iap, wzyβ, Oac	Serotype converting unit, which alters O-antigen structure to prevent phage absorption
Salmonella anatum	D3112	Tip	Inhibits type IV pilus biosynthesis by disrupting PilB localization
	ε15	iap, wzyβ	Serotype converting unit, which alters O-antigen structure to prevent phage absorption
	ε34	gtrA, gtrB, gtr(type)	Adds a side-chain Glc residue to LPS, preventing phage adsorption
Shigella flexneri	SfII	gtrA, gtrB, gtr(type)	Adds a side-chain Glc residue to LPS preventing phage adsorption
Escherichia coli	KplE1	gtrA, gtrB, gtr(type)	Adds a side-chain Glc residue to LPS preventing phage adsorption
	phiV10	Acetyltransferase	Acetylation of the O-antigen prevents phage binding to LPS
	HK97	gp15	Inner membrane protein that prevents DNA entry
Streptococcus thermophillus	TP-778	ltp	Surface lipoprotein which prevents phage entry

A Gtr locus observed in other *Salmonella, Shigella*, and *E. coli* phages is responsible for synthesizing a lipid-linked glucose residue, which is subsequently translocated to the periplasmic region and ligated to the elongating O-antigen chain (Lehane et al., 2005). This modification inhibits additional phage infection. Phages can similarly achieve side-group blocking by expressing a site-specific acetyltransferase that adds an acetyl group to the O-antigen. The second line of defense is initiated when a phage circumvents the modification of cell surface receptors. A study by Cumby et al. reported the presence of another phage-encoded protein, gp15, in the *E. coli* phage HK97 (Cumby et al., 2015). The gp15 protein is associated with the inner membrane and engages with the tape-measure protein of the infecting phages, thereby inhibiting the entry of the phage genome. In Gram-positive *Streptococcus thermophilus*, superinfection exclusion is mediated by the phage-encoded lipoprotein (lpt), likely due to the interaction between lpt's negatively charged residues and the positively charged tape measure protein (Ali et al., 2014). Various phages, including *Salmonella* phage P22, *V. cholerae* phage K139, and *Lactococcus* phage Tuc2009 encode superinfection exclusion proteins (McGrath et al., 2002). Superinfection exclusion proteins are characterized by targeting the inner membrane through lipid anchors or membrane-spanning helices.

Casas and Maloy assert that numerous pathogenic bacteria acquire their virulence, either partially or entirely, through the presence of prophages, as the similarity of prophage-derived sequences with pathogenic bacteria is a prevalent phenomenon (Casas & Maloy, 2011). The virulence of pathogenic bacteria is significantly impacted by their prophages, which are implicated in various functions, including toxin release, bacterial infectivity, and the regulation of virulence. Certain toxins encoded inside the prophage sections of the bacteria enhance its infectiousness. *Vibrio cholerae*, in the absence of its prophage, is a benign microorganism. However, when the temperate phage CTXφ integrates its genome into *V. cholerae*, it begins to secrete cholera exotoxin owing to the presence of the CTX gene in the prophage region (Brüssow et al., 2004). The development of the Shiga toxin-producing *E. coli* (STEC) pathogen transpires following the lysogenization of *E. coli* cells by Stx-encoding lambdoid phages (Koudelka et al., 2018). The action of Shiga toxin impedes the production of eukaryotic proteins. Furthermore, the causal agents of neurological paralysis, specifically neurotoxins types C1 and D, arise from the integration of the temperate phages CEβ and Deβ into *Clostridium botulinum*, respectively (Sakaguchi et al., 2005).

Besides toxin generation, certain prophage-containing bacteria have been documented to affect host motility. Bacterial motility is facilitated by type IV pili (T4P) and flagella located on the cell surface (Craig et al., 2019). These structures facilitate several functions, including bacteriophage attachment, nutrient acquisition, and host tissue invasion. For example, the shiga toxin generated by the *E. coli* temperate phage Sp5 diminishes the expression of fliA, fliC, and flhCD (Mitsunaka et al., 2018). A decline in swimming motility has been noted due to the overall downregulation of flagellin.

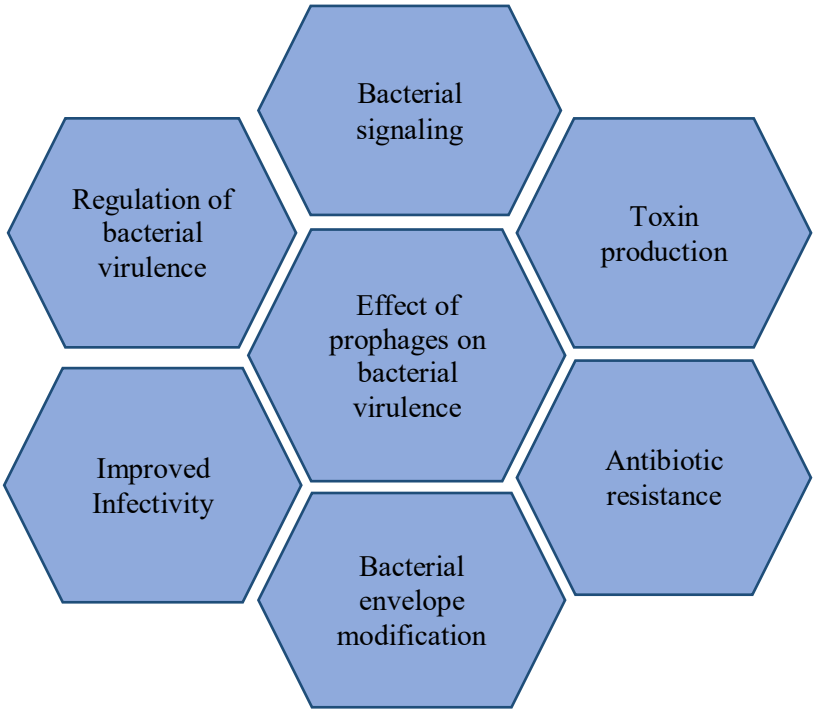


Figure 7: *Diverse Effect of prophages on bacterial virulence*

Conversely, the prophage phi4 protein EC958_1546 facilitates the motility of uropathogenic *E. coli* (UPEC) (Agarwal et al., 2012). The overexpression of this putative protein, possibly via an altered chemotactic response, leads to hypermotile *E. coli* strains exhibiting increased flagellin production. Prophages have been shown to obstruct the operational mechanism of T4P in various infections. Bondy-Denomy et al. assert that the temperate phage JBD26 can inhibit the twitching motility of *P. aeruginosa* without altering the visibility of the pili structures on the surface (Bondy-Denomy & Davidson, 2014). Prophages have also been shown to inhibit the assembly of T4P on the bacterial surface. Conversely, there are three mechanisms by which the presence of prophages might hinder bacterial infectivity. The first level comprises whole virions that facilitate bacterial infectivity. The second stage entails the secretion of effectors

and the encoding of structural proteins by temperate phages, which are crucial in bacterial infection. Phage-encoded regulators, which govern the activation of bacterial infection factors, constitute a third tier of interference. Phages may promote bacterial adhesion, colonization, and invasion of eukaryotic cells while concurrently causing immune suppression (Bille et al., 2017). The bacteriophage itself may serve as the virulence factor encoded within the prophage. Numerous data indicate that the bacterium *Pseudomonas aeruginosa* facilitates colonizing tissue via virion generation (Ridley & Thornton, 2018). The biofilm's development stimulates the synthesis of the filamentous prophage Pf4. The resulting filamentous particles augment bacterial cell adherence to mucin, a crucial component of the mucus barrier, so entrapping the bacteria within the pulmonary environment. Research has revealed that the filamentous prophage φMDA in *Neisseria meningitidis* engages with the epithelial cell layer, facilitating bacterial cell adhesion (Bille et al., 2017). The virions facilitate bacterial cell aggregation, enhancing mucosal colonization due to their resemblance to type IV pili structures. Research has revealed that phages encode many characteristics that enhance bacterial colonization and invasion. Numerous prophage proteins can enhance the adhesion of human platelets to their host cells. The phage coat proteins in *Streptococcus mitis* are PblA and PblB, encoded by φSM1 (Mitchell et al., 2007). Nonetheless, the bacterial cell may also excrete these proteins, enhancing its contact with platelets. PblA and PblB function as platelet adhesins by engaging with choline residues in the bacterial cell wall, thereby enhancing the capacity of bacterial cells to adhere to platelets. *Enterococcus faecalis* pseudotypes pp1, pp4, and pp6 produce phage tail proteins that facilitate human platelet attachment (Matos et al., 2013). In contrast, specific *V. cholerae* and *S. aureus* bacteriophages possess genes for colonization that are manifested on the bacterial surface (Bondy-Denomy & Davidson, 2014). The CTXφ phage in *V. cholerae* encodes the cholera toxin and also produces a gene that induces the production of toxin-coregulated pili (TCP). The pili structures serve as adherence receptors for CTX φ and are essential for bacterial adhesion to epithelial cells. Likewise, *S. aureus* possesses a bacteriophage akin to φSPβ that harbors the sasX gene, which encodes a surface protein that promotes attachment to nasal epithelial cells (Mulcahy & McLoughlin, 2016). Third, by regulating surface-expressed proteins, bacteriophages can influence bacterial adherence. For instance, following the integration of the temperate phage φCD38–2, the conserved cell wall protein CwpV in *C. difficile* is increased (Fortier, 2018). The host exhibits increased cell aggregation and biofilm development due to heightened CwpV expression. Additionally, CwpV serves as a superinfection exclusion mechanism in *C. difficile* owing to its anti-phage characteristics.

Prophages modulate bacterial pathogenicity either directly or by producing regulatory proteins and RNAs. Viral integration into the bacterial genome facilitates site-specific disruption of host activities without the host's awareness. Traditionally, temperate phages are integrated at critical loci, resulting in the loss of a regulatory region or bacterial gene. The incorporation of the prophage φ13 into the hlb gene of *S. aureus* leads to the pathogen's failure to synthesize β-toxin (Coleman et al., 1991). Likewise, the loss of the pilB gene leads to diminished virulence of *B. bronchiseptica*, as demonstrated by the integration of phage PHB09 into its genome (Chen et al., 2020). In this context, a complex link observed between a pathogen and phage is the phage-regulatory switch (phage-RS). This switch induces the excision of the prophage in reaction to a specific stimulus, without leading to the lysis of the host cell or the production of new virions (Feiner et al., 2015). This type of switch was first observed in the Gram-positive pathogen *Listeria monocytogenes*. The comK gene in this bacterium is disrupted by the temperate prophage φ10403S (Rabinovich et al., 2012). Macrophages phagocytose *L. monocytogenes* during pathogenesis, resulting in the excision of the temperate phage from the host chromosome. Consequently, the bacterial comK gene is synthesized intact, facilitating the production of the ComK protein and the consequent activation of the Com system. Consequently, the bacteria successfully escape the phagosome and infiltrate the macrophage's cytoplasm to proliferate. The temperate phage φ10403S subsequently reinserts itself into the comT gene.

In addition to the regulation by whole prophages, bacteriophages can actively modulate bacterial behavior by producing regulatory proteins encoded by the phages. Certain phages have been shown to regulate the expression of alternative sigma factors. In response to stressful environmental conditions, the host cell typically activates these sigma factors, which govern the expression of supplementary virulence-associated bacterial genes. Phages can actively modulate quorum sensing (QS), which governs cell-to-cell communication and the regulation of virulence factors, as well as the production of alternative sigma factors (Schroven et al., 2021). Table 3 delineates the diverse functions of prophage-encoded proteins in virulence.

Table 3. *Some examples of prophage-encoded proteins*

Bacteria name	Prophage name	Gene name	Function
Toxins			
Escherichia coli	H-19B	stx1, stx2	Prophage induction disrupts the repressor and produces Shiga-like toxins
Vibrio cholerae	CTX ϕ	ctx	Prophage encodes cholera toxin
Clostridium difficile	ϕCD38-2	tcdA	Prophage-increased production of host toxins
Staphylococcus aureus	80α	Tst	Toxic-shock syndrome toxin-1 production
Bacterial signaling			
Clostridium difficile	ϕCD38-2	Unknown	Lysogeny by this prophage increases the release of exotoxins
	phiCDHMI	agrD, agrB, agrC	Produces a quorum-sensing precursor
Bacillus subtillis	Phi3T	Arbitrium	A quorum sensing molecule capable of promoting the lytic pathway
Clostridium tyrobutyricum	ϕCTP1	luxR homolog	This protein may provide the phage a way to alter host processes
Escherichia coli	Unknown	torI	Inhibit the host response regulator for low pH
Salmonella typhimurium	Gifsy-2	sodC	Expression of this superoxide dismutase neutralizes reactive oxygen species
		gogB	Inhibits degradation of inflammatory agents thereby attenuating ROS
Bacterial infectivity			
Pseudomonas aeruginosa	Pf4	Entire phage	Promotion adhesion of bacterial cells to mucin
Neisseria meningitidis	φMDA	Entire phage	Boost adhesion of bacterial cells to epithelial cells
Pseudomonas aeruginosa	Pf	Entire phage	Reduced inflammation and phagocytosis
Staphylococcus mitis	φSM1	PblA, PblB	Mediation bacterial interaction with platelets
Vibrio cholera	CTXφ	Tcp	Expression of toxin-coregulated pili
Clostridium difficile	φCD38–2	cwpV	Increase in bacterial cell aggregation
Salmonella enterica	Gifsy-1	gogB	Increased inflammatory response, enhanced tissue damage and increased bacterial colonization
Antibiotic resistance			
Escherichia coli	Unknown	blaTEM	B-lactamase
	Unknown	blaCTX-M9	B-lactamase
	Unknown	mecA	Penicillin binding protein (PBP)
	Unknown	qnrA	Inhibits quinolone
	Unknown	qnrS	Inhibits quinolone
	933W	tet	Tetracyclin resistance
Staphylococcus xylosis	ΦJW4341	erm	Erythromycin ribosome methylase (MLSB resistance)

Bacterial virulence regulator			
Escherichia coli	Sp5	agxR	Increase in iron uptake
Staphylococcus aureus	φ11, φ80α	cI	Induction SigB regulon, impacting the development of the bacterial infection
Vibrio cholera	VP882	vqmA	Interfering with the quorum sensing system
Staphylococcus aureus	φ13	Entire phage	Loss of β-toxin production due to disruption hlb
Listeria monocytogenes	φ10403S	Entire phage	Interruption comK gene
Clostridium difficile	φCD2, φCD6,	Unknown	Increase in β-toxin expression
	φCD8 φCD27	Unknown	Decrease in β-toxin expression

2.3 Self-Targeting Spacer Acquisition and Their Controversial Role

Conversely, bacterial cells possess specific mechanisms to defend against foreign invaders, such as viruses and plasmids. Similar to eukaryotes, the molecular immunology of bacteria can be broadly categorized into two types: innate and adaptive immunity. The innate response includes the utilization of restriction modification systems (RM) and superinfection exclusion systems, among others. Conversely, the adaptive immune system is characterized by the multifaceted function of CRISPR-Cas mediated immunity. Interestingly, during spacer acquisition in the adaptation phase (in both naïve acquisition and primed adaptation), bacteria may inadvertently obtain spacers from their own genome. This results in the integration of self-targeting spacers (STS), a phenomenon referred to as autoimmunity. A self-targeting spacer refers to a spacer that targets the "self" genome rather than the genomes of external invaders such as phages or plasmids. Bacteria can assault their own genome, leading to cell death. This section of the literature review will examine the acquisition of STSs.

During spacer acquisition, the CRISPR system distinguishes between self and non-self through two separate mechanisms. The protospacer adjacent motif (PAM) sequence is located within the foreign genome, near to the protospacer. PAM is essentially a 2-5 nucleotide sequence that is specifically identified by the CRISPR cascade complex or Cas9 during target degradation (Shiimori et al., 2018). Conversely, Levi and a colleague have documented the existence of “chi” sites inside the chromosomal DNA of the host. This chi site designates the host chromosomal DNA as self and occurs at a frequency of once every 5kb in the *E. coli* genome(Levy et al., 2015). Nonetheless, chi sites are also found in extrachromosomal DNA and phage DNA, albeit at a minimal frequency. It can now be predicted that any failure of the

system to differentiate self from nonself may inadvertently result in the acquisition of spacers from the host's own genome. Franziska and Chase have proposed five distinct methods for the insertion of STSs into the CRISPR array (Wimmer & Beisel, 2019).

The initial method may be referred to as naive acquisition. STSs are obtained independently of the current collection of spacers. Cas1 and Cas2 proteins are largely conserved across all CRISPR types, but Cas3, Cas4, Csn2, or Cas9 may also be required for this acquisition (Heler et al., 2015; Yosef et al., 2012). However, the precise process remains to be investigated. Levy and others have posited that spacers may be preferentially extracted from the replication fork (Levy et al., 2015). During replication, the helicase enzyme RecBCD in Gram-negative bacteria and AddAB in Gram-positive bacteria unwinds the double-stranded DNA. It also destroys linear double-stranded DNA in many contexts. Levy suggested that segments of chromosomal DNA may be inadvertently integrated, resulting in STS when the system fails to distinguish between self and non-self.

The second form of STS acquisition is facilitated by prophages or mobile genomic elements. As previously discussed, phages can integrate their genome into their bacterial host and possess the ability to excise themselves in response to certain signals. Before the lysogenic phase, spacers from the same prophage may be integrated into the CRISPR array as a defense against the bacteriophage. Consequently, this type of mechanism will target the prophage region embedded inside its own chromosomal DNA, resulting in cellular demise. Shmakov revealed in their study that 83% of the self-targeting spacers exhibited similarity to their integrated prophages (Shmakov et al., 2017).

The third type of STS acquisition involves primed adaptation. During the excision of the phage genome from the bacterial genome, imprecise cutting may result in the incorporation of prophage-flanking bacterial chromosomal segments alongside the phage genome. It can subsequently be transferred to other bacteria. This phenomenon is called viral transduction and serves as the primary mechanism for horizontal gene transfer. In this scenario, a portion of the bacterial chromosome is incorporated into the phage particles, resulting in the formation of defective phage or heterogenote. Stern et al. have documented a nearly equal distribution of STSs from MGE and non-MGE, with proportions of 47% and 53%, respectively (Stern et al., 2010). It can be suggested that the defective phage may serve as a potential source of spacers originating from the bacterial chromosomal DNA. Finally, spacers that have developed to target foreign DNA may initiate with similar chromosomal DNA sequences. Prior studies have

shown that priming can occur even when the target location exhibits thirteen mutations relative to the original spacer.

