이학박사 학위논문

세균 비교유전체 연구를 위한 생명정보 분석 시스템 개발

Development of bacterial tools for comparative genomics

2019년 2월

서울대학교대학원

생명과학부

이 임 창

세균 비교유전체 연구를 위한 생명정보 분석 시스템 개발

지도 교수 천 종 식

이 논문을 이학박사 학위 논문으로 제출함

2018 년 12 월

서울대학교 대학원

생명과학전공

이 임 창

이임창의 이학박사 학위논문을 인준함

2018년 12월

위 원 장 _____ 이 병 재 _____ (인)

부위원장 _____ 천 종 식 _____ (인)

위 원 _____ 김 동 욱 _____ (인)

위 원 _____ 송 만 기 _____ (인)

위 원 _____ 임 영 운 _____ (인)

Ph.D. Dissertation

Development of tools for bacterial comparative genomics

by Imchang Lee

Advisor: Professor Jongsik Chun. Ph. D.

A Thesis Submitted for the Degree of Doctor of Philosophy

February 2019

School of Biological Sciences

Seoul National University

Abstract

Due to the recent rapid advancement DNA sequencing technologies, genomics has played a significant role in various microbiological disciplines. Adequate algorithms and bioinformatics tools must be developed to analyze large-scale genomic data. The general procedure for analyzing the genome of a bacterium consists of assembly, gene-finding, and functional annotation. Two or more genomes can be compared in various ways, which is called comparative genomics. The objectives of comparative genomics are to predict biological implications and biomarkers by comparing genomic features of multiple genomes. In this study, three bioinformatics tools were developed that can be used for bacterial and comparative genomics.

The bacterial species concept has been changed to adopt genomic relatedness, which is more objective than previously used phenotypic methods. Pairwise genome sequence similarity, called the Overall Genomic Relatedness Index (OGRI), is used in bacterial taxonomy for identification. The most widely used algorithm to calculate the OGRI is average nucleotide identity (ANI). However, conventional ANI using BLAST may produce different similarity values from reciprocal calculations depending on the query sequence selected. To minimize this discrepancy, a new algorithm, OrthoANI, was devised to incorporate the concept of orthology. Both query and subject sequences were fragmented instead of fragmenting only the query in the original ANI algorithm. The pairwise similarity values were included when two fragments were considered orthologous. The values provided by OrthoANI show a good correlation with the original ANI values, and the reciprocal values were almost identical. OrthoANI is readily available for taxonomic purposes without the functional annotation or gene-finding processes. It allows for simple, reproducible and reliable taxonomy.

As the use of next-generation sequencing (NGS) becomes more routine in microbiology, there is growing concern about quality assurance of the sequencing data produced, including contamination. This issue is of particular importance in clinical laboratories as contamination events can lead to false diagnostic results. Development of a system to detect such cases, as a quality control process, is of primary importance in routine microbial genomics labs. In this context, a novel algorithm to detect possible biological contamination from prokaryotic genome assemblies using 16S rRNA gene sequences was proposed in this study and called ContEst16S.

Predictive tools for the *Vibrio cholerae* phenotype were newly developed in this study. The programs are useful to predict the O antigen serotype, the presence of cholera

toxin phage elements, and antibiotic resistance of the *V. cholerae* strain using genomic data. Predicting O antigen serotype provides visualization of the structure of the O antigen gene cluster in the genome data. The tool for predicting cholera toxin phage elements reveals the categorized genetic elements of CTX phage. Antibiotic resistance of *V. cholerae* can also be predicted by the program developed in this study. The process to predict antibiotic resistance uses the RGI (CARD-The Comprehensive Antibiotic Resistance Database) program.

A simple text from sequencing data may not provide a decisive answer to a biological issue. Without biochemical verification, it is only a prediction of the question. However, a prediction produced by bioinformatics has a powerful impact, and the programs developed in this study can help advance microbiology. OrthoANI provides standardized procedures for the taxonomic field, and ContEst16S allows researchers to consult information about contaminated microbial genome assembly data. The tool for predicting the *V. cholerae* phenotype offers species-driven genomic insight, including identifying the O antigen and virulence factors, as well as predicting antibiotic resistance.

Keywords: *Vibrio cholerae*, O serogroup, O serotype, Cholera Toxin (CT), Antibiotic Resistance (AR) of *V. cholerae*, Bacterial genomics, Comparative genomics, OrthoANI, ContEst16S.

Student Number: 2012-20323

Table of Contents

Abbreviation

AGC: Antigen gene cluster

ANI: Average nucleotide identity

AR: Antibiotic resistance

BPGN: Bacterial polysaccharide gene nomenclature

CDS: Coding sequence

CT: Cholera Toxin

CTX: Cholera Toxin Phage

DDH: DNA-DNA hybridization

ELISA: The enzyme-linked immunosorbent assay

G + C: Guanine plus cytosine

HGP: Human Genome Project

LPS: Lipopolysaccharide

MDR: Multiple drug resistance

MLSA: Multi-locus sequencing analysis

NGS: Next-generation sequencing

OGRI: Overall genomic relatedness index

ORF: Open reading frame

OrthoANI: Orthology-based average nucleotide identity

RS1: Repetitive sequences 1

UPGMA: Unweighted pair group method with arithmetic mean

WGS: Whole genome sequencing

LIST of FIGURES

LIST of TABLES

Chapter1. Introduction

Genomics has revolutionized biological and pathological research over the last 20 years. The speed and quantity by which genomics has abandoned the disciplines from which it formerly developed is astonishing. The rapid developments in the field have left enough of the early history of genomics behind, and many essential issues have not been documented correctly.

