Assessing the Interspecies and Intraspecies Diversity of Cas1 and Cas3 Proteins of the CRISPR-Cas System in *Vibrio cholerae* and *Escherichia coli* Strains

By

Tahani Tabassum

ID: 18136067

Sagarika Shahriar

ID: 18136087

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

Department of Mathematics and Natural Sciences

Brac University

December 2021

Declaration

It is hereby declared that,

1. The thesis submitted is our own original work in order to complete a degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

TAHANI TABASSUM

Candidate 1

SAGARIKA SHAHRIAR

Candidate 2

Approval

The thesis/project titled "Assessing the Interspecies and Intraspecies Diversity of Cas1 and Cas3 Proteins of the CRISPR-Cas system in *Vibrio cholera* and *Escherichia coli* Strains" submitted by Tahani Tabassum (ID: 18136067) and Sagarika Shahriar (ID: 18136087) of Spring 2018 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Biotechnology on 30th December, 2021.

Examining Committee:

Supervisor

(Member)

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

Program Coordinator

Member

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

Department Head

(Chair)

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

Ethics Statement

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

Abstract:

Cholera is an extremely virulent waterborne disease caused by the ingestion of food or water contaminated with pathogenic strains of *Vibrio cholera*e. On the other hand, pathogenic strains of *E. coli*, particularly Shiga-toxin producing *E. coli*, are most commonly responsible for diarrheagenic illness, urinary tract infections, as well as life-threatening complications such as Hemolytic Uremic Syndrome (HUS). The presence of the CRISPR-Cas system in both of the bacterial species have raised major global concerns regarding the enhanced chances of pathogenicity or virulence in the bacterial strains due to this adaptive immune system. Besides, the presence of the CRISPR-Cas system within these species can interfere with the emerging phage therapy treatment approaches against drug-resistant bacteria. In this study, we aimed to assess the diversity of cas1 and cas3 protein sequences in the CRISPR locus of several CRISPR confirmed *V. cholerae* and *E. coli* strains, and characterize this diversity across the functional domains of the reference cas proteins. Moreover, we established the interspecies relatedness of both species in terms of their cas1 and cas3 sequences.

Keywords: Cas1, Cas3, Phylogenetic tree, SNP, Protein Functional Domains.

Dedication

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

Acknowledgement

Praise be to Almighty Allah whose blessing and compassion have guided us in our studies and works throughout our academic career including this internship.

We offer our utmost gratitude to Professor A F M Yusuf Haider, Ph.D Chairperson, Department of Mathematics and Natural Sciences, Brac University for allowing us and constantly motivating us to complete our Undergraduate thesis properly..

We convey our sincere gratitude to Dr. Iftekhar Bin Naser, Assistant Professor and Coordinator of Biotechnology program, Department of Mathematics and Natural Sciences, Brac University for his continuous guidance, patience, motivation, enthusiasm, scholarly inputs, and immense knowledge throughout the process.

We would like to express our gratitude and love to our parents who always give us courage and support. We would also love to thank all our dearest friends and well-wishers for their constant support and motivation.

Sincerely,

Tahani Tabassum

Sagarika Shahriar

06 January, 2022.

List of Acronyms

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

Cas: CRISPR-associated

NCBI: National Center for Biotechnology Information

BLAST: Basic Local Alignment Search Tool

SNP: Single-nucleotide polymorphism

V. cholerae: Vibrio cholerae

E. coli: Escherichia coli

ETEC: Enterotoxigenic *E. coli*

EPEC: Enteropathogenic *E. coli*

EAEC: Enteroaggregative *E. coli*

DAEC: Diffusely-adherent *E. coli*

EIEC: Enteroinvasive *E. coli*

STEC: Enterohemorrhagic *E. coli*

bp: Base Pair

ORF: Open Reading Frame

HUS: Hemolytic Uremic Syndrome

List of Figures:

List of Tables:

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Introduction

Clustered Regularly Interspaced Short Palindromic Repeats, popularly referred to as CRISPR, is an array of short direct repeat DNA sequences as a part of the natural adaptive immune defense system of bacteria and archaea against bacteriophages, plasmids, and infections with other mobile genetic elements. In association with different CRISPR-associated proteins, also known as cas proteins, the system recognizes and destroys foreign invader genomes that significantly enhance the infectivity and virulence of certain bacterial strains. A number of cas proteins with distinct function and significance in the types of CRISPR system have been identified, however, most of these cas sequences are highly diverged owing to the fast evolution of this adaptive immune system. Amongst the distinct types of cas proteins, cas1 and cas3 evolve comparatively slower, making both of these proteins better for phylogenetic relatedness assessment of bacterial strains.

Vibrio cholerae and *Escherichia coli* are two of the most abundantly distributed bacteria in nature. Both of these bacteria have been associated with major outbreaks as well as complicated infections of the gastrointestinal tract and the urinary tract of the human body, raising global health concerns. In this study, we aimed to access the divergence of cas proteins within several *Vibrio cholerae* and *Escherichia coli* strains. We expected the cas proteins to be quite conserved across their functional domains. For both *Vibrio cholera* and *Escherichia coli*, there were fewer divergences across the functional domain of cas1 but significant SNPs were distributed against the core functional domain of cas3. While accessing the interspecies relatedness, one V. cholerae strain showed to have a common ancestry for cas1 with the rest of the E. coli strains and two V. cholerae strains showed to have a common ancestry with the rest of the E. coli strains.

1.2 Objective

The main objective of our research was to assess the diversity of cas1 and cas3 proteins of the CRISPR system across *V. cholerae* and *E. coli* strains, as well as establishing phylogenetic trees based on their divergence through different bioinformatics software.

1.3 Brief Methodology

- Retrieving the *V. cholerae* and *E. coli* cas1 and cas3 protein sequence from the NCBI Protein database.
- Conducting tblastn search with the protein query sequences against the nucleotide sequences of subject strains.
- Extracting the cas1 and cas3 sequences from those subject strains.
- Isolating strains that contained both cas1 and cas3 regions within their sequences.
- Based on the percentage of positives, 40 strains were chosen to conduct multiple-sequence alignment using the Clustal-Omega software tool and establish phylogenetic trees for both *V. cholerae* and *E. coli*.
- Highly conserved strains from among *V. cholerae* and *E.coli* were chosen to construct an interspecies phylogenetic tree using Clustal-Omega software.
- Detection of SNPs from the cas1 and cas3 sequences and categorizing the probability of such SNPs across protein functional domains. Protein domain function prediction was done using InterPro software.

Chapter 2

Literature Review

2.1 Introduction to *Vibrio cholerae* and *Escherichia coli*

2.1.a *Vibrio cholerae*

Vibrio cholerae, the aetiological agent of a profound secretory diarrhoea, is a natural member of the aquatic environment. It is a comma-shaped, gram-negative aerobic or facultatively anaerobic bacillus. This bacterium is ubiquitously distributed in the aquatic environment, but only a small portion of the environmental strains are capable of causing cholera (Faruque et al., 1998). In 1884, the bacterium was first described as the cause of cholera by Robert Koch and in 1959, Sambhu Nath De first isolated the core virulence factor of the species, the cholera toxin, and demonstrated that the toxin was responsible for cholera.

2.1.b *Escherichia coli*

Escherichia coli, the most common aetiological agent of urinary tract infections and urinary tract sepsis, is a common intestinal microbiota of most warm-blooded organisms. It is a non-spore forming, gram-negative, facultatively anaerobic, coliform bacteria. This bacterium is ubiquitous in the human gastrointestinal tract, but it exists in a very small proportion and does not negatively impact host health except for immunocompromised conditions (Köhler & Dobrindt, 2011). In 1885, the bacterium was first introduced by Theodor Eschrich as a common intestinal inhabitant of the neonate and infants, termed as "bacterium coli commune".

2.1.1 Global health concerns induced by *V. cholerae* and *E. coli*

The most notorious enteric pathogen, *Vibrio cholerae*, has been responsible for causing many cholera outbreaks in history. Although previously the outbreaks were common throughout the world, infection is now mainly centered around under-developed or developing countries with poor sanitation and hygiene measures. According to WHO reports, there are an estimated 1.3-4 million cases of cholera with around 21,000-143,000 deaths globally every year (Ali et al., 2015). During the 19th century, cholera spread across the world from its natural reservoir in the Ganges delta, subsequently causing six deadly pandemics across all continents. The disease is now endemic in many countries. At present around 200 serogroups of *Vibrio cholerae* have been characterized, however, principally only *V. cholerae* O1 and *V. cholerae* O139 strains have been

associated with epidemic cholera ("Cholera", 2020). Since transmission occurs through fecal-oral routes, the humanitarian crisis such as inadequate sanitation and overcrowded camps increases the frequency of outbreaks and epidemics.

Escherichia coli is one of the most common causes of diarrheagenic illness globally, as well as the most common causative agent of uncomplicated and complicated urinary tract infections. Although most of the strains are harmless, six pathogenic strains namely enterotoxigenic E. coli (ETEC), enteropathogenic E. coli (EPEC), enteroaggregative E. coli (EAEC), diffusely-adherent E. coli (DAEC), enteroinvasive E. coli (EIEC), and enterohemorrhagic E. coli (STEC), are often associated with food-borne infections. The annual incidence of 31 pathogenic strains in the USA in the year 2011 was estimated to be around 6.6-12.7 million with around 700-2300 mortality cases (Heiman et al., 2015). The pathogenic strains are responsible for food-borne intoxication by creating toxins, the deadliest of which is the Shiga-like toxin produced by STEC (Poole, 2015). Besides foodborne infections, *E. coli* is the most widespread agent causing meningitis in the neonatal period that displays a very high mortality rate worldwide. The mortality rates in neonatal meningitis varies between 15-50% and the survivors are reported to retain neurological damage for the rest of their lives (Donnenberg, 2017).

2.2 CRISPR-Cas system

2.2.1 CRISPR anatomy and mode of action

There are mainly three types of genome-editing tools that are being used in recent time, which are – zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and the RNA-guided CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR associated) nucleases systems (Miller et al., 2007). However, the most popular tool used is the CRISPR nucleases, due to their simple anatomy, higher efficiency, low cost, and having good repeatability with short cycles (Komor et al., 2017). CRISPR-Cas is an adaptive immune system found in prokaryotes such as bacteria and archaea, which serves to eliminate any phage or invader foreign DNA attacking them. The CRISPR system is composed of repeat-spacer arrays, which can be transcribed into CRISPR RNA (crRNA), trans-activating CRISPR

RNA (tracrRNA), and a set of CRISPR-associated (cas) genes encoding proteins with multifaceted activities (Xu & Li, 2020). The repeated sequences detect and help destroy invaders and the spacers between the repeated arrays are the genetic codes retained from the past invaders through the assistance of certain cas proteins. If the same pathogen attacks again, the crRNA recognizes and pairs with the foreign DNA that guides cas protein with endonuclease activity to cleave target sequences of foreign invader DNA, thereby protecting the host (Makarova et al., 2011).

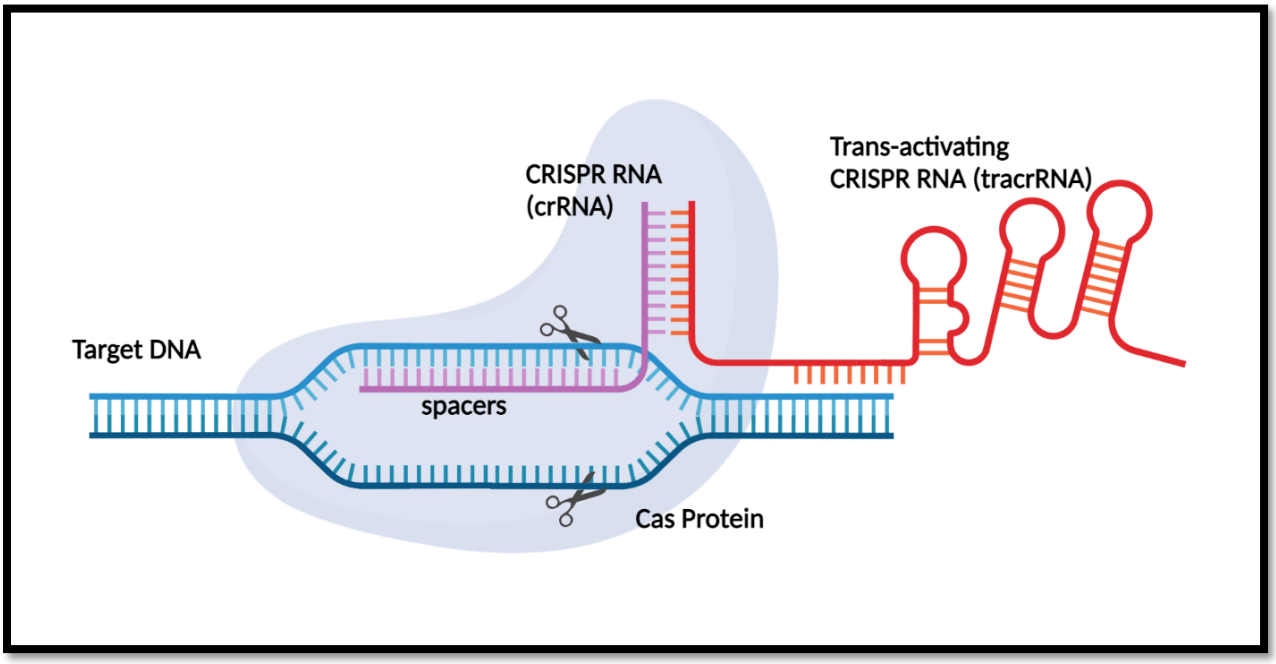

Figure 1: CRISPR anatomy and mechanism.