The subsequent type of STS acquisition involves RNA mediated naïve acquisition. This unique approach to acquisition involves the rare Cas proteins, which specifically recognize RNA rather than DNA. Included in this group of proteins is an RT that is often translationally fused to Cas1, as well as a fusion that combines Cas1 with the Cas6 protein, responsible for crRNA synthesis. This specific RNA-acquiring apparatus is primarily associated with type III CRISPR-Cas systems, although it can also be found in type I-E and type VI-A systems (Kojima & Kanehisa, 2008; Toro et al., 2019). In the limited cases studied, these RTs reverse transcribe an acquired RNA into DNA, resulting in the formation of an acquisition substrate. The linked type III CRISPR-Cas systems have the potential to target the host, leading to autoimmunity, particularly when the RNA-derived spacers are sourced from host RNA. It is noteworthy that three strains incorporating a RT within their type III CRISPR-Cas systems have demonstrated the presence of self-targeting spacers.

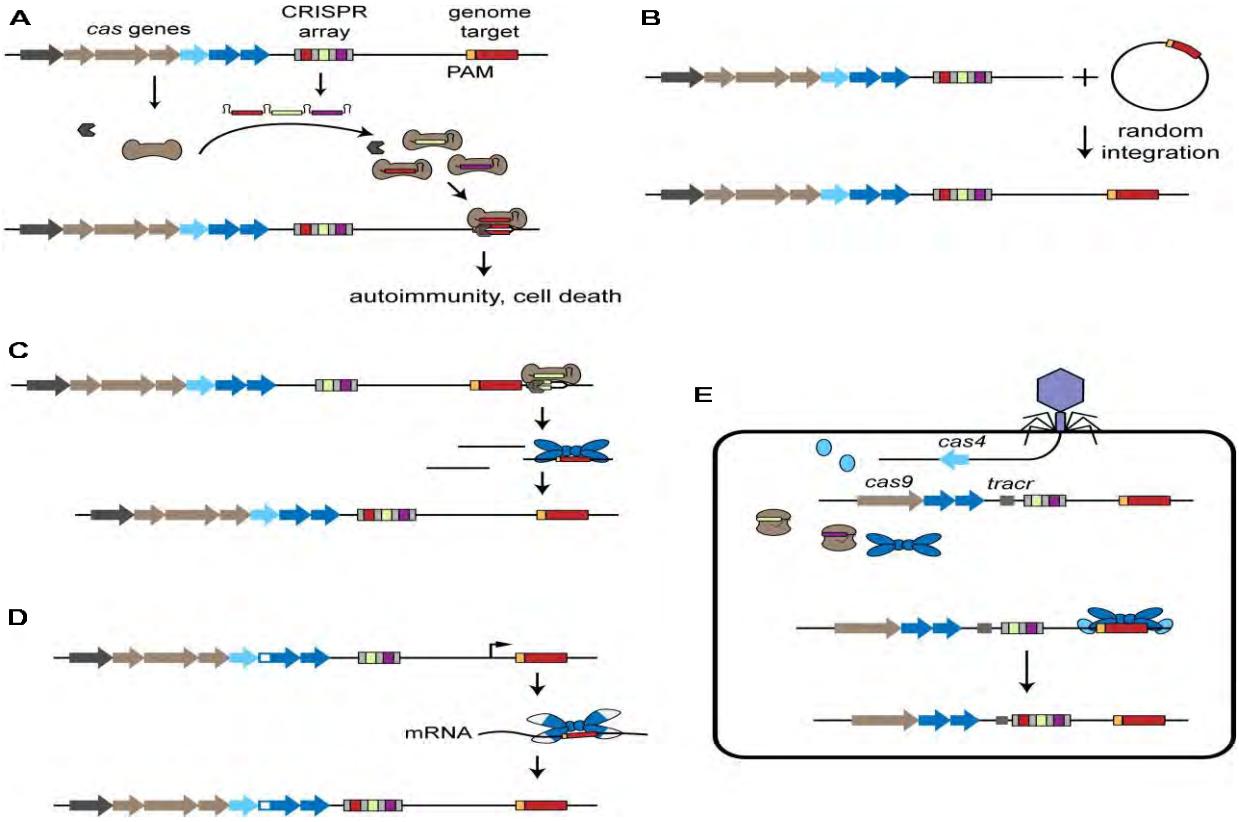


Figure 8: *Different methods of* STS *acquisition*

(Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC6990116/)

Foreign genetic elements shape the fifth and final type of STS acquisition. Certain phages have demonstrated the ability to encode Cas proteins, leading to the selection and preferential uptake of self-targeting spacers by endogenous CRISPR-Cas systems. Investigating the origin of spacers encoded in the CRISPR array of *Campylobacter jejuni* PT14, characterized by a limited type II-C CRISPR-Cas system, yields the first direct evidence (Hooton & Connerton, 2015). Several of the spacers exhibited partial matches with chromosomal regions in the PT14 genome, even though none of the spacers demonstrated a 100% sequence similarity with any known sequences. PT14 cells were co-cultured with the CP8/CP30A phage, and monitoring the spacer content revealed that all spacers, including those newly acquired from the phage, originated from the host chromosome. The endogenous type II-C CRISPR-Cas system typically does not contain this gene; however, the phage has encoded a copy of the cas4 gene, which plays a crucial role in protospacer maturation within various CRISPR-Cas systems (Zhang et al., 2018). Consequently, the authors attributed the unexpected self-targeting acquisition events to the phage-encoded Cas4.

Nonetheless, the function of STSs has been a subject of debate since the initial discovery of CRISPR due to their dual roles. In certain instances, reports highlight their diverse regulatory functions within the host cell, while other studies have focused on the autoimmunity driven by STSs. Several studies have reported that one-fifth of the bacteria containing CRISPR are harboring self-targeting spacers. Types IB and IF exhibit a higher susceptibility to STS acquisition, whereas type IIIA STSs are nearly nonexistent. Noberga and coworkers have shown that STSs mediated CRISPR-Cas autoimmunity is strongly associated with the presence of prophages in the host genome (Nobrega et al., 2020). It has been reported that STS maps to prophage regions in more than half of the cases, and genomes containing STS almost always possess a prophage. The remaining portion is aimed at the endogenous genome. It is noteworthy that STS hits within the endogenous genome are more prevalent near prophages, displaying a pattern that aligns with primed adaptation from a protospacer originally situated on the prophage. Furthermore, in bacteria harboring multiple STS, the prophage was primarily targeted by the STS located farthest from the leader sequence, while subsequent STS engaged with both endogenous regions and the prophage itself (Tangarife et al., 2023). The results bolster a framework where acquiring new spacers from both prophage and nearby regions, through primed adaptation, enhances STS. It is intriguing to consider an important question: given that STS can induce cell death, how do we observe the continued presence of bacteria containing STS in nature? Bacteria have developed various mechanisms to escape the harmful

impact of STS-mediated autoimmunity. Numerous strategies utilized by bacteria have been documented to date, such as the mutation or deletion of Cas genes, spacers, repeats, or protospacer adjacent motifs (PAM), as well as the action of anti-CRISPR (Acr) proteins (Arroyo-Olarte et al., 2021; Wu et al., 2022). Numerous severely compromised CRISPR systems harboring cas pseudogenes are posited to result from the necessity for the CRISPR–Cas system to become inactivated following the incorporation of an STS. This phenomenon may explain the widespread occurrence of these systems. Recent experimental evolution studies have shown that auto-immunity to prophages can lead to extensive genomic deletions along the CRISPR–Cas locus. Nobrega et al. reported that partial or nonexistent CRISPR-Cas systems are found in 12% of the genomes of STS-containing organisms, while complete CRISPR-Cas systems are present in 88%. The implication is that self-targeting can be circumvented by numerous means, although the deletion of CRISPR-Cas may occur as a strategy to endure STS (Nobrega et al., 2020). Conversely, the same study indicates that most STS locations within the CRISPR array are situated proximal to the leader sequence. In contrast, a minority are located at the distal end. It can be postulated that following the acquisition of STS, certain CRISPR systems may become inactivated in some CRISPR-containing bacteria. When STSs are positioned distal to the leader end in certain bacteria, the conflict between CRISPR-Cas and phage persists actively. Consequently, they employ alternative mechanisms to prevent cellular apoptosis. Table 4 summarizes some findings that have been adopted by bacteria in response to STS acquisition.

Table 4. *Strategies by some species to evade STS-mediated autoimmunity*

Bacteria name	Phage	Gene	Function
Pseudomonas aeruginosa	JBD88a	*acrE1*	Inhibits the host's type IE CRISPR/CAS system through Cas3 inhibition
	JBD30	*acrF1*	Inhibits the hosts' type IF CRISPR/Cas system by preventing DNA-binding
	HK97	gp15	Inner membrane protein that prevents DNA entry
Streptococcus thermophillus	D4276	*acrIIA5*	Inhibits the host's type II CRISPR/Cas through Cas9 interaction
Vibrio cholerae	phiCDHMI	*agrD*, *agrB*, *agrC*	Contains a CRISPR locus capable of shutting down the hosts' repressor
Neisseria meningitidis	Unknown	*acrIIC1*	Inhibits Cas9 by inactivating the HNH domain required for DNA cleavage
	Unknown	*acrIIC3*	Inactivates Cas9 by locking it into an unfavorable oligomeric state

Furthermore, Anti-CRISPR proteins, which suppress the CRISPR Cas system, are encoded by prophage genes (morons). These individuals were initially found to inhibit the type I-E and I-F CRISPR-Cas systems in closely related phages that infect *P. aeruginosa* (Pawluk et al., 2014). Subsequently, anti-CRISPRs encoded by phages were identified as effective against various type I and type II CRISPR-Cas systems in bacteria, as well as a type I-D CRISPR-Cas system in archaea (He et al., 2018). These Acrs function through several means. Type I anti-CRISPRs have been shown to inhibit Cas3 endonuclease activity and prevent the Cascade complex from binding to DNA (Choudhary et al., 2023). Cas9, a singular protein that identifies and cleaves exogenous DNA pieces, is the central component of the type II CRISPR–Cas system. Anti-CRISPR proteins inhibit this mechanism by obstructing DNA recognition, sequestering Cas9 into an unfavorable oligomeric state, or inactivating the HNH endonuclease activity essential for cleavage (Song et al., 2023). A phage derived from *Vibrio cholerae* ICP1 utilizes its inherent capabilities to encode the CRISPR-Cas system (Naser et al., 2017).

Despite the detrimental impact of STS, it is evident that the host bacterium continues to sustain the active CRISPR system. This may be attributable to the additional functions of CRISPR-Cas that confer advantages to the bacterium. Certain roles derived from prophages have been elucidated in the preliminary section of the literature review. This section will overview the advantages of CRISPR-Cas mediated mechanisms within host cells. Two strains of *Enterococcus faecalis* were examined utilizing a murine urinary tract model in a study investigating the pathogenicity of isolates: one possessing a type II-A CRISPR-Cas system and the other without it. When equal inocula of both strains were administered, the strain containing CRISPR-Cas induced a more rapid mortality rate in mice (Wu et al., 2022). Initially, the virulence of the type II CRISPR-Cas-bearing strain seemed diminished, as shown by a greater 50% fatal dose (LD50). Histological investigations demonstrated that the CRISPR-Cas-containing strain colonized the mouse organs more efficiently than the isolate devoid of the system due to its superior biofilm formation capability. In summary, the research on *Enterococcus* suggests that CRISPR-Cas systems, in conjunction with other genetic changes, may influence bacterial pathogenicity via two non-exclusive mechanisms: In scenarios where, mobile elements may introduce exogenous DNA harboring putative virulence factors (toxins or antibiotic resistance genes), CRISPR-Cas protection may mitigate bacterial pathogenicity. Conversely, CRISPR-Cas regulation of gene expression may enhance bacterial pathogenicity, for as by facilitating host colonization (Louwen et al., 2014).

Self-targeting spacer-mediated toxicity has been documented in archaea, similar to its occurrence in bacteria. The findings indicated that targeting the beta-galactosidase region using a small CRISPR array resulted in the *Sulfolobus solfataricus* M18 strain exhibiting markedly slower growth compared to the control group (Stachler et al., 2017). The micro CRISPR locus in the transfected cells exhibited substantial alterations, presumably to avert autoimmunity, as indicated by the polymerase chain reaction data. Notably, transformants derived from the beta-galactosidase-deficient *S. solfataricus* PBL2025 strain displayed no alterations in the micro CRISPR locus. The modifications noted in the micro CRISPR locus of *S. solfataricus* M18 strains are ascribed to the bacterial approach designed to prevent self-targeting and toxicity. Unlike the *S. solfataricus* M18 strain, self-targeting of the redundant gene in *Haloferax volcanii*, an archaeon, was comparatively well tolerated (Wimmer & Beisel, 2019). However, the strategy employed by *H. volcanii* to circumvent autoimmune remains mostly unknown.

Self-targeting by CRISPR-Cas systems can facilitate bacterial or archaeal evolution by disrupting a vital gene and inducing changes in the organism. Specifically, chromosomal targeting may lead to minor deletions or mutations in the target gene (Dy et al., 2013). These deletions may also encompass numerous adjacent non-targeted genes and can be significantly bigger. The loss of any important gene would be lethal; but more extensive deletions may also provide a fitness advantage by generating novel phenotypes or reducing the overall genome size. Furthermore, the reorganization of pathogenicity islands may modify the bacterium's virulence. Besides producing active mutations, self-targeting CRISPR-Cas systems can selectively target a small subpopulation that already lacks the target. Furthermore, a significant example arises from the bacterium *Pelobacter carbinolicus. P. carbinolicus* cannot decrease Fe(III) throughout its metabolic processes, unlike other members of the Geobacteraceae family (Richter et al., 2007). A spacer complementary to a segment of the histidyl-tRNA synthetase gene hisS is present within the endogenous type I-E CRISPR-Cas system and may be responsible for this trait. Proteins containing many adjacent histidines would exhibit reduced translation efficiency without histidyl-tRNA synthetase (Aklujkar & Lovley, 2010).

Self-targeting spacers may induce mRNA degradation and modify gene expression instead of necessarily provoking autoimmunity when RNA, rather than DNA, is the target (Hale et al., 2009; Strutt et al., 2018). An instance of a CRISPR-Cas system that degrades mRNA is the type III-B system identified in *Myxococcus xanthus*. However, the precise process remains unidentified. The authors undertook a study to do a transposon screen in a pilA strain lacking

the type IV pilus essential for exopolysaccharide production. A transposon introduced into the CRISPR3 array was utilized to isolate a mutant that resumed exopolysaccharide production and inhibited the formation of fruiting bodies. Wallace et al. (2014) proposed a mechanism wherein the transposon enhanced precrRNA processing, leading to crRNA-mediated regulation of exopolysaccharide production and fruiting body formation (Wielgoss & Julien, 2023). Alternative situations involve the inclusion of certain transposons within the repertoire of crRNAs, altering the array's targeting potential. Alternative processes may be involved in *M. xanthus*, including transposon insertion, considering recent reports of type III-B systems targeting transcriptionally active DNA (Guzmán et al., 2021).

Table 5. STS *and gene regulation*

Name of the organism	**Description**	**Function**
Aggregatibacter actinomycetemcomitans	Target metabolic glycogen, phosphorylase enzyme encoding gene	Potentially affecting bacterial metabolism
Pseudomonas aeruginosa	Partial complementary binding of the STS with lasR gene.	mRNA degradation and subsequent immune evasion
		Induces the SOS response, biofilm formation, and impairing swarming motility
Neisseria meningitidis	Unknown	To cleave RNA in vitro
Staphylococcus aureus	Unknown	To cleave RNA in vitro and in vivo
Campylobacter jejuni	Unknown	To cleave RNA in vitro and in vivo
Lactococcus lactis.	STS targets the prophage region	Preventing the reactivation of lysogenic phages

Chapter 3
MATERIALS AND METHODS

3.1 Sequence Retrieval

All the reference sequences of *Vibrio cholerae* and *Staphylococcus aureus* were downloaded from the NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/) in fasta format(Tatusova et al., 2016). The NCBI Reference Sequence (RefSeq) database comprises a taxonomically varied, non-redundant, and extensively annotated collection of sequences that reflect naturally occurring DNA, RNA, and protein molecules. Sequences from plasmids, organelles, viruses, archaea, bacteria, and eukaryotes are included. Each RefSeq is entirely derived from sequence data supplied to the International Nucleotide Sequence Database Collaboration (INSDC). A RefSeq, akin to a review article, is a synthesis of knowledge amalgamated from several sources at a specific point in time. RefSeqs establish a basis for integrating sequencing data with genetic and functional information. They are created to serve as reference standards for many purposes, including genome annotation and documenting sequence variation locations in medical records. The RefSeq collection is accessible without restrictions and can be obtained through various methods, including searches or links in NCBI resources such as PubMed, Nucleotide, Protein, Gene, and Map Viewer, sequence searches via BLAST, and downloads from the RefSeq FTP site.

All the sequences were checked for quality, and low-quality sequences (more than 5% gaps or N) were discarded in the Ubuntu terminal (2022.4)(Zhou et al., 2013) . Quality control in sequencing entails evaluating the precision and dependability of the DNA sequences produced by sequencing technology. Considering that contemporary sequencing generates extensive data, detecting and eliminating sequences likely to have errors or artifacts that may jeopardize subsequent investigations is crucial. A sequence exhibiting over 5% of its bases as gaps or N's indicates considerable uncertainty or absent data, rendering it untrustworthy for precise biological interpretation. This threshold ensures that only high-quality sequences are preserved for subsequent analysis, which is essential for achieving accurate results in investigations such as variant calling, phylogenetic analysis, or metagenomics.

Contig lengths were determined using the faidx function of Samtools (H. Li et al., 2009). Samtools is a prevalent software suite for processing high-throughput sequencing data. The faidx function generates an index for FASTA files, facilitating rapid access to sequence data. This function, when executed on a FASTA file, produces an index containing details regarding the length and position of each contig within the file.

Contigs more than 5000 bases were kept using the subseq function of the seqtk tool (Shen et al., 2016). Contigs are continuous DNA sequences built from small bits during sequencing. The subseq function in seqtk is utilized to extract particular sequences from these files according to specified identifiers or criteria. The subseq function would be employed to filter contigs according to their length. The command probably entails designating a list of contig identifiers that satisfy the length criterion or analyzing the full dataset to preserve only those over 5000 bases. We have established the threshold point to anticipate a CRISPR locus at a minimum of 5000 bp. These filtering resulted in 1697 *Vibrio cholerae* and 16286 *Staphylococcus aureus* sequences.