The term “genetics” began to be used in the late 1970s. According to the first article in ELSEVIER *Genomics*, the term was coined by TH Roderick from the Jackson Laboratory (Bar Harbor, ME, USA) (Lalley *et al.* 1987), September 1987 in discussions with editors VA McKusick and FH Ruddle, who were looking for recommendations to name their new journal (McKusick and Ruddle 1987).

There is no all-encompassing definition of genomics, and the term is used for various purposes. When McKusick and Ruddle wrote their editorial, they regarded genomics to be mapping and sequencing to analyze the arrangement and construction of genomes (McKusick and Ruddle 1987). When the *Genomics* journal was established, only three years had passed since automatic DNA sequencers dominated the initial stages of genomics. Hence, a description, such as McKusick and Ruddle’s concept of genomics, should be interpreted in the context of that era.

Genomics is now subdivided into structural genomics, which considers complete sequences of genomes (DNA sequencing), or the discovery and noting of all the sequences in the entire genome of a particular complete set of proteins in an organism (proteomics). Also, functional genomics investigates the function of genes and metabolic pathways or the gene expression patterns in an organism.

The term genomics has a broader meaning in a genomic study that includes bioinformatics and other research on a genome or proteome to understand the outline and function of an organism. Various technologies that are part of today’s genomics toolkit were developed and automated to apply in large-scale, high-throughput environments. Some scientists, such as the Canadian Nobel laureate Michael Smith, have contended that they were doing genomics all along, considering its apparent origins in molecular biology. Thus, by this definition, genomics started when Watson and Crick discovered the structure of DNA.

The introduction of bioinformatics into molecular biology was a significant factor in the advancement of genomics. Laboratory automation led to the production of enormous quantities of data, and the necessity to analyze, link and understand these results has led to the advancement of bioinformatics, a new discipline at the interface of some traditional methods. Bioinformatics is the adhesive that combines the various

aspects of genomics.

1.1. Bacterial genomics

Bacterial genomics is the discipline that studies bacterial genomes. The field of bacterial genomics reveals the structure and function of genomes encompassing all hereditary information of bacteria. Genomics methods include DNA sequencing, and recombination to manipulate DNA. However, most of the experiments described above, cannot be performed in the vast majority of microbes such as unculturable microbes. The majority of current studies on bacterial genomics are conducted using the computational science called bioinformatics. Bioinformatics data allows scientists to make predictions about bacterial physiology and evolution, even if the microbes cannot be cultured in the laboratory. The achievements of bioinformatics are highly appreciated that less than 1-% of microorganisms can be cultured (Albertsen *et al.* 2013). Bioinformatics has forever changed how the study of microbiology is performed.

1.2. Comparative genomics

Comparative genomics is a broad field that has moved beyond simply comparing two or more genomes. The field of comparative genomics encompasses both NGS-based studies and non-sequencing technologies such as microarrays (Willenbrock *et al.* 2007, Yuan-Hai *et al.* 2010), function-targeted studies, pathway analysis, and whole-genome comparisons through sequence alignment (Hay and Docherty 2006, Sone *et al.* 2007, Iyer *et al.* 2008, Chun *et al.* 2009).

Comparative genomics focuses on the use of genome sequence data to answer biological questions about bacterial evolution, physiology, and pathogenicity (Prentis *et al.* 2004). One of the major goals of comparative studies is to understand the evolution of a bacterial species. For example, genome rearrangement has an important influence on bacterial evolution, including reduction process of genome and creation of new DNA regions (Sun *et al.* 2012). Genome arrangements with genomics tools allow us to determine if a change in genes derived from an evolutionary mechanism contains insertions, deletions, selective sweep, and change by mobile elements on the bacterial genome. Comparative genomics allows us to understand the broad evolutionary history of the microbial world.

Whole genome sequencing technology permits the reconstruction of robust taxonomic trees including pan-genome trees, core-genome trees, super-trees, and universal trees with all genes based on whole genome sequencing alignment (Brown

et al. 2001, Daubin and Gouy 2001, Wolf *et al.* 2002). A good example is the 16S rRNA gene sequence based tree of all bacterial species (Daubin and Gouy 2001). Although the 16S rRNA genes is relatively short sequence, it is a huge amount of work to target all bacterial species and it would have been impossible without the development of NGS and the advance of bioinformatics. Another good part of comparative genomics is the field of overall genome relatedness indices (OGRI). OGRI algorithms are generally used to calculate similarity between two genome sequences without gene-finding and functional annotation steps (Chun and Rainey 2014). In 2006, Konstantinidis et al. suggested the average nucleotide identity (ANI) and the average amino acid identity (AAI) that were could be used to distinguish between prokaryotic species (Konstantinidis and Tiedje 2005, Konstantinidis *et al.* 2006, Goris *et al.* 2007). As the correlation between DDH and ANI was well established (Goris *et al.* 2007) than AAI, ANI was reassuring for the more traditional microbial taxonomists. Because the ANI was computational mimics of DDH, Richter & Rossello-Mora suggested that ANI is the best alternative for a gold standard to delineate microbial species (Richter and Rosselló-Móra 2009, Kim *et al.* 2014, Beaz-Hidalgo *et al.* 2015, Li *et al.* 2015, Yi and Chun 2015).

A second major focus of comparative genomics is understanding the distinction between pathogen and non-pathogenic species. Comparing multiple genomes could enrich our knowledge of the variations and relatedness between pathogen and non-pathogenic organisms. Comparisons between different pathogen genomes can lead to faster identification of distinct mechanisms underlying pathogenicity. Genomic islands have been found between pathogens and non-pathogens and even in closely related species of the same genus (Perna *et al.* 2001, Chain *et al.* 2004, Dobrindt *et al.* 2004).