2.2.2 Classification of CRISPR-Cas system

CRISPR-Cas systems can be classified majorly into two classes – Class 1, and Class 2, which can be further divided into six subtypes from type I to type VI. Class 1 system includes subtype I, III, and IV, on the other hand, Class 2 system includes subtype II, V, and VI (Jiang & Doudna, 2017). Various Cas proteins play significant roles in the CRISPR system, and are also classified into several types depending on their set of functions. These are the cas1, cas2, cas3, cas3″, cas4, cas5, cas6, cas7, cas8, cas9, cas10, and some other small subunits. Cas1 protein is abundantly seen and can form a complex with cas2, this cas1-cas2 complex is required for the adaptation

and spacer acquisition process. The cas1-cas2 complex is present in the great majority of the known CRISPR-Cas systems and represents the highly conserved "information processing" module of the CRISPR-Cas system and is usually less divergent compared to the other cas proteins. Among the different cas proteins, the endonuclease activity of cas1 is required for spacer integration of the system (Nuñez et al., 2014). Cas3 proteins, on the other hand, are single-stranded DNA nuclease (HD domain) and ATP dependent helicase, which is required for the interference of the CRISPR system (Brouns et al., 2008).

2.2.3 Conserveness of cas proteins across species

There are four proposed names for the most conserved and abundant cas genes on the CRISPR repeats, which are – cas1, cas2, cas3, and cas4. From these four, cas1 is the most conserved protein that is present in most of the CRISPR-Cas systems and evolves slower than other cas proteins. Cas3 protein is the next abundant conserved protein, which also has a slower evolution than is typical protein of defense systems. Cas3 helicases are the central component of most CRISPR-mediated adaptive immune systems and also closely related to the DEAH/RHA and NS3/NPH-II protein families (Jackson et al., 2014).

2.3 Evidence of CRISPR system in *V. cholerae* and *E. coli*

CRISPR-mediated adaptive immune system against bacteriophages was first documented in *Streptococcus thermophilus* in 2007 (Barrangou et al., 2007). Since then, several studies have revealed the detailed molecular understanding of these sophisticated adaptive immune defenses in bacterial species. Being responsible for major outbreaks and alarming global concerns, *Vibrio cholerae* and *Escherichia coli* have been repeatedly assessed for the presence of the CRISPR system to unleash the role such bacterial immune systems may play in causing deadly infections. Analysis of the CRISPR-Cas system within the species *Vibrio cholerae* has revealed the presence of majorly type I-E and type I-F subtypes, with highly diverse CRISPR type distributions across the species (Naser et al., 2017). Analysis across Vibrio species have revealed the presence of type I-C, I-E, I-F, II-B, III-A, III-B, III-D, and the rare type IV systems

(McDonald et al., 2019). Similarly, the species *E. coli* usually contains the type I-E subtype, with the reported presence of both I-E and I-F subtypes across the Enterobacteriaceae family (Brouns et al., 2008). The greater evidence of CRISPR-Cas system presence across the Vibrio and Escherichia species suggests there might be a probable role of this defense mechanism in the pathogenicity of certain species. More virulent species might have a positive influence of the CRISPR-Cas system presence or have better chances of establishing infection in hosts.

Chapter 3

Materials and Methods

3.1 Bacterial Genome source

A total of 208 *Vibrio cholerae* strains and 225 *Escherichia coli* strains having confirmed CRISPR were utilized for the purpose of cas sequence analysis within the genus. The fasta format of both the whole genome sequences and the contigs/nodes/scaffolds containing CRISPR sequences were retrieved from the NCBI database.

3.2 Extraction of cas1 and cas3 reference protein sequences

Reference cas1 and cas3 proteins were extracted from the NCBI Protein database. From 262 bacterial cas1 protein entries on *V. cholerae*, one reference cas1 protein sequence with 329 amino acids was chosen. From 303 bacterial cas3 protein entries on *V. cholerae*, one reference cas3 protein sequence with 837 amino acids was chosen. Similarly, from 10,215 bacterial cas1 protein entries on *E. coli*, one sequence with 307 amino acids was chosen. From 11,349 bacterial cas3 protein entries on *E. coli*, one sequence with 888 amino acids was chosen. For ensuring the reference protein is representative for the entire genus, blastn search with the reference proteins were performed.

3.3 Analyzing cas1 and cas3 protein sequence presence through tBLASTn

The cas1 and cas3 protein query sequences were analyzed through tBLASTn against the extracted genome sequences for both *V. cholerae* and *E. coli* strains. While performing the search, the default parameters of tBLASTn were retained.

3.4 Extracting cas1 and cas3 encoding nucleotide sequences from bacterial genome

After performing tBLASTn search, the nucleotide regions within the whole genome sequences that aligned to reference cas protein query sequences were extracted in fasta format from the tBLASTn site. For further analysis, the strains that showed match for both cas1 and cas3 proteins were isolated among *V. cholerae* and *E. coli* strains.

3.5 Analysis of cas1 and cas3 protein sequence divergence

Among the *V. cholerae* strains with both cas1 and cas3 sequences, 20 strains for cas1 protein and 20 strains for cas3 protein were chosen. Similarly, among the *E. coli* strains with both cas1 and cas3 sequences, 20 strains for cas1 and 15 strains for cas3 protein were chosen. While selecting these strains priority was given to the query coverage, percentage of identity, and percentage of positives. In every case, sequences with 100% query coverage were given priority. For percentage of identity and percentage of positives highly diverse sequences were chosen for further analysis. The cas protein sequences extracted from these chosen strains were compiled within one file and multiple-sequence alignment was performed through Clustal-Omega online tool. The alignment files were viewed using the M view tool and SNPs were detected from these alignments. The phylogenetic trees of the chosen strains were also constructed using the Clustal-Omega online tool.

3.6 Protein functional domain prediction

The domains of the reference cas1 and cas3 proteins for both *V. cholerae* and *E. coli* were functionally characterized using the InterPro online tool. There were multiple predictions for some sequence regions. In such cases, the predictions spanning the majority of the protein sequence were selected as functional domains.

3.7 Analysis of SNP distribution across protein functional domains

The positions of the SNPs were merged with the predicted cas protein functional domains in order to assess the prevalence of SNPs across different domains of cas1 and cas3 proteins.

Chapter 4

Results

4.1 Distribution of cas1 and cas3 proteins

4.1.a cas1 and cas3 proteins across *Vibrio cholerae* strains

Out of 208 *V. cholerae* strains, 187 showed match for cas1 protein, 94 showed match for cas3 protein, and 93 showed match for both cas1 and cas3. For the chosen reference proteins, the strains that showed a match for cas1 were most likely to show a match for cas3. Only one strain was found (Vibrio cholerae_strain_TSY216) that showed a match for cas3 only. The position of the cas1 and cas3 protein sequences were in proximity within the bacterial genomes, with cas3 sequences a few thousand bp downstream from the cas1 sequences.

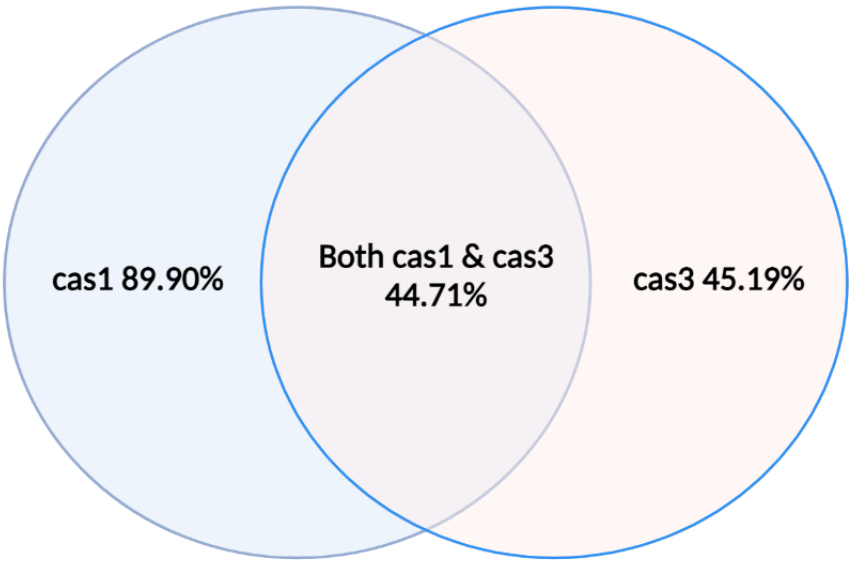

Figure 2: Venn diagram of the statistical percentages of cas1, cas3 and both cas1-cas3 matches in 208 *V. cholerae* strains.

4.1.b cas1 and cas3 proteins across *Escherichia coli* strains

Out of 225 *E. coli* sequences, 115 showed match for cas1 protein, 133 showed match for cas3 protein, and 103 showed match for both cas1 and cas3.

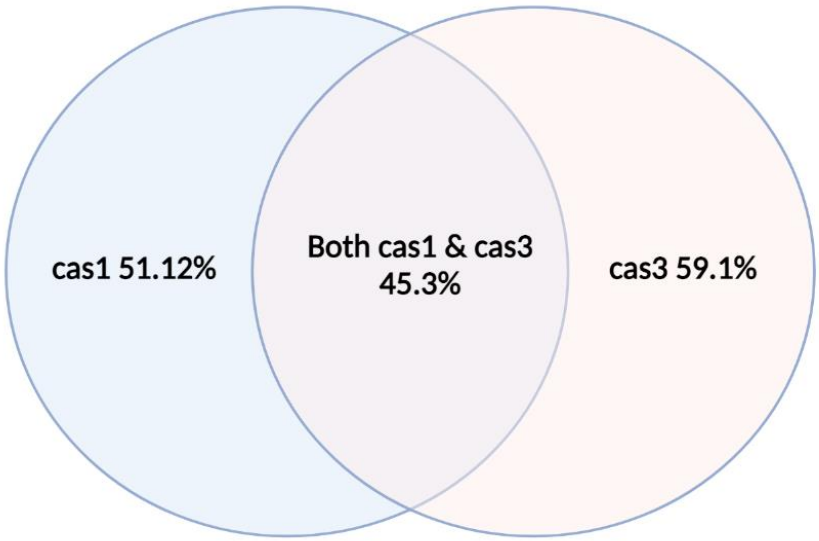

Figure 3: Venn diagram of the statistical percentages of cas1, cas3 and both cas1-cas3 matches in 208 *Escherichia coli* strains.

4.2 Diversity of cas sequences

4.2.a Diversity of cas sequences among *V. cholerae* strains and Phylogenetic relationship

The chosen 20 strains with 100% query coverage but very different percentages of identities and positives were aligned using Clustal-Omega. For cas1, one of the strains (Vibrio cholerae_strain_M1457) was most diverse compared to the rest 19 strains. In respect to that strain, 1 strain showed 95.2% similarity, 6 strains showed 95.3% similarity, 1 strain showed 95.4% similarity, 3 strains showed 95.5% similarity, 2 strains showed 95.6% similarity, 1 strain showed 95.7% similarity, and the rest 5 strains showed 96% similarity. For cas3, two strains (Vibrio cholerae_strain_A12JL36W75 and Vibrio cholerae_strain_920008-15) showed more divergence compared to the rest 18 strains. With respect to these strains, 5 strains showed 99.0% similarity, 5 strains showed 99.1% similarity, 4 strains showed 99.2% similarity, 1 strain showed 99.7% similarity, and 3 strains showed 99.8% similarity.