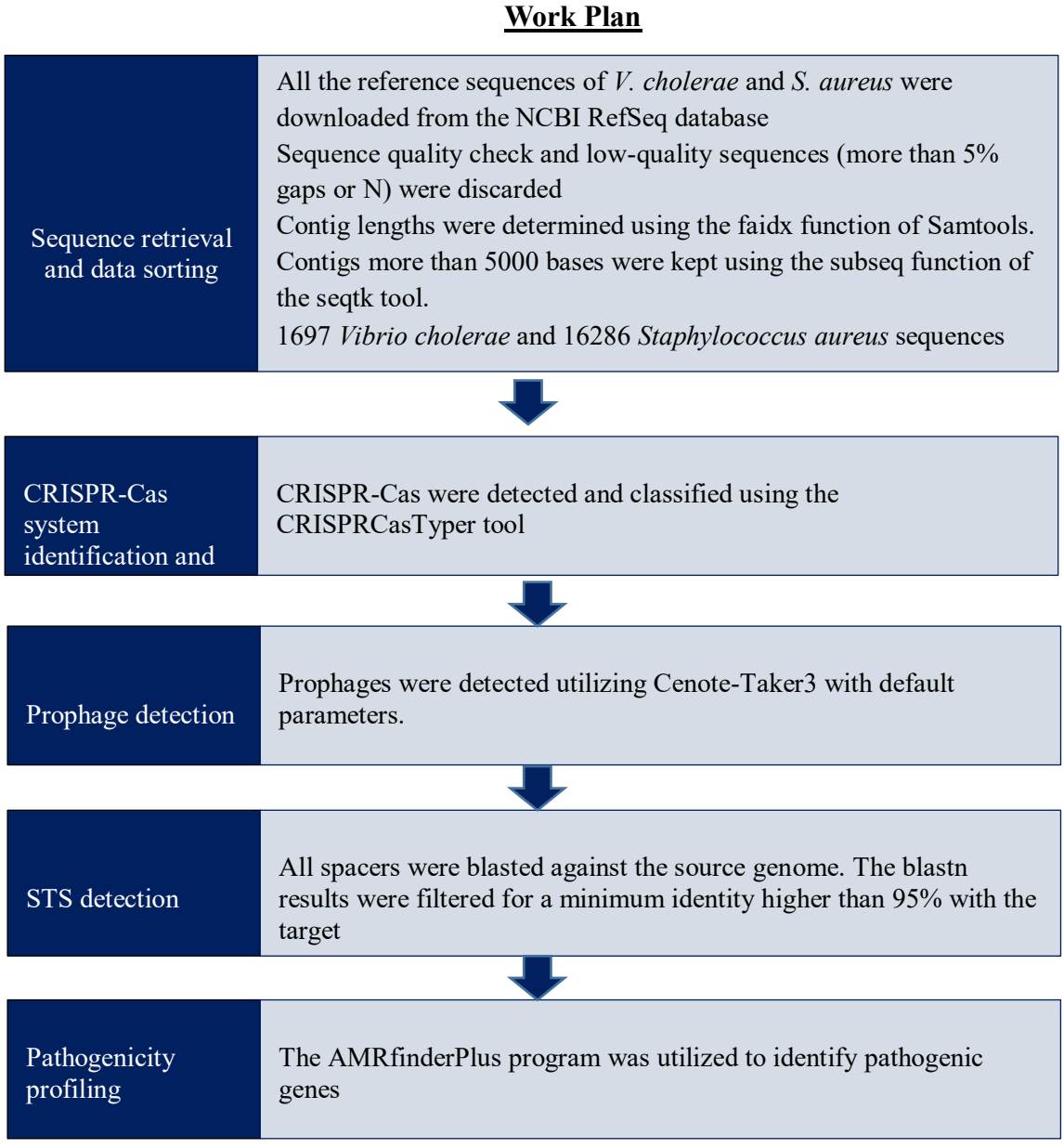


Figure 9: *Work plan of the study*

3.2 CRISPR-Cas system identification and classification

CRISPR-Cas were detected and classified using the CRISPRCasTyper tool (Russel et al., 2020). CRISPRCasTyper Identify CRISPR-Cas genes and arrays and classify the subtype based on the Cas genes and the CRISPR repeat sequence. This software identifies Cas genes using a comprehensive set of HMMs, subsequently organizes these HMMs into operons, and predicts the operon subtype utilizing a scoring system. Additionally, it identifies CRISPR arrays by analyzing minced data and employing BLAST against a comprehensive collection of known repeats. A Kmer-based machine learning method, specifically extreme gradient boosting trees, is utilized to predict the subtype of the CRISPR arrays based on the consensus repeat. The process subsequently links the Cas operons and CRISPR arrays, yielding the output: CRISPR-Cas loci exhibit subtype predictions derived primarily from Cas genes and CRISPR consensus repeats. Additionally, orphan Cas operons, their predicted subtypes, orphan CRISPR arrays, and their associated predicted subtypes are discussed.

3.3 Prophage identification

Prophages were identified using Cenote-Taker3 using default parameters (Tisza et al., 2021). Cenote-Taker 3 is a software application developed for identifying, annotating, and analyzing viral sequences and integrated viral genomes (prophages) in microbial and metagenomic assemblies. It is very beneficial for identifying and characterizing viral components from assembled genomic data. Cenote-Taker 3 enhances prior iterations by augmenting speed, precision, and the scope of viral detection across diverse datasets.

3.4 Self-targeting spacers

All spacers were blasted against the source genome (Altschul et al., 1990). The blastn results were filtered for a minimum identity higher than 95% with the target. Any hit on the genome was considered a self-target, except for those within all of the predicted CRISPR arrays. Hits closer than 500 bp from each end of the predicted arrays were also ignored to avoid considering spacers from the array that were possibly not identified. Finally, STS from CRISPR arrays of two or fewer spacers were excluded, except when the associated repeat belonged to a known CRISPR repeat family, as identified by CRISPRDetect (Biswas et al., 2016). Duplicates were removed by searching of similar genomes, contigs, and arrays.

3.5 Pathogenicity-related gene identification

AMRfinderPlus tool was used to find the pathogenic genes (Feldgarden et al., 2021). AMRFinderPlus is a sophisticated bioinformatics tool created by the National Center for Biotechnology Information (NCBI). The system is intended to identify antimicrobial resistance genes, virulence factors, and particular resistance mutations within bacterial genomes. AMRFinderPlus expands upon the earlier AMRFinder tool by incorporating features such as improved detection of point mutations and the identification of more virulence genes.

3.6 Statistical analysis

Done using Excel and Linux terminal. Some figures were generated using Stats Kingdom website (https://www.statskingdom.com/violin-plot-maker.html).

Chapter 4
RESULT

4.1 Functional CRISPR-Cas System

A total of 1,794 and 16,286 genomes (Refseq) for *V. cholerae* and *S. aureus* were studied from the NCBI database. Among 1,794 *V. cholerae* sequences, Cas operons were identified in 402 strains, whereas CRISPR arrays were detected in 408 strains. Conversely, among 16,286 strains of *S. aureus*, the Cas operon was identified in just 87 strains, whereas the CRISPR array was present in 4,107 strains. Of the 402 Cas-containing strains of *V. cholerae*, 346 possess Cas operons that are associated with CRISPR arrays, rendering them functional. According to our research, 19.29% of the strains had a functional CRISPR system. Fifty-six strains were identified to possess cas operons that lack accompanying CRISPR arrays, rendering them orphan cas operons. Of the 346 organisms that possess both Cas and CRISPR, 41 were identified as containing additional Cas operons absent of a corresponding CRISPR array, indicating that these 41 strains harbor both a functional CRISPR-Cas system and orphan Cas operons. Among the 87 *S. aureus* strains possessing Cas operons, 83 include Cas operons that are associated with CRISPR arrays. Four strains were identified to possess Cas operons that lack accompanying CRISPR arrays, rendering them orphan cas operons. Among all the strains containing both Cas and CRISPR, none were discovered to possess an orphan Cas operon. Data indicates that 0.51% of the sequencing strains possess a functional CRISPR system.

Table 6. *CRISPR-Cas distribution*

Name of the bacteria	Total sequence number	Cas operon presence	CRISPR array presence	Functional CRISPR Cas system	Percentage of functional CRISPR Cas system
Vibrio cholerae	1794	402	408	346	19.29%
Staphylococcus aureus	16286	87	4107	83	0.51%

4.1.1 Diverse Cas operon number

Among the 402 *V. cholerae* strains, 296 contained single cas operons, 90 contained 2 cas operons, 15 contained 3 cas operons, and one strain contained 4 cas systems. On the contrary, among these 87 *S. aureus* strains, all of them were found to contain a single cas operon.

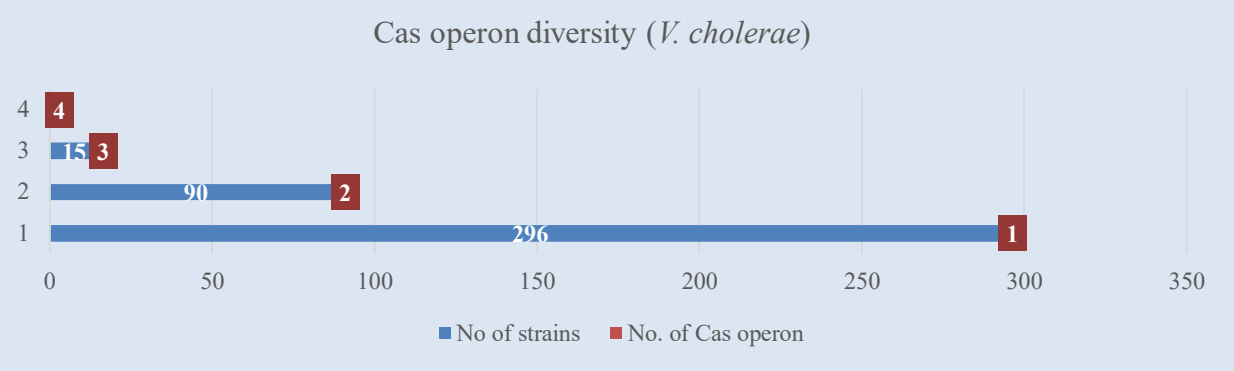


Figure 10: *Diversity in Cas operon number of V. cholera*

4.1.2 Diversity in CRISPR Array Number

Among the 408 *V. cholerae* strains containing the CRISPR array, 255 contained a single array. Several strains contained multiple arrays with a maximum number of 11 arrays in a single strain. In addition, among the 4107 *S. aureus* strains containing the CRISPR array, 3147 contained a single array, and 917 contained two arrays. The maximum number of arrays present was 4 in a single strain.

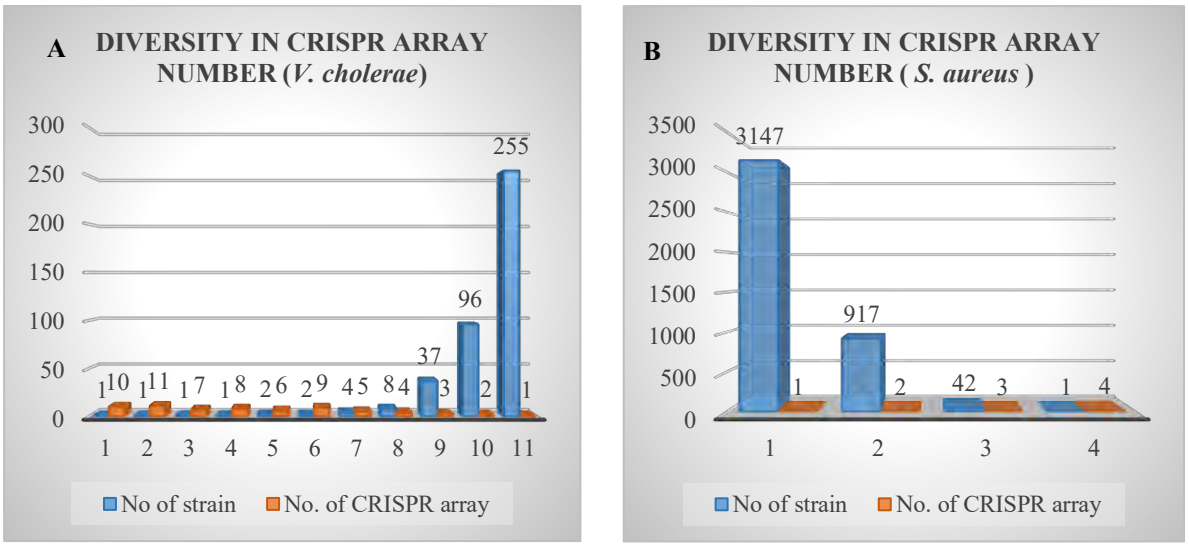


Figure 11: *CRISPR array diversity. (A) V. cholerae CRISPR array diversity. (B) S. aureus CRISPR array diversity*

4.2 Cas Operon Distribution

4.2.1 Single operon completeness: Among the 296 strains of *V. cholerae*, containing one Cas operon, the operon completeness is shown in figure 12. On other hand, among the 87 *S. aureus* strains containing Cas operon, all the Cas operons were 100% complete except one strain.

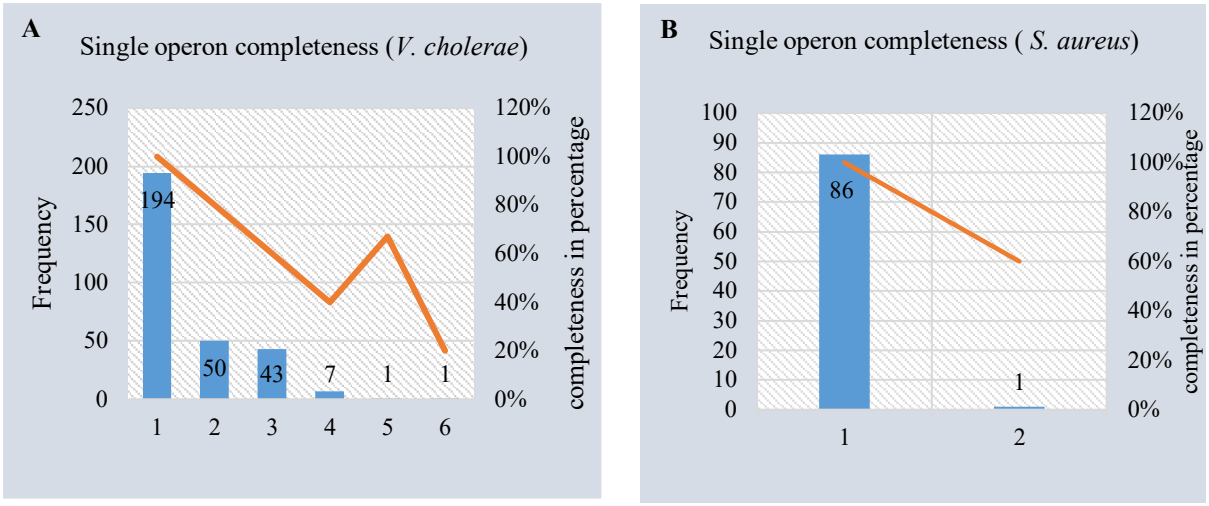


Figure 12: *Single operon (Cas) completeness. (A) V. cholerae operon completeness. (B) S. aureus operon completeness*

4.2.2 Single operon types: Among these 296 single operon coantaining *V. cholerae* strains, mostly (157) were carrying type IF system. On contrary, among 87 strains of *S. aureus*, 86 strains were carrying type IIIA system while one strain were reported to carry IIA system.

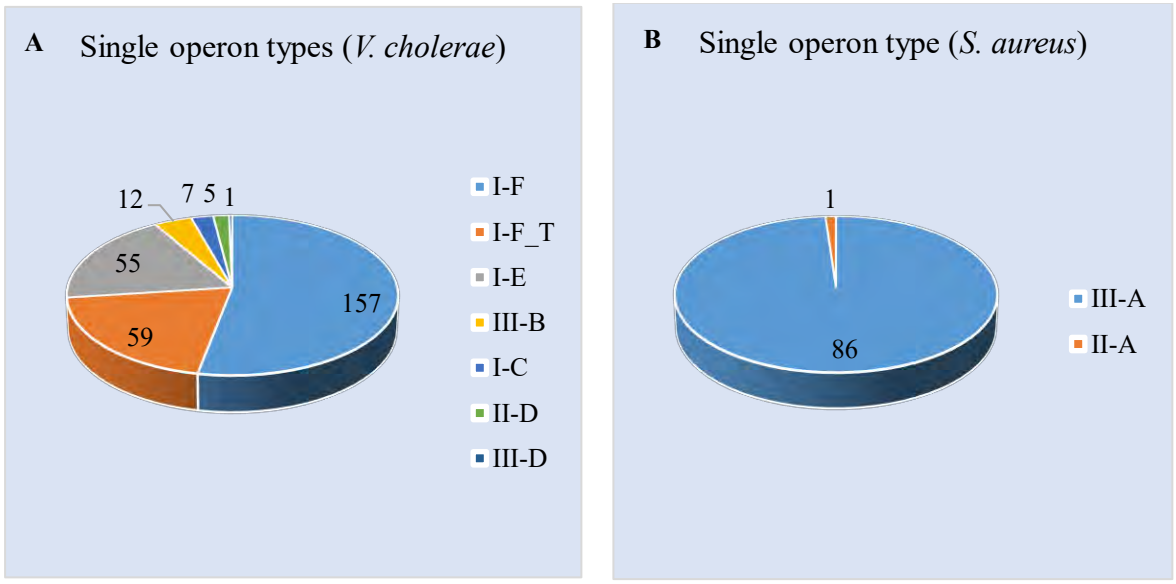


Figure 13: *Single operon types. (A) V. cholerae operon types. (B) S. aureus operon types*

4.2.3 Single Operon Completeness according to Types: The completeness of the single operon is presented on table 7 according to percentage in different types.