Differences in genome sequences alone may not provide a decisive answer to which sequences are responsible for a specific phenotype, but genome comparisons generate manageable lists of genomic regions and gene candidates for further study. It is fascinating to make a definitive declaration of gene or protein function based on the computer analysis in DNA sequences. However, without biochemical verification of function in an organism, it is only a prediction. Nevertheless, the predictions provided by bioinformatics have a powerful effect on a scientist's ability to reveal biological issues.

1.3. Objective of this study

The fundamental goal of bioinformatics for microbial genomics and comparative genomics is to build a genomics-related system to make it easier to access genomics

data. In this context, this study developed genomics tools for building a system related to genomics and conducting research.

A program related to taxonomy was first developed because one of the most important disciplines in genomics is taxonomy. ANI has been the most widely used among several overall genome relatedness indices used to calculate the similarity between two genomic sequences. However, ANI values between two strains are often not the same; therefore, they are not symmetrical. Thus, the ANI discrepancy problem was investigated in this study, and a new algorithm, called OrthoANI, was proposed.

As the use of NGS becomes more routine in microbiology, there is growing concern about quality assurance of the sequence data produced, including contamination. This issue is of particular importance in clinical laboratories as contamination events can lead to false results. Developing a system to detect such cases as a quality control process is of primary importance in routine microbial genomics laboratories. In this context, a novel algorithm to detect possible biological contamination from prokaryotic genome assemblies using 16S rRNA gene sequences was proposed in this study and called ContEst16S.

Finally, new tools and algorithms to predict the phenotypes of the bacterium *Vibrio cholerae* were developed. The 67 distinctive O antigen serotypes were predicted, CTX elements of 798 *V. cholerae* genomes were investigated, and the antibiotic resistance status of all *V. cholerae* genome datasets was investigated using the phenotype prediction tools. The programs and algorithms are easy to use for researchers studying *V. cholerae*, and the results from this study were stored in a database.

The objective of this study was to develop helpful, easily usable, fast, and reliable tools to advance microbiology.

Chapter 2. OrthoANI: An improved algorithm and software for calculating average nucleotide identity

2.1. Introduction

The genome is the ultimate source of information for taxonomic purposes and its use has been accelerated significantly thanks to advances in high-throughput DNA sequencing technologies (Chun and Rainey 2014). Currently, the major application of genome sequence data in bacterial taxonomy is to measure overall genomic relatedness between two strains, which also serves as the framework for the species concept (Rosselló-Móra and Amann 2015). The DNA–DNA hybridization (DDH) method has been regarded as the gold standard for the last few decades (Krichevsky *et al.* 1987), despite the fact that it is only an indirect measure of genome sequence similarity, error-prone and labor-intensive (Johnson and Whitman 2007). Since whole-genome sequencing is readily available for general microbiology laboratories, several overall genome relatedness indices (OGRI) have been developed to replace the problematic DDH methods. In general, OGRI algorithms are used to calculate similarity between two genome sequences without gene-finding and functional annotation steps, therefore they tend to be more objective, reproducible, fast and easy-to-implement.

Among various OGRI, average nucleotide identity (ANI) has been the most widely used (Stropko *et al.* 2014, Beaz-Hidalgo *et al.* 2015, Li *et al.* 2015, Rosselló-Móra and Amann 2015, Yi and Chun 2015). ANI was first introduced to mimic the process of experimental DDH and thereby also called as digital version of DDH (Konstantinidis and Tiedje 2005, Goris *et al.* 2007). ANI values can be obtained using either BLASTn or mummer software (Richter and Rosselló-Móra 2009) and the former is much widely used for taxonomic purposes (Kim *et al.* 2014, Stropko *et al.* 2014, Rosselló-Móra and Amann 2015, Yi and Chun 2015). Recently, Li *et al*. (2015) suggested that mummer is not suitable for ANI calculation. Therefore, I use the term ANI for the technique based on BLASTn in this study.

ANI is calculated from two genome sequences (of the query and subject strains) as follows: First, the genome sequence of the query strain is divided into 1020 bp-long sequences (fragments). Second, each fragment is searched against the whole genome sequence of the subject strain using NCBI's BLASTn program (Altschul *et al.* 1997). In this process, the BLASTn program calculates nucleotide identity values between fragments of the query strain and the genome of the subject strain. Average nucleotide identity is the mean of these nucleotide identity values.

It has been known that reciprocal DDH values between two strains are often not the same, therefore not symmetrical, when DDH methods use labelled DNA (Johnson and Whitman 2007, Tindall *et al.* 2010). Since the theoretical concept of ANI derives from DDH, this may be also true for ANI. In other words, ANI of strain A (as query) to strain B (as subject) may be different from that of strain B (as query) to strain A (as subject). A reasonable practice would be to use the mean of two reciprocal ANI values, even though there is no theoretical basis for this, or for choosing either value. In this context, I investigate this problem and propose a new algorithm, called OrthoANI (Average Nucleotide Identity by Orthology), which can replace the original ANI.

2.2. Method

Dataset

A total of 14,745 genome sequences representing members of 10 genera (*Acinetobacter, Bacillus, Enterococcus, Escherichia, Mycobacterium, Pseudomonas, Salmonella, Staphylococcus, Streptococcus* and *Vibrio*) were selected from the EzBioCloud Genome database (Yoon *et al.* 2017) in which low quality and potentially contaminated genomes were checked and excluded. These genera were chosen as they contain the largest numbers of genomes.

Calculation of the original ANI values

Since calculating all possible pairs in our dataset was not computationally possible, I randomly selected genome pairs belonging to the same genus. The final dataset contained 63 690 genome pairs. For the ANI calculation, I used the previously described algorithm (Richter and Rosselló-Móra 2009) except that NCBI blastn+ was used instead of the legacy BLASTn package. The reciprocal ANI values were obtained for each of the genome pairs.