For cas1,

Percentage Identity	**Number of Strains**
100	1
95.2	1
95.3	6
95.4	1
95.5	3
95.6	2
95.7	1
96	5

Table 1: Cas1 similarity percentages among *V. cholerae* strains.

For cas3,

Percentage Identity	**Number of Strains**
100	2
99	5
99.1	5
99.2	4
99.7	1
99.8	3

Table 2: Cas3 similarity percentages among *V. cholerae* strains.

Using the neighbor-joining method, a phylogenetic tree was constructed for both cas1 and cas3 nucleotide sequences. The cas1 protein sequences are initially divided into three branches. One strain, Vibrio cholerae_strain_N2750, serves as the outgroup as it displayed maximum dissimilarity with the rest of the strains. The cas3 protein sequences are also initially divided into three branches, where both Vibrio cholerae_strain_234 and Vibrio cholerae_strain_5369-93 serve as outgroups due to their maximum dissimilarities with the rest of the strains.

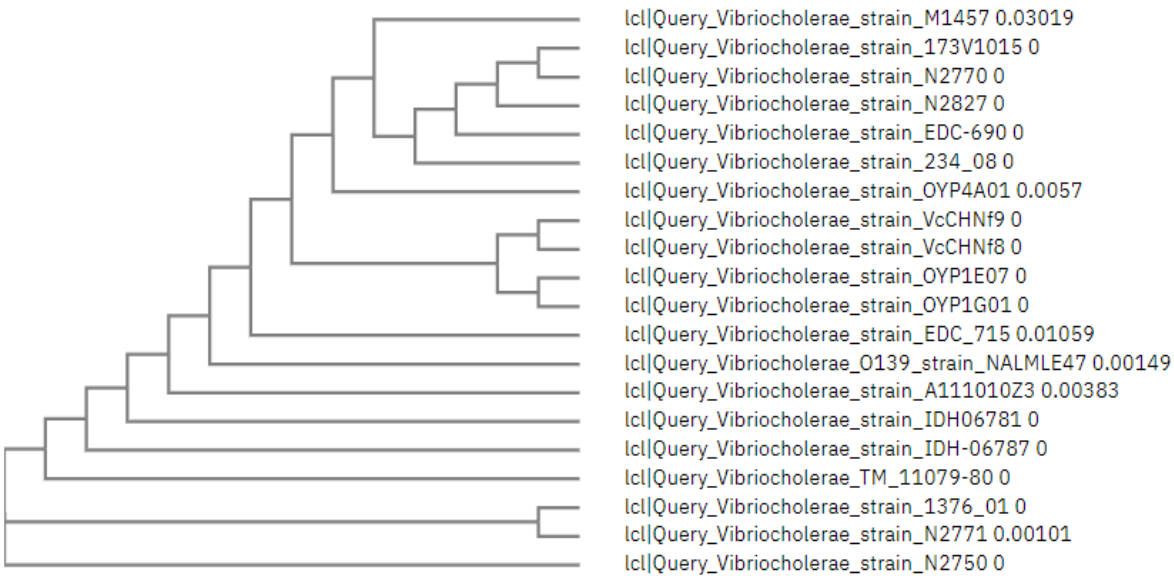

Figure 4: Phylogenetic relatedness between cas1 sequences extracted from 20 *Vibrio cholerae* strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. The strains initially branch off into three internal nodes from one common ancestor. One strain, Vibrio cholerae_strain_N2750, serves as the outgroup as it displayed maximum dissimilarity with the rest of the strains.

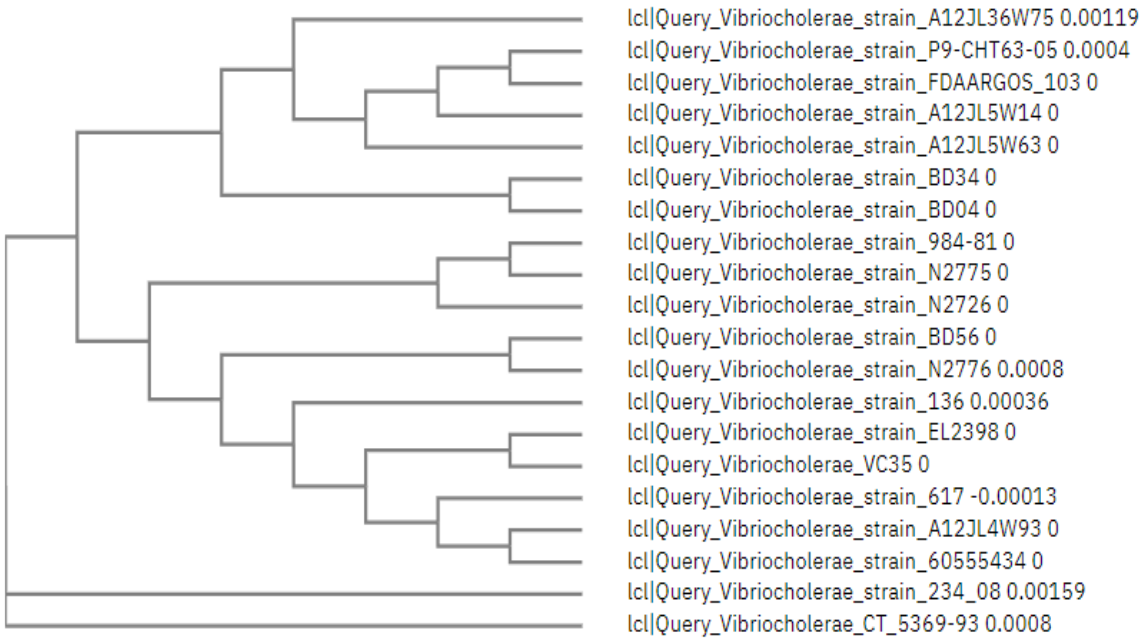

Figure 5: Phylogenetic relatedness between cas3 sequences extracted from 20 *Vibrio cholerae* strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. The strains initially branch off into three internal nodes from one common ancestor. Two of the strains, Vibrio cholerae_strain_234 and Vibrio cholerae_strain_5369-93, serve as outgroups due to their maximum dissimilarities with the rest of the strains.

4.2.b Diversity of cas sequences among *E. coli* strains and Phylogenetic relationship

The chosen strains had 100% query coverage but very different percentages of identities and positives in the tblastn result. For cas1, 20 strains of 100% query coverage were taken, where six different percentage identities were observed. 4 strains showed 93.81% similarity, 8 strains showed 93.49% similarity, 1 strain showed 92.83% similarity, 2 strain showed 92.51% similarity, 4 strain showed 92.18% similarity and only 1 strain showed 91.53% similarity. The most similarity 93.49% was observed by 8 strains among the 20.

Percentage Identity	Number of Strains
93.81%	4
93.49%	8
92.83%	1
92.51%	2
92.18%	4
91.53%	1

Table 3: Cas1 similarity percentages among *E. coli* strains.

For cas3, 15 strains were taken, where six different percentage identities were observed. 5 strains showed 99.89% similarity, 2 strains showed 99.77% similarity, 2 strains showed 98.87% similarity, 1 strain showed 98.37% similarity, 3 strains showed 96.85% similarity, and the rest of the 2 strains showed 96.75% similarity. The most similarity 99.89% was observed by 5 strains among the 15.

Percentage Identity	Number of Strains
99.89%	5
99.77%	2
98.87%	2
98.37%	1
96.85%	3
96.75%	2

Table 4: Cas3 similarity percentages among *E. coli* strains.

Using the neighbor-joining method, a phylogenetic tree was constructed for both cas1 and cas3 nucleotide sequences. The cas1 protein sequences are initially divided into three branches. Two strains, Escherichia.coli_Zam_UTH_26 and Escherichia.coli_HE-MDREc48, served as the outgroup as they displayed maximum dissimilarity with the rest of the strains. The cas3 protein sequences also initially divided into three branches, where another two strains Escherichia.coli_WCHEC050606 and Escherichia.coli_KCJK7164 served as outgroup due to their maximum dissimilarities with the rest of the strains.

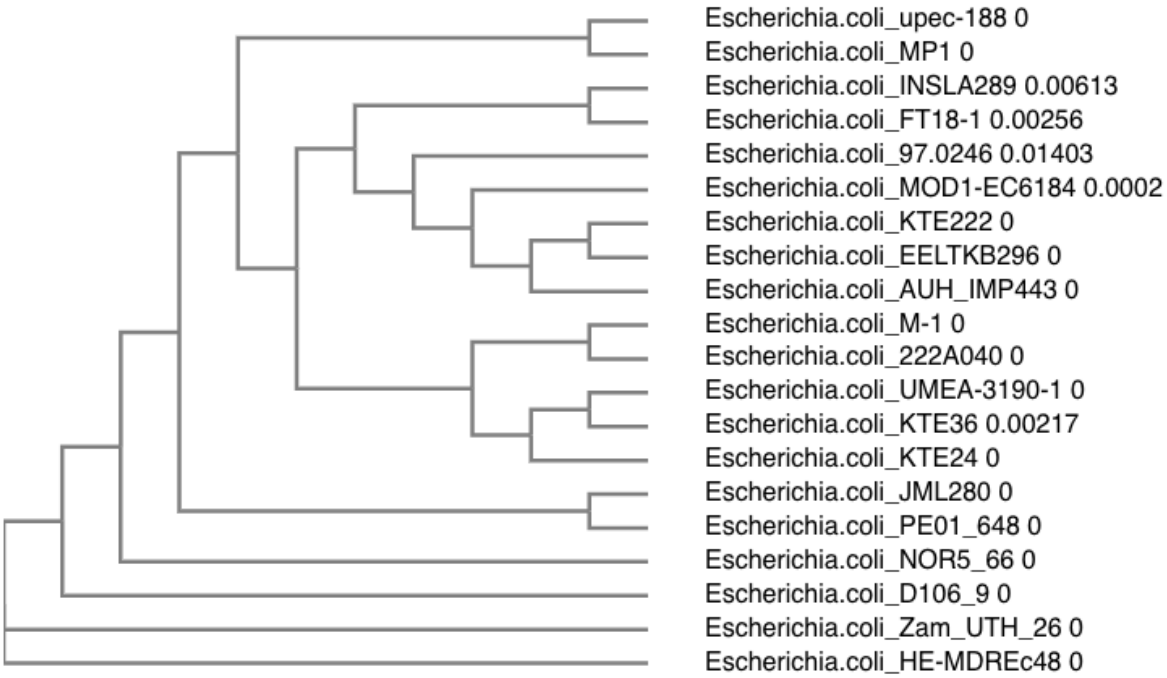

Figure 6: Phylogenetic relatedness between cas1 sequences extracted from 20 Escherichia coli strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. The strains initially branch off into three internal nodes from one common ancestor. Two strains, Escherichia.coli_Zam_UTH_26 and Escherichia.coli_HE-MDREc48, served as the outgroup as they displayed maximum dissimilarity with the rest of the strains.

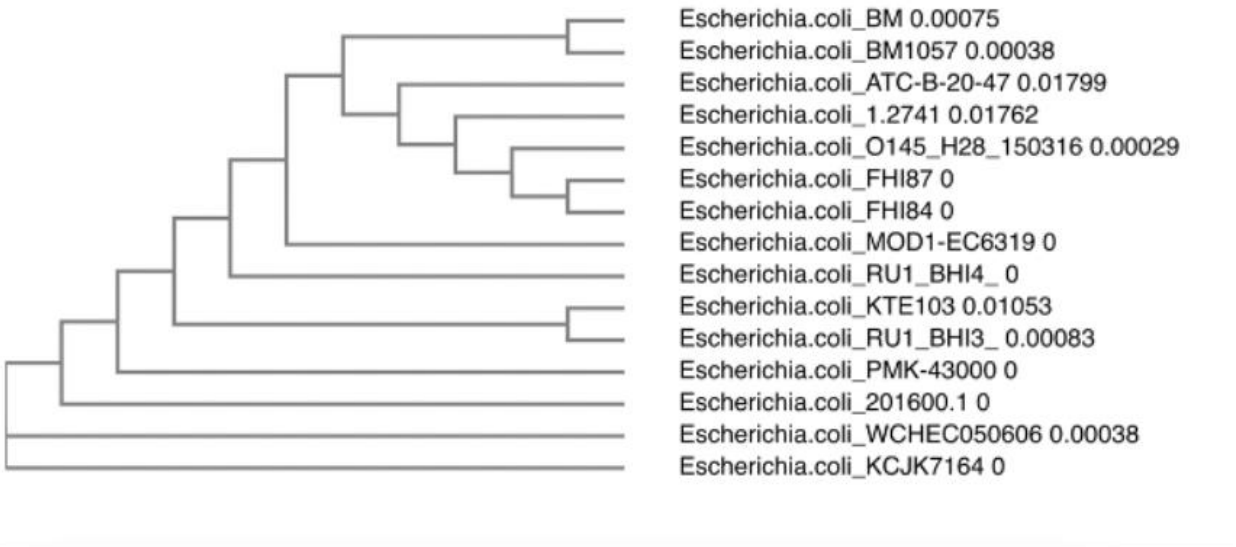

Figure 7: Phylogenetic relatedness between cas3 sequences extracted from 15 *Escherichia coli* strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. The strains initially branch off into three internal nodes from one common ancestor. Two strains, Escherichia.coli_WCHEC050606 and Escherichia.coli_KCJK71, served as the outgroup as they displayed maximum dissimilarity with the rest of the strains.