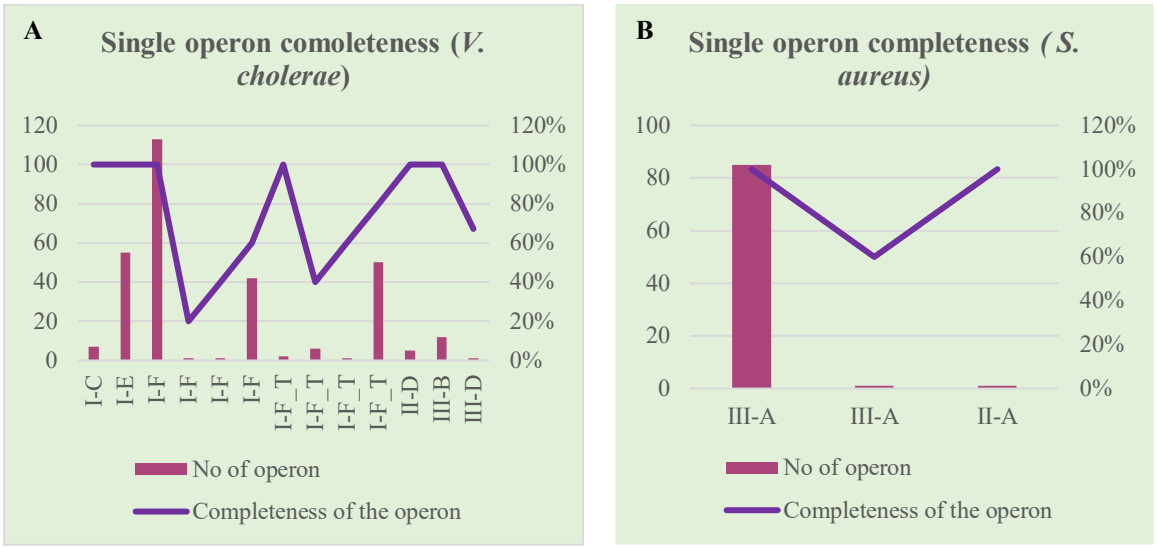


Figure 14: *Single operon completeness according to types. (A) V. cholerae single operon completeness. (B) S. aureus single operon completeness*

4.2.4 Two Operons Distribution: Among 90 strains of *V. cholerae* containing double Cas operon, only 4 contained the same operon twice, while 86 contained different operons. Each of these 90 strains contained two cas operons, accounting for 180 operon total in 90 strains (Figure 15). The four strains containing the same operons, all of them contained I-F Operons and in none of them both operons were found complete. Among the rest 86 strains, the combination of Cas operons are depicted on Figure

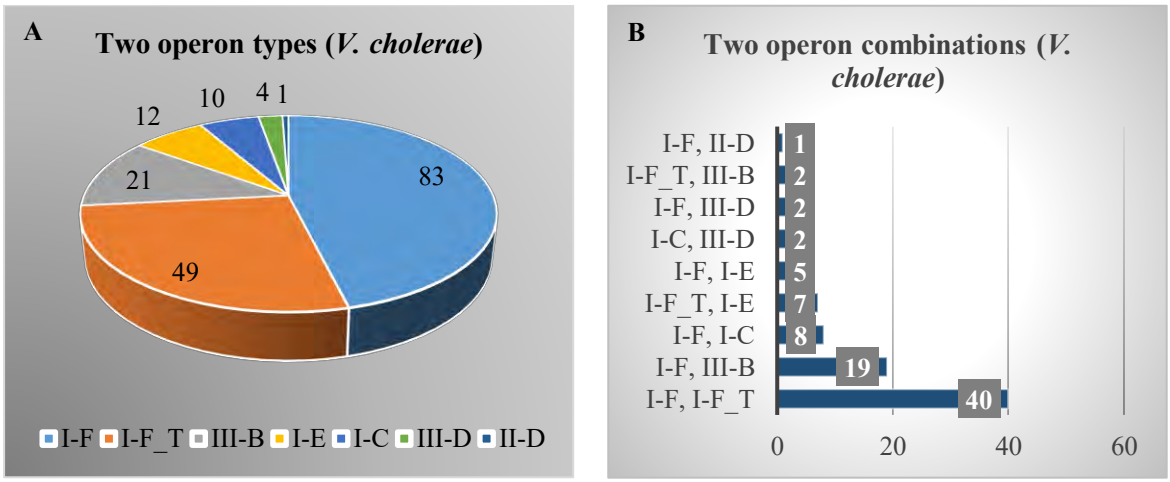


Figure 15: *Two operon types and combinations of V. cholerae. (A) Two operon types. (B) Two operon combinations*

4.2.5 Two Operon Completeness according to Types: Among the 90 strains, 19 were found to contain both operon 100% complete while 50 contained 1 complete operon and one incomplete one. Rest 21 were found to have both the cas operons incomplete.

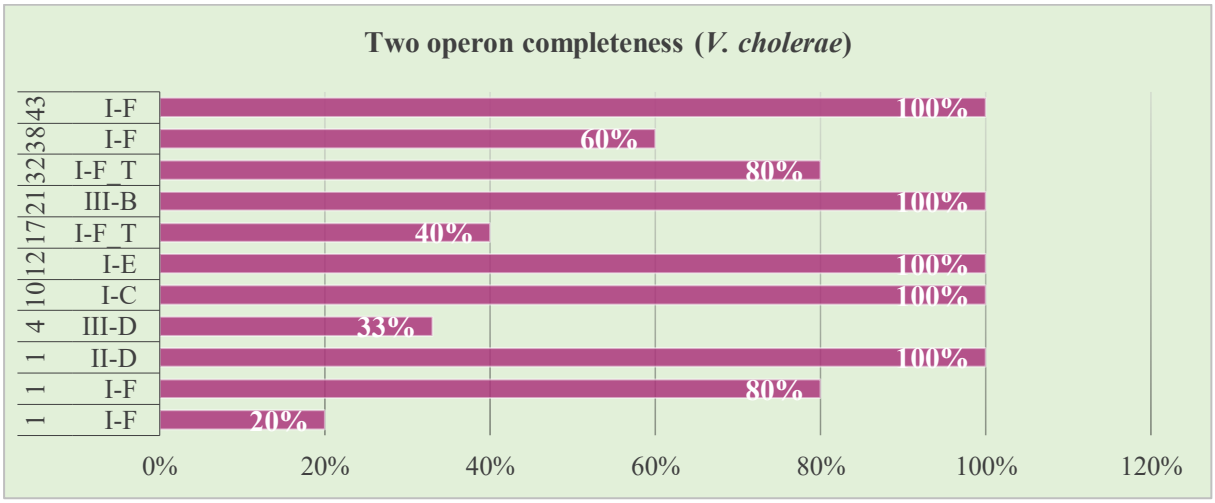


Figure 16: *two operon completeness according to types*

4.2.6 Three Operons Distribution: Among 15 strains having 3 Cas operons, none of them were found to contain 3 same operons. 8 strains were found to contain 2 same operon and one different operon. While 7 strains were found to contain all 3 different Cas operons. Among the 15 strains, each of them contained 3 operons, which accounts for 45 operons in total. The types and combination of Cas operons is depicted in figure 17.

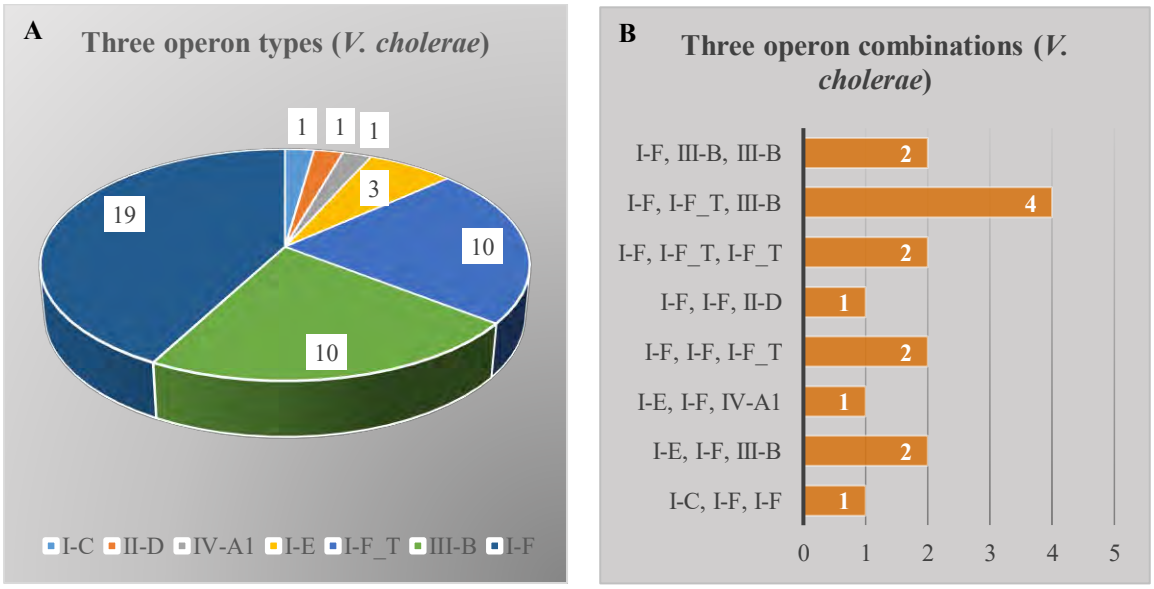


Figure 17: *(A) Three operon types. (B) Three operon combinations of V. cholerae*

4.2.7 Three Operon Completeness according to Types: Among 15 strains, only two were found to contain all operons 100% complete, while 4 were found to contain 2 operons complete and one incomplete. Then, 6 strains were found to contain 1 operon complete and 2 in complete.

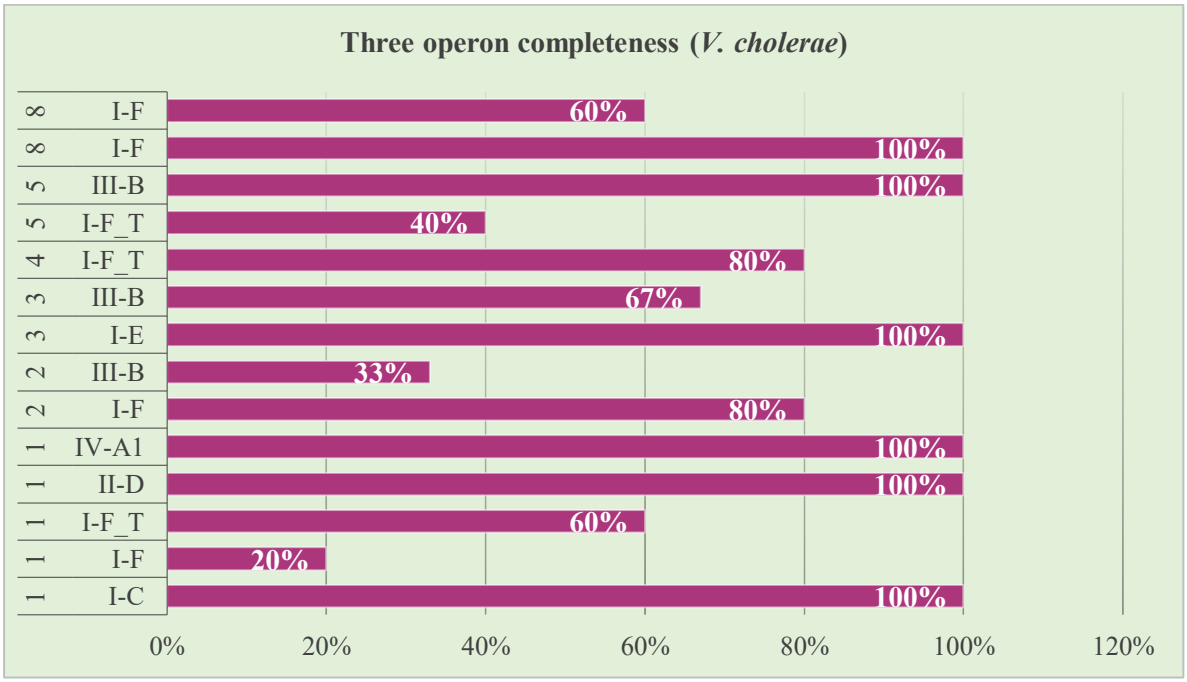


Figure 18: *Three operon completeness according to types*

Overall completeness of the operons are shown in table 7 according to percentage.

Table 7. *Overall completeness of the operons according to types*

Vibrio cholerae									***Staphylococcus aureus*** operon		
Single operon			Two operons			Three operons					
Types	Count	Completeness	Types	Count	Completeness	Types	Count	Completeness	Types	Count	Completeness
I-C	7	100%	1	1	20%	I-C	1	100%	III-A	85	100%
I-E	55	100%	I-F	1	80%	I-F	1	20%	III-A	1	60%
I-F	113	100%	I-F	1	100%	I-F_T	1	60%	II-A	1	100%
I-F	1	20%	II-D	4	33%	II-D	1	100%			
I-F	1	40%	III-D	10	100%	IV-A1	1	100%			
I-F	42	60%	I-C	12	100%	I-F	2	80%			
I-F_T	2	100%	I-E	17	40%	III-B	2	33%			
I-F_T	6	40%	I-F_T	21	100%	I-E	3	100%			
I-F_T	1	60%	III-B	32	80%	III-B	3	67%			
I-F_T	50	80%	I-F_T	38	60%	I-F_T	4	80%			
II-D	5	100%	I-F	43	100%	I-F_T	5	40%			
III-B	12	100%				III-B	5	100%			
III-D	1	67%				I-F	8	100%			
						I-C	8	60%			

4.3 CRISPR Repeat Diversity in the Array

4.3.1 Repeat number variation: The highest and lowest number of repeats present in a single CRISPR array of *V. cholerae* was 141 and 3, respectively. The most abundant number of repeats present in a single CRISPR array was 4, and the second highest was 6 repeats. Among more than 10 repeats, 40 were the most abundant in 26 CRISPR arrays. On the contrary, the highest and lowest number of repeat present in a single CRISPR array of *S. aureus* was 25, 3 respectively. The most abundant number of repeats present in a single CRISPR array was – 4, and the second highest was 5 repeats. Among more than 10 repeats, 16 were the most abundant in 17 CRISPR arrays.

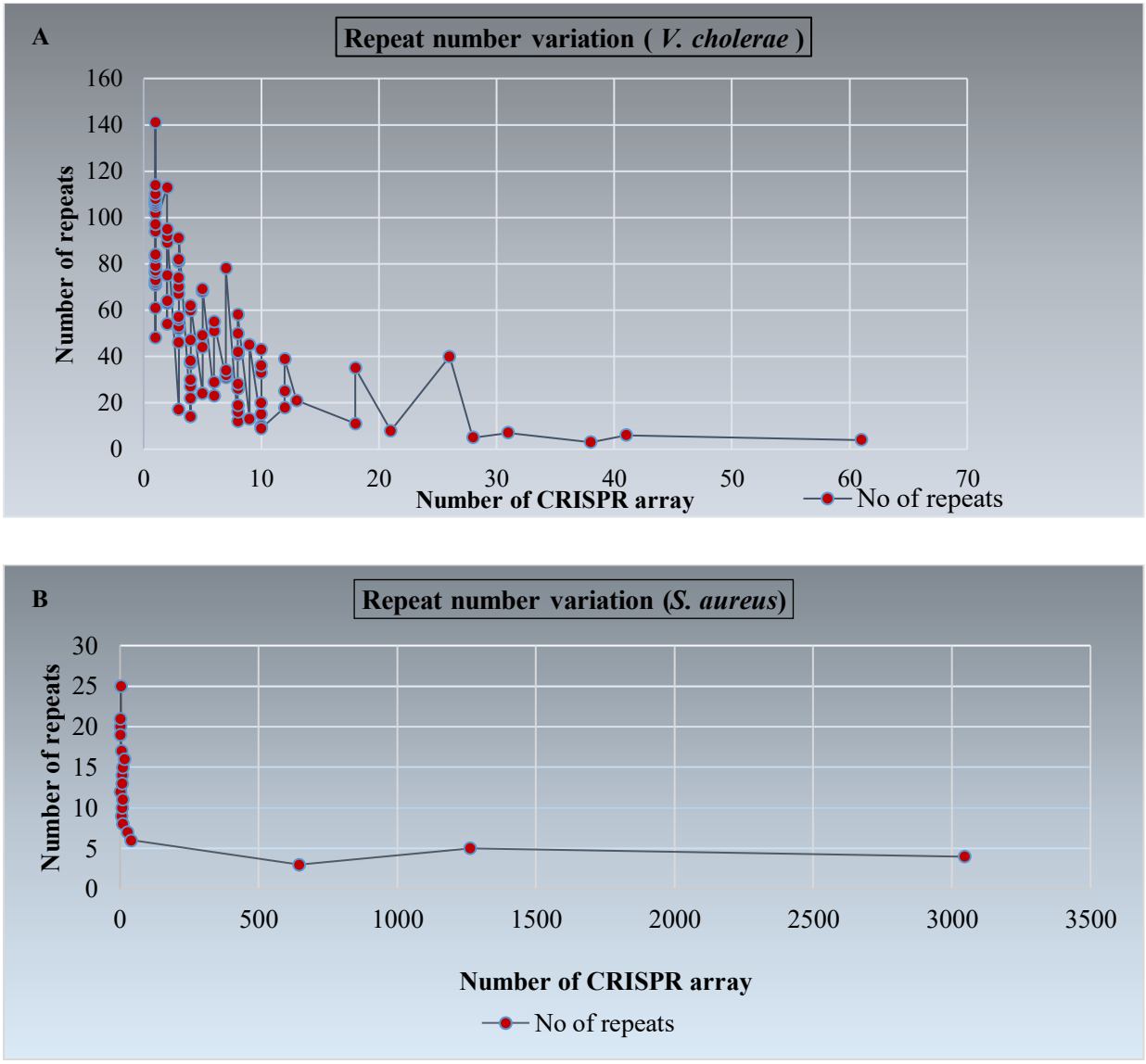


Figure 19: *Repeat number variation according to CRISPR array. (A) V. cholerae repeat number variation. (B) S. aureus repeat number variation*

4.3.2 Repeat length variation: The shortest CRISPR repeat found in *V. cholerae* was 23 basepair, and the longest CRISPR repeat was 38 basepairs. The most abundant repeat length present was 28 basepairs, found in 533 CRISPR arrays. The second highest was 29 basepairs present in only 52 arrays. In *S. aureus* the shortest CRISPR repeat found was 23 basepairs, while the longest CRISPR repeat 43 basepairs. The most abundant length of repeat present was 23 basepairs, found in 2653 CRISPR arrays. The second highest was 30 basepairs present in only 976 arrays.

