Distribution of Complete CRISPR-Cas Systems in *Vibrio cholerae* and its Effect on Presence of Plasmid Derived Contigs in Complete Genome Assemblies

By

Ratul Reza
18236002

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

Department of Mathematics and Natural Sciences
Brac University
May 2023

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name and Signature:

Student's Name: Ratul Reza

Approval

The thesis/project titled “Distribution of Complete CRISPR-Cas Systems in *Vibrio cholerae* and its Effect on Presence of Plasmid Derived Contigs in Complete Genome Assemblies” submitted by **Ratul Reza** with student ID **18236002** of Fall, 2018 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Biotechnology on 23rd May, 2023.

Examining Committee:

Supervisor:
(Member)

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

Co-Supervisor:
(Member)

Tushar Ahmed Shishir
Lecturer
Department of Mathematics and Natural Sciences
Brac University

Program Director:
(Member)

Dr. Munima Haque
Associate Professor
Department of Mathematics and Natural Sciences
Brac University

Departmental Head:
(Chairperson)

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

Abstract

Cholera is a water-borne disease caused by *Vibrio cholerae* that causes severe diarrhea and dehydration. Plasmids disseminate antibiotic resistance and have the potential to play critical roles in epidemic outbreaks. Understanding the distribution and coexistence of *V. cholerae* plasmids and CRISPR-Cas systems is critical for investigating pathophysiology and developing effective control methods. The NCBI database yielded a total of 5873 genomic assemblies. PlasForest, a machine learning-based classifier, was used to predict plasmid sequences, while CRISPRCasTyper was utilized to identify Cas operons and CRISPR arrays. The results demonstrate a statistically significant decrease in %PDC between groups with no CCS and groups with 1 CCS and 2 CCS, although the difference in %PDC across groups with multiple CCS was not significant.

Dedication

This thesis is dedicated to my mother, Irin Sultana, whose unwavering love, encouragement, and sacrifices have been the driving force behind my academic pursuits. Her selflessness, determination, and resilience have been a constant source of inspiration and motivation, and I am eternally grateful for her unyielding support throughout my educational journey.

Acknowledgement

I would like to convey my heartfelt thanks to Dr. Iftekhar Bin Naser, Associate Professor, Department of Mathematics and Natural Sciences, Brac University, for his tremendous assistance and support throughout my thesis study. His knowledge, patience, and unshakable dedication to my academic achievement have all played a role in determining the course of my work and assisting me in reaching my objectives.

I would also want to express my deepest gratitude to Tushar Ahmed Shishir, Lecturer, Department of Mathematics and Natural Sciences, Brac University, for his unselfish devotion and readiness to assist me even outside of usual office hours. His ideas, support, and technical skills helped me overcome many challenges and complete my thesis in a timely and effective manner.

I am deeply indebted to both Dr. Naser and Mr. Shishir for their guidance and support, and I am truly grateful for their contributions to my academic and personal growth. Their mentorship and friendship will be cherished for years to come.

Ratul Reza

Table of Contents

List of Tables

List of Figures

List of Acronyms

%PDC	Percentage of Plasmid Derived Contigs
AMR	Anti-Microbial Resistance
CCS	Complete CRISPR-Cas System
CDS	Coding Sequence
CT	Cholera Toxin
DR	Highly Conserved regions in the CRISPR array
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
MGE	Mobile Genetic Elements
MMC	Matthew's Correlation Coefficient
ORFs	Open Reading Frames
PDC	Plasmid Derived Contigs

CHAPTER 1

INTRODUCTION

1. What is Cholera and what causes it?

Vibrio cholerae is a bacterium that causes cholera, a water-borne disease that can lead to severe diarrhea, dehydration and that may result in renal failure. Leaving severe dehydration untreated can result in shock, a coma, and death within hours. According to the CDC, the following symptoms are prevalent in the early stages of cholera and affect around one in every ten people:

- Thirst
- Excessive watery stools ("rice-water stools")
- Leg cramps
- Restlessness, or irritation.

High quantities of the pathogenic *Vibrio cholerae* bacteria are found in the copious amounts of diarrhea that cholera patients expel, which can infect others if consumed. This could happen if the bacteria come in touch with food or water.

In order to prevent the pathogen from spreading, the feces (human waste) from unwell persons should be appropriately disposed of to avoid infecting everything around. Those caring for cholera patients need to wash their hands thoroughly after handling anything that could be contaminated with patient feces.

1.2. Literature Review

1.2.1. Epidemiological Review of *Vibrio cholerae*

Vibrio cholerae is a bacterium that causes cholera, a water-borne disease that can lead to severe diarrhea, dehydration and death if left untreated. Cholera is an ancient disease, Barua in his extensive article on the history of cholera (Barua, 1992), states report of diseases with cholera like symptoms dating back to the periods of Hippocrates and Buddha. Prior to 1817, when the first pandemic is said to have originated in India, real cholera, caused by *Vibrio cholerae* 01, was already present in Europe. According to CBC news, from 1817, there were seven distinct pandemics of cholera, which is summarized in the table below.

Table 1: *Brief summary of Past Vibrio cholerae epidemics*

Pandemic	*Period*	*Origin*	*Responsible biotypes (Siddique & Cash, 2014)*
1st	1817-1823	Bengal region of India	Not recognized
2nd	1829-1849	India	Not recognized
3rd	1852-1859	India	Not recognized
4th	1863-1879	Bengal region of India	Not recognized
5th	1881-1896	Bengal region of India	O1: Classical
6th	1899-1923	India	O1: Classical
7th	1963-1991	Indonesia	O1: El Tor

Since the El Tor biotype still persists and infects, The World Health Organization (WHO) continues to classify the 7th pandemic as currently ongoing in a cholera factsheet dated 30 March 2022 (“Cholera”, 2022)

1.2.2. Pathogenicity of *Vibrio Cholerae*

According to Reidl & Klose's study from 2002, extensive molecular and genomic research has helped to clarify the main virulence components and the regulatory mechanisms of *V. cholerae* that contribute to virulence. Despite the fact that no animal model can accurately simulate the human infection with *V. cholerae*, several of them have shown to be highly useful in the research of cholera pathogenesis. The rabbit ileal loop model has been effective in measuring the fluid accumulation in the intestines caused by CT in vivo.

Additionally, the article refers to other studies where it has been shown that a number of *V. cholerae* gene products are essential for colonization of the small intestine, notably in cholera models in adult rabbits and young mice. Among these are the lipopolysaccharide (LPS) O-antigen, accessory colonization factors (ACFs), regulatory proteins (such as ToxR/ToxS, TcpP/TcpH, and ToxT), outer membrane porins, genes for biotin and purine biosynthesis, an iron-regulated OMP protein called IrgA, and features of the LPS core region. A type IV pilus known as TCP is regarded as the most important colonization component since it has been experimentally shown to be required for colonization of the intestines in animal models as well as research with human volunteers. While CT is essential for producing cholera symptoms, but it doesn't appear to have any real effect on colonization of the intestines.

The article (Reidl & Klose, 2002) also refers to an experimental study which produced evidence from the deletion of ctxAB mutants, pointing to the possibility that CT activity can promote development in the gut environment by destroying epithelial cells, but whether or not CT activity outside human host provides any added advantage to pathogenic *V. cholerae* lacks experimental confirmation. The genes for CT are actually found in the single-stranded filamentous phage known as CTXφ. The phage has both the "RS2" and "core" portions. The ace, zot, cep, and orfU genes, as well as the CT operon (ctxAB), make up the core area, while

RS2 encodes genes that are involved in CTX integration, replication, and regulation (Waldor et al., 1997).

The most frequent exposed molecule on the outer membrane of Gram-negative bacteria is LPS, which also provides barrier function. The acidity, osmolarity, temperature and exposure to antibacterial agents and elements of the innate immune system are only a few of the external changes that *V. cholerae* cells are subjected to during infection. The cell's outer membrane efficiently blocks the passage of harmful substances, aids in avoiding detection by host agents, and could even make colonization easier.