4.3 Interspecies conservedness among *V. cholerae* and *E. coli* cas protein sequences

Using the neighbor-joining method, an interspecies phylogenetic tree was constructed for both cas1 and cas3 nucleotide sequences. 10 *V. cholerae* sequences with 100% query coverage and 10 *E. coli* sequences with 100% query coverage were chosen for the construction of the tree. For cas1, the most divergent sequence of Vibrio.cholerae_M1457 showed to have a common ancestor with the *E. coli* cas1 sequences. For cas3, two strains, Vibrio.cholerae_BD34 and Vibrio.cholerae_BD04 showed to have a close relationship to the rest of the *E. coli* cas3 sequences.

The Mview results of cas1 alignment between *Vibrio cholerae* strains and *Escherichia coli* strains shows several similarities among Vibrio.cholerae_M1457 with the rest of the *E. coli* strains. The sequences of cas1 were quite similar among the species groups but highly different between both species groups. For cas3, the downstream sequences appeared to have quite a lot

similarity between both the species with a few exceptions. The cas3 sequences had quite a lot of SNPs across E. coli sequences that did not resemble Vibrio strains. Two of the strains Vibrio.cholerae_BD34 and Vibrio.cholerae_BD04 had some similarities with the rest of the *E. coli* strains, resulting in their close position within the phylogenetic tree. For both the cas proteins, *E. coli* strains showed much SNPs within their sequences, that resulted in comparatively more branches of *E. coli* inside the phylogenetic trees.

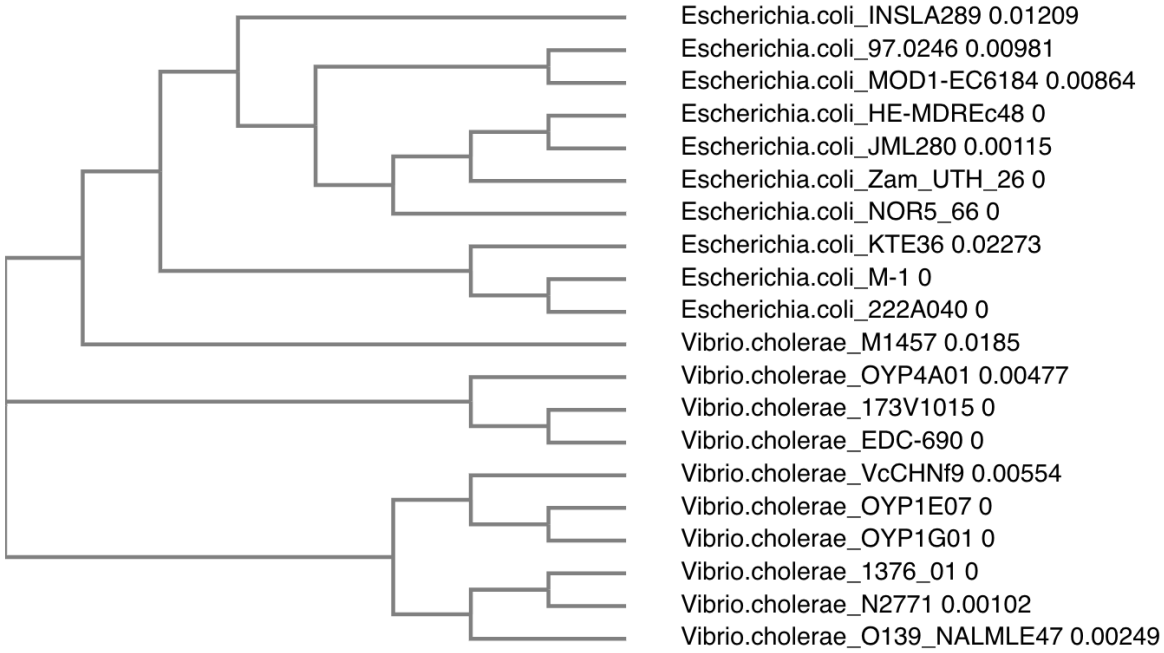

Figure 8: Phylogenetic relatedness between cas1 sequences extracted from 10 *Vibrio cholerae* and 10 *Escherichia coli* strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. One of the *Vibrio cholerae* strains, that had the highest number of unique SNPs across the cas1 sequence, appear to have common ancestry for the cas1 nucleotide sequence with the rest of the *Escherichia coli* strains.

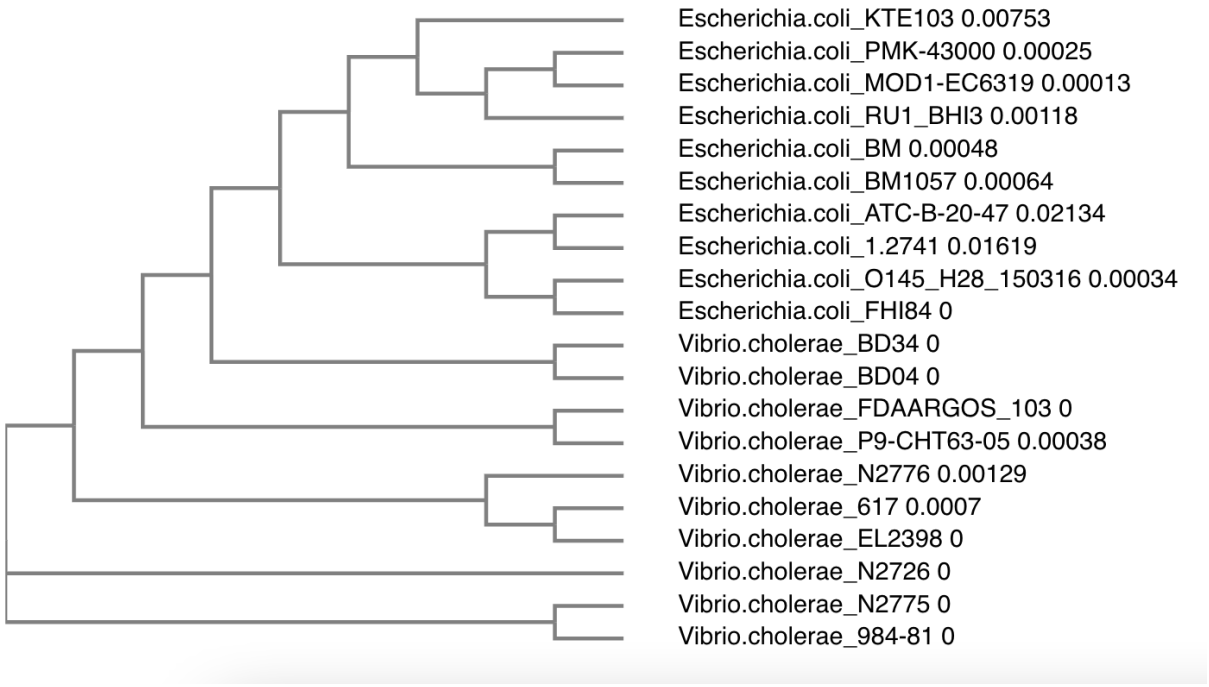

Figure 9: Phylogenetic relatedness between cas3 sequences extracted from 10 *Vibrio cholerae* and 10 *Escherichia coli* strains with confirmed CRISPR and a match of both cas1 and cas3 proteins. Two of the *Vibrio cholerae* strains, that had the highest number of unique SNPs across the cas3 sequence, appear to be closely associated with the ancestor of cas3 nucleotide sequence with the rest of the *Escherichia coli* strains.

4.4 SNP distribution

The distribution of SNPs across the cas sequences of *V. cholerae* and *E. coli* were detected from the Clustal-Omega M view result viewer.

4.4.a SNP detection across *V. cholerae* genomes

For cas1, a total of 61 polymorphisms were detected within the 987 bp sequence. For cas3, a total of 52 polymorphisms were detected within the 2511 bp sequence. Comparing the ratio of SNP occurrence, cas1 sequences appear to be less prone to divergence than cas3 sequences in *Vibrio cholerae* strains.

Position of SNP	Change in nucleotide base	Percentage (within 20 strain)
21	T→C	5%
27	T→G	5%
84	G→T	5%
124	G→A	5%
186	T→C	5%
207	G→C	20%
216	C→T	5%
243	T→C	10%
261	C→T	10%
264	T→C	10%
270	A→G	5%
291	C→G	30%
321	T→C	20%
364	G→T	5%
408	G→A	20%
426	A→C	30%
454	C→G	40%
489	G→A	10%
535	C→A	10%
536	C→A	5%
537	G→A	35%
549	T→C	25%
555	T→G	45%
564	G→A	5%
570	T→C/C→T	50%
585	G→A	5%
610	G→C	10%
621	G→C	5%
651	C→T	35%

675	T→C	5%
705	G→A	5%
729	C→T	40%
747	T→C	5%
768	C→A	30%
771	A→T	5%
774	C→T	5%
807	A→G	5%
819	G→A	30%
831	T→G	30%
849	G→A	5%
876	A→G	45%
883	C→T	25%
891	C→T	5%
915	C→T	30%
917	T→A	30%
918	C→T	30%
930	A→G	5%
933	G→A	30%
937	C→T	25%
942	C→T	20%
950	A→C	45%
954	C→T	45%

957	A→G	20%
960	G→A	5%
961	G→A	20%
966	A→T	5%
967	G→C	5%
968	A→G	5%
969	G→C	5%
970	C→G	5%
979	C→T	20%

Table 5: Distribution of SNPs across cas1 protein of *Vibrio cholerae* strains.

Position of SNP	**Change in nucleotide base**	**Percentage (within 20 strain)**
9	T→C	10%
80	C→T	10%
195	T→A	10%
288	A→G	5%
297	G→C	10%
313	C→T	5%
412	C→T	5%
473	C→T	5%
546	T→C	20%
549	C→T	30%
756	G→A	10%
759	A→T	10%
770	A→T	10%
786	G→A	20%
832	G→A	40%

913	A→C	40%
924	T→C	40%
927	A→C	40%
960	T→G	40%
981	C→A	30%
1056	C→T	10%
1113	G→C	30%
1143	C→T	10%
1155	C→T	10%
1201	G→A	10%
1222	T→C	40%
1250	T→A	5%
1256	A→G	30%
1272	T→C	40%
1356	G→A	10%
1392	C→T	10%
1617	C→T/T→C	50%
1662	C→T	5%
1688	A→G	30%
1713	G→A	10%
1747	G→A	10%
1775	A→C	10%
2005	G→T	5%
2031	T→A	25%
2204	T→C	30%
2266	C→T	30%
2289	C→T	30%
2307	C→A	45%
2326	C→T	15%
2340	T→C	10%

2352	C→T	25%
2394	G→A	35%
2409	C→A	10%
2415	G→A	35%
2418	G→A	20%
2492	G→T	5%
2493	G→A	35%

Table 6: Distribution of SNPs across cas3 protein of *Vibrio cholerae* strains.

4.4.b SNP detection across *E. coli* genomes

For cas1, a total of 89 polymorphisms were detected within the 921 bp sequence. For cas3, a total of 212 polymorphisms were detected within the 2664 bp sequence. Comparing the ratio of SNP occurrence, cas1 sequences appears to be 9.67%, and cas3 sequences appears to be 7.97% in *Escherichia coli* strains, which means cas1 is less prone to divergence than cas3.