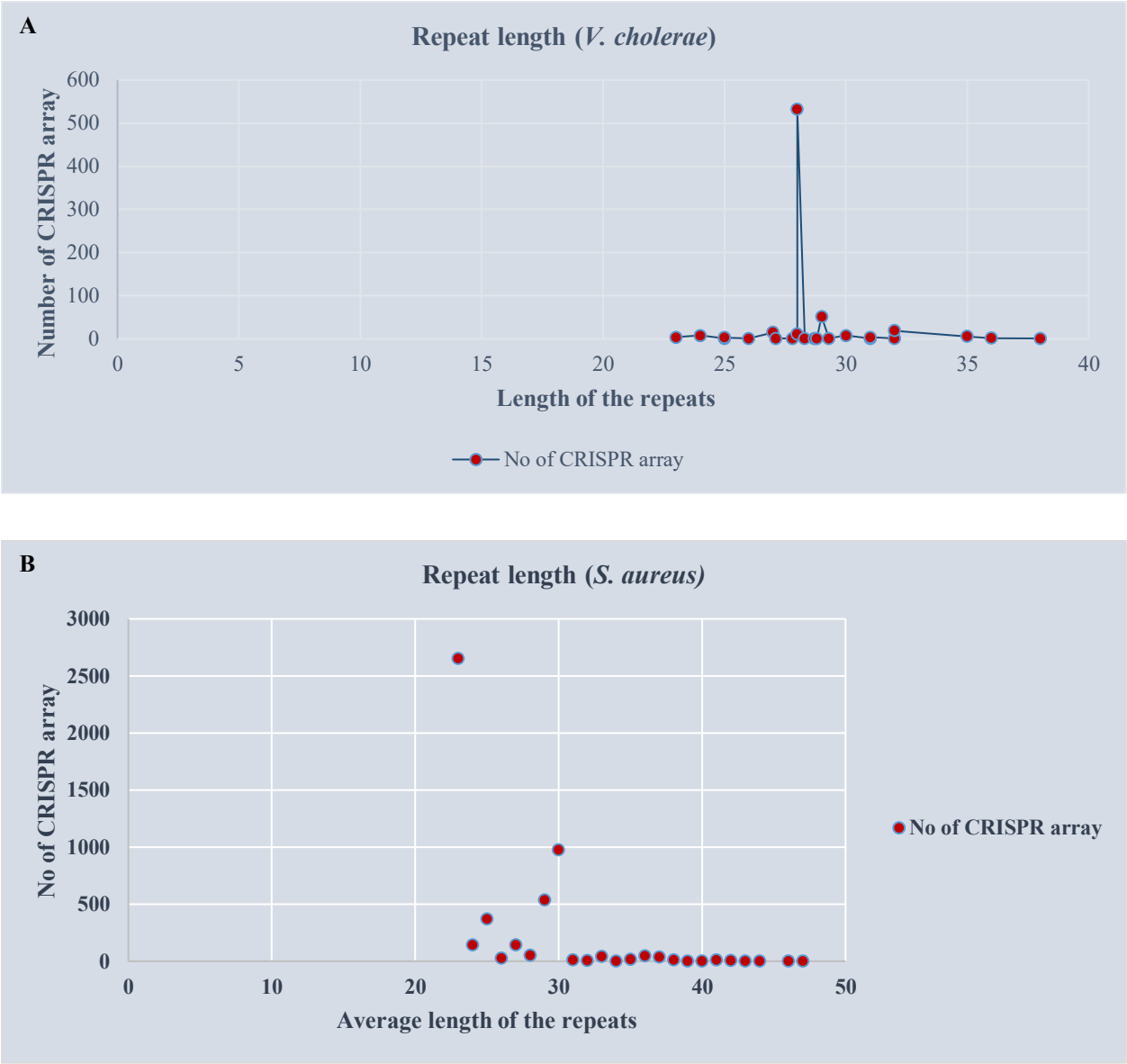


Figure 20: *Repeat length variation within CRISPR array. (A) V. cholerae repeat length variation. (B) S. aureus repeat lenthg variation*

4.3.3 Average Identity of the repeats within the CRISPR array: Maximum and minimum identity of the repeats in the *V. cholerae* strains, in CRISPR array 100% and 76%, respectively. Most of the arrays had repeats that were more than 95% identical. Among them 226 arrays harbored repeats those are 100% identical. 98 arrays with 98% identical, 100 arrays with 97% identical repeats. On the other hand, less than 90% identical repeats were present only in very least number of arrays. In addition, Maximum and minimum identity of the repeats of *S. aureus*, present in array are 100% and 66%, respectively. Most abundant number of repeats found in the CRISPR arrays were 82% identical. Besides 649 arrays harbored repeats those are 100% identical.

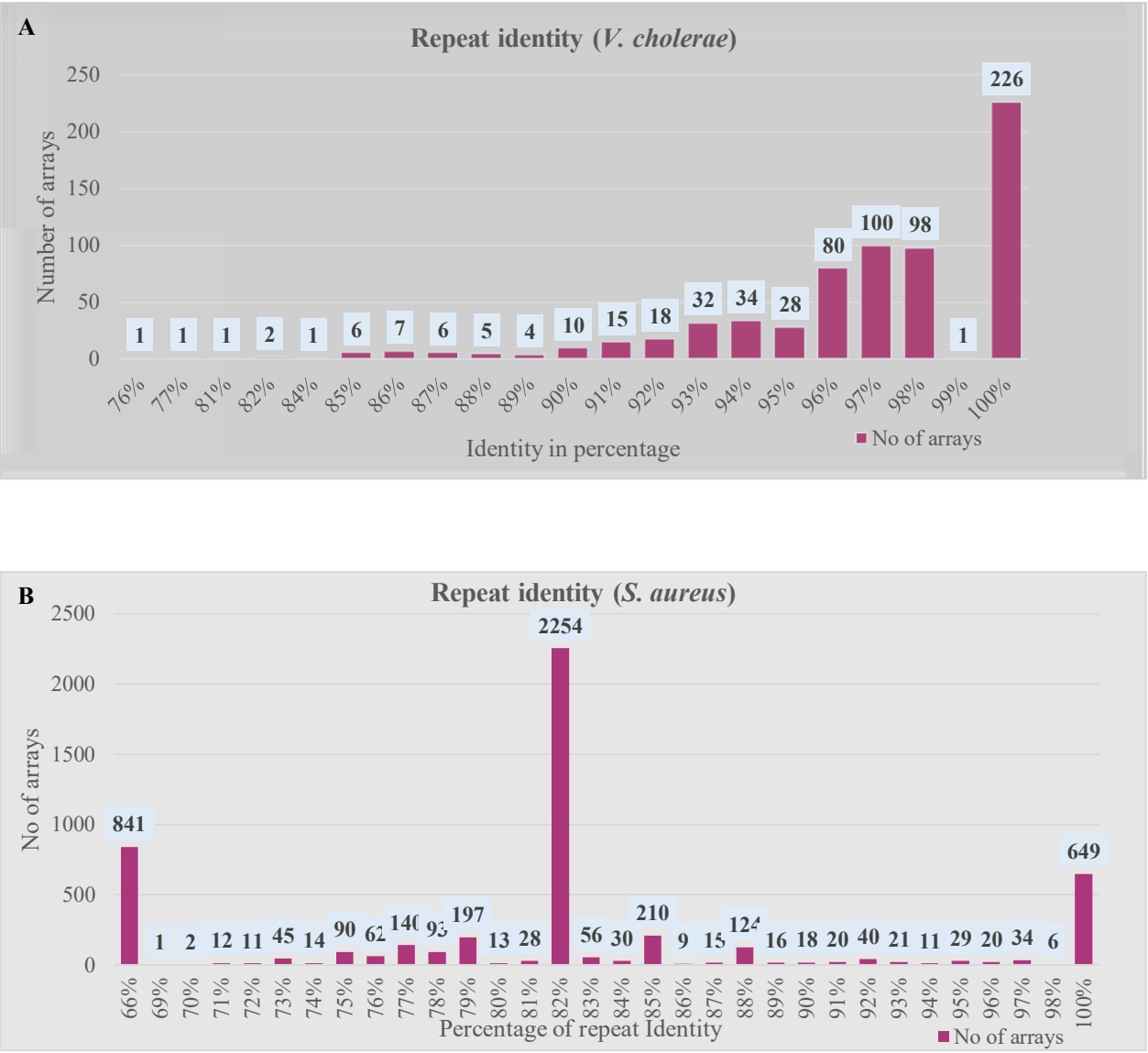


Figure 21: *Average Identity of the repeats within the CRISPR array. (A) V. cholerae repeat identity. (B) S. aureus repeat identity*

4.4 CRISPR Spacer Diversity in the Array

4.4.1 Spacer number variation: The highest and lowest number of spacers present in a single sequence of *V. cholerae* were 150 and 2, respectively. The most abundant number of spacers present in a single sequence was 39. Similarly, in *S. aureus*, highest number of spacers found was 27 while the lowest was 2. The most abundant number of spacers present in a single sequence was 3. Overall, the distribution is shown in figure

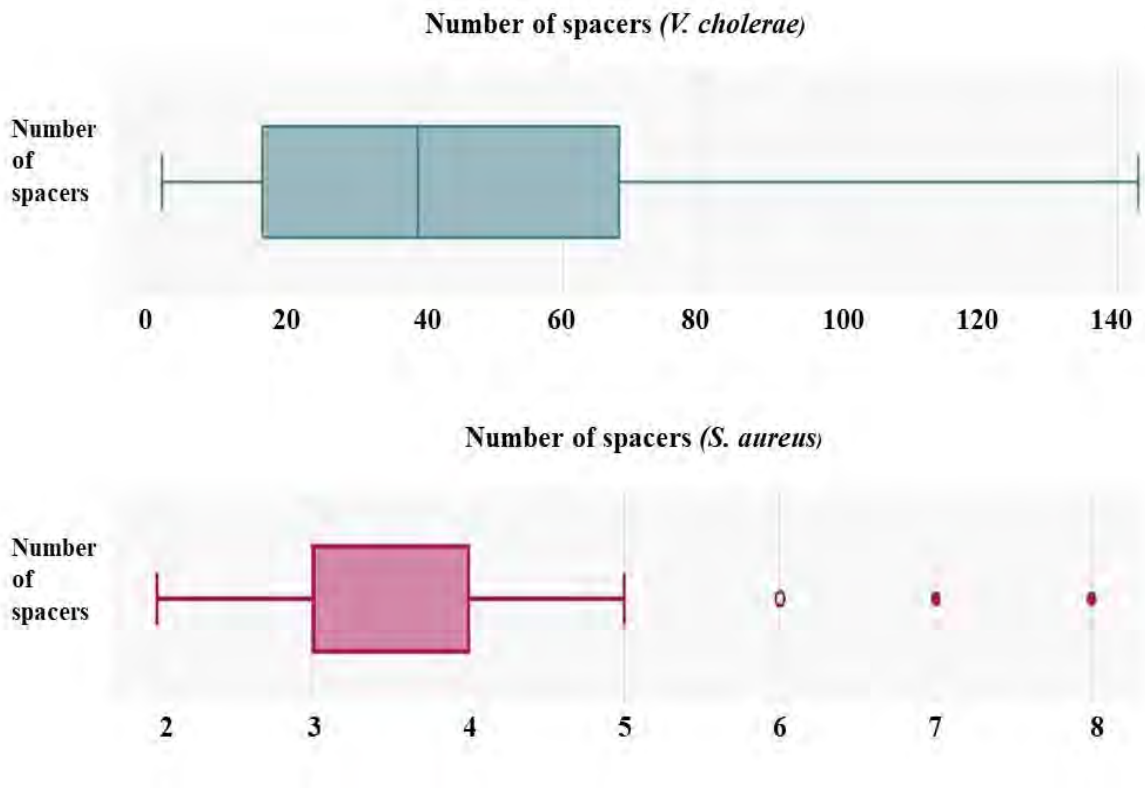


Figure 22: *Spacer number variation according to CRISPR array. (A) V. cholerae spacer number variation. (B) S. aureus spacer number variation*

4.4.2 Spacer length variation: In *V. cholerae* shortest and the longest CRISPR spacer was 29 bp and 41 bp, respectively. On the other hand, it was 26 bp and 82 bp in *S. aureus*. The most abundant length of repeat present in *V. cholerae* was 32 bp (462 CRISPR arrays). While in *S. aureus*, the most frequent spacer was 34 bp long and occurred 2242 times in the arrays.

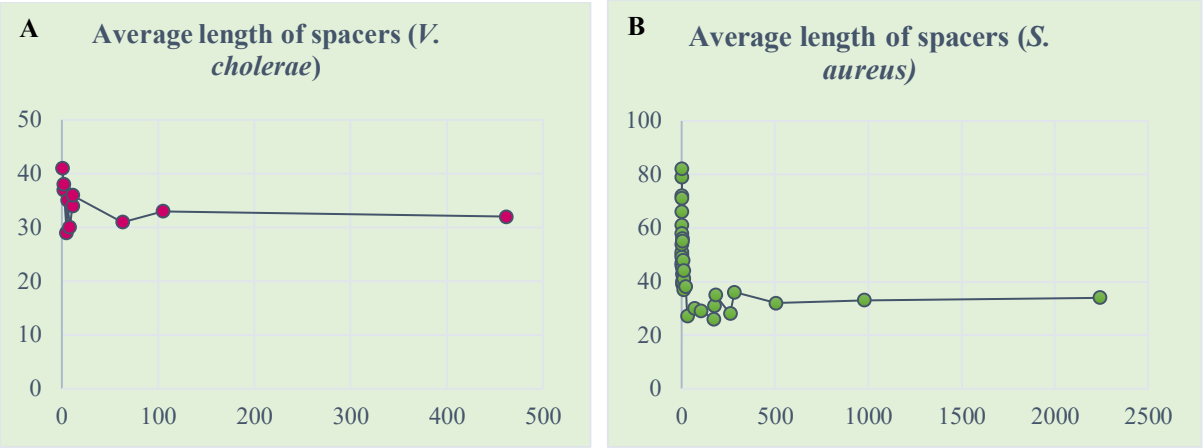


Figure 23: *Spacer length variation among CRISPR arrays. (A) V. cholerae spacer length variation. (B) S. aureus spacer length variation*

4.4.3 Average Identity of the Spacers within the CRISPR array: Maximum and minimum identity of the spacers present in *V. cholerae* array were 57% and 33% respectively. 88.31% of the arrays had spacers that were 40% to 45% Identical. Among them 190 arrays harbored spacers those are 41% identical, which is the highest prevalence. 98.81% of the arrays contained spacers those shared identity lower than 50% among themselves in the CRISPR array. In *S. aureus*, maximum and minimum identity of the spacers present in an array were 81%, and 15%, respectively. 88.31% of the arrays had spacers that were 45% to 70% Identical. Among them 2187 arrays harbored spacers those are 59% identical, which is the highest prevalence.

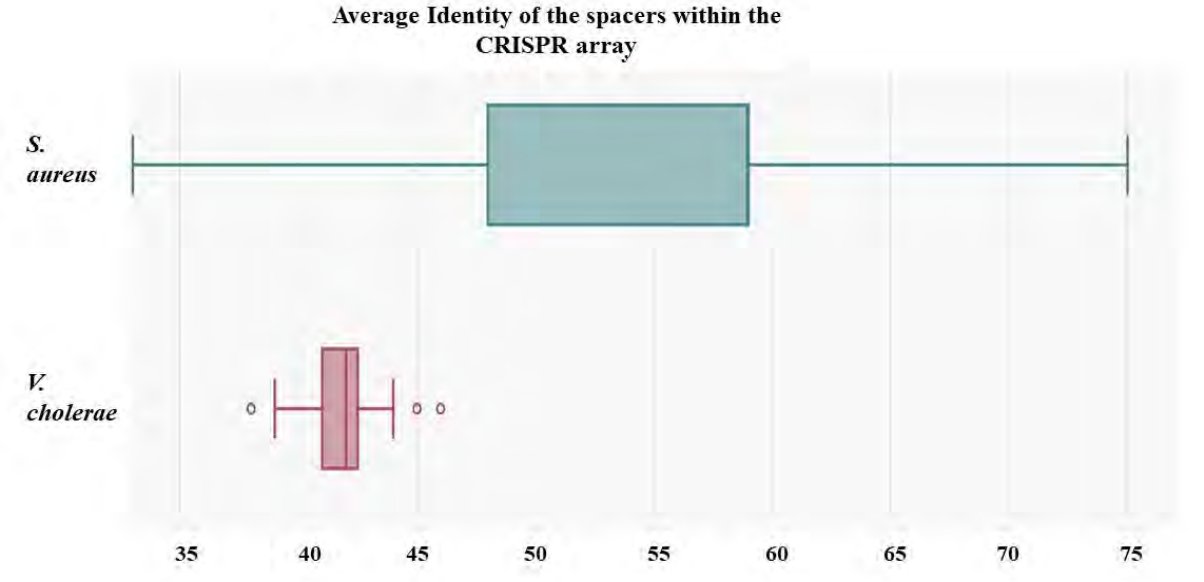


Figure 24: *Average Identity of the spacers within the CRISPR array*

4.5 Self-targeting spacers (STS) Diversity

Among 1794 strains of *V. cholerae*, 33 strains were found to harbour 68 STS which is on average 1.84% of the total genome. On the other hand, 16286 *S. aureus* strain were found to harbour STS in 3089 strains, which is around 18.97% of the total genome.

Table 8. *Self-targeting spacers (STS) summary*

Name of the species	Total genome	Total CRISPR array containing genome	Total spacers	Total STS containing genome	Total STS found	Average CRISPR array containing genome in percentage	Average STS containing genome in percentage
V. cholerae	1794	408	18223	33	68	22.74%	1.84%
S. aureus	16286	4107	16967	3205	4484	25.22%	19.68%

Figure 25 represents the STS summary we found in both *V. cholerae* and *S. aureus*.

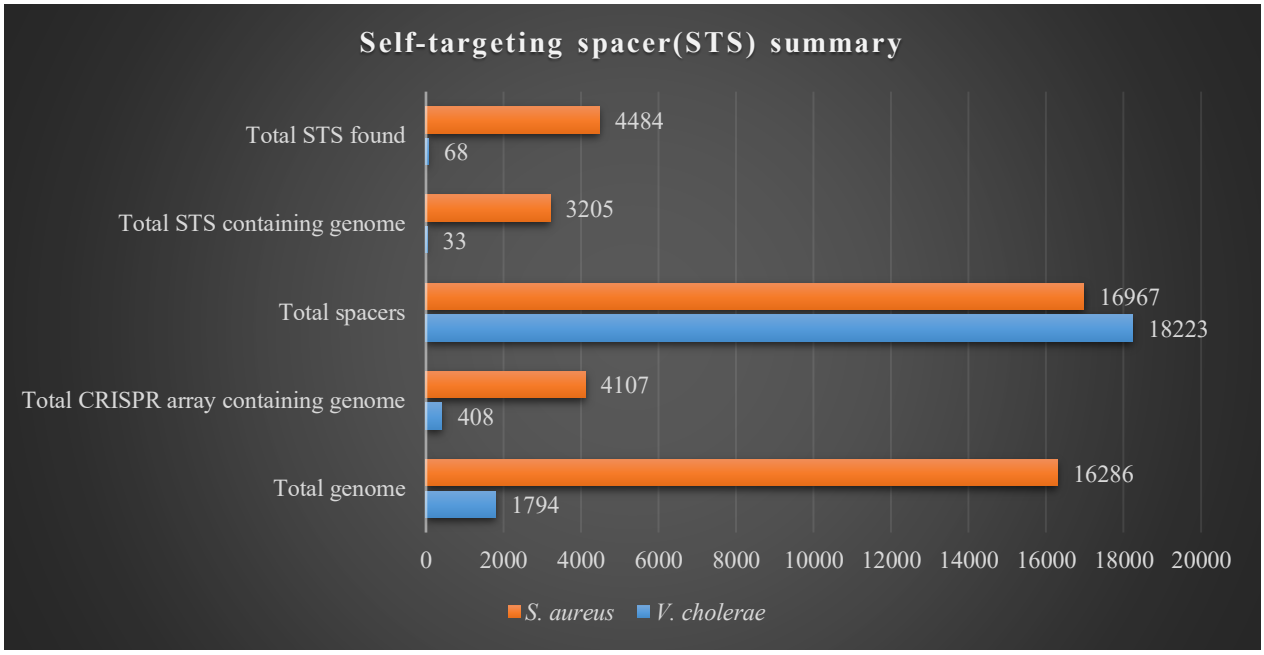


Figure 25: *Self-targeting spacers (STS) summary*

In this part, a quantitative analysis was performed and visualized among total genomes, CRISPR-containing genomes, and STS-containing genomes to see the percentage in which the CRISPR and STS are occurring in both species.