1.2.3. Significance of Plasmids

Plasmids are circular pieces of extrachromosomal DNA that can replicate on their own and are essential for HGT and the evolution of microorganisms. Virulence and drug resistance in *V. cholerae* have been linked to plasmids. Since the 1960s, it has been common to find plasmids in the genomes of environmental and clinical *V. cholerae* isolates from the serogroups O1, O139 and non-O1, non-O139 (Amaro et al., 1988). A single plasmid can cause resistance to more than six different antibiotics, including tetracycline, streptomycin, ampicillin, chloramphenicol, gentamicin, and SXT, in Cases of *V. cholerae* and other intestinal infections (Carattoli, 2013). Antibiotic resistance-inducing plasmids are frequently large (>40-Kb), conjugative, and have a low copy number in the host bacteria. The IncA mega plasmid (pVC1447) of *V. cholerae* O139, which was discovered in China between 2000 and 2006, had genes for resistance to erythromycin, aminoglycosides, chloramphenicol, tetracycline, and SXT (Luo et al., 2022).

There have also been reports of a variety of Gram-negative pathogens containing the plasmid genes qnrA, qnrB, qnrC, qnrD, qnrS, and qnrVC. These genes encode pentapeptide repeat family proteins that protect topoisomerase IV and DNA gyrase from quinolone inhibition. Both qnrVC1 and qnrVC3 were found in a clinical strain of *V. cholerae* that was isolated from Bangladesh in 2005 and Brazil in 1998, respectively (Fonseca & Vicente, 2013).

In an article by Carraro et al., (2016), it states that SXT/R391 integrative and conjugative elements were principally responsible for the spread of the multidrug resistance genes that *V. cholerae*. There have been infrequent reports of IncA/C conjugative plasmids mediating antibiotic resistance in clinical and environmental isolates of *V. cholerae*. Their findings demonstrated that, while being uncommon in *V. cholerae* populations, IncA/C plasmids play a crucial yet stealthy function by particularly spreading a novel family of genomic islands that confer resistance to several drugs. Their findings also imply a reservoir of transmissible resistance genes in non-epidemic *V. cholerae* non-O1/non-O139 isolates that may be passed on by IncA/C plasmids to virulent *V. cholerae* serotypes in epidemic geographic areas as well as to other pathogenic Enterobacteriaceae species.

1.2.4. Significance of the CRISPR-Cas System

The DNA sequences known as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) can be found in the genomes of prokaryotic organisms like bacteria and archaea (Barrangou, 2015). Invasive genome fragments and mobile genetic elements (MGEs) like plasmids, phages, and transposons are incorporated into the CRISPR locus as spacers, showing that the prokaryotic immune system is adaptable and can recall previous infections. The spacers between the repeats are created as small guide CRISPR RNAs (crRNAs) when an infection recurs, and Cas proteins use these crRNAs to target invaders in a sequence-specific way (Hille & Charpentier, 2016). The workings of several CRISPR-Cas sub-systems are depicted in the schematic figure below by Hille & Charpentier.

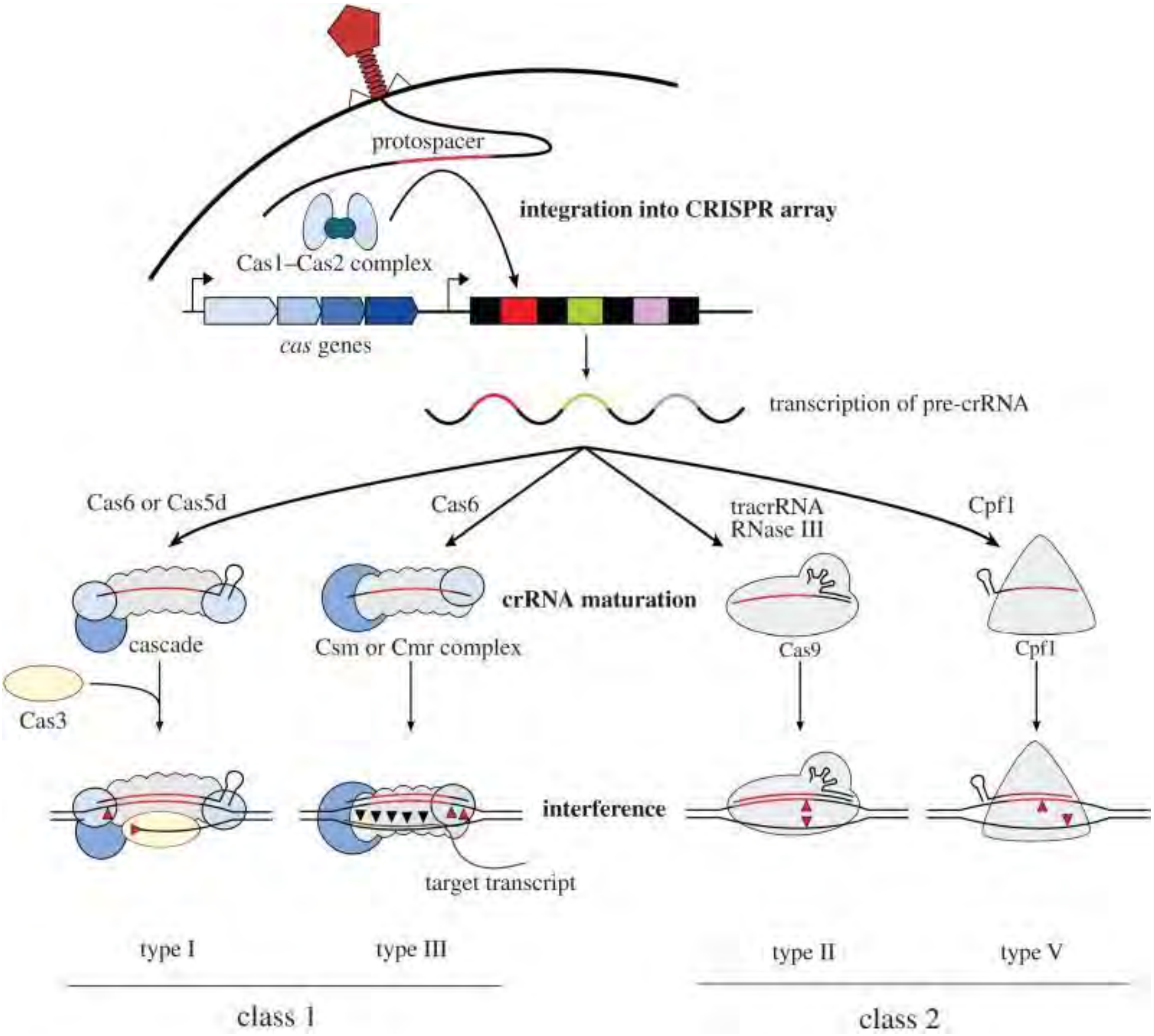


Figure 1: Schematic diagram of the working mechanism of CRISPR-Cas sub-systems (Hille & Charpentier, 2016).

The two halves of the CRISPR-Cas systems include a Cas operon (blue arrows) and a CRISPR array, which comprises of identical repeat sequences (black rectangles) that are separated by spacers acquired from phages (colored rectangles). After phage infection, the Cas1-Cas2 complex inserts a protospacer sequence from the invasive DNA into the CRISPR array. Cas6 further processes the long precursor CRISPR RNA (pre-crRNA) generated by the CRISPR array in type I and III systems (Cas5d processes it in type I-C CRISPR-Cas systems).

In type II CRISPR-Cas systems, TracrRNA, RNase III, and Cas9 are necessary for crRNA maturation, but in type V-A systems, Cpf1 alone is sufficient. CrRNA tells Cascade proteins to bind the foreign DNA in a sequence-specific way when type I systems are in the interference state. The misplaced strand is subsequently broken down by Cas3 using its 3′–5′ exonucleolytic activity. In Type III-A and Type III-B CRISPR-Cas systems, respectively, Csm and Cmr complexes are employed to cleave DNA (red triangles) and its transcripts (black triangles). A ribonucleoprotein complex composed of Cas9 and a tracrRNA:crRNA duplex detects and destroys invading DNA in type II CRISPR-Cas systems. Targeted cleavage in type V systems is carried out by the crRNA-guided effector protein Cpf1. The interference machinery's cleavage points are depicted by red triangles.