Position of SNP	**Change in nucleotide base**	**Percentage (within 20 strain)**
15	A→G	30%
21	T→C	30%
30	G→A	15%
45	C→T	15%
47	T→G/G→T	50%
72	A→G	45%
73	A→G	10%
114	C→A	15%
120	C→A	15%
129	C→T	15%
135	G→A	15%

142	T→G	15%
152	C→T	15%
174	C→A	15%
189	G→A	15%
192	G→A	15%
195	G→C	15%
201	G→A	20%
216	G→A	25%
219	G→A	25%
222	A→G	25%
246	T→C	45%
255	C→T	45%
261	T→C	35%
276	T→A	45%
285	T→G	45%
288	T→A	45%
294	G→A	45%
336	T→C	25%
337	C→T	35%
378	G→A	35%
384	G→C	15%
387	A→C	15%
405	G→T	10%
409	C→T	35%

414	T→G	35%
420	T→C	15%
429	C→T	25%
432	C→T	10%
441	C→T	15%
444	G→C	35%
462	A→G	10%
528	G→C	25%
549	C→T	25%
558	C→T	25%
570	T→A/A→T	50%
582	A→G	10%
600	T→A	10%
607	G→A	5%
615	A→G	25%
621	C→T	10%
624	T→A	5%
680	C→T	25%
708	A→G	30%
711	G→A	30%
714	G→A	30%
717	A→C	30%
720	C→G	30%
727	C→T	5%

729	T→A	40%
741	C→G	25%
744	A→G	40%
747	G→A	35%
750	C→T	10%
756	T→C	35%
761	T→A	5%
765	C→T	10%
769	A→G	5%
773	G→C	35%
777	G→A	25%
795	T→A	35%
804	T→C	40%
813	C→G	40%
819	C→T	30%
822	A→C	25%
846	C→T	35%
864	G→A	25%
867	T→A	25%
870	C→T	25%
876	T→G	25%
886	A→T	25%
887	C→T	25%
891	G→A	25%

894	T→C	30%
900	T→C	40%
907	C→T	5%
912	G→A	30%
914	G→A	10%
916	G→A	25%

Table 7: Distribution of SNPs across cas1 protein of *Escherichia coli* strains.

Position of SNP	**Change in nucleotide base**	**Percentage (within 15 strain)**
24	C→T	13.33%
36	A→G	13.33%
52	T→C	33.33%
62	G→A	6.67%
81	A→G	13.33%
95	T→G	13.33%
105	T→A	13.33%
144	G→A	13.33%
151	A→G	13.33%
185	G→A	13.33%
192	G→A	13.33%
216	T→C	13.33%
228	T→C	13.33%

295	C→T	6.67%
316	A→G	6.67%
350	C→A	6.67%
372	G→A	26.67%
394	T→C	33.33%
423	C→T	20%
453	A→T	6.67%
464	C→T	6.67%
538	C→T	20%
594	G→C	33.33%
612	G→A	6.67%
618	G→C	33.33%
658	C→T	6.67%
696	C→T	26.67%
705	T→C	26.67%
712	A→G	26.67%
723	G→A	26.67%
726	T→C	26.67%
729	T→A	33.33%
732	G→A	26.67%

738	G→A	33.33%
744	C→A	26.67%
745	G→A	6.67%
746	A→G	26.67%
747	C→A	20%
751	A→C	26.67%
753	T→A	26.67%
756	T→G	26.67%
757	C→G	20%
759	G→A	33.33%
764	C→A	6.67%
765	G→A	26.67%
771	C→T	26.67%
777	C→T	26.67%
778	C→A	26.67%
783	G→A	6.67%
792	A→T	26.67%
794	C→T	26.67%
802	T→C	26.67%
806	A→C	20%

808	T→C	20%
813	T→C	6.67%
816	A→C	33.33%
834	A→T	26.67%
835	T→C	20%
843	A→C	26.67%
846	T→C	26.67%
849	T→A	26.67%
855	A→C	6.67%
856	C→T	26.67%
879	C→A	26.67%
891	G→A	6.67%
897	A→G	26.67%
900	T→C	20%
912	A→T	26.67%
913	G→A	26.67%
914	T→C	26.67%
918	T→G	6.67%
933	A→G	26.67%
936	A→T	26.67%

948	A→G	33.33%
951	C→A	26.67%
954	C→A	33.33%
963	G→C	26.67%
969	A→T	6.67%
984	T→C	33.33%
999	T→C	33.33%
1006	A→C	33.33%
1014	T→G	20%
1035	A→T	33.33%
1050	G→A	33.33%
1062	T→C	33.33%
1064	C→G	33.33%
1065	G→C	33.33%
1079	G→A	33.33%
1088	A→G	33.33%
1089	C→T	33.33%
1096	T→A	33.33%
1098	A→C	33.33%
1125	C→T	33.33%

1134	G→T	33.33%
1146	C→T	33.33%
1170	G→A	33.33%
1171	A→T	33.33%
1179	A→G	33.33%
1197	G→A	6.67%
1212	T→C	6.67%
1215	G→A	6.67%
1257	C→A	6.67%
1260	C→A	6.67%
1269	G→T	20%
1377	C→G	13.33%
1401	A→T	13.33%
1428	A→G	13.33%
1455	C→T	13.33%
1459	C→T	13.33%
1480	C→G	20%
1514	T→A	13.33%
1539	C→T	20%
1566	G→A	6.67%

1582	C→T	13.33%
1591	C→T	6.67%
1617	G→A	6.67%
1622	A→G	6.67%
1625	C→T	6.67%
1664	C→G	6.67%
1685	C→T	6.67%
1695	C→T	6.67%
1707	A→G	13.33%
1713	T→C	20%
1716	T→C	20%
1719	T→C	6.67%
1734	C→T	6.67%
1743	A→G	6.67%
1744	G→C	26.67%
1751	T→C	6.67%
1752	A→T	20%
1759	C→T	6.67%
1767	G→A	26.67%
1774	A→G	20%

1806	A→G	40%
1830	A→T	40%
1845	T→A	20%
1848	T→C	20%
1855	G→A	40%
1863	A→C	20%
1873	G→A	20%
1875	G→T	13.33%
1890	A→G	20%
1902	C→A	13.33%
1905	G→C	40%
1907	A→G	40%
1923	G→A	26.67%
1929	C→T	20%
1932	T→A	20%
1953	C→T	20%
1962	T→C	6.67%
1965	C→T	13.33%
1968	A→T	20%
1974	G→A	6.67%

1989	G→A	20%
1992	T→C	46.67%
1998	A→T	6.67%
1999	C→G	6.67%
2000	A→T	6.67%
2001	T→A	6.67%
2002	C→G	6.67%
2003	G→T	6.67%
2004	C→G	6.67%
2005	C→G	6.67%
2006	A→C	6.67%
2007	T→A	6.67%
2010	T→C	6.67%
2011	C→T	6.67%
2012	G→A	6.67%
2013	C→T	20%
2015	A→T	6.67%
2016	A→T	6.67%
2018	A→G	6.67%
2019	T→G	6.67%

2021	G→C	6.67%
2022	T→A	6.67%
2025	C→T	6.67%
2028	T→C	20%
2038	A→G	6.67%
2044	G→C	26.67%
2049	C→A	20%
2053	A→G	26.67%
2073	G→A	26.67%
2145	A→G	13.33%
2172	T→C	6.67%
2178	C→T	20%
2214	G→T	20%
2238	C→T	26.67%
2239	A→G	20%
2240	T→G	20%
2244	A→T	20%
2256	G→A	6.67%
2278	T→C	6.67%
2280	G→A	6.67%

2295	C→T	6.67%
2325	T→C	26.67%
2346	C→A	6.67%
2398	G→T	6.67%
2455	C→T/T→C	50%
2474	C→T	6.67%
2476	C→T	13.33%
2478	T→A	13.33%
2519	G→A	6.67%
2576	A→T	13.33%
2597	T→G	13.33%
2601	C→T	13.33%
2603	G→A	13.33%
2619	C→T	33.33%
2622	T→A	33.33%
2623	C→G	33.33%
2641	A→C	6.67%
2642	T→G	6.67%
2643	A→G	6.67%

Table 8: Distribution of SNPs across cas3 protein of *Escherichia coli* strains.

4.5 Reference protein functional domain predictions

The functional domains of cas1 and cas3 reference proteins were predicted from the InterPro database.

4.5.a *V. cholerae*

For cas1, 11-109 amino acid sequence was predicted to form the N-terminal domain and 110-319 amino acid was predicted to form the C-terminal domain of the protein (Figure 10). For cas3, 1-177 amino acid was predicted to form the cas3 HD domain, 220-445 amino acid was predicted to form the Helicase ATP binding domain, 492-581 amino acid was predicted to form the Helicase C terminal (Figure 11). Within cas3, 227-409 amino acid sequence was predicted to form the DEAD/DEAH box motif.