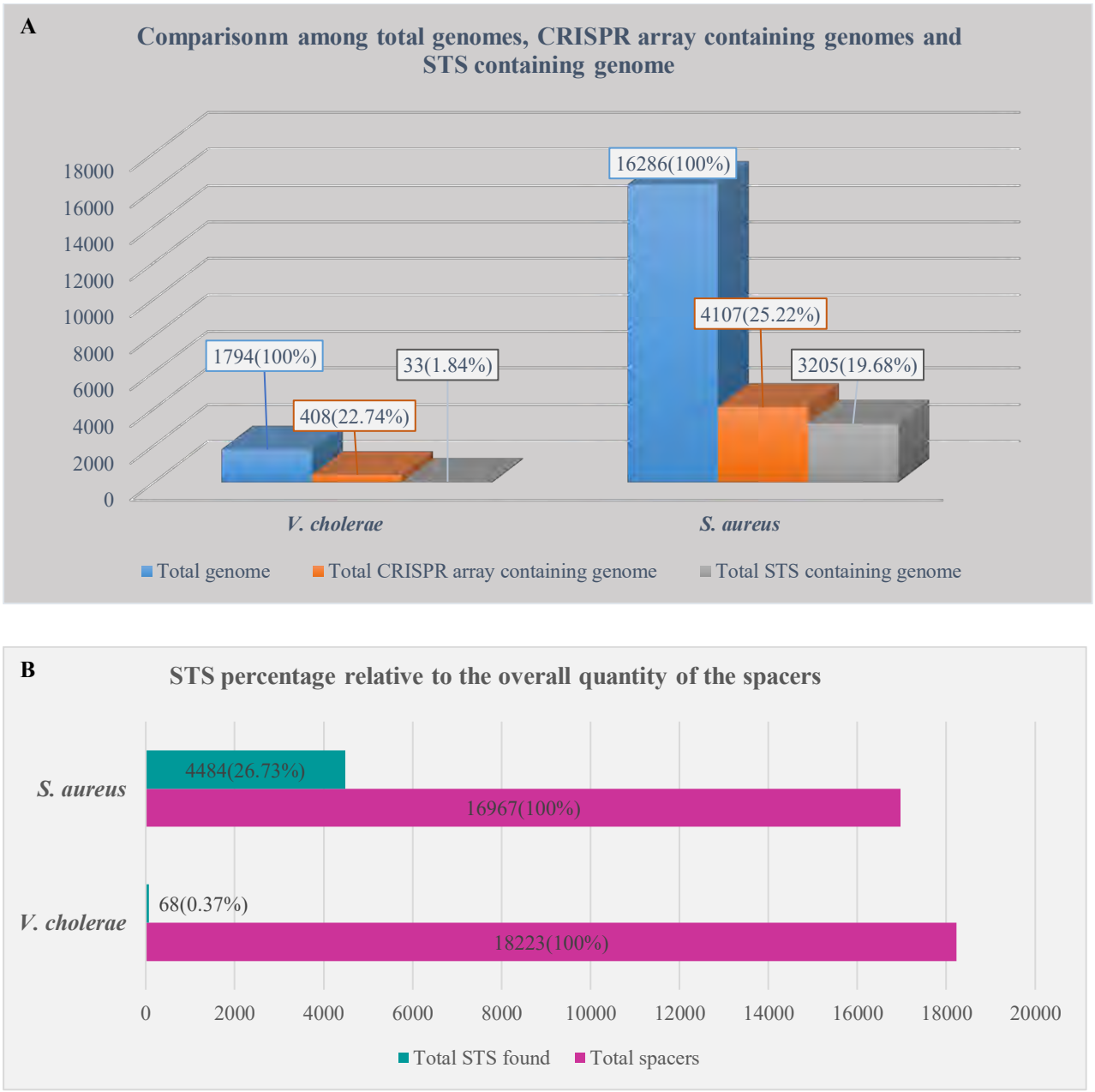


Figure 26: *(A) Comparison among total genome, CRISPR-containing genomes, and STS-containing genome. (B)* STS *percentage relative to the overall quantity of the spacers*

We further mapped the position of the STSs within their CRISPR array to see the position. Surprisingly, the STSs were found to occur in any position within the array for *V. cholerae*. But for *S. aureus* they were found to concentrate around the 1st to 5th position mostly.

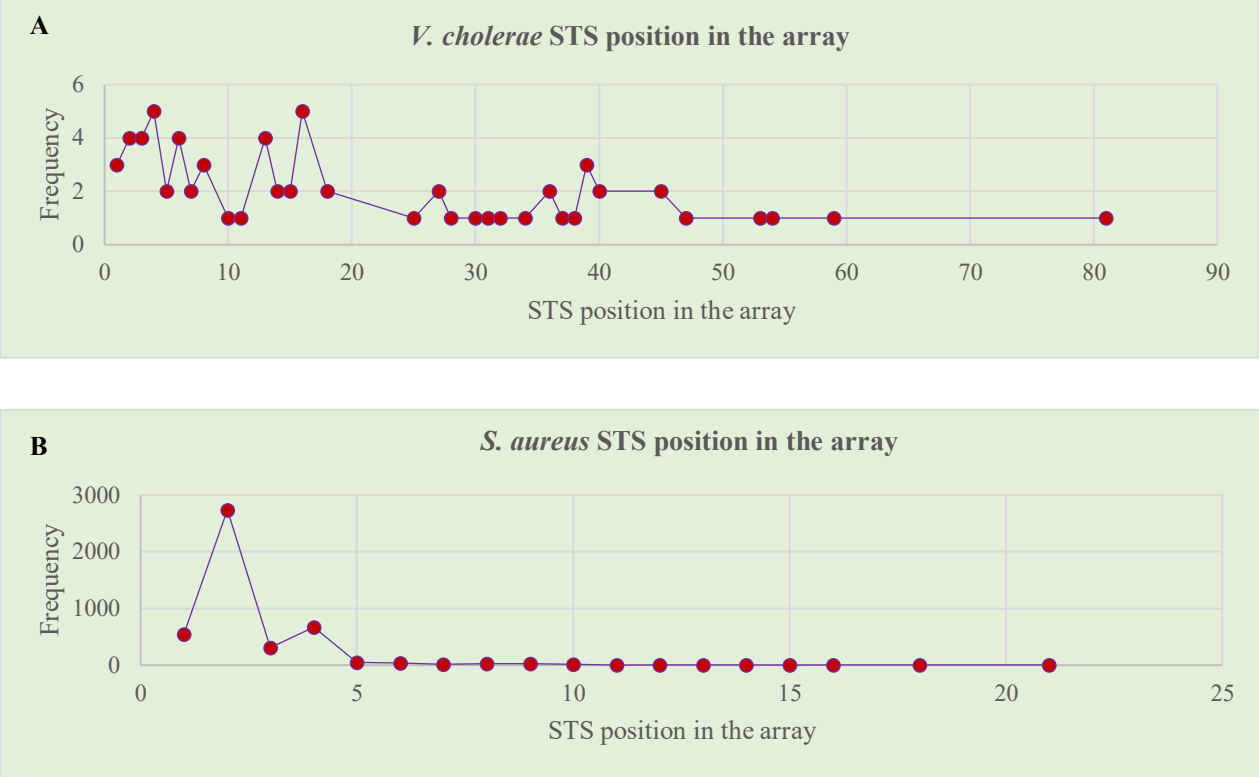


Figure 27: STS *mapping within the CRISPR array. (A) V. cholerae* STS *position in the array. (B) S. aureus* STS *position in the array*

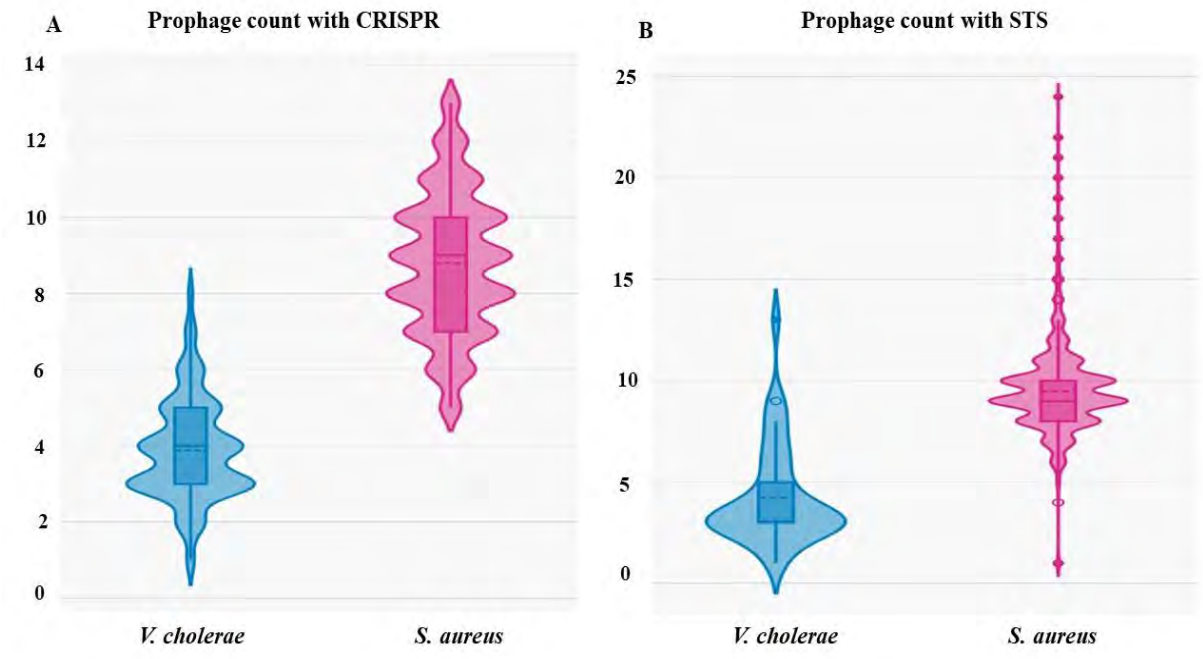
4.6 Prophage count

When a bacteriophage viral genome gets integrated into its hosts bacterial genome, it is termed as a prohage. Prophage presence in the host genome is a very common phenomena. And pathogenicity is extensively aasociated with the presence of prophage. As our aim is to see the comparative pathogenic profile, in the next step we tried to measure the association of prophages according to different CRISPR-Cas immunity context scenario.

4.6.1 Prophage count with CRISPR: Total 366 *V. cholerae* and 901 *S. aureus* CRISPR containing strains were analyzed for prophage count. All of the strains in both species had been found to contain at least one prophage. While majority (74 strains) of *V. cholerae* were found to contain 4 prophages, 231 strains of *S. aureus* were containing 8 prophages. The maximum number of prophages incorporated into *V. cholerae* genome is 12 prophages in GCF_001857545 strain and for *S. aureus* it was carrying 47 prophages in GCF_000683395 strain. For *V. cholerae* 366 strains were containing total 1499 prophages that is on average 4.096 per strain. On the other hand, 901 *S. aureus* strains were carrying total 6922 prophages which is on average 7.683 prophages per strain.

4.6.2 Prophage count with STS: 42 and 3205 STS containing *V. cholerae* and *S. aureus* strains respectively were analyzed for prophage count. All of the STS containing strains are found to contain at least one prophage. 42 *V. cholerae* strains were carrying 177 prophages that is on average 4.214. The maximum number of prophages were carrying by the *V. cholerae* strain GCF_006802685, and it was carrying total 13 prophages. In addition, 3205 *S. aureus* strains were carrying 30368 prophges and 9.475 in average. One strains of *S. aureus*, GCF_024454375 were carrying the maximum number of prophages and that is 24.

4.6.3 Prophage count without CRISPR: Total 1385 *V. cholerae* and 12173 *S. aureus* strains without CRISPR in their system were analyzed. Again, it has been found that all of them were carrying atleast one integrated prophage. Majority of the *V. cholerae* strains (total 473 strains) were containing 3 prophages, while it is 8 prophages by 2262 strains of *S. aureus*. Maximum number of prophages carrying by *V. cholerae* strain GCF_019780645 and the prophage count is 19. On the other hand, two *S. aureus* strains GCF_000247275, GCF_014638515 respectively were carrying 27 prophages in their genome. 1385 *V. cholerae* strains were carrying total 5465 prophages which stands for 3.946 on average. On the other hand, 12173 *S. aureus* strains were carrying 111767 prophages, which is 9.182.



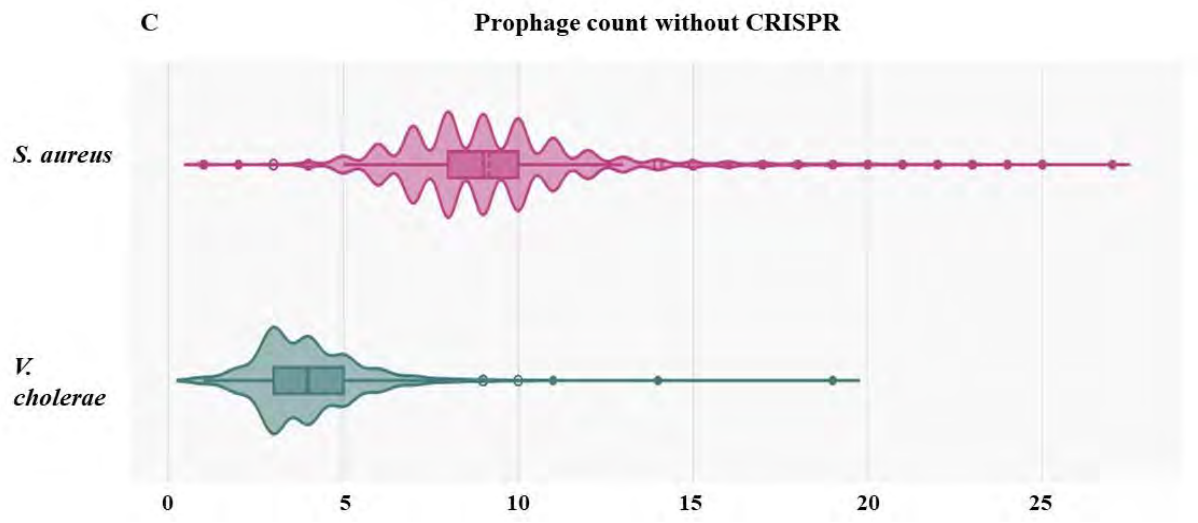


Figure 28: *Prophage count. (A) Prophage count with CRISPR. (B) Prophage count with STS. (C) Prophage count without CRISPR*

4.7 Pathogenic Gene Count

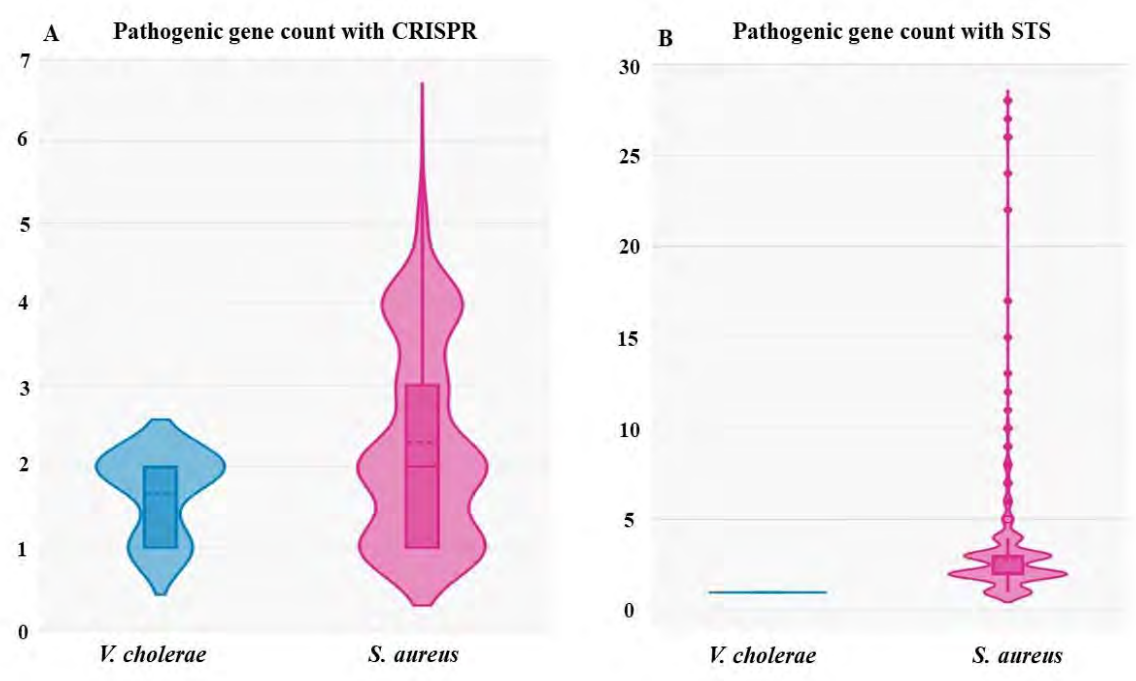
For the count of Pathogenic gene, we classified the all the strains into three distinct groups and they are PF gene count with CRISPR, PF gene count with STS and PF gene count without CRISPR.

4.7.1 Pathogenicity with CRISPR: Total 365 *Vibrio cholerae* and 901 *Staphylococcus aureus* CRISPR containing strains were analyzed for the count of pathogenicity factors. Interestingly, out of 365 strains, very low number of strains i.e, 22 *V. cholerae* strains were carrying total 48 pathogenic factors which is 0.315 on average. On contrary, pathogenic factors were higher in CRISPR containing *Staphylococcus aureus* strains. Out of 901 strains, 569 strains were containing total 1399 pathogenic factors. That makes the average pathogenic factors of CRISPR containing *S. aureus* strains 1.553. The highest number of pathogenic factors for *V. cholerae* strains are GCF_001471455, GCF_001525525 (containing 5 factors). On the other hand, maximum 34 pathogenic factors were found in one *S. aureus* GCF_000361625 strains.