1.3. Objective

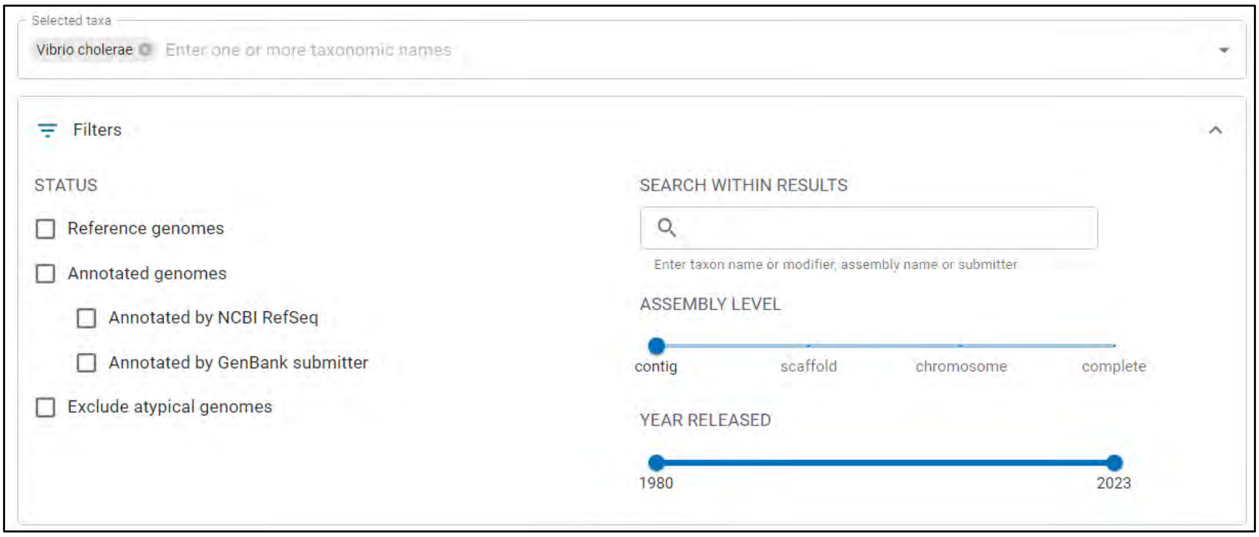
From the review of literature in the above sections, it can be hypothesized that the presence and diversity of CRISPR-Cas would lead to a decreased diversity of plasmid derived contigs in *V. cholerae* genome assemblies. The objective of this research is to identify plasmid derived contigs and CRIPSPR-Cas systems from all available genomes from the NCBI database using alignment and machine learning techniques and look into their distribution and co-existence in order to test this hypothesis.

CHAPTER 2

METHODS

2.1. Collection of Genome Assemblies

All submitted genome assemblies were downloaded from NCBI genome database (https://www.ncbi.nlm.nih.gov/data-hub/genome/). When asked for taxa, "*Vibrio cholerae*" was given as input with the following additional filters. At the time of downloading, a total of 5873 genome assemblies were available.



Selected taxa

Vibrio cholerae ⊗ Enter one or more taxonomic names

Filters

STATUS

□ Reference genomes

□ Annotated genomes

□ Annotated by NCBI RefSeq

□ Annotated by GenBank submitter

□ Exclude atypical genomes

SEARCH WITHIN RESULTS

Enter taxon name or modifier, assembly name or submitter

ASSEMBLY LEVEL

● contig — scaffold — chromosome — complete

YEAR RELEASED

● 1980 — 2023 ●

Figure 2: New NCBI Genome page from where V. cholerae assemblies were downloaded

2.2. Plasmid prediction via machine learning using PlasForest

A homology-based Random Forest classifier called PlasForest (Pradier et al., 2021) is used to recognize bacterial plasmid sequences in incompletely assembled genomes. With an F1 score of 0.950 PlasForest reads contigs from FASTA files and classifies them as plasmids or chromosomes without knowing the samples' taxonomic origin. Importantly, it has a 2.8% false positive rate for plasmid contigs under 1 kb and a 99.9% false positive rate for contigs exceeding 50 kb. The PlasForest pipeline's workflow is illustrated by the following figure.

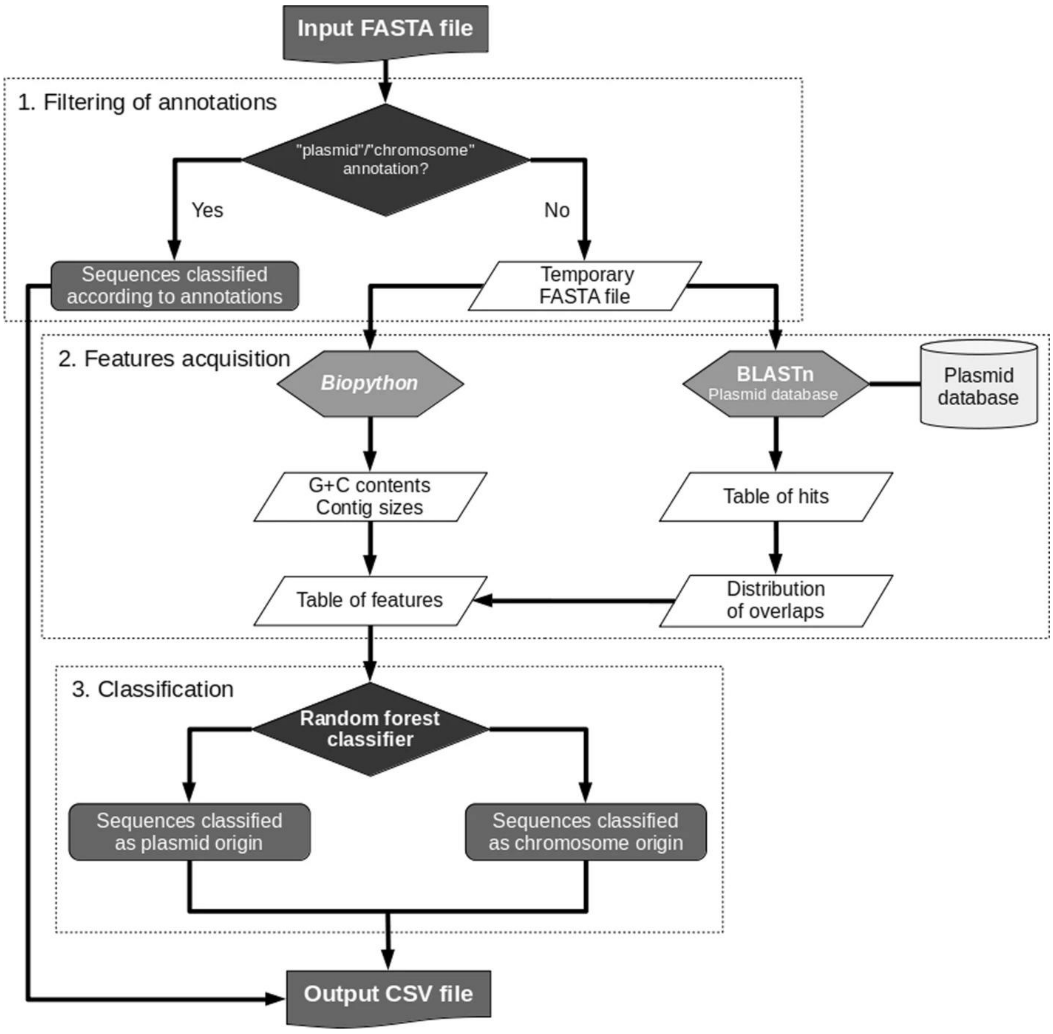


Figure 3: PlasForest Workflow (Pradier et al., 2021).

The NCBI RefSeq Genomes FTP site was used to retrieve all of the bacterial plasmid sequences. A total of 36,450 sequences from this database with an e-value of 10-3 were utilized as the basis for finding homology.

Strong differentiation between plasmids and chromosomes was established by combining homology search with measurements of nucleotide content. The features listed below were chosen for categorization.

- The maximum overlap among hits in the relevant database was evaluated since it is anticipated that the plasmid derived contigs will form longer alignments with sequences from the plasmid database than contigs derived from chromosomes.
- Short contigs match with the query database more frequently than large contigs, so contig size was taken into account.
- In order to distinguish between chromosomes and plasmids, other metrics of the distribution of overlaps among hits, such as the quantity of hits, average overlap, median overlap, and variance of overlaps, may be utilized. In reality, it can also be expected that query chromosomal contigs show hits in the plasmid database due to recombination.
- Plasmid nucleotide composition differs from that of chromosomes, so the G + C content was also added.

The discrepancies between the properties of plasmid contigs and chromosomal contigs were found using a Random Forest classifier. By employing a vast number of distinct decision trees, this strategy lowers individual error. In order to train and test the classifier, 10,152 complete bacterial genomes were randomly picked from the NCBI Refseq Genomes FTP site.