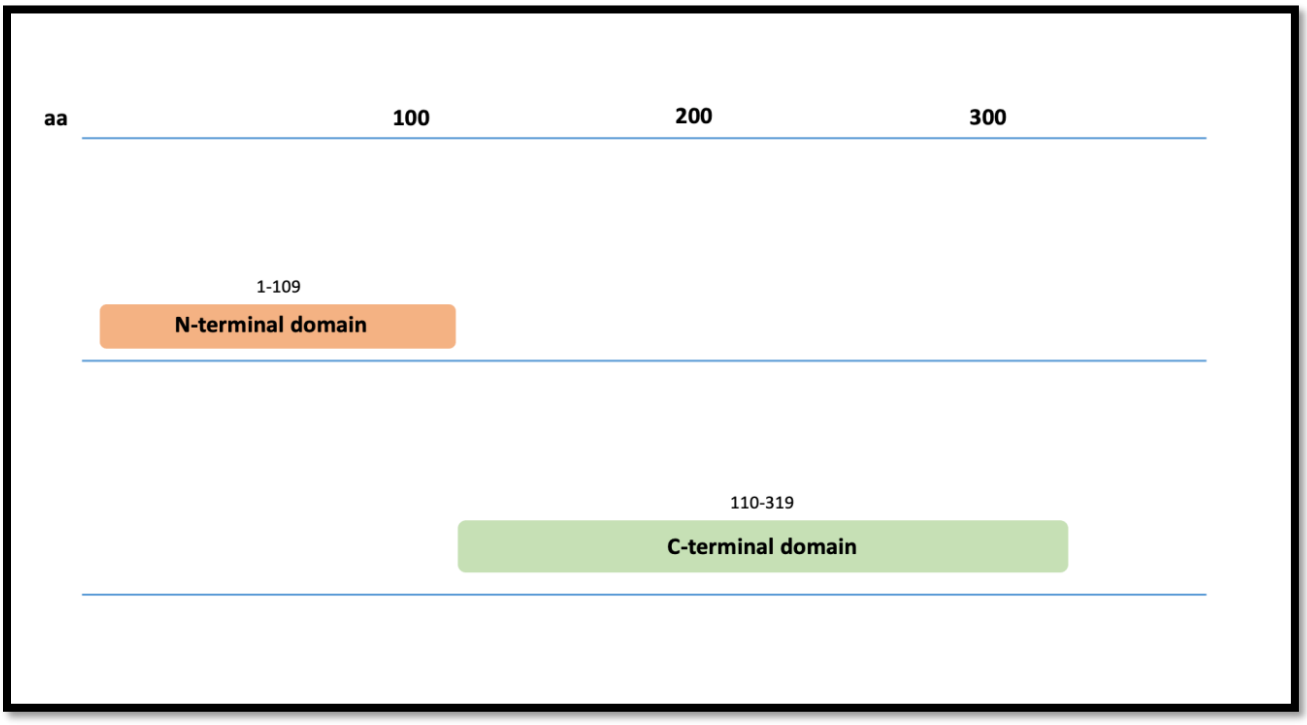

Figure 10: Protein domain prediction of the cas1 protein of *Vibrio cholerae*

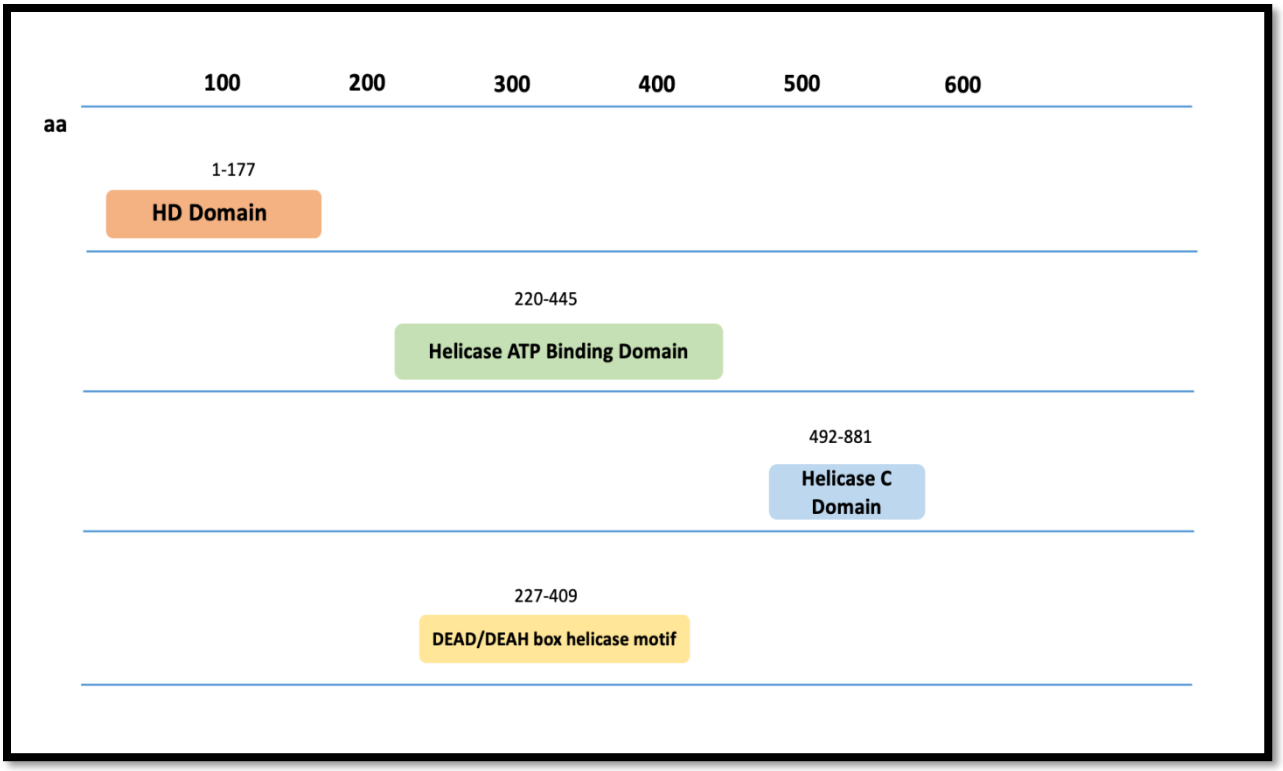

Figure 11: Protein domain prediction of cas3 protein of *Vibrio cholera*e

4.5.b *E. coli*

For cas1, 1-93 amino acid was predicted to form the N-terminal domain and 95-291 amino acid was predicted to form the C-terminal domain of the protein (Figure 12). For cas3, 5-286 amino acid sequences were predicted to form the cas3 HD nuclease domain and 309-715 amino acid sequences were predicted to form the Helicase core domain (Figure 13). Cas3 also was predicted to form the DEAD/DEAH box helicase motif from its 308-493 amino acid sequence.

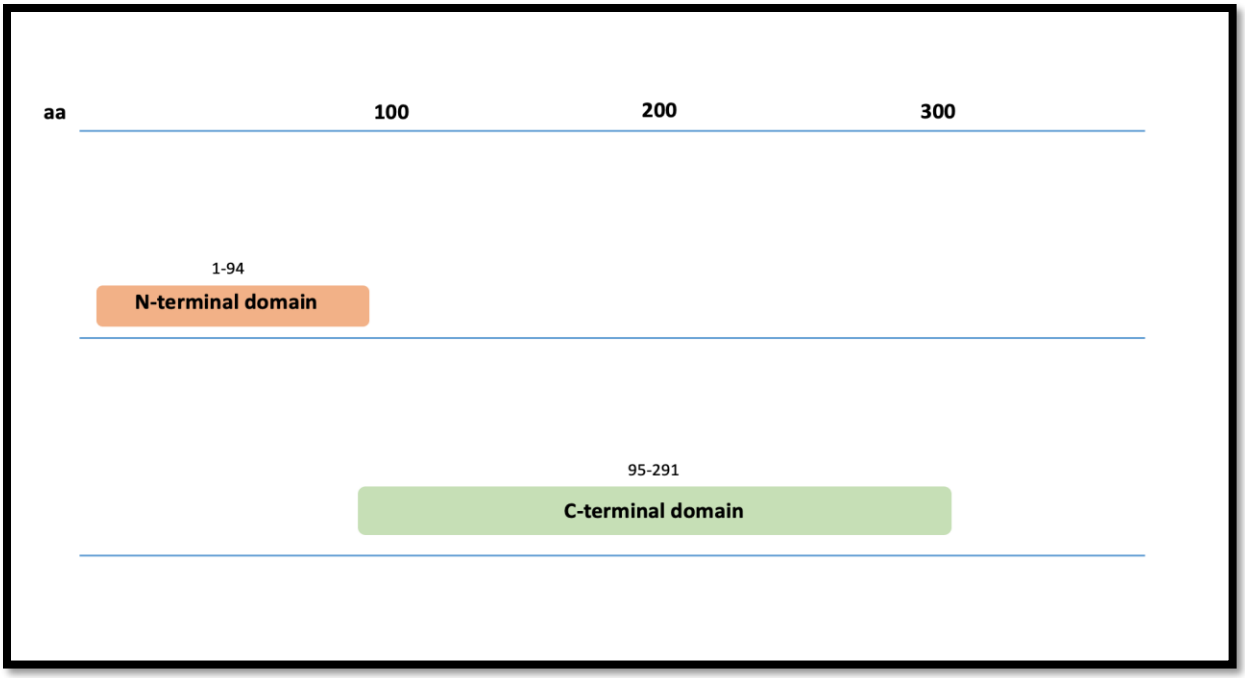

Figure 12: Protein domain prediction of cas1 of *Escherichia coli*

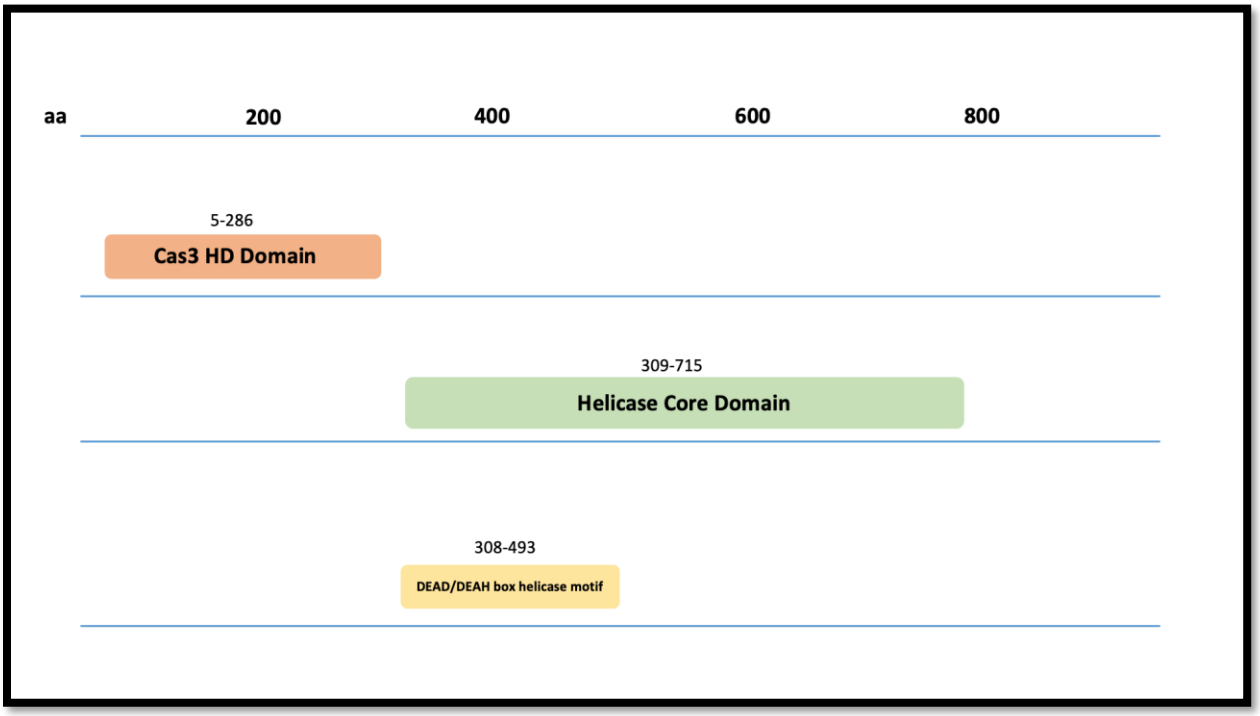

Figure 13: Protein domain prediction of cas3 of *Escherichia coli*

4.6 Distribution of SNPs across protein domains

4.6.a *Vibrio cholerae*

Out of the 61 polymorphisms detected for cas1, 11 SNPs were found across the N-terminal domain and 38 SNPs were found across the C-terminal domain (Figure 14). On the other hand, out of 52 polymorphisms detected for cas3, 8 SNPs were found across the cas3 HD domain, 19 SNPs were found across the Helicase ATP-binding domain, and 4 SNPs were found across the Helicase C terminal domain (Figure 14). For cas3, DEAD/DEAH Box Helicase was the one predicted protein motif. This region was predicted to span from 227-409 amino acids and there were 16 SNPs across this region.

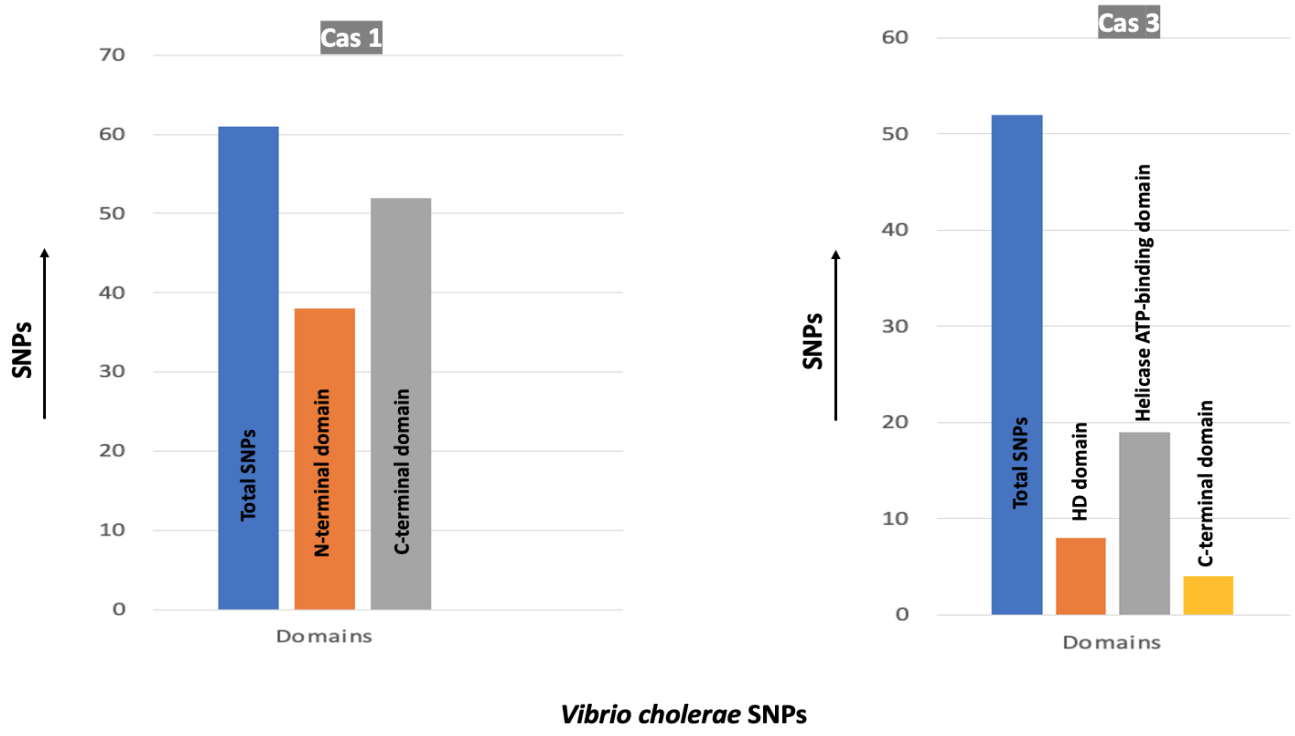

Figure 14: Categorization of SNPs across protein functional domain for *Vibrio cholerae* strains.