4.7.2 Pathogenicity with STS: 43 *V. cholerae* and 3205 *S. aureus* STS containing strains were analyzed to see whether the pathogenic factors are increased after the acquisition of STS. Surprisingly only one *V. cholerae* strain was found to one pathogenic factor (strain GCF_011750775). That makes the average pathogenic factor count 0.0238. On the other hand, among 3205 STS containing *S. aureus* strains, 2817 strains were containing total 8022

pathogenic factors which is 2.503 on average. Most *S. auresus* (total 1064 strains) were reported to conatin 2 factors and the highest pathogenic factors containing strains are GCF_000562145 and GCF_001717685 (containing 28 Pathogenic factors).

4.7.3 Pathogenicity without CRISPR: Pathogenicity of *V. cholerae* is significantly increased in the absence of CRISPR in their system. Among 1386 non CRISPR containing CRISPR strains, 757 strains were found to contain 1547 pathogenic factors i.e., on average 1.117 per strain. In addition, most of the strains (total 711) were containing 2 pathogenicity factors and one strain (GCF_019704235) was found to have 14 pathogenic factors. On contrary, 12173 *S. aureus* strains those were not containing CRISPR system were analyzed and we found that 8665 strains were carrying total 20023 pathogenic factors which is 1.665 on average per strain. Most strains hav been found to carry one (2923 strains) or two (2749 strains) pathogenic factors. Strikingly one strain GCF_000361385 was carrying the maximum 36 pathogenic factors.



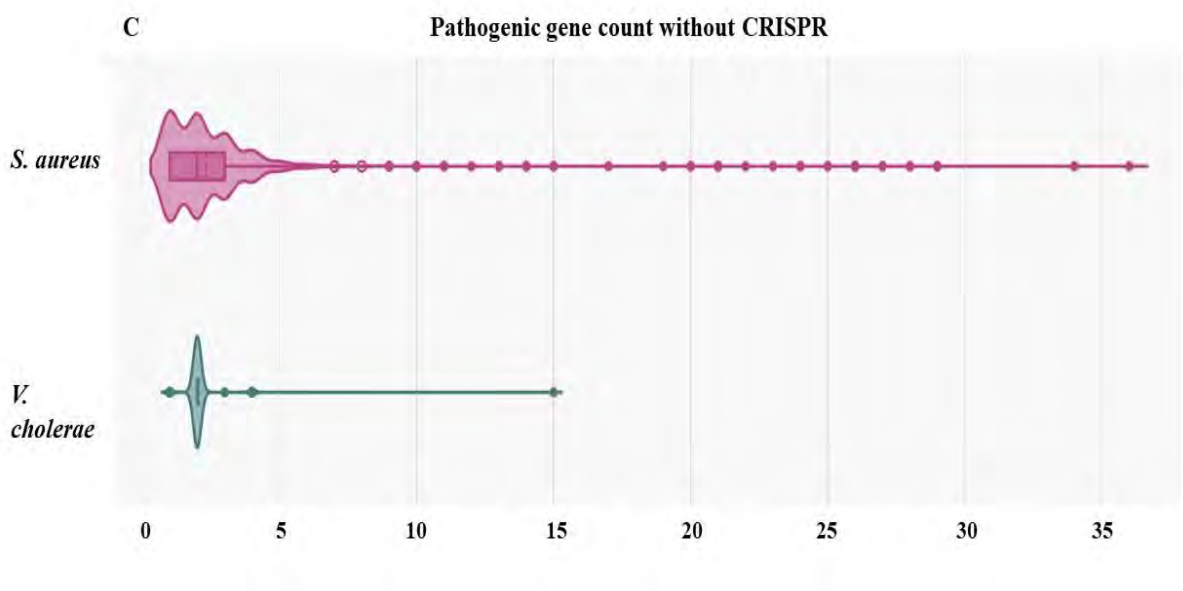


Figure 29: *Pathogenic gene count. (A) Pathogenic gene count with CRISPR. (B) Pathogenic gene count with STS. (C) Pathogenic gene count without CRISPR*

Chapter 5
DISCUSSION

Total 1794 *V. cholerae* and 16286 *S. aureus* genome had been found to harbour 19.29% and 0.51% functional CRISPR system, respectively. The previous finding was 0.83% for *S. aureus* in 761 genomes (Cruz-López et al., 2021). This small difference may be due to the vast number of genomes we have taken. In addition, in both cases, we found orphan Cas operon, operons those are not complete or lacking CRISPR array and the number were 56 and 4 for *V. cholerae* and *S. aureus* accordingly. On the other hand, in 6 *V. cholerae* and 4024 *S. aureus* genome, CRISPR arrays were not accompanied by Cas genes. There can arise few reasons behind this scenario. First, CRISPR genes might have undergone deletion evolutionarily after the incorporation of self-targeting spacers (STSs) to inactivate the CRISPR-Cas mediated immunity (Koonin & Makarova, 2019). Else it can be proven to be lethal. As Cas gene deletion is a common phenomenon. Other possible reasons can be, de novo array formation through integration of off-target spacers into sequences that resemble the appropriate repetitions; and transfer via mobile genetic elements such as transposons or plasmid, where they didn't carry the Cas genes (Shmakov et al., 2020).

However, STSs have been found in 33 *V. cholerae* (1.84%) and 3205 *S. aureus* (19.68%) genomes. That is STSs are much higher occurring scenario in *S. aureus*. The potential reason can be, in most of the *S. aureus* strain we have found type 3 CRISPR Cas system. As the type III-A system can attack both DNA and RNA (Samai et al., 2015), it can generate a lot of substrate for spacer selection.

V. cholerae had been found to contain multiple Cas operon ranging from 1 to 4, while aureus had been found to carry single Cas operon. On the other hand, CRISPR array distribution is much diverse in *V. cholerae* (ranging from 1 to 11). While in *S. aureus* the array ranges from 1 to 4, mostly are single array. The probable reason can be that merely a small percentage (approximately 0.83%) of *S. aureus* strains possess CRISPR-Cas systems, and when these systems are present, they generally have a reduced number of spacers. In contrast, *V. cholerae* exhibits multiple CRISPR arrays across various strains, indicating an evolutionary adaptation to its pathogenic lifestyle and environmental pressures. Besides, *V. cholerae* displays a range of CRISPR-Cas types, encompassing various variant systems like type I-F and type III-B, whereas *S. aureus* primarily shows a restricted occurrence of type III-A systems. On the other hand, *Vibrio cholerae*, an aquatic pathogen, engages with diverse phages in its surroundings, resulting in the frequent acquisition of spacers from these contacts. Conversely, *S. aureus*

predominantly resides in mammalian hosts, where the selective pressure for varied CRISPR systems may be less significant due to differing dynamics of phage exposure.

In the question of completeness, the *S. aureus* Cas operons are much complete than the *V. cholerae* operons. Though most of the *V. cholerae* strains are complete, its completeness ranges from 20% to 100%. On the other hand, in *V. cholerae* not only the CRISPR types and their distribution is much diverse but also, they many types had been found to occur in combination. In the contrary, *S. aureus* is basically found to harbor type IIIA system.

The average repeat number for *V. cholerae* and *S. aureus* are 28 and 4 respectively. That is the repeat number is almost 7 times higher in the *V. cholerae*. On the other hand, the repeat length found in both species is almost the same i.e, 28 for *V. cholerae* and 26 for *S. aureus*. And this finding is aligned with the statement that repeat length ranges between 18-52bp depending on the types (Alkhnbashi et al., 2021). In addition, the *V. cholerae* repeats are found to be more identical (96.67%) than the *S. aureus* (81.76%). The proposed reason can that, almost all *S. aureus* strain is containing type III-A CRISPR cas system. And unlike with the other types, type III-A system recruit its "repeats" to differentiate between "self" vs "non-self" due to the absence of PAM recognizing machinery in their system (Kolesnik et al., 2021). However, this variability in the repeats will provide them advantage by restraining from attacking their own genome. On the other hand, Research indicates that bacterial species possessing highly active CRISPR systems are likely to retain more conserved repeats, which is essential for the functionality of the system (Makarova et al., 2011). Consequently, the increased conservation observed in *V. cholerae* could indicate a more significant or vital function for its CRISPR system in comparison to *S. aureus*.

Both species greatly vary in term of spacer number. In *V. cholerae* the average spacer number is 39 per CRISPR array containing strain while in *S. aureus* the number is much lower and it is 4 per strain. Besides, *V. cholerae* had been found to contain a higher number of spacers. As we have mentioned previously, the CRISPR types in *V. cholerae* is much diverse and due to its common habituation in the aquatic environment it might encounter with broad spectrum of MGEs, which in turn will lead to the higher number of spacers. On the other hand, 75% of the CRISPR containing strains of *S. aureus* were bearing self-targeting spacers. In a study, STSs are mostly mapped on the 1st or 2nd place on the CRISPR array, adjacent to the leader sequence(Nobrega et al., 2020). But in our data we found STS, randomly distributed.

In term of spacer length, both strains are having almost the same size of spacer i.e., 32bp for *V. cholerae* and 33bp for *S. aureus* aligned with the average spacer lengths documented in previous studies (Shmakov et al., 2017). In addition, the spacers are 41.75% identical in *V. cholerae* and 56.07% identical in *S. aureus*.

In terms of prophage count both species showed variation in number. However, carrying an integrated phage is a very common phenomenon for prokaryotes. In lieu with that statement, we found both bacteria containing at least one prophage. The average prophage count *in V. cholerae*, in the presence of CRISPR is higher (Table 9) than the in the absence of CRISPR. On the contrary, the average prophage count is getting increased in both of the species, in the presence of STSs.

Table 9. *Prophage count summary*

Name of bacteria	Average prophage counts without CRISPR	Average prophage counts with CRISPR	Average prophage counts with STSs
V. cholerae	3.94584837545	4.10684931507	4.21428571429
S. aureus	9.18154933049	7.68257491676	9.4751950078

While the average pathogenic factor count is increasing significantly in *V. cholerae* in the absence of CRISPR, in *S. aureus* it is not remarkably higher. On contrary, in *V. cholerae*, only one STS containing strain being found to carry one pathogenic factor. A significant decrease in *V. cholerae* psthogenic gene count was recorded with CRISPR (from 1.12 to 0.13) and with STS (to 0.024), suggesting that the active CRISPR-Cas system may inhibit the expression of genes linked to pathogenicity. This may happen by directly targeting pathogenic genes or by broader modulation of mobile genomic elements that often contain virulence genes. Self-targeting spacers (STS) indicate probable disruption of critical or pathogenicity-related genes, leading to a substantial decrease in PF count, possibly due to autoimmunity.

Table 10. Pathogenic gene count summary

Name of bacteria	Average pathogenic gene count without CRISPR	Average pathogenic gene count with CRISPR	Average pathogenic gene count with STSs
V. cholerae	1.11696750903	0.13150684931	0.0238095238
S. aureus	1.64486979381	1.55271920089	2.50296411856

In *S. aureus*, the CRISPR-Cas system may demonstrate a diminished inhibition of pathogenic gene expression relative to *V. cholerae*. The reduction in pathogenic count with CRISPR (from 1.65 to 1.55) signifies a little regulatory effect on virulence factors. The significant increase in pathogenic gene count with STS (to 2.50) indicates that STS may promote horizontal gene transfer or activate pathogenicity islands, leading to enhanced virulence. This may happen if STS interferes with regulatory elements that control the integration of virulence genes or if the bacteria adapt to STS by obtaining supplementary virulence factors. Furthermore, *S. aureus* may utilize an unidentified mechanism to coexist with STS-containing systems, augmenting this bacterium's overall pathogenic gene count of this bacterium.

Chapter 6
CONCLUSION

This work sought to examine the alterations in pathogenicity of two particular bacteria (*V. cholerae* and *S. aureus*) resulting from the presence of CRISPR-Cas and self-targeting spacer (STS) within their genomes. Analysis of 1794 *V. cholerae* and 16286 *S. aureus* genomes (RefSeq) revealed that 22.74% of *V. cholerae* and 25.22% of *S. aureus* possess CRISPR arrays in their genomes. The functional CRISPR-Cas system (Cas operon supported by CRISPR array) is present at 19.29% in *V. cholerae* and 0.51% in *S. aureus*, respectively. Furthermore, 1.84% of *V. cholerae* and 19.68% of *S. aureus* genomes were identified to possess STS inside their CRISPR array. The standard the pathogenic gene count without CRISPR, was 1.12 for *V. cholerae* and 1.65 for *S. aureus*, respectively.

A notable reduction in *V. cholerae* pathogenic gene count was seen with CRISPR (from 1.12 to 0.13) and with STS (to 0.024), indicating that the active CRISPR-Cas system may proficiently target and suppress the expression of genes associated with pathogenicity. This may occur by directly targeting pathogenic genes or by broader modulation of mobile genomic elements that frequently harbor virulence genes. Self-targeting spacers (STS) suggest a possible disruption of critical or pathogenicity-related genes, resulting in a significant reduction in average pathogenic gene count, potentially attributable to autoimmunity or self-destructive mechanisms. In *Staphylococcus aureus*, the CRISPR-Cas system may exhibit a diminished efficacy in suppressing pathogenic gene expression compared to *Vibrio cholerae*. The noted decrease in average pathogenic gene count with CRISPR (from 1.65 to 1.55) indicates a negligible regulatory impact on virulence factors. The substantial rise in pathogenic gene count with STS (to 2.50) suggests that STS may facilitate horizontal gene transfer or activate pathogenicity islands, resulting in increased virulence. This may occur if STS disrupts regulatory elements governing the integration of virulence genes, or if the bacteria adapt for STS by acquiring supplementary virulence components. The *S. aureus* CRISPR-Cas system may employ an unexplained mechanism to cohabit with STS-containing systems, hence augmenting the overall PF count of this bacteria.

Nonetheless, the cause remains unidentified and may be numerous. This facilitates the need of additional inquiries. We have utilized two types of bacteria: gram-positive and gram-negative. Further inquiry may be conducted to ascertain the pathogenic gene count scores in more Gram-positive and Gram-negative bacteria. Should the outcome correspond with our findings, a hypothesis may be formulated. Conversely, as previously noted in the preceding section, a type IIIA system may possess an unidentified mechanism that coexists with STSs, warranting

additional investigation. CRISPR Cas typing is conducted based on variations in Cas proteins and repetitive sequences. Our findings revealed that *S. aureus* exhibited a diversification of CRISPR-Cas types inside its system. However, the elevated prevalence of STS in *S. aureus* may have led to the deletion of certain Cas operons as an adaptive method to evade lethality.

In conclusion, this study will provide thoughtful insight into the pathogenicity profiling in bacteria in the numerous ongoing research on CRISPR-Cas territory, which might be helpful in many sectors of biological sciences.

Chapter 7
REFERENCES

Abedon, S. T., Thomas-Abedon, C., Thomas, A., & Mazure, H. (2011). Bacteriophage prehistory: Is or is not Hankin, 1896, a phage reference? *Bacteriophage*, *1*(3), 174–178. https://doi.org/10.4161/bact.1.3.16591

Agarwal, J., Srivastava, S., & Singh, M. (2012). Pathogenomics of uropathogenic Escherichia coli. *Indian Journal of Medical Microbiology*, *30*(2), 141–149. https://doi.org/10.4103/0255-0857.96657

Aklujkar, M., & Lovley, D. R. (2010). Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of Pelobacter carbinolicus. *BMC Evolutionary Biology*, *10*(1), 230. https://doi.org/10.1186/1471-2148-10-230

Ali, Y., Koberg, S., HeÃŸner, S., Sun, X., Rabe, B., Back, A., Neve, H., & Heller, K. J. (2014). Temperate Streptococcus thermophilus phages expressing superinfection exclusion proteins of the Ltp type. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00098

Alkhnbashi, O. S., Mitrofanov, A., Bonidia, R., Raden, M., Tran, V. D., Eggenhofer, F., Shah, S. A., Öztürk, E., Padilha, V. A., Sanches, D. S., de Carvalho, A. C. P. L. F., & Backofen, R. (2021). CRISPRloci: Comprehensive and accurate annotation of CRISPR-Cas systems. *Nucleic Acids Research*, *49*(W1), W125–W130. https://doi.org/10.1093/nar/gkab456

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Arroyo-Olarte, R. D., Bravo Rodríguez, R., & Morales-Ríos, E. (2021). Genome Editing in Bacteria: CRISPR-Cas and Beyond. *Microorganisms*, *9*(4), 844. https://doi.org/10.3390/microorganisms9040844

Bernal-Bernal, D., Abellón-Ruiz, J., Iniesta, A. A., Pajares-Martínez, E., Bastida-Martínez, E., Fontes, M., Padmanabhan, S., & Elías-Arnanz, M. (2018). Multifactorial control of the expression of a CRISPR-Cas system by an extracytoplasmic function σ/anti-σ pair and a global regulatory complex. *Nucleic Acids Research*, *46*(13), 6726–6745. https://doi.org/10.1093/nar/gky475

Bille, E., Meyer, J., Jamet, A., Euphrasie, D., Barnier, J.-P., Brissac, T., Larsen, A., Pelissier, P., & Nassif, X. (2017). A virulence-associated filamentous bacteriophage of Neisseria meningitidis increases host-cell colonisation. *PLOS Pathogens*, *13*(7), e1006495. https://doi.org/10.1371/journal.ppat.1006495

Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C., & Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, *17*(1), 356. https://doi.org/10.1186/s12864-016-2627-0

Bondy-Denomy, J., & Davidson, A. R. (2014). When a virus is not a parasite: The beneficial effects of prophages on bacterial fitness. *Journal of Microbiology*, *52*(3), 235–242. https://doi.org/10.1007/s12275-014-4083-3

Brüssow, H., Canchaya, C., & Hardt, W.-D. (2004). Phages and the Evolution of Bacterial Pathogens: From Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews*, *68*(3), 560–602. https://doi.org/10.1128/MMBR.68.3.560-602.2004

Casas, V., & Maloy, S. (2011). Role of Bacteriophage-Encoded Exotoxins in the Evolution of Bacterial Pathogens. *Future Microbiology*, *6*(12), 1461–1473. https://doi.org/10.2217/fmb.11.124

Chen, Y., Yang, L., Yang, D., Song, J., Wang, C., Sun, E., Gu, C., Chen, H., Tong, Y., Tao, P., & Wu, B. (2020). Specific Integration of Temperate Phage Decreases the Pathogenicity of Host Bacteria. *Frontiers in Cellular and Infection Microbiology*, *10*, 14. https://doi.org/10.3389/fcimb.2020.00014