As a result, a machine learning model was created that can determine whether a contig originates from a plasmid or a chromosome based on the attributes retrieved.

PlasForest generates a CSV file for each assembly, therefore a python script (Appendix 1B) was written to counting plasmid and chromosome contigs from all genome assemblies into a single CSV file.

2.3. CRISPR-Cas system identification via machine learning using CRISPRCasTyper

CRISPRCasTyper (Russel et al., 2020) is a bioinformatics tool used to identify Cas operons and associated CRISPR arrays from an input DNA sequence.

The program works by initially using Prodigal (Hyatt et al., 2010), a quick and accurate protein-coding gene prediction tool for prokaryotic genomes to find ORFs and then running HMMER3 (*HMMER v3.3.2*, n.d.) against 680 manually curated hidden Markov models (HMMs) to find Cas and other genes that are functionally associated to CRISPR-Cas systems. Class 2 effectors and the III-E gRAMP fusion protein matches are filtered using E-value and coverage cutoffs that are tuned for each effector. Overall cutoffs are used to filter the remaining HMM matches. Synteny is used to link adjacent Cas and associated genes into operons. These operons are then classified using a score system.

Following that, minced is used to detect CRISPR arrays. To reduce false-positive arrays, an additional step is included, which filters and eliminates arrays with non-similar repeat sequences, identical spacer sequences, or differing lengths of spacers. If the mean repeat sequence identity is less than 70%, the mean spacer sequence identity is greater than 55%, or the mean spacer length standard error is larger than 3.5, CRISPR arrays are quarantined. CRISPRs, on the other hand, are invariably conserved close to Cas operons. CRISPR-Cas systems where the subtype is predicted with probability greater than 0.9 are always retained. To compute sequence identity, pairwise2.align.globalxx from the Biopython package is used with default penalties. A CRISPR repeat classification algorithm called repeatTyper was built to classify distant and orphan CRISPR arrays and predict the subtype of CRISPR-Cas operons that could not be accurately categorized on the basis of Cas ORFs. To foreCast the subtype, the model employs extreme gradient-boosting decision trees fitted to counts of canonical tetramers, regardless of their order. The workflow is summarized in the figure below.

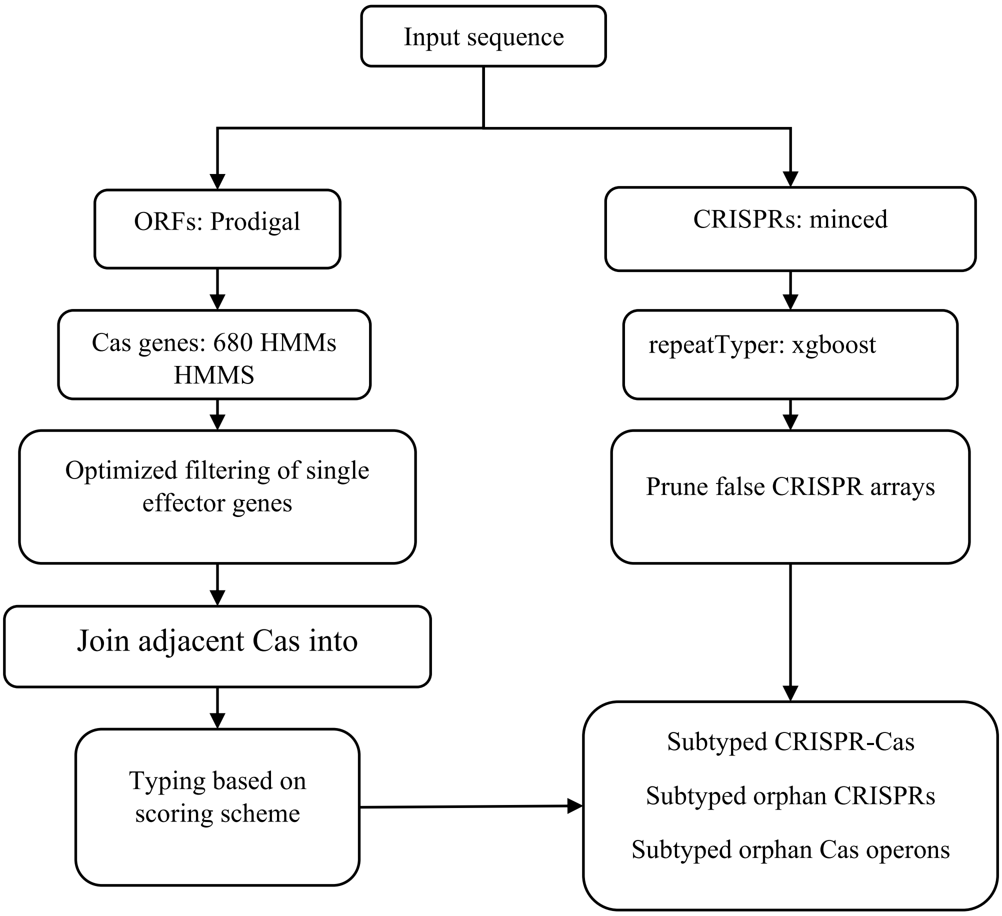


Figure 4: CRISPRCasTyper Workflow

During the development period. CRISPRCasTyper has a median accuracy of 98.6% when tested against a carefully chosen collection of 31 subtypes. The version used in this study can detect 44 subtypes/variants, which, along with the complete typing scheme is provided in Appendix 2A.

2.4. Statistical Analysis and Visualization

All statistical analyses used in this study was performed using Analysis Toolpak provided with Microsoft® Excel® LTSC MSO (16.0.14332.20492), Microsoft Corporation. (2019).

All visualization was performed with the default charts provided with Microsoft® Excel® LTSC MSO (16.0.14332.20492), Microsoft Corporation. (2019).

CHAPTER 3

RESULTS

3.1. Plasmid distribution

5873 *V. cholerae* genome assemblies from NCBI GenBank were used in this study. The downloaded genomes were assembled as scaffolds (13.45%), contigs (84.27%), chromosomes (0.19%) and complete genomes (2.09%).

PlasForest was able to identify plasmid derived contigs in >99% of the scaffold and contig level assemblies. The following figure visualizes the frequency of the different assembly levels and the assemblies with successfully identified plasmids.

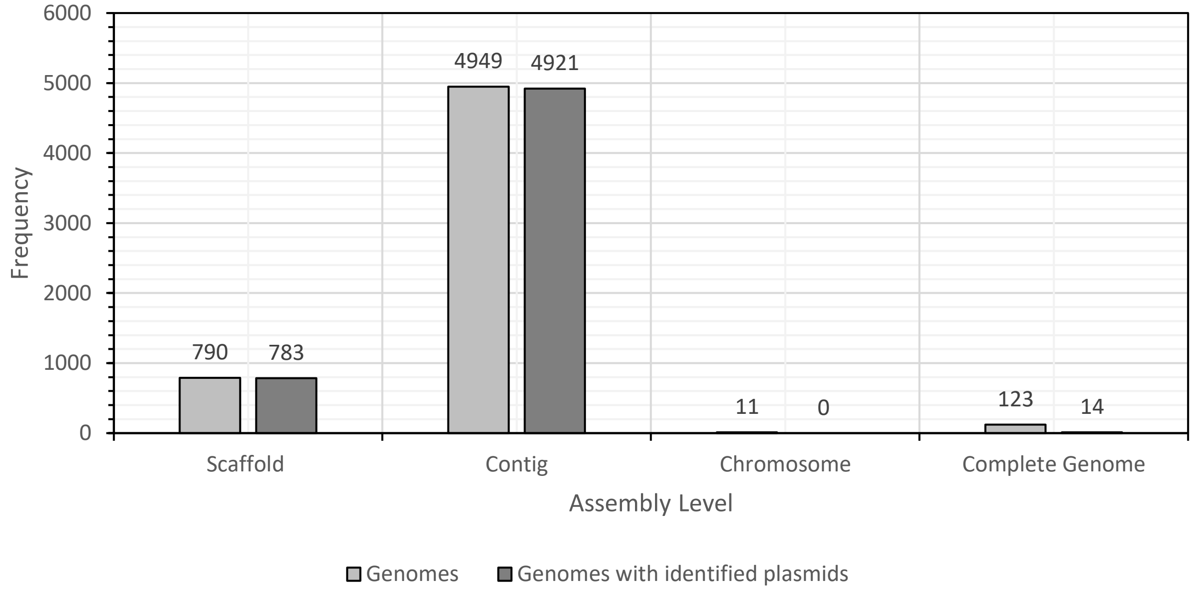


Figure 5: Frequency of assembly levels in the downloaded genomes.