4.6.b *Escherichia coli*

Out of the 89 polymorphisms detected for cas1, 25 SNPs were found across the N-terminal domain and 55 SNPs were found across the C-terminal domain. On the other hand, out of the

212 polymorphisms detected for cas3, 64 SNPs were found across the cas3 HD nuclease domain and 113 SNPs were found across the Helicase core domain. Moreover, for cas3, DEAD/DEAH Box Helicase motif was the one predicted protein motif, where 40 SNPs were predicted to occur.

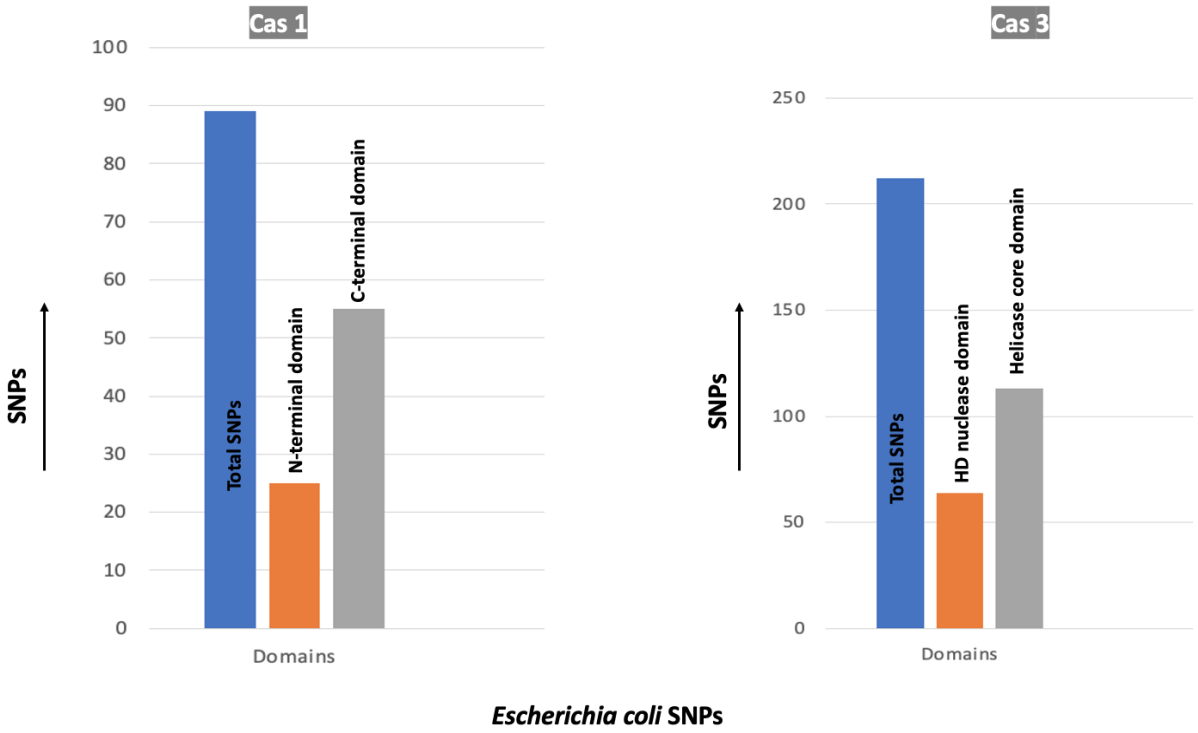

Figure 15: Categorization of SNPs across protein functional domain for *Escherichia coli* strains.

Chapter 5

Discussion

5.1 CRISPR-type classification

All the reference genomes for both *V. cholerae* and *E. coli* were confirmed CRISPR containing strains. Since *V. cholerae* and *E. coli* are classified as type-I CRISPR-Cas systems, it was primarily expected that all of the strains will display a match for both cas1 and cas3 sequences, as these cas genes are signature sequences of CRISPR type-I. However, only 44.71% of the *Vibrio cholerae* strains and 45.77% of the *E. coli* strains showed match for both cas1 and cas3 sequences, which was far below the expected threshold.

According to the recent classification of CRISPR-Cas subtypes, the type-I CRISPR-Cas system can further be divided into I-A to I-F subtypes. One of the characteristic features of the exception I-F variant 1 is the absence of cas1-cas2-cas3 genes. This variant is the only group within type-I CRISPR-Cas system to lack both of the cas genes. This variant consists of three potentially mobile effector complexes: csy1/csy2 fusion, csy3, and cas6f instead of cas1-cas2-cas3 genes (Makarova and Koonin, 2015). The strains of *V. cholerae* and *E. coli* that did not show match for either cas1 or cas3 proteins might belong to this I-F variant 1 subtype.

Out of the six subtypes of type-I CRISPR-Cas system, except for I-A, I-F and I-F variant 2, all the subtypes have cas1 sequence downstream of cas3 sequence. Besides, only type I-A appeared to have cas3 split into two domains, as in the case of our reference protein sequence (Makarova and Koonin, 2015). All the *V. cholerae* strains that showed match for both cas1 and cas3 protein in this analysis had cas3 downstream of cas1 sequence, as well as, cas3 sequence split into two functional domains. So all of the *V. cholerae* strains are most likely to belong to the I-A subtype. On the other hand, all the *E. coli* strains that showed match for both cas1 and cas3 protein in this analysis had cas1 downstream of cas3. This indicates that the *E. coli* strains might belong to any one of the type-I subtypes, but it is more feasible that they belong to the type-I E subtype that is most common across *E. coli* species.

The majority of the *V. cholerae* strains did not show a match for cas3 but showed a match for cas1. In case of *E. coli* strains, comparatively more strains showed a match for cas3 than cas1. So far no CRISPR-Cas type-I classification has displayed the presence of cas1 protein in absence of cas3 protein within the CRISPR locus. The discrepancy in these results may be attributed to the selection of reference proteins for translated BLAST analysis. Although the cas1 and cas3

reference protein sequences displaying maximum homogeneity with most other *V. cholerae* and *E. coli* strains were chosen, query coverage and sequencing error might have contributed to such discrepancy. Since, type-II CRISPR-Cas system shows the presence of cas1 in absence of cas3 protein, where the function of cas3 is most likely substituted by cas9 protein, the strains showing a match for cas1 but lacking a match for cas3 protein can be searched for match with cas9 protein sequence. The presence of cas9 within such strains might suggest a novel classification of CRISPR-Cas subtypes.

One interesting result was the Vibrio cholerae_strain_TSY216 that showed a match for cas3 but no match for cas1 protein. There were similar cases in *Escherichia coli* results, where some strains showed a match for cas3 protein but not for cas1 protein. These strains might belong to a different or even unclassified CRISPR subtype. Recently it has been found that many CRISPR systems might function in a cas1-independent fashion, such as the type-IV system and the I-F variant 1 subtype (Makarova and Koonin, 2015), that might explain the different cas protein patterns in these strains.

5.2 Diversity of cas sequences

Core cas sequences are conserved across most of the CRISPR-Cas subtypes (Jansen et al., 2002). Among the core cas proteins, cas1 and cas3 sequences are comparatively more conserved across species. Other than cas1 and cas3, the sequences of other cas genes are highly diverged, presumably due to the rapid evolution of this adaptive defense system (Makarova and Koonin, 2015). Cas1 evolves slower compared to all the other cas genes (Takeuchi *et al.*, 2012), hence cas1 phylogeny is the ideal guide for CRISPR-Cas divergence analysis. In this analysis, cas1 and cas3 phylogenetic trees were used to access the probable regions of divergence across the Vibrio and Escherichia genus.

5.2.a *V. cholerae*

For *V. cholerae* strains, cas1 sequences showed the presence of 61 polymorphisms within the 987 bp sequence and cas3 sequences showed the presence of 52 polymorphisms within the 2511 bp sequence. The probability of occurrence of SNPs in cas1 appears to be much higher than in

cas3 for *V. cholerae*. It was expected that the cas sequences, especially the cas1 sequence, will be more conserved.

Out of the 20 analyzed sequences, one strain Vibrio cholerae_strain_M1457 showed the highest number of polymorphisms across the cas1 sequence against our reference protein sequence. According to NCBI BioSample database, the strain was isolated from a water source in 2009 at Russia. There were a number of unique SNPs across another strain, Vibrio cholerae_strain_EDC_715, which according to NCBI BioSample database was isolated from environmental samples in 2015 at Bangladesh. Some unique SNPs were observed in the same location for a group of strains, for instance, Vibrio cholerae_strain_OYP1E07 and Vibrio cholerae_strain_OYP1G01. For cas3, two strains Vibrio cholerae_strain_A12J36W75 and Vibrio cholerae_strain_920008-15 showed the highest number of polymorphisms. According to NCBI BioSample database, Vibrio cholerae_strain_A12J36W75 was isolated from a water source in 2012 at Austria and Vibrio cholerae_strain_920008-15 was isolated at Austria in 2015 but the isolation source is undefined. Two groups of strains; Vibrio cholerae_BD34 and Vibrio cholerae_strain_BD04, and Vibrio cholerae_strain A12JL4W93 and Vibrio cholerae_strain_60555434; showed similar SNPs at some locations that were less prevalent among the other strains. All the strains were much conserved across 2040-2200 bp regions. Some less frequent yet unique SNPs were distributed in some other strains.

5.2.b *E. coli*

For E. coli strains, cas1 sequences showed the presence of 89 polymorphisms within the 921 bp sequence and cas3 sequences showed the presence of 212 polymorphisms within the 2664 bp sequence. The probability of occurrence of SNPs in cas1 appears to be a bit higher than in cas3 for E. coli. It was expected that the cas sequences, especially the cas1 sequence, will be more conserved.

Out of the analyzed sequences, one strain Escherichia.coli_97.0246 showed the highest number of unique polymorphisms across the cas1 sequence against our reference protein sequence. According to NCBI BioSample database, this strain was first described in the year 2012 and its host organism is cow. There was a pattern of similar SNPs being present in the same location for a group of strains; for instance, three strains Escherichia.coli_UMEA-3190-1, Escherichia.coli_KTE24, and Escherichia.coli_KTE36 were more likely to show similar SNPs.

Similarly, two other groups Escherichia.coli_M-1 and Escherichia.coli_222A040, as well as, Escherichia.coli_INSLA289 and Escherichia.coli_FT18-1 showed similar regions of SNPs. Interestingly, one strain Escherichia.coli_INSLA289 had a lot of unique SNPs across the 700-730 nucleotide region. This strain, according to the NCBI BioSample database, was isolated from poultry carrying twelve acquired antibiotic resistance genes in 2015 at Portugal. For cas3, a number of unique SNPs were found across two strains, Escherichia.coli_ATC-B-20-47 and Escherichia.coli_1.2741. According to NCBI BioSample database, the strain Escherichia.coli_ATC-B-20-47 was isolated from raw meat-based diets for companion animals in 2020 at Switzerland, whereas, Escherichia.coli_1.2741 was isolated from cow in 2011. Even in this case, the pattern of observing SNPs in clusters across strain Escherichia.coli_ATC-B-20-47 and Escherichia.coli_1.2741 were observed.

5.3 Interspecies conservedness among *V. cholerae* and *E. coli* cas protein sequences

Significant differences exist between the CRISPR array and cas gene sequences in different species. The highly conserved sequence of cas1, and in some cases cas3, serves as the most suitable guide for interspecies divergence analysis of cas genes (Haft *et al.,* 2005; Makarova *et al.,* 2006). In this analysis, it was expected that the cas1 and cas3 sequences will vary for Vibrio and Escherichia genus, resulting in two distinct branches in the phylogenetic tree. However, for both cas1 and cas3, the E. coli strains appeared to evolve from an internal common ancestor of the *V. cholerae* strains.

For cas1, the most divergent *Vibrio cholerae* strain (Vibrio cholerae_strain_M1457) is displayed to evolve from the same ancestor as the rest of the E. coli cas1 sequences. This observation can explain the distribution of so many novel SNPs across the genome of this strain.

For cas3, the *E. coli* strains appeared to be most related to two *Vibrio cholerae* strains (Vibrio cholerae_strain_BD34 and Vibrio cholerae_strain_BD04). These *Vibrio cholerae* strains appeared to evolve from the same ancestor as the rest of the *E. coli* cas3 sequences. Analysis of the multiple-sequence alignment of these two *V. cholerae* strains showed several novel SNPs across their cas3 sequences. This observation can explain the common ancestry of these two strains with the rest of the *E. coli* strains.