OrthoANI algorithm

The algorithmic schema to calculate OrthoANI between two genomes is given in Figure 1., which consists of three steps. First, both genome sequences were cut into consecutive 1020 bp-long fragments. Any fragments less than 1020 bp in size were omitted and ignored. Second, all fragments were searched and nucleotide identities were calculated using the BLASTn program. In this study, NCBI- blastn+ (version 2.2.30) was used with the following parameters: -task = blastn, -dust = no, -xdrop_gap = 150, -penalty = − 1, -reward = 1 and -evalue = 1.0e − 15; the rest of the parameters that could affect the result were set to default. Third, orthologous

fragments between two genomes were identified when they showed reciprocal best hit in BLASTn searches. Because BLASTn is based on local alignment, I chose local alignments (also called HSP) with at least 35 % of the total length of the fragment (i.e. 357 bp out of 1020 bp); this cut-off value is set to match the value of 70 % suggested by Goris et al. (2007) in which only one genome sequence is fragmented. In contrast, both genome sequences are fragmented for OrthoANI. Since nucleotide identities can be obtained reciprocally, these were averaged to give average nucleotide identity of an orthologous fragment pair. The genome-wide nucleotide identity value was finally calculated as the average of identity values among all orthologous fragment pairs between two genomes.

Statistical analysis

Statistical analysis was performed to investigate the correlation between the original ANI and OrthoANI values using the R package (https://www.r-project.org).

Implementation and availability

The OrthoANI algorithm is implemented in JAVA programming language and is provided as two different software types: OAT (Orthologous ANI Tool) is a graphical user interface program that can be used interactively on personal computer environments and provides the functionality of performing UPGMA clustering. OAT_cmd is a command-line program that can be integrated into the user's own bioinformatics pipeline. The software tool is freely available at (https://www.ezbiocloud.net/tools/orthoani).

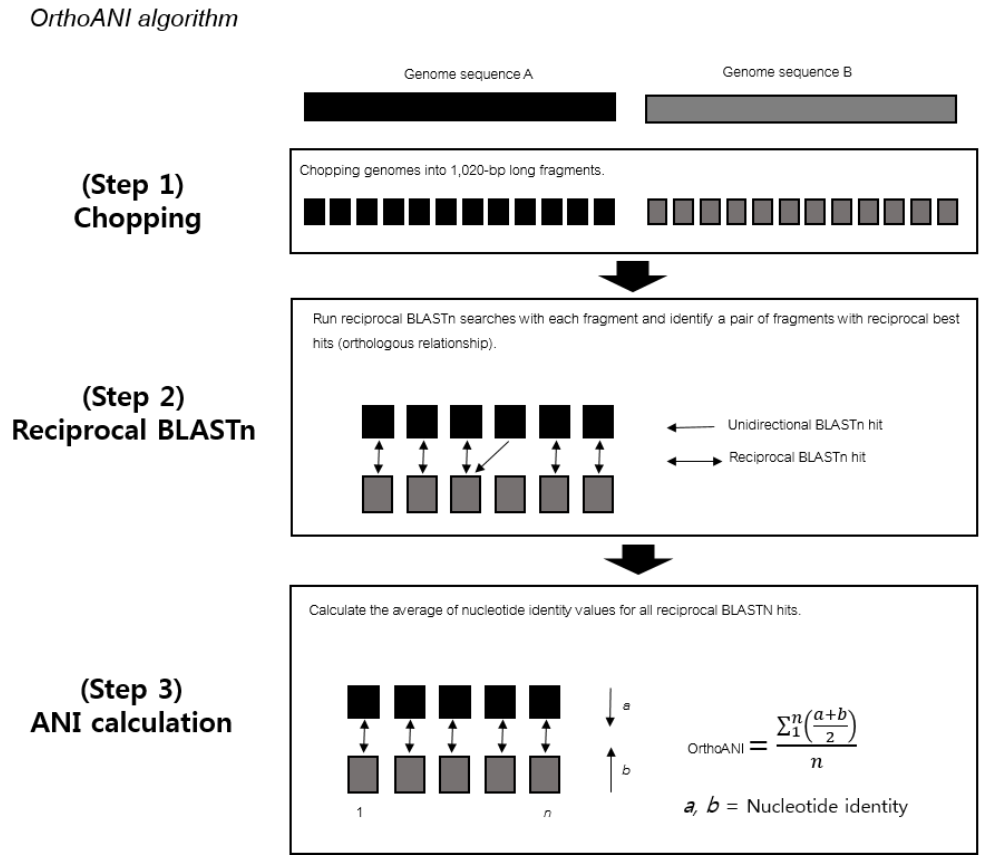

Figure 1. Schematic diagram for the OrthoANI algorithm.

Figure 1-legend. The major differences between ANI and OrthoANI are: (1) in OrthoANI, both genomes are fragmented in silico, (2) OrthoANI does not use fragments of less than 1,020 bp, and (3) in OrthoANI, only when two fragments are reciprocally searched as best hits using BLASTn program are their nucleotide identity values included in the subsequent computation.

2.3. Results and discussion

Like DDH methods based on labelled DNA, ANI is not symmetrical. Indeed, 55 % of 63,690 genome pairs examined in this study exhibited over 0.1 % discrepancy between reciprocal ANI values. Moreover, 1,101 pairs showed more than 1 % discrepancy with the highest being 4.15 % (Fig 2.). Given that approximately 95–96 % ANI values are considered as the species boundary (Goris *et al.* 2007, Richter and Rosselló-Móra 2009, Chun and Rainey 2014), this level of discrepancies is significant enough to affect subsequent taxonomic interpretation. I also obtained reciprocal nucleotide identities values using ANI calculator (http://enve-omics.ce.gatech.edu/ani/) and JSpecies (http://imedea.uib-csic.es/jspecies/) for 100 genome pairs. In general, all software tools do not provide exactly identical values, albeit they provide very similar values.