The genomes assemblies where plasmids derived contigs (PDC) were successfully identified, the assemblies contain varying %PDC, with majority containing 25-30% of PDC (mean 27.43 ± 14.02% PDC). The distribution of %PDC is positively skewed (S_{KP} = 1.6, right tailed) and highly leptokurtic (k = 6.18). The figure below shows the frequency distribution of %PDC among genome assemblies.

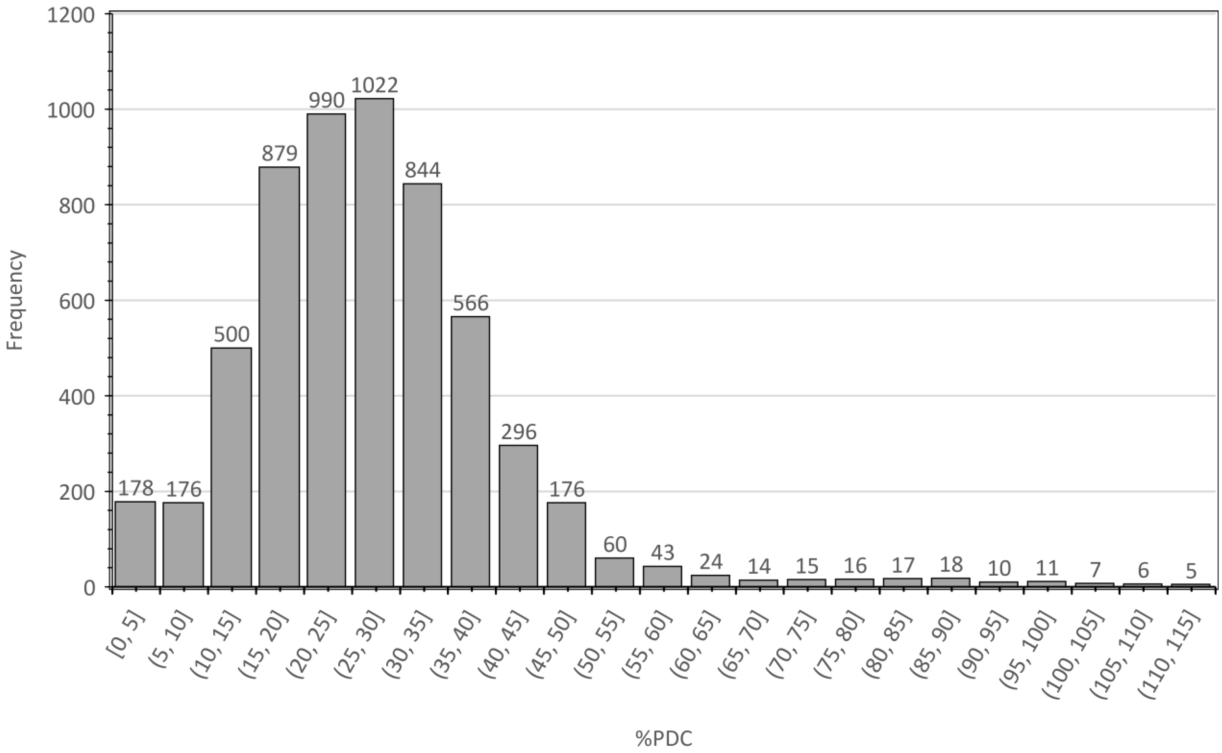


Figure 6: Frequency distribution of %PDC in genome assemblies

3.2. CRISPR-Cas distribution

CRISPRCasTyper was able identify orphan CRISPR arrays and Cas operons alongside complete CRISPR-Cas systems (CCS) and their subtypes. Out of the 5873 genome assemblies, 1073 contained CCS. Figure 7 shows the frequency distribution of number of CCS in genome assemblies.

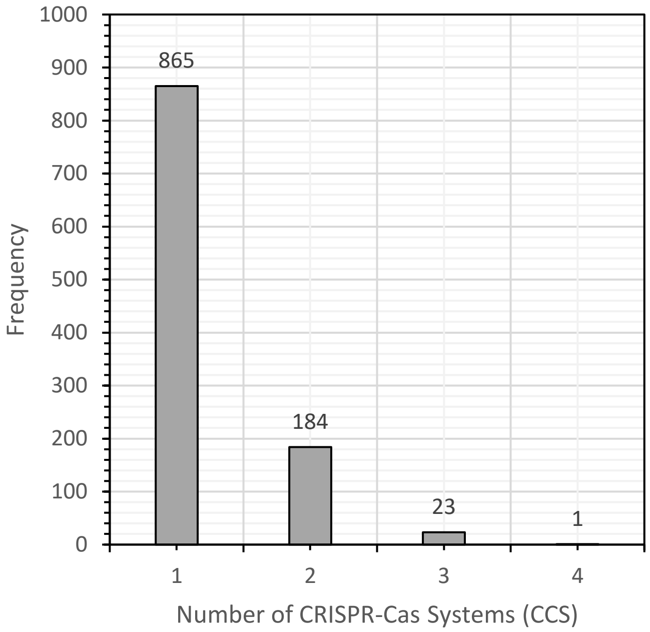


Figure 7: Frequency distribution of number of complete CRISPR-Cas Systems

The different subtypes of CCS identified were: I-C, I-E, I-F, I-F_T, III-B, III-D, IV-A1 and II-D.

All the other subtypes except II-D was found to exist singularly or co-exist with other identified subtypes. II-D was identified in only one assembly where it co-existed with I-F and I-F_T. Figure 8 shows the frequency distribution of CCS subtypes in genome assemblies.