5.4 Reference protein functional domain predictions for *V. cholerae* and *E. coli*

Analysis of the cas1 and cas3 protein sequence through the InterPro database predicted the functional domains of the reference proteins.

5.4.a *Vibrio cholerae*

For cas1, 11-109 amino acid sequence was predicted to form the N-terminal domain and 110-319 amino acid sequence was predicted to form the C-terminal domain. The N-terminal domain of cas1 is considered to be sufficient enough for the characteristic feature of this cas protein. This N-terminal domain can function as a metal-dependent DNA-specific endonuclease that can undergo dimerization with cas2 protein to mediate spacer acquisition for the CRISPR systems (Nuñez *et al.,* 2014). The function of the C-terminal domain of the cas1 protein has not been properly described yet.

For cas3, the first 1-177 amino acid sequence was predicted to form the HD domain. This domain tends to be nearest to the CRISPR repeats and can be found separately in some CRISPR subtypes. This domain mediates nuclease activity against ssDNA and ssRNA (Beloglazova *et al*., 2011; Sinkunas *et al.,* 2011). In our reference protein, the HD domain was linked to a helicase-containing domain. The first 225 amino-acid sequence of the helicase, spanning from 220-445 bp, formed the ATP-binding domain. Whereas, the 89 amino acid sequence of the helicase, spanning from 492-581 bp, formed the Helicase-C terminal domain. This helicase domain represents members of the classical helicase superfamily 1 and 2 DNA-binding domain, in particular the DEAD/DEAH box helicases. This whole domain functions as ATP-dependent RNA helicases and is involved in various aspects of RNA metabolism (Tanner and Linder, 2001; Schütz *et al.,* 2010).

5.4.b *E. coli*

For cas1, 1-93 amino acid sequences were predicted to form the N-terminal domain, and 95-291 amino acid sequences were predicted to form the C-terminal domain of the protein. Cas1 proteins are asymmetrical homodimers with each monomer having an N-terminal β-sheet domain and C-terminal α-helical domain (Nuñez et al., 2014). The C-terminal tail of cas1 is not essential for spacer acquisition, although it may supplement the critical interactions at the interface.

For cas3, 5-286 amino acid sequences were predicted to form the cas3 HD nuclease domain and 309-715 amino acid sequences were predicted to form the Helicase core domain. Within the cas3 Helicase core domain, there was a predicted DEAD/DEAH box helicase motif from 308-493 amino acid sequence. Cas3 HD nucleases working together with the cas3 helicases can completely degrade invasive DNAs through the combination of endo- and exonuclease activities (Beloglazova *et al.*, 2011). A recent work has revealed that the *Pseudomonas aeruginosa* cas3 protein functions downstream of CRISPR RNA processing and both the Cas3 HD and helicase domains are required for the CRISPR function in the suppression of biofilm formation by phage-infected cells (Cady and O'Toole, 2011). *E. coli* contains 5 DEAD-box genes and 13 DEAH-box genes, among which one (hrpA) participates in RNA metabolism, and most others in DNA metabolism (Iost and Dreyfus, 2006).

5.5 Distribution of SNPs across protein domains

5.5.a *V. cholerae*

A total of 11 SNPs were detected across the N terminal domain of cas1 protein that represents 18.03% of the total SNP detected within the cas1 sequences. Whereas, a total of 38 SNPs were detected across the C terminal domain, that represents 62.29% of the total SNP detected within the cas1 sequence. Since the N terminal domain is significant for the characteristic function of cas1 protein, it is expected that there will be fewer SNPs across this domain. Our result complied with our expectations as the majority of the SNPs were found across the domain with uncharacterized functions.

A total of 8 SNPs were detected across the HD domain of cas3 protein that represents 15.38% of the total SNP detected within the cas3 sequences. The largest SNP distribution was across the helicase ATP-binding domain. A total of 19 SNPs were detected across this domain, representing 36.53% of the total SNPs detected. Among these 19 SNPs, around 16 (30.76%) were found inside the DEAD/DEAH box helicase motif, which is considered to have a core function in RNA metabolism. A total of 4 SNPs were detected across the Helicase C terminal domain, representing 7.69% of the total SNPs detected. We expected SNPs to be least common within the Helicase C domain considering it constitutes the main function of cas3 endonucleases in the

CRISPR type-I system, and as per our hypothesis SNPs were less common within the Helicase domain. The ATP-binding domain showed the least conservedness, whereas, the N terminal HD domain showed comparatively more conservedness than the ATP-binding domain.

5.5.b *E. coli*

A total of 25 SNPs were detected across the N terminal domain of cas1 protein that represents 28.08% of the total SNP detected within the cas1 sequences. Whereas, a total of 55 SNPs were detected across the C terminal domain, that represents 61.8% of the total SNP detected within the cas1 sequence. Since the N terminal domain is significant for the characteristic function of cas1 protein, it was expected that there will be fewer SNPs across this domain. Our result complied with our expectations as the majority of the SNPs were found across the domain with uncharacterized functions.

A total of 64 SNPs were detected across the cas3 HD domain that represents 30.2% of the total SNP detected within the cas3 sequences. A total of 113 SNPs were detected across the Helicase core domain, representing 53.3% of the total SNPs detected. Among this 113 SNPs from this domain, 40 SNPs were detected within the DEAD/DEAH box helicase motif, representing 18.9% of the total SNPs. Just like the *V. cholerae* result, the Helicase domain of *E. coli* had least SNPs compared to the rest of the domains, which complied with our hypothesis. We expected SNPs to be least common within the Helicase domain considering it constitutes the main function of cas3 endonucleases in the CRISPR type-I system. The cas3 HD domain showed less conservedness compared to the Helicase core domain.

Chapter 6

Conclusion

6.1 Limitations

1. Although the bacteria strains were confirmed CRISPR positive, all of the strains did not show a match for cas1 and cas3 reference protein sequences. One probable explanation for this observation might be the absence of the cas proteins within the CRISPR type found in those strains or a poor choice of cas1 and cas3 protein selection. The latter case appears more feasible in this case.

2. In terms of SNP detection, the selection of a certain percentage of strains could have led to incorrect mutation characterization. In each case, the nucleotide appearing more frequent was considered to be a reference nucleotide, whereas the one appearing less frequent was detected as SNP. The better option here would be to focus on the locations where such a probable shift in nucleotide can occur.

3. Search for diversity using cas protein sequences can often result in false positives or false negatives due to the similarity of such proteins or their domains with many other proteins of the prokaryotic living sytem.

4. Although all the strains were chosen according to 100% query coverage retrieved from tblastn, one strain of *Escherichia coli* showed 76% query coverage in the clustal omega alignment software. Therefore, the results were not fully focused on 100% query coverage.

5. One major drawback of this project can be the conversion of protein sequences through the tblastn tool to perform queries on nucleotide subject sequences. The change in ORFs might have altered the nucleotide sequences coding for different amino acids, thereby generating false SNP results.

6.2 Recommendations

1. If the origin of the bacterial strains can be confirmed, their origin and geographic location can be correlated with the probability of SNPs and relatedness with other species.

2. Using a better reference nucleotide sequence of cas1 and cas3 protein similar search should be performed within subject nucleotide sequences.

3. While constructing the phylogenetic trees a more accurate reference protein must be selected so that the locations and probabilities of SNPs could be accurately defined.

4. Large scale analysis with more bacterial sequences can help access the conservedness of all the other cas proteins and define the functionality of such proteins in the CRISPR-cas systems.

References

Ali, M., Nelson, A., Lopez, A., & Sack, D. (2015). Updated Global Burden of Cholera in Endemic Countries. *PLOS Neglected Tropical Diseases*, *9*(6), e0003832.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., & Moineau, S. et al. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, *315*(5819), 1709-1712.

Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A., & Yakunin, A. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *The EMBO Journal*, *30*(22), 4616-4627.

Brouns, S., Jore, M., Lundgren, M., Westra, E., Slijkhuis, R., & Snijders, A. et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, *321*(5891), 960-964.

Cady, K., & O'Toole, G. (2011). Non-Identity-Mediated CRISPR-Bacteriophage Interaction Mediated via the Csy and Cas3 Proteins. Journal Of Bacteriology, 193(14), 3433-3445. https://doi.org/10.1128/jb.01411-10

Cholera. (2020). Retrieved 4 January 2022, from https://www.who.int/news-room/fact-sheets/detail/cholera

Donnenberg, MS. (2017). Escherichia coli Pathotypes and Principles of Pathogenesis. Baltimore, Maryland, USA: International Encyclopedia of Public Health, 585-593.

Faruque, S., Albert, M., & Mekalanos, J. (1998). Epidemiology, Genetics, and Ecology of Toxigenic Vibrio cholerae. *Microbiology And Molecular Biology Reviews*, *62*(4), 1301-1314.

Haft, D., Selengut, J., Mongodin, E., & Nelson, K. (2005). A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *Plos Computational Biology*, *1*(6), e60.

Heiman, K., Mody, R., Johnson, S., Griffin, P., & Gould, L. (2015). Escherichia coliO157 Outbreaks in the United States, 2003–2012. *Emerging Infectious Diseases*, *21*(8).

Iost, I., & Dreyfus, M. (2006). DEAD-box RNA helicases in Escherichia coli. Nucleic Acids Research, 34(15), 4189-4197.

Jackson, R., Lavin, M., Carter, J., & Wiedenheft, B. (2014). Fitting CRISPR-associated Cas3 into the Helicase Family Tree. Current Opinion In Structural Biology, 24, 106-114.

Jansen, R., Embden, J., Gaastra, W., & Schouls, L. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, *43*(6), 1565-1575.

Jiang, F., & Doudna, J. (2017). CRISPR–Cas9 Structures and Mechanisms. Annual Review Of Biophysics, 46(1), 505-529.

Komor, A., Badran, A., & Liu, D. (2017). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell, 168(1-2), 20-36.

Köhler, C., & Dobrindt, U. (2011). What defines extraintestinal pathogenic Escherichia coli?. *International Journal Of Medical Microbiology*, *301*(8), 642-647.

Makarova, K., & Koonin, E. (2015). Annotation and Classification of CRISPR-Cas Systems. *Methods In Molecular Biology*, 47-75.

Makarova, K., Grishin, N., Shabalina, S., Wolf, Y., & Koonin, E. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*, *1*(1).

Makarova, K., Haft, D., Barrangou, R., Brouns, S., Charpentier, E., & Horvath, P. et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology, 9(6), 467-477.

McDonald, N., Regmi, A., Morreale, D., Borowski, J., & Boyd, E. (2019). CRISPR-Cas systems are present predominantly on mobile genetic elements in Vibrio species. *BMC Genomics*, *20*(1).

Miller, J., Holmes, M., Wang, J., Guschin, D., Lee, Y., & Rupniewski, I. et al. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. Nature Biotechnology, 25(7), 778-785.

Naser, I., Hoque, M., Nahid, M., Tareq, T., Rocky, M., & Faruque, S. (2017). Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of Vibrio cholerae in Bangladesh. *Scientific Reports*, *7*(1).

Nuñez, J., Kranzusch, P., Noeske, J., Wright, A., Davies, C., & Doudna, J. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Structural & Molecular Biology*, *21*(6), 528-534.

Okada, K., Na-Ubol, M., Natakuathung, W., Roobthaisong, A., Maruyama, F., & Nakagawa, I. et al. (2014). Comparative Genomic Characterization of a Thailand–Myanmar Isolate, MS6, of Vibrio cholerae O1 El Tor, Which Is Phylogenetically Related to a "US Gulf Coast" Clone. *Plos ONE*, *9*(6), e98120.

Poole, TL. (2007). In: Simjee S, editor. Foodborne Diseases. 1st ed. Totowa, New Jersey: Humana Press, 535.

Schütz, P., Karlberg, T., van den Berg, S., Collins, R., Lehtiö, L., & Högbom, M. et al. (2010). Comparative Structural Analysis of Human DEAD-Box RNA Helicases. *Plos ONE*, *5*(9), e12791.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *The EMBO Journal*, *30*(7), 1335-1342.

Takeuchi, N., Wolf, Y., Makarova, K., & Koonin, E. (2012). Nature and Intensity of Selection Pressure on CRISPR-Associated Genes. *Journal Of Bacteriology*, *194*(5), 1216-1225. Tanner, N., & Linder, P. (2001). DExD/H Box RNA Helicases. *Molecular Cell*, *8*(2), 251-262.

Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. Computational And Structural Biotechnology Journal, 18, 2401-2415.