To resolve this problem, a new ANI algorithm was developed, named 'OrthoANI', to include the concept of orthology (Figure 1). Unlike the original ANI, reciprocal OrthoANI values are always identical because of its algorithmic nature. The correlation between the original ANI and OrthoANI is very high (R 2=0.9998 for whole range and $R^2 = 0.9995$ for > 90 % OrthoANI range; Fig 3.). OrthoANI values are slightly higher (approximately 0.1 %) than the original ANI values in the range of approximately 95–96 %.

The computing time required for calculating OrthoANI between two genomes is 1.3–4-fold less than reciprocal original ANI, when tested on a desktop personal computer (Table 1). The degree of speed-up depends on the number of threads, length of the contigs and the overall genome sizes. In general, more threads and longer contigs result in a higher speed-up while the overall size of the genome is inversely proportional to the speed-up. Therefore, OrthoANI should be better suited to large scale comparison studies.

Several early studies recommended ANI value of approximately 95–96 % as cut-off for species demarcation (Goris *et al.* 2007, Richter and Rosselló-Móra 2009). Since OrthoANI in this range is only slightly higher than original ANI, I also recommend a similar range of cut-offs. It is also worth noting that ANI and OrthoANI do not provide good measures for distantly related genomes (Kim *et al.* 2014). For example, they should not be used to compare genomes belonging to different genera.

In conclusion, a modified version of ANI is proposed, named OrthoANI, to solve the problem of reciprocal inconsistency of the original ANI algorithm. Moreover, this new measure of genomic relatedness correlates well with the original ANI and can be readily used for taxonomic purposes. Like original ANI, it does not require gene-finding and functional annotation processes, allowing simple, reproducible and

standardized procedures for taxonomic uses. With the easy-to-use GUI version and command-line version for large-scale computation, the algorithm should be accessible to all levels of microbiologists and students.

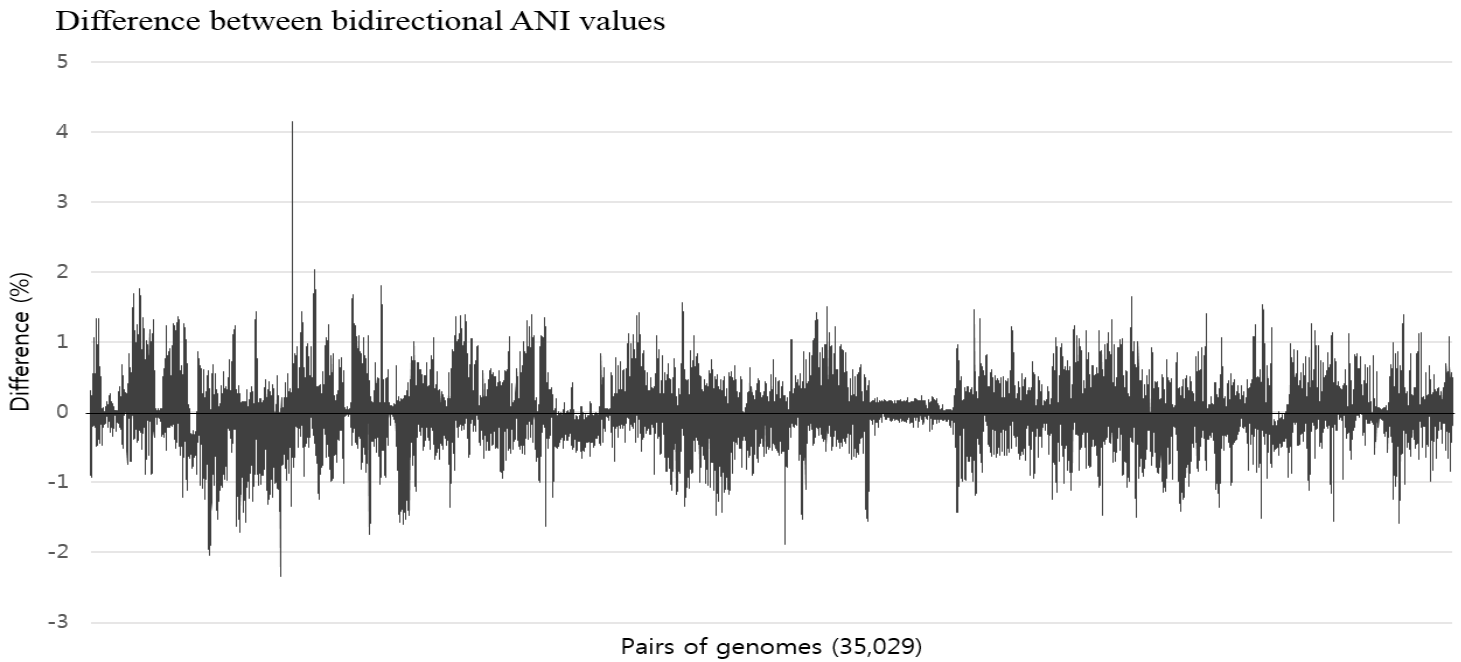

Figure 2. Differences between reciprocal ANI values on the basis of 63,690 pairs of genome sequences