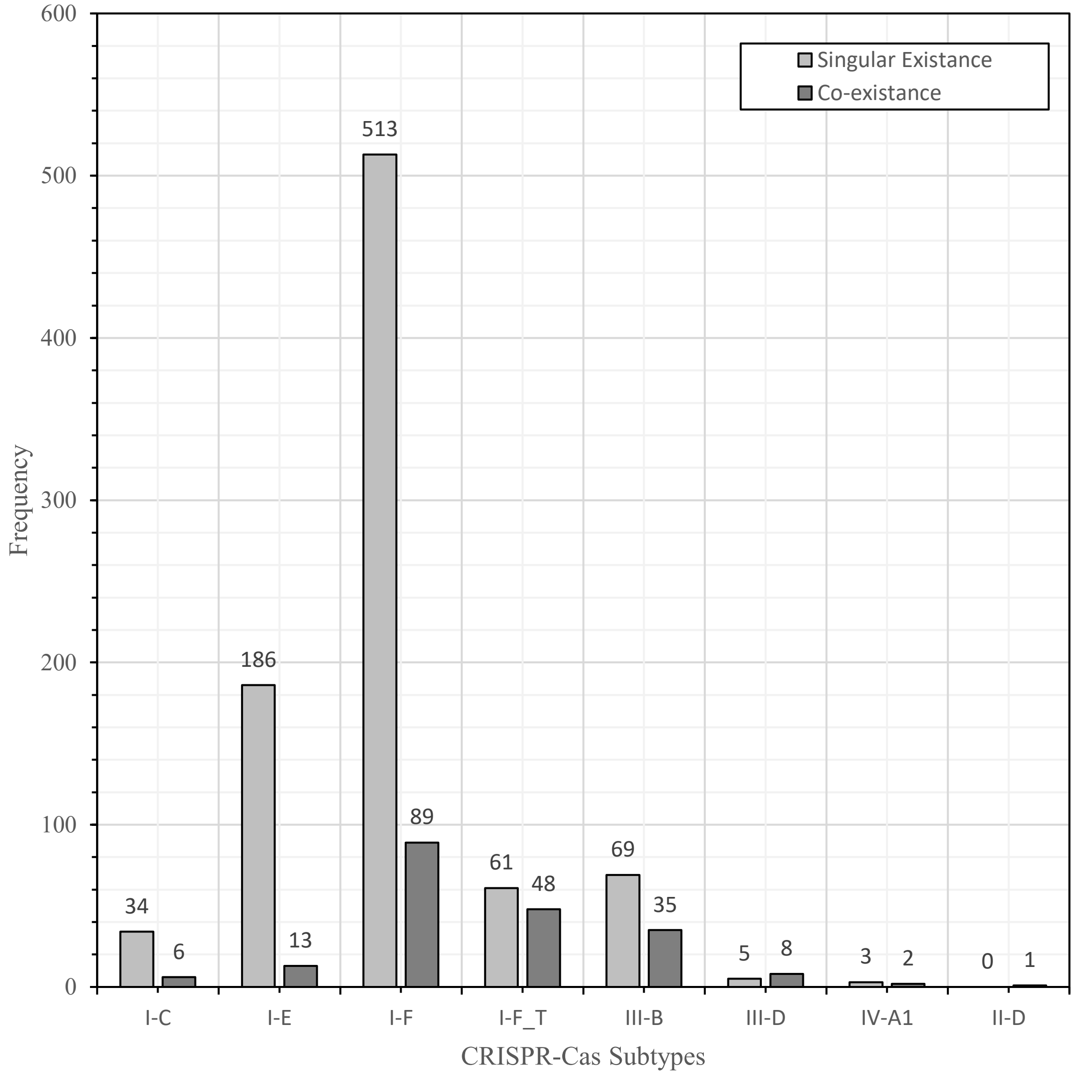


Figure 8: Frequency Distribution of CRISPR-Cas Subtypes

3.3. Comparative analysis of the co-existence of plasmid and CRISPR-Cas systems

In order to get an initial idea of the relationship between number of CCS and %PDC in genomes, a Pearson's Correlation test was performed. No significant correlation found between number of complete CRISPR-Cas systems in genomes and % of plasmid contigs (r = -0.06645, p = 0.03255).

The data was then divided into four groups based on the number of CCS in the assemblies: 0 CCS, 1 CCS, 2 CCS and ≥3 CCS. The assemblies with no PDC were ignored. Figure 9 shows the mean %PDC across the four groups.

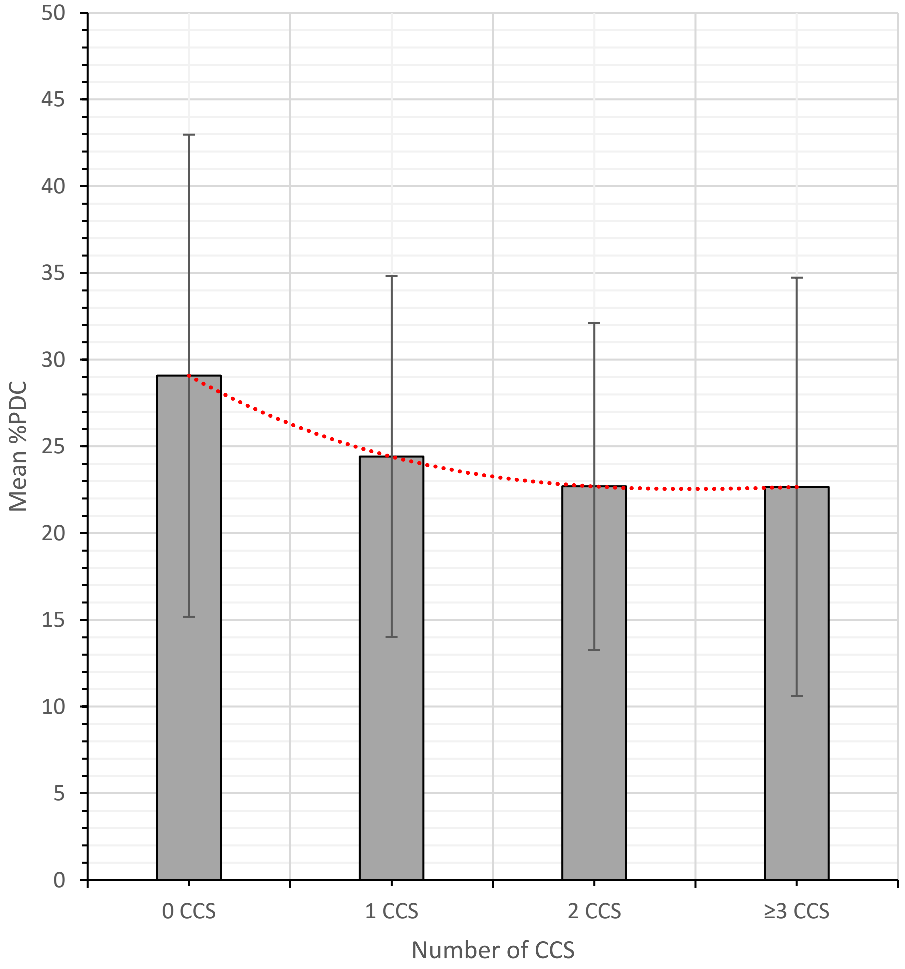


Figure 9: Mean %PDC against Number of CCS

Mean %PDC for the group with 0 CCS was 29.08±13.89% which was higher than the groups with 1 CCS, 2 CCS and ≥3 CCS with mean %PDC 24.41±10.41%, 22.69±9.43% and 22.67±12.06%. Table 2 shows the results of some descriptive statistics.

***Table 2:** Descriptive Statistics*

	%PDC 0 CCS	*%PDC 1 CCS*	*%PDC 2 CCS*	*%PDC ≥3 CCS*
Mean	29.08054495	24.41296879	22.69420943	22.66825776
Standard Error	0.203063947	0.359882504	0.710641485	2.515620653
Median	27.35042735	22.79591384	21.70289855	20
Mode	33.33333333	20	20	#N/A
Standard Deviation	13.89616311	10.40552202	9.427724667	12.06449283
Sample Variance	193.1033492	108.2748885	88.88199239	145.5519872
Kurtosis	7.161842106	1.328306793	1.312932724	0.46826614
Skewness	1.975925588	0.82998283	0.841880767	0.680128025
Range	113.3063154	74.74189676	52.38095238	50.95238095
Minimum	1.587301587	4.081632653	5.194805195	3.333333333
Maximum	114.893617	78.82352941	57.57575758	54.28571429
Sum	136184.192	20409.24191	3994.18086	521.3699284
Count	4683	836	176	23

One-way ANOVA was performed at 5% significance level and it was determined that there was a significant difference in mean %PDC among the groups (p<0.01). ANOVA results are shown in the table below.

***Table 3:** One-way ANOVA*

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0 CCS	4683	136184.192	29.08054495	193.1033492
1 CCS	836	20409.24191	24.41296879	108.2748885
2 CCS	176	3994.18086	22.69420943	88.88199239

≥3 CCS	23	521.3699284	22.66825776	145.5519872		

ANOVA

Variation Source	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	21656.25035	3	7218.750115	40.70750913	5.07E-26	2.60646322
Within Groups	1013275.905	5714	177.33215			
Total	1034932.156	5717				

Further post-hoc analysis was done using T-test with Bonferroni correction and it was determined that there was a significant difference in the mean %PDC when the test was performed between 0 CCS and 1 CCS (p<0.01), and 0 CCS and 2 CCS (p<0.01). Comparisons between other groups show no significant differences in mean %PDC between them.

Number of post-hoc tests made, k = 6

Bonferroni Corrected α = $\frac{\alpha}{k} = \frac{0.05}{6}$ = 0.008333333

Table 4: Post Hoc T-test w/ Bonferroni Correction

Groups	*P-value*	*Significant?*
0CCS VS 1CCS	2.1314E-28	Yes
0CCS VS 2CCS	1.60001E-15	Yes
0CCS VS 3CCS	0.018522621	No
1CCS VS 2CCS	0.031825403	No
1CCS VS 3CCS	0.499250253	No
2CCS VS 3CCS	0.992155848	No

CHAPTER 4

DISCUSSION

The study analyzed 5873 *V. cholerae* genome assemblies from NCBI GenBank and used PlasForest and CRISPRCasTyper tools to identify plasmid derived contigs (PDC) and CRISPR-Cas systems (CCS) respectively.

PlasForest's classification approach was trained on a diverse range of simulated draft genomes, resulting in robust and exact findings. It has a high sensitivity (92.7%) and accuracy (97.3%) in distinguishing between plasmid and chromosomal sequences. While other classifiers have disadvantages such as low accuracy for short contigs or limited sensitivity, PlasForest consistently outperformed them. It had the greatest MCC (up to 0.988) and F1 score (up to 0.988) for contigs larger than 50 kb, demonstrating its unrivaled accuracy. PlasForest was able to identify PDC in >99% of the scaffold and contig level assemblies, with the majority containing 25-30% of PDC. The distribution of %PDC is positively skewed and highly leptokurtic.