Chevallereau, A., Pons, B. J., van Houte, S., & Westra, E. R. (2022). Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology*, *20*(1), 49–62. https://doi.org/10.1038/s41579-021-00602-y

Choudhary, N., Tandi, D., Verma, R. K., Yadav, V. K., Dhingra, N., Ghosh, T., Choudhary, M., Gaur, R. K., Abdellatif, M. H., Gacem, A., Eltayeb, L. B., Alqahtani, M. S., Yadav, K. K., & Jeon, B.-H. (2023). A comprehensive appraisal of mechanism of anti-CRISPR proteins: An advanced genome editor to amend the CRISPR gene editing. *Frontiers in Plant Science*, *14*, 1164461. https://doi.org/10.3389/fpls.2023.1164461

Chylinski, K., Makarova, K. S., Charpentier, E., & Koonin, E. V. (2014). Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Research*, *42*(10), 6091–6105. https://doi.org/10.1093/nar/gku241

Coleman, D., Knights, J., Russell, R., Shanley, D., Birkbeck, T. H., Dougan, G., & Charles, I. (1991). Insertional inactivation of the *Staphylococcus aureus* β-toxin by bacteriophage φ13 occurs by site-and orientation-specific integration of the φ 13 genome. *Molecular Microbiology*, *5*(4), 933–939. https://doi.org/10.1111/j.1365-2958.1991.tb00768.x

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, *339*(6121), 819–823. https://doi.org/10.1126/science.1231143

Craig, L., Forest, K. T., & Maier, B. (2019). Type IV pili: Dynamics, biophysics and functional consequences. *Nature Reviews Microbiology*, *17*(7), 429–440. https://doi.org/10.1038/s41579-019-0195-4

Cruz-López, E. A., Rivera, G., Cruz-Hernández, M. A., Martínez-Vázquez, A. V., Castro-Escarpulli, G., Flores-Magallón, R., Vázquez, K., Cruz-Pulido, W. L., & Bocanegra-García, V. (2021). Identification and Characterization of the CRISPR/Cas System in Staphylococcus aureus Strains From Diverse Sources. *Frontiers in Microbiology*, *12*, 656996. https://doi.org/10.3389/fmicb.2021.656996

Cumby, N., Reimer, K., Mengin-Lecreulx, D., Davidson, A. R., & Maxwell, K. L. (2015). The phage tail tape measure protein, an inner membrane protein and a periplasmic chaperone play connected roles in the genome injection process of *E . Coli* phage HK 97. *Molecular Microbiology*, *96*(3), 437–447. https://doi.org/10.1111/mmi.12918

Dimitriu, T., Szczelkun, M. D., & Westra, E. R. (2020). Evolutionary ecology of prokaryotic innate and adaptive immune systems and their interplay. *Current Biology : CB*, *30*(19), R1189–R1202. https://doi.org/10.1016/j.cub.2020.08.028

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096. https://doi.org/10.1126/science.1258096

Dy, R. L., Pitman, A. R., & Fineran, P. C. (2013). Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. *Mobile Genetic Elements*, *3*(5), e26831. https://doi.org/10.4161/mge.26831

F, H., H, R., Sp, W., M, B., S, R., & E, C. (2018). The Biology of CRISPR-Cas: Backward and Forward. *Cell*, *172*(6). https://doi.org/10.1016/j.cell.2017.11.032

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., & Herskovits, A. A. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. *Nature Reviews Microbiology*, *13*(10), 641–650. https://doi.org/10.1038/nrmicro3527

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports*, *11*(1), 12728. https://doi.org/10.1038/s41598-021-91456-0

Fortier, L.-C. (2018). Bacteriophages Contribute to Shaping Clostridioides (Clostridium) difficile Species. *Frontiers in Microbiology*, *9*, 2033. https://doi.org/10.3389/fmicb.2018.02033

Gao, Z., & Feng, Y. (2023). Bacteriophage strategies for overcoming host antiviral immunity. *Frontiers in Microbiology*, *14*, 1211793. https://doi.org/10.3389/fmicb.2023.1211793

Garriss, G., & Henriques-Normark, B. (2020). Lysogeny in Streptococcus pneumoniae. *Microorganisms*, *8*(10), 1546. https://doi.org/10.3390/microorganisms8101546

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., & Randau, L. (2018). PAM identification by CRISPR-Cas effector complexes: Diversified mechanisms and structures. *RNA Biology*, *16*(4), 504–517. https://doi.org/10.1080/15476286.2018.1504546

Guzmán, N. M., Esquerra-Ruvira, B., & Mojica, F. J. M. (2021). Digging into the lesser-known aspects of CRISPR biology. *International Microbiology*, *24*(4), 473–498. https://doi.org/10.1007/s10123-021-00208-7

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M., & Terns, M. P. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell*, *139*(5), 945–956. https://doi.org/10.1016/j.cell.2009.07.040

Hay, I. D., & Lithgow, T. (2019). Filamentous phages: Masters of a microbial sharing economy. *EMBO Reports*, *20*(6), e47427. https://doi.org/10.15252/embr.201847427

He, F., Bhoobalan-Chitty, Y., Van, L. B., Kjeldsen, A. L., Dedola, M., Makarova, K. S., Koonin, E. V., Brodersen, D. E., & Peng, X. (2018). Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype I-D immunity. *Nature Microbiology*, *3*(4), 461–469. https://doi.org/10.1038/s41564-018-0120-z

Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D., & Marraffini, L. A. (2015). Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature*, *519*(7542), 199–202. https://doi.org/10.1038/nature14245

Hille, F., & Charpentier, E. (2016). CRISPR-Cas: Biology, mechanisms and relevance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1707), 20150496. https://doi.org/10.1098/rstb.2015.0496

Hochschild, A., & Lewis, M. (2009). The bacteriophage λ CI protein finds an asymmetric solution. *Current Opinion in Structural Biology*, *19*(1), 79–86. https://doi.org/10.1016/j.sbi.2008.12.008

Hooton, S. P. T., & Connerton, I. F. (2015). Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00744

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

Jiang, F., & Doudna, J. A. (2017). CRISPR–Cas9 Structures and Mechanisms. *Annual Review of Biophysics*, *46*(1), 505–529. https://doi.org/10.1146/annurev-biophys-062215-010822

Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F., & Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages 1 1Edited by M. Gottesman. *Journal of Molecular Biology*, *299*(1), 27–51. https://doi.org/10.1006/jmbi.2000.3729

Kieper, S. N., Almendros, C., & Brouns, S. J. J. (2019). Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiology Letters*, *366*(11), fnz129. https://doi.org/10.1093/femsle/fnz129

Kojima, K. K., & Kanehisa, M. (2008). Systematic Survey for Novel Types of Prokaryotic Retroelements Based on Gene Neighborhood and Protein Architecture. *Molecular Biology and Evolution*, *25*(7), 1395–1404. https://doi.org/10.1093/molbev/msn081

Kolesnik, M. V., Fedorova, I., Karneyeva, K. A., Artamonova, D. N., & Severinov, K. V. (2021). Type III CRISPR-Cas Systems: Deciphering the Most Complex Prokaryotic Immune System. *Biochemistry (Moscow)*, *86*(10), 1301–1314. https://doi.org/10.1134/S0006297921100114

Koonin, E. V., & Makarova, K. S. (2019). Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1772), 20180087. https://doi.org/10.1098/rstb.2018.0087

Koudelka, G. B., Arnold, J. W., & Chakraborty, D. (2018). Evolution of STEC virulence: Insights from the antipredator activities of Shiga toxin producing E. coli. *International Journal of Medical Microbiology*, *308*(7), 956–961. https://doi.org/10.1016/j.ijmm.2018.07.001

Krylov, S. V., Kropinski, A. M., Shaburova, O. V., Miroshnikov, K. A., Chesnokova, E. N., & Krylov, V. N. (2013). New Temperate *Pseudomonas aeruginosa* Phage, phi297: Specific Features of Genome Structure. *Генетика*, *49*(8), 930–942. https://doi.org/10.7868/S0016675813080079

Lehane, A. M., Korres, H., & Verma, N. K. (2005). Bacteriophage-encoded glucosyltransferase GtrII of *Shigella flexneri*: Membrane topology and identification of critical residues. *Biochemical Journal*, *389*(1), 137–143. https://doi.org/10.1042/BJ20050102

Levy, A., Goren, M. G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., & Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*, *520*(7548), 505–510. https://doi.org/10.1038/nature14302

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, M., Gong, L., Zhao, D., Zhou, J., & Xiang, H. (2017). The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. *Nucleic Acids Research*, *45*(8), 4642–4654. https://doi.org/10.1093/nar/gkx229

Louwen, R., Staals, R. H. J., Endtz, H. P., van Baarlen, P., & van der Oost, J. (2014). The role of CRISPR-Cas systems in virulence of pathogenic bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *78*(1), 74–88. https://doi.org/10.1128/MMBR.00039-13

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., Van Der Oost, J., & Koonin, E. V. (2011). Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, *9*(6), 467–477. https://doi.org/10.1038/nrmicro2577

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., … 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

Marraffini, L. A., & Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews. Genetics*, *11*(3), 181–190. https://doi.org/10.1038/nrg2749

Matos, R. C., Lapaque, N., Rigottier-Gois, L., Debarbieux, L., Meylheuc, T., Gonzalez-Zorn, B., Repoila, F., Lopes, M. D. F., & Serror, P. (2013). Enterococcus faecalis Prophage Dynamics and Contributions to Pathogenic Traits. *PLoS Genetics*, *9*(6), e1003539. https://doi.org/10.1371/journal.pgen.1003539

McGinn, J., & Marraffini, L. A. (2019). Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nature Reviews Microbiology*, *17*(1), 7–12. https://doi.org/10.1038/s41579-018-0071-7

McGrath, S., Fitzgerald, G. F., & Sinderen, D. V. (2002). Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Molecular Microbiology*, *43*(2), 509–520. https://doi.org/10.1046/j.1365-2958.2002.02763.x

Mitchell, J., Siboo, I. R., Takamatsu, D., Chambers, H. F., & Sullam, P. M. (2007). Mechanism of cell surface expression of the *Streptococcus mitis* platelet binding proteins PblA and PblB. *Molecular Microbiology*, *64*(3), 844–857. https://doi.org/10.1111/j.1365-2958.2007.05703.x

Mitsunaka, S., Sudo, N., & Sekine, Y. (2018). Lysogenisation of Shiga toxin-encoding bacteriophage represses cell motility. *The Journal of General and Applied Microbiology*, *64*(1), 34–41. https://doi.org/10.2323/jgam.2017.05.001

Mulcahy, M. E., & McLoughlin, R. M. (2016). Host–Bacterial Crosstalk Determines Staphylococcus aureus Nasal Colonization. *Trends in Microbiology*, *24*(11), 872–886. https://doi.org/10.1016/j.tim.2016.06.012

Naser, I. B., Hoque, M. M., Nahid, M. A., Tareq, T. M., Rocky, M. K., & Faruque, S. M. (2017). Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of Vibrio cholerae in Bangladesh. *Scientific Reports*, *7*(1), 14880. https://doi.org/10.1038/s41598-017-14839-2

Nobrega, F. L., Walinga, H., Dutilh, B. E., & Brouns, S. J. J. (2020). Prophages are associated with extensive CRISPR–Cas auto-immunity. *Nucleic Acids Research*, *48*(21), 12074–12084. https://doi.org/10.1093/nar/gkaa1071

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., & Doudna, J. A. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Structural & Molecular Biology*, *21*(6), 528–534. https://doi.org/10.1038/nsmb.2820

Pawluk, A., Bondy-Denomy, J., Cheung, V. H. W., Maxwell, K. L., & Davidson, A. R. (2014). A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. *mBio*, *5*(2), e00896-14. https://doi.org/10.1128/mBio.00896-14

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., & Herskovits, A. A. (2012). Prophage Excision Activates Listeria Competence Genes that Promote Phagosomal Escape and Virulence. *Cell*, *150*(4), 792–802. https://doi.org/10.1016/j.cell.2012.06.036

Ramamurthy, T., Pragasam, A. K., Taylor-Brown, A., Will, R. C., Vasudevan, K., Das, B., Srivastava, S. K., Chowdhury, G., Mukhopadhyay, A. K., Dutta, S., Veeraraghavan, B., Thomson, N. R., Sharma, N. C., Nair, G. B., Takeda, Y., Ghosh, A., Dougan, G., & Mutreja, A. (2022). Vibrio cholerae O139 genomes provide a clue to why it may have failed to usher in the eighth cholera pandemic. *Nature Communications*, *13*(1), 3864. https://doi.org/10.1038/s41467-022-31391-4

Richter, H., Lanthier, M., Nevin, K. P., & Lovley, D. R. (2007). Lack of Electricity Production by *Pelobacter carbinolicus* Indicates that the Capacity for Fe(III) Oxide Reduction Does Not Necessarily Confer Electron Transfer Ability to Fuel Cell Anodes. *Applied and Environmental Microbiology*, *73*(16), 5347–5353. https://doi.org/10.1128/AEM.00804-07

Ridley, C., & Thornton, D. J. (2018). Mucins: The frontline defence of the lung. *Biochemical Society Transactions*, *46*(5), 1099–1106. https://doi.org/10.1042/BST20170402

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *The CRISPR Journal*, *3*(6), 462–469. https://doi.org/10.1089/crispr.2020.0059

Sakaguchi, Y., Hayashi, T., Kurokawa, K., Nakayama, K., Oshima, K., Fujinaga, Y., Ohnishi, M., Ohtsubo, E., Hattori, M., & Oguma, K. (2005). The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proceedings of the National Academy of Sciences*, *102*(48), 17472–17477. https://doi.org/10.1073/pnas.0505503102

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., & Marraffini, L. A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell*, *161*(5), 1164–1174. https://doi.org/10.1016/j.cell.2015.04.027

Schroven, K., Aertsen, A., & Lavigne, R. (2021). Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation. *FEMS Microbiology Reviews*, *45*(1), fuaa041. https://doi.org/10.1093/femsre/fuaa041

Sharma, V., Hünnefeld, M., Luthe, T., & Frunzke, J. (2023). Systematic analysis of prophage elements in actinobacterial genomes reveals a remarkable phylogenetic diversity. *Scientific Reports*, *13*(1), 4410. https://doi.org/10.1038/s41598-023-30829-z

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Shiimori, M., Garrett, S. C., Graveley, B. R., & Terns, M. P. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Molecular Cell*, *70*(5), 814-824.e6. https://doi.org/10.1016/j.molcel.2018.05.002

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2017). The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio*, *8*(5), e01397-17. https://doi.org/10.1128/mBio.01397-17

Shmakov, S. A., Utkina, I., Wolf, Y. I., Makarova, K. S., Severinov, K. V., & Koonin, E. V. (2020). CRISPR Arrays Away from cas Genes. *The CRISPR Journal*, *3*(6), 535–549. https://doi.org/10.1089/crispr.2020.0062

Song, G., Tian, C., Li, J., Zhang, F., Peng, Y., Gao, X., & Tian, Y. (2023). Rapid characterization of anti-CRISPR proteins and optogenetically engineered variants using a versatile plasmid interference system. *Nucleic Acids Research*, *51*(22), 12381–12396. https://doi.org/10.1093/nar/gkad995

Stachler, A.-E., Turgeman-Grott, I., Shtifman-Segal, E., Allers, T., Marchfelder, A., & Gophna, U. (2017). High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. *Nucleic Acids Research*, *45*(9), 5208–5216. https://doi.org/10.1093/nar/gkx150

Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. (2010). Self-targeting by CRISPR: Gene regulation or autoimmunity? *Trends in Genetics: TIG*, *26*(8), 335–340. https://doi.org/10.1016/j.tig.2010.05.008

Strutt, S. C., Torrez, R. M., Kaya, E., Negrete, O. A., & Doudna, J. A. (2018). RNA-dependent RNA targeting by CRISPR-Cas9. *eLife*, *7*, e32724. https://doi.org/10.7554/eLife.32724

Tangarife Cardona, M. A., & Arboleda Rivera, J. C. (2023). *Self-targeting spacers reveal new functions of CRISPR systems*. https://doi.org/10.1101/2023.07.26.550751

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569

Tisza, M. J., Belford, A. K., Domínguez-Huerta, G., Bolduc, B., & Buck, C. B. (2021). Cenote-Taker 2 democratizes virus discovery and sequence annotation. *Virus Evolution*, *7*(1), veaa100. https://doi.org/10.1093/ve/veaa100

Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L., & Fowler, V. G. (2015). Staphylococcus aureus infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clinical Microbiology Reviews*, *28*(3), 603–661. https://doi.org/10.1128/CMR.00134-14

Toro, N., Martínez-Abarca, F., Mestre, M. R., & González-Delgado, A. (2019). Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. *RNA Biology*, *16*(10), 1486–1493. https://doi.org/10.1080/15476286.2019.1639310

Tsui, T. K. M., & Li, H. (2015). Structure Principles of CRISPR-Cas Surveillance and Effector Complexes. *Annual Review of Biophysics*, *44*(1), 229–255. https://doi.org/10.1146/annurev-biophys-060414-033939

Wang, X., & Wood, T. K. (2016). Cryptic prophages as targets for drug development. *Drug Resistance Updates*, *27*, 30–38. https://doi.org/10.1016/j.drup.2016.06.001

Wei, Y., & Zhou, C. (2024). Bacteriophages: A Double-Edged Sword In the Gastrointestinal Tract. *Frontiers in Microbiomes*, *3*. https://doi.org/10.3389/frmbi.2024.1450523

Wielgoss, S., & Julien, B. (2023). Whole-Genome Sequence of Myxococcus Phage Mx9. *Microbiology Resource Announcements*, *12*(6), e0022123. https://doi.org/10.1128/mra.00221-23

Wimmer, F., & Beisel, C. L. (2019). CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. *Frontiers in Microbiology*, *10*, 3078. https://doi.org/10.3389/fmicb.2019.03078

Wu, Q., Cui, L., Liu, Y., Li, R., Dai, M., Xia, Z., & Wu, M. (2022). CRISPR-Cas systems target endogenous genes to impact bacterial physiology and alter mammalian immune responses. *Molecular Biomedicine*, *3*(1), 22. https://doi.org/10.1186/s43556-022-00084-1

Yosef, I., Goren, M. G., & Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*, *40*(12), 5569–5576. https://doi.org/10.1093/nar/gks216

Zhang, F., Zhao, S., Ren, C., Zhu, Y., Zhou, H., Lai, Y., Zhou, F., Jia, Y., Zheng, K., & Huang, Z. (2018). CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions. *Communications Biology*, *1*(1), 180. https://doi.org/10.1038/s42003-018-0184-6

Zhou, Q., Su, X., Wang, A., Xu, J., & Ning, K. (2013). QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE*, *8*(4), e60234. https://doi.org/10.1371/journal.pone.0060234