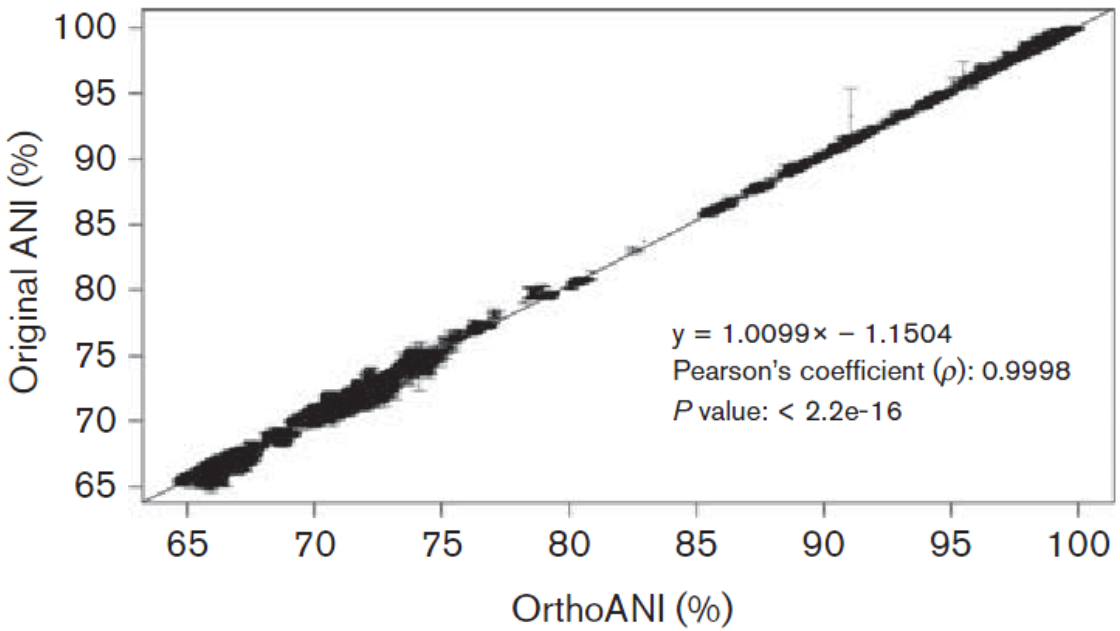

Figure 3. Correlation between original ANI and OrthoANI identities. For the original ANI both reciprocal values were plotted.

Table 1. Speed-up of OrthoANI algorithm over reciprocal ANI.

Strain 1	Genome Accession	Size (bp)	Strain 2	Genome Accession	Size (bp)	ANIb (%)	Time (sec)	OrthoANI (%)	Time (sec)	Speed-up
Candidatus *Hodgkinia cicadicola* TETULN	GCA_000699475.1	150,297	Candidatus *Hodgkinia cicadicola* Dsem	GCF_000021505.1	143,795	65.95	0.4	64.53	0.3	1.3
Mycoplasma genitalium G37	GCA_000027325.1	580,076	*Mycoplasma genitalium* M2321	GCA_000292405.1	579,977	99.44	10.8	99.45	4	2.7
Borrelia hermsii YBT	GCA_000568775.1	919,983	*Borrelia hermsii* YOR	GCA_000568675.1	919,292	96.41	31.7	96.45	9	3.5
Streptococcus pneumoniae SPN994039	GCA_000211055.2	2,026,505	*Streptococcus pneumoniae* SPN994038	GCA_000211035.2	2,026,239	100	48.8	100	13.8	3.5
Brucella suis bv. 2 Bs364CITA	GCA_000698325.1	3,328,972	*Brucella suis* bv. 2 Bs396CITA	GCA_000698345.1	3,328,458	99.98	61.4	99.99	28.8	2.1
Escherichia coli ER2796	GCA_000800215.1	4,558,663	*Escherichia coli* K-12 strain ER3413	GCA_000800765.1	4,558,660	100	105.1	100	34.2	3.1
Burkholderia thailandensis MSMB121	GCA_000385525.1	6,731,379	*Burkholderia thailandensis* str. 2003015869	GCA_000808035.2	6,728,980	93	3552.3	93.47	1477.4	2.4
Sorangium cellulosum So0157-2	GCA_000418325.1	14,782,125	*Sorangium cellulosum* So ce56	GCA_000067165.1	13,033,779	86.61	28481.9	88.14	7126.4	4

Chapter 3. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences

3.1. Introduction

In recent years, the cost and time of genome sequencing have been decreased dramatically thanks to the development of new DNA sequencing techniques, called next-generation sequencing (NGS). At present, the number of prokaryotic genome sequences in public databases reaches almost 70000. It has been suggested that the use of large-scale genome data greatly facilitates our knowledge and understanding of the microbial world (Ward and Fraser 2005, Chun and Rainey 2014). Also, its application to clinical microbiology should pave the way to the better diagnosis of infectious diseases (Pak and Kasarskis 2015).

As the use of NGS becomes more routine in microbiology, there is increasing concern regarding quality assurance of the sequence data generated, including contamination (Alkan *et al.* 2010, Longo *et al.* 2011, Gargis *et al.* 2012, Mukherjee *et al.* 2015). Contamination in DNA sequence data may result from either biological sources (cells) or DNA present in reagents or instruments. Because NGS produces much more raw data than the conventional Sanger method (>10-fold), there is more chance of contamination. This issue is of particular importance in clinical laboratories as contamination events can lead to false diagnostics. Development of a method to detect such cases as a quality control process is of primary importance in routine microbial genomics laboratories.

A few algorithms and software tools are available to detect contamination in draft genome assemblies. DeconSeq (Schmieder and Edwards 2011) requires a pre-built database of potential contaminants that is specialized to detect human DNA in genome or metagenome assemblies. ProDeGe (Tennessen *et al.* 2016) and CheckM (Parks *et al.* 2015) use the single-copy protein-coding genes that are highly conserved across the domains Bacteria and Archaea. These methods are useful in detecting possible contaminations in draft genome assemblies in public databases. However, in principle, they cannot differentiate contamination from lateral gene transfer, which often takes place in many bacterial species (Chun *et al.* 2009). In contrast to single-copy protein-coding genes, rRNA genes are present in multiple copies and are known to be less prone to horizontal gene transfer events (Kitahara *et al.* 2012). Here I propose a novel algorithm to detect possible biological contamination from prokaryotic genome assemblies using 16S rRNA gene sequences, which I have named

ContEst16S. The method developed here successfully identified potentially contaminated genome assemblies in public databases and proved to be useful in complementing the existing bioinformatics tools based on protein-coding genes.