In terms of accuracy, CRISPRCasTyper exceeds CRISPRCasFinder, with a median accuracy of 99.5% for tested subtypes and 98.6% for all 31 subtypes, compared to CRISPRCasFinder's median accuracy of 93.9%. It identifies Cas operons more thoroughly, including additional genes inside operons such as Cas1, Cas2, Cas4, Cas10, and Cas6. The technique has a 0.4% false-positive rate and may find operons that curated data misses, exposing unexpected subtypes and variations. CRISPRCasTyper includes features including gene maps for displaying operonic organization, precise resolution of loci spanning circular sequences, and quick processing times. It can perform a deep metagenome assembly (60-100 Mbp) in less than 10 minutes and evaluate a normal sized genome (2-6 Mbp) in under a minute. These benefits make CRISPRCasTyper a dependable, thorough, and efficient tool for this research. CRISPRCasTyper was able to identify 1073 CCS and their subtypes, with I-C, I-E, I-F, I-F_T, III-B, III-D, IV-A1, and II-D being the identified subtypes.

A Pearson's Correlation Test showed no significant correlation between the number of CCS and %PDC. The data was then divided into four groups based on the number of CCS in the assemblies, and a one-way ANOVA test was conducted. The one-way ANOVA (Analysis of Variance) method is used to assess whether or not there are significant differences in the means of three or more groups or treatments. It enables researchers to study numerous groups at the same time and determine if any detected differences are statistically significant or merely coincidental. The test showed that there was a significant difference in %PDC and number of CCS.

However, ANOVA does not specify between which groups the differences occur and so further post-hoc analysis of pairwise T-test with Bonferroni correction among all of the four groups was performed. The Bonferroni correction was necessary as multiple hypothesis testing was performed, the probability of making type I error (rejection of null hypothesis when it is actually true) increases. The adjusted α-level value (0.00833) is lower than that used in ANOVA (0.05) in order to reduce the significance level to counteract the effect of increased type I error probability.

Although there has been extensive research within the recent years on the mechanism and properties of CRISPR-Cas systems in bacteria, the novelty in this research is the determination of sub-species wide CRISPR-Cas distribution in all publicly available genome assemblies of *V. cholerae* available in NCBI genome was used. According to Garneau et al., 2010, the CRISPR-Cas system may also spontaneously acquire spacers from a self-replicating plasmid harboring an antibiotic-resistance gene, resulting in plasmid loss. The ability of CRISPR-Cas systems to cleave plasmid sequences is also confirmed by many other articles available, however, no confirmatory evidence was produced whether or not presence of CRISPR-Cas in V. cholera itself is responsible for the decrease in plasmid content, therefore, further research is necessary. A possible approach is to assess whether any of the spacer sequences match with an extensive plasmid database, and it is important to look into why presence of more than one CCS had no statistically significant change in %PDC.

CHAPTER 5

CONCLUSION

The objective of this research was to investigate the presence and diversity of plasmid-derived contigs and CRISPR-Cas systems in *Vibrio cholerae* genome assemblies. The study utilized alignment and various machine learning techniques to analyze genome assemblies from the NCBI database. In order to identify plasmid distribution PlasForest was used and CRISPRCasTyper was used to determine the distribution of CRISPR-Cas subtypes. The results show there is a statistically significant decrease in %PDC between groups where no CCS was present and the group which had 1 CCS and 2 CCS, but the difference in %PDC was not significant between the groups with multiple CCSs. However, there is no causative evidence the presence of CCS is the primary reason for the decrease in %PDC and why multiple CCS had no effect on decrease in %PDC for which further research is necessary. What is unique in this study is the discovery of sub-species wide distribution of plasmid derived contigs and CRISPR-Cas subtypes in *Vibrio cholerae*.

REFERENCES

1. Abby, S. S., Néron, B., Ménager, H., Touchon, M., & Rocha, E. P. C. (2014, October 17). MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLoS ONE, 9(10), e110726. https://doi.org/10.1371/journal.pone.0110726
2. Amaro, C., Aznar, R., Garay, E., & Alcaide, E. (1988, November). R plasmids in environmental *Vibrio cholerae* non-O1 strains. Applied and Environmental Microbiology, 54(11), 2771–2776. https://doi.org/10.1128/aem.54.11.2771-2776.1988
3. Barrangou, R. (2015, February). The roles of CRISPR–Cas systems in adaptive immunity and beyond. Current Opinion in Immunology, 32, 36–41. https://doi.org/10.1016/j.coi.2014.12.008
4. Barua, D. (1992). History of Cholera. Cholera, 1–36. https://doi.org/10.1007/978-1-4757-9688-9_1
5. Carattoli, A. (2013, August). Plasmids and the spread of resistance. International Journal of Medical Microbiology, 303(6–7), 298–304. https://doi.org/10.1016/j.ijmm.2013.02.001
6. Carraro, N., Rivard, N., Ceccarelli, D., Colwell, R. R., & Burrus, V. (2016, September 7). IncA/C Conjugative Plasmids Mobilize a New Family of Multidrug Resistance Islands in Clinical *Vibrio cholerae* Non-O1/Non-O139 Isolates from Haiti. MBio, 7(4). https://doi.org/10.1128/mbio.00509-16
7. Cholera. (2022, March 30). Cholera. Retrieved March 30, 2023, from https://www.who.int/news-room/fact-sheets/detail/cholera
8. Cholera's Seven Pandemics. (n.d.). Cholera's seven pandemics - Technology & Science - CBC News. Retrieved March 30, 2023, from https://web.archive.org/web/20160302113041/http://www.cbc.ca/news/technology/cholera-s-seven-pandemics-1.758504
9. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009, March 20). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163
10. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018, May 22). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research, 46(W1), W246–W251. https://doi.org/10.1093/nar/gky425

11. Fonseca, E. L., & Vicente, A. C. P. (2013, October 1). Epidemiology of qnrVC alleles and emergence out of the Vibrionaceae family. Journal of Medical Microbiology, 62(10), 1628–1630. https://doi.org/10.1099/jmm.0.062661-0
12. Garneau, J. E., Dupuis, M. V., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H., & Moineau, S. (2010, November 3). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 468(7320), 67–71. https://doi.org/10.1038/nature09523
13. Grissa, I., Vergnaud, G., & Pourcel, C. (2007, May 8). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research, 35(Web Server), W52–W57. https://doi.org/10.1093/nar/gkm360
14. Hille, F., & Charpentier, E. (2016, November 5). 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
15. HMMER v3.3.2. (n.d.). HMMER: biosequence analysis using profile hidden Markov models. Retrieved April 14, 2023, from http://hmmer.org/
16. Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010, March 8). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1). https://doi.org/10.1186/1471-2105-11-119
17. Luo, Y., Ye, J., Payne, M., Hu, D., Jiang, J., & Lan, R. (2022, November). Genomic Epidemiology of *Vibrio cholerae* O139, Zhejiang Province, China, 1994–2018. Emerging Infectious Diseases, 28(11), 2253–2260. https://doi.org/10.3201/eid2811.212066
18. Neron, B., Denise, R., Coluzzi, C., Touchon, M., Rocha, E. P. C., & Abby, S. S. (2022, September 4). MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. https://doi.org/10.1101/2022.09.02.506364
19. Pradier, L., Tissot, T., Fiston-Lavier, A. S., & Bedhomme, S. (2021, June 26). PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. BMC Bioinformatics, 22(1). https://doi.org/10.1186/s12859-021-04270-w
20. Reidl, J., & Klose, K. E. (2002, June). *Vibrio cholerae* and cholera: out of the water and into the host. FEMS Microbiology Reviews, 26(2), 125–139. https://doi.org/10.1111/j.1574-6976.2002.tb00605.x
21. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020, December 1). 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

22. Waldor, M. K., Rubin, E. J., Pearson, G. D. N., Kimsey, H., & Mekalanos, J. J. (1997, June). Regulation, replication, and integration functions of the *Vibrio cholerae* CTXφ are encoded by region RS2. Molecular Microbiology, 24(5), 917–926. https://doi.org/10.1046/j.1365-2958.1997.3911758.x