3.2. Method

Algorithm

The overall scheme of the ContEst16S algorithm is provided in Fig. 4. The 16S rRNA gene fragments were extracted from genome sequences using the infernal software (Nawrocki and Eddy 2013), with the following parameters: cmsearch -g - -noali -E 1.0E-5. The data model used was Rfam 12.1 (Nawrocki *et al.* 2014). Only fragments of at least 500 bp were selected for subsequent analysis.

If one or no fragments are detected, the genome is classified as 'Undecided'. Otherwise, all possible pairwise sequence similarities are calculated among the extracted fragments using the algorithm of Myers and Miller (Myers and Miller 1988). If a pair of fragments is not aligned at all or by at least 400 bp, the calculation is ignored. If two fragments meet all of the following two conditions, the genome is classified as 'Contaminated': (i) two fragments differ by >5% in sequence similarity, and (ii) the best search hits of two fragments show >97% similarity to the known sequences in the EzBioCloud 16S rRNA database containing type strains and representatives of phylotypes (Yoon *et al.* 2017). They should also belong to different genera. Otherwise, it is considered 'Undecided'. I did not use the term 'Not Contaminated' for the cases of 'Undecided', as this may lead to the notion that the genome is free of contamination. It should be emphasized that our algorithm can only detect undoubtedly contaminated genome sequences, but cannot guarantee that it is free of contamination.

Implementation

The algorithm was implemented using the JAVA programming language (www.java.com) and MySQL database (www.mysql.com/) on a Linux operating system. Searching against the reference 16S rRNA database was carried out using the combination of blastn and pairwise sequence alignment (Yoon *et al.* 2017). A multiple sequence alignment was generated from all the extracted 16S rRNA gene fragments, their best hits to the reference database and selected representative sequences (*Escherichia coli* , *Bacillus cereus* , *Flavobacterium aquatile* , *Micrococcus luteus* , *Nostoc punctiforme*) using muscle software (Edgar 2004). Aligned nucleotide positions with >50% of bases (non-gaps) in the resultant multiple sequence alignment were then selected to generate a maximum-likelihood phylogenetic tree using the

RaxML software (Stamatakis 2006). A web-based service to detect contamination from a whole genome assembly (as FASTA format) using the ContEst16S algorithm is provided at www.ezbiocloud.net/tools/contest16s.