APPENDIX

Appendix 1: Code Snippets

1A: Bash shell script for plasmid identification using PlasForest.

```
#!/bin/bash

cd /mnt/e/Thesis/Database/vibrio_cholerae/data/
ls>/mnt/e/Thesis/Tools/PlasForest/PlasForest/filenames.txt
cd /mnt/e/Thesis/Tools/PlasForest/PlasForest/
awk '{ print "python3 PlasForest.py -i
/mnt/e/Thesis/Database/vibrio_cholerae/data/"$1"
-o /mnt/e/Thesis/Results/Plasmids/"substr($1, 1, length($1)-4".csv}'
filenames.txt | bash
```

1B: Python script for counting plasmid and chromosome contigs from all genome assemblies into a single .csv file

```python
import os
import csv
import pandas as pd

# Set the directory containing the CSV files
directory = input("Path to Directory: ")

# Initialize an empty list to store the results
results = []

# Loop through all CSV files in the directory
for filename in os.listdir(directory):
    if filename.endswith(".csv"):
        filepath = os.path.join(directory, filename)

        # Initialize a counter for the number of plasmid entries
        num_plasmid = 0
        num_chromosome = 0

        # Read the CSV file and count the number of "Plasmid" entries
        with open(filepath, "r") as csvfile:
            reader = csv.DictReader(csvfile)
            for row in reader:
                if row["Prediction"] == "Plasmid":
                    num_plasmid += 1

        # Read the CSV file and count the number of "Chromosome" entries
        with open(filepath, "r") as csvfile:
            reader = csv.DictReader(csvfile)
            for row in reader:
                if row["Prediction"] == "Chromosome":
                    num_chromosome += 1

        # Append the results to the list
        results.append({"Assembly Name": filename, "#Plasmid_Contigs":
num_plasmid, "#Chromosome_Contigs": num_chromosome})

# Write the results to a new CSV file
with open("statistics.csv", "w", newline="") as csvfile:
    fieldnames = ["Assembly Name", "#Plasmid_Contigs", "#Chromosome_Contigs"]
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)

    # Write the header row
    writer.writeheader()

    # Write the results for each CSV file
    for result in results:
        writer.writerow(result)

#Read statistics.csv and remove duplicates
for filename in os.listdir(directory):
    if filename == "statistics.csv":
        # read the csv file into a pandas DataFrame
        df = pd.read_csv(filename)

        # remove duplicate entries
        df.drop_duplicates(inplace=True)
```

1C: Python script for gathering CRISPR-Cas data from all genome assemblies into a single .csv file

```
import os
import csv

# initialize the list and count variable
l = []
n = 0

# create the csv file and write the headers
with open('results.csv', mode='w', newline='') as results_file:
    writer = csv.writer(results_file)
    writer.writerow(['Assembly name', '#Complete_CRISPRCas', 'Subtype'])

# iterate through all subfolders in the parent folder
for root, dirs, files in os.walk("."):
    for dir in dirs:
        # check if the CRISPR_Cas.tab file exists in the current subfolder
        if os.path.isfile(os.path.join(root, dir, 'CRISPR_Cas.tab')):
            # read the tab file and extract the data under the "Prediction"
column header
            with open(os.path.join(root, dir, 'CRISPR_Cas.tab'), mode='r') as
tab_file:
                reader = csv.DictReader(tab_file, delimiter='\t')
                for row in reader:
                    l.append(row['Prediction'])
                    n += 1

            # append the subfolder name, count variable, and list to the csv
file
            with open('results.csv', mode='a', newline='') as results_file:
                writer = csv.writer(results_file)
                writer.writerow([dir, n, l])

            # reset the list and count variable for the next subfolder
            l = []
            n = 0
```

Appendix 2: Supplementary Tables

2A: repeatTyper typing scheme

Class	*Type*	*Subtype*	*Variant*	*RepeatTyper*	*Genes*				
1	I	I-A		Included	*Cas3 Cas3HD Cas5 Cas7 Cas8a/csaX csa5*	*Cas1 Cas2 Cas4*		*Cas6*	
1	I	I-B		Included	*Cas3 Cas5 Cas7 Cas8b*	*Cas1 Cas2 Cas4*		*Cas6*	
1	I	I-C		Included	*Cas5 Cas7 Cas8c*	*Cas1 Cas2 Cas4*			
1	I	I-D		Included	*Cas3 Cas5/csc1 Cas7/csc2 Cas10d*	*Cas1 Cas2 Cas4*		*Cas6*	
1	I	I-E		Included	*Cas3 Cas5 Cas7 Cas8e/cse1 Cas11/cse2*	*Cas1 Cas2*		*Cas6*	
1	I	I-F		Included	*Cas5f/csy2 Cas7f/csy3 Cas8f/csy1*	*Cas1*	*Cas3-Cas2*	*Cas6f*	
1	I	I-F	_T	Included	*Cas5f/csy2 Cas7f/csy3 Cas8f/csy1 tniQ*			*Cas6f*	
1	I	I-G		Included	*Cas3 Cas5/csb2 Cas7/csb1 Cas8u/csb3*	*Cas1/Cas4 Cas2*			
1	III	III-A		Included	*Cas10 csm2 csm3 csm4 csm5*				*csm6*
1	III	III-B		Included	*Cas10 cmr1 cmr3 cmr4 cmr5 cmr6*				
1	III	III-C		Included	*Cas10 cmr1 cmr3 cmr4 cmr5 cmr6*				
1	III	III-D		Included	*Cas10 Cas11/csm2 csm3 csm5 csx10 csx19*				
1	III	III-E		Included	*gRAMP*				
1	III	III-F		Included	*SSgr11 Cas10 Cas5 csm3*				
1	IV	IV-A	1	Included	*csf1 csf2 csf3 csf4*			*Cas6*	
1	IV	IV-A	2	Included	*csf2 csf3 csf4*			*Cas6*	
1	IV	IV-A	3	Included	*csf1 csf2 csf3 csf4*			*Cas6*	
1	IV	IV-B		NA	*Cas11 csf1 csf2 csf3*				*(cysH)*
1	IV	IV-C		Not enough data	*csf2 csf3 Cas10-like Cas11*			*(Cas6)*	
1	IV	IV-D		Included	*csf1 csf2 csf3 recD*			*Cas6*	

1	IV	IV-E		Included	*csf3-csf1 csf2 csf4*		*Cas6*	
2	II	II-A		Included	*Cas9*	*Cas1 Cas2 csn2*		
2	II	II-B		Included	*Cas9*	*Cas1 Cas2 Cas4*		
2	II	II-C		Included	*Cas9*	*Cas1 Cas2*		
2	II	II-C	2	Not enough data	*Cas9*			
2	II	II-D		Not enough data	*Cas9*			
2	V	V-A		Included	*Cas12a*	*Cas1 Cas2 Cas4*		
2	V	V-B	1	Included	*Cas12b1*	*Cas1 Cas2 Cas4*		
2	V	V-B	2	Included	*Cas12b2*	*Cas1 Cas2 Cas4*		
2	V	V-C		Not enough data	*Cas12c*	*Cas1*		
2	V	V-D		Not enough data	*Cas12d*			
2	V	V-E		Included	*Cas12e*	*Cas1 Cas2 Cas4*		
2	V	V-F	1	Included	*Cas12f1*	*Cas1 Cas2 Cas4*		
2	V	V-F	2	Included	*Cas12f2*	*Cas1 Cas2 Cas4*		
2	V	V-F	3	Included	*Cas12f3*	*Cas1 Cas2 Cas4*		
2	V	V-F		Not included	*Cas12f6/Cas12f7*			
2	V	V-G		Included	*Cas12g*			
2	V	V-H		Not enough data	*Cas12h*			
2	V	V-I		Included	*Cas12i*			
2	V	V-J		Included	*Cas12j (Cas-phi)*			
2	V	V-K		Included	*Cas12k tniQ tnsB tnsC*			*merR*
2	V	V-L		Not enough data	*Cas12l*	*Cas1 Cas2 Cas4*		
2	V	V-M		Included	*Cas12m*			
2	VI	VI-A		Included	*Cas13a*			
2	VI	VI-B	1	Included	*Cas13b1*			*(csx27)*
2	VI	VI-B	2	Included	*Cas13b2*			*csx28*

2	VI	VI-C	Included	*Cas13c*
2	VI	VI-D	Included	*Cas13d*
2	VI	VI-X	Not enough data	*Cas13X*
2	VI	VI-Y	Not enough data	*Cas13Y*

Table 5: repeatTyper typing scheme (Russel et al., 2020).