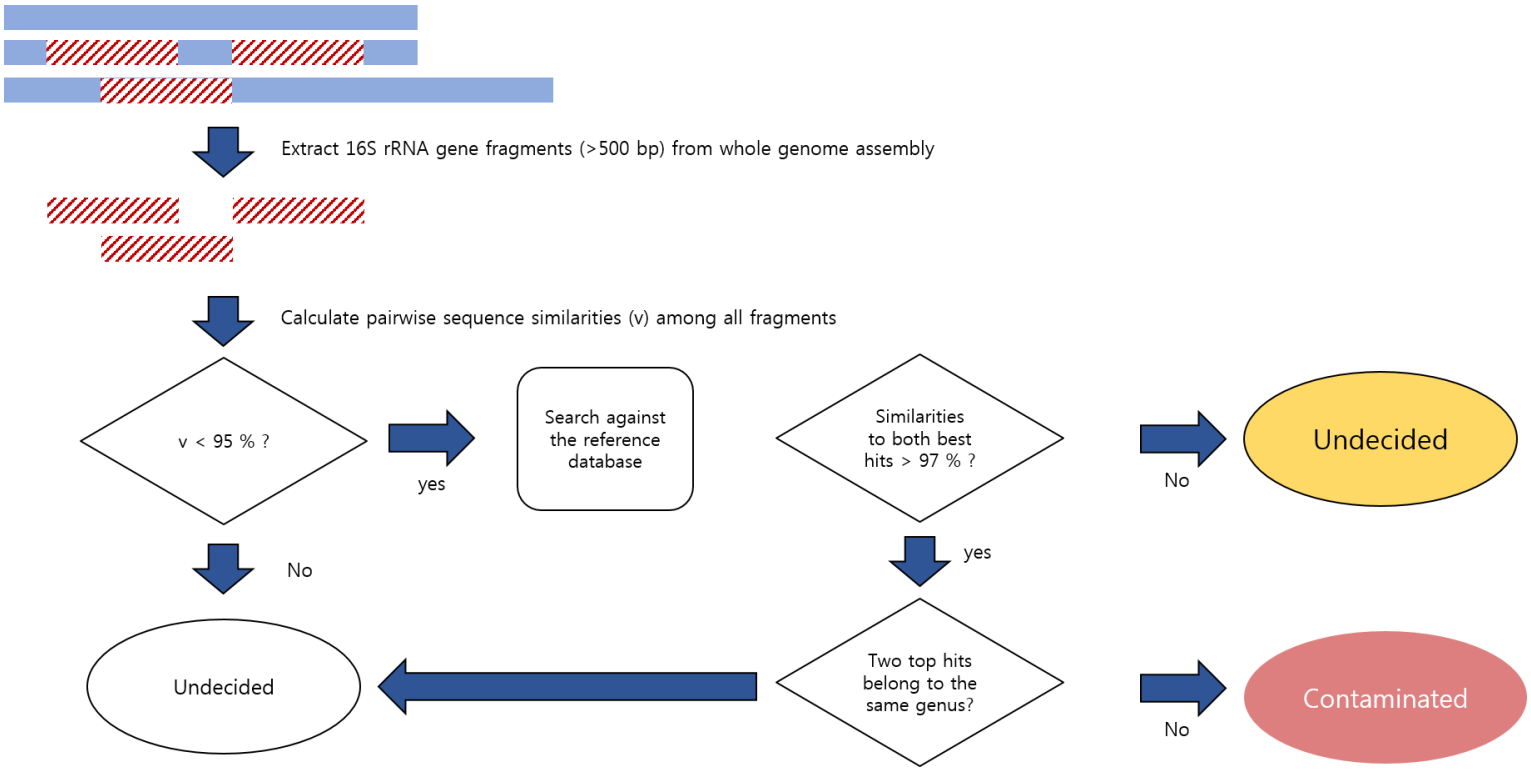

Figure 4. Schematic diagram of the ContEst16S algorithm.

3.3. Results and discussion

The algorithm was applied to the currently available entries of the NCBI Assembly Database (www.ncbi.nlm.nih.gov/assembly/), which is a primary public depository of prokaryotic genomes. The first step of the ContEst16S algorithm is to extract 16S rRNA gene fragments from whole genome sequences. Only fragments of >500 bases are considered as they are used for taxonomic identification in the next step. These extracted fragments do not necessarily represent operons or whole genes in draft whole genome assemblies.

Of 69,745 genomes, 44,933 contain a single 16S rRNA gene fragment, whereas 4285 contain no fragments. These cases are not considered further because the ContEst16S algorithm could not be applied. The remaining genomes contain two or more 16S rRNA gene fragments (>500 bp). *Streptococcus agalactiae* 18RS21 (NCBI assembly accession: GCF_000167715.1) contains the highest number of 16S rRNA gene fragments among the tested genomes. Since all the 62 fragments were matched to the type strain of *Streptococcus agalactiae* with >97% similarity, it was not predicted as 'Contaminated'.

Nucleotide sequence differences between a pair of the extracted 16S rRNA gene fragments in 1662 genomes were found to be 5% or higher, among which 1,068 did not have best hits against the reference 16S rRNA database with the 97% similarity cutoff. In these cases, I reasoned that (i) one or both fragments contain substantial sequencing errors, (ii) one or both fragments are pseudogenes or (iii) there is no similar entry in the reference 16S rRNA database that matched the extracted fragments.

A probable case of pseudogenes was found for the complete genome of *Borrelia afzelii* strain PKo (GCF_000222835.1), and is discussed further. Two 16S rRNA gene fragments, namely Fragment #1 (1536 bp) and Fragment #2 (1509 bp), were extracted from the *Borrelia afzelii* genome sequence which differed by 18.1%. In this case, I was able to rule out the possibility of high sequencing error as both fragments were also found in other complete genomes (*Borrelia afzelii* HLJ01; GCF_000304735.1 and *Borrelia afzelii* Tom3107; GCF_000741005.1). The presence of these two fragments in different strains can be only explained by vertical evolution, not sequencing errors. Fragment #2, a potential pseudogene, was found between two tRNA genes within an rRNA operon where Fragment #1 was also located. In the maximum-likelihood phylogenetic tree (Fig. 5-a), Fragment #2 shared the common ancestor with Fragment #1, with a high rate of mutations, which resulted in a very long branch; this phenomenon is a typical characteristic of recently duplicated pseudogenes.

Although it is rare, two 16S rRNA gene fragments on the same genome can have significantly different sequences while both remain functional (Mylvaganam and Dennis 1992, Yap *et al.* 1999). In our reference 16S rRNA database, all major sequence types of these known special cases are included to make sure that they are not predicted as 'Contaminated'. Also, I relaxed the criteria for calling for contamination by excluding the cases where two best hits belong to the same genus when compared against the reference 16S rRNA database.

Using our algorithm, 594 genomes (0.85% of the total analyzed genomes) out of 69,745 in the NCBI Assembly Database were predicted as 'Contaminated'. None of 5412 complete genomes were detected as 'Contaminated' by the ContEst16S algorithm. This is because sequencing reads derived from minor contaminants are probably ignored during the process of genome assembly.

A typical case of contaminated genome assemblies, namely *Acinetobacter baumannii* strain 45057_1 (GCF_000682075.1), is given in Fig. 5-b. It contains ten 16S rRNA gene fragments among which eight were correctly matched to *Acinetobacter baumannii* but the remaining two to *Enterococcus faecium* and *Escherichia coli* group with >99 % sequence similarity, respectively. This case can be only explained by contamination, rather than double events of lateral gene transfer.

CheckM (Parks *et al.* 2015) is a widely used software tool that can be used to determine if a genome assembly is contaminated. I used the CheckM tool to screen the 594 genomes that were already predicted as 'Contaminated' by ContEst16S. CheckM was not able to detect 42 genomes (7 %) that are clearly recognized as 'Contaminated' by ContEst16S. CheckM could not extract single-copy protein-coding genes from seven genomes.

The 16S rRNA gene fragments that are considered to have originated from potential contaminants were taxonomically identified against the quality-controlled reference database (Yoon *et al.* 2017). The frequencies of the biological contaminants in the 594 genomes are summarized in Table 2. The most frequent contaminant is the Bacillus cereus group in which *Bacillus thuringiensis* and *Bacillus anthracis* are also included as they are indistinguishable by 16S rRNA gene sequences. These organisms are commonly found in soil (Logan and De Vos 2009), implying that the cultures used in the DNA extraction step were probably contaminated. Contamination by human, Arabidopsis and soybean DNA probably occurred in the laboratories where DNA of these organisms is also handled for NGS library preparation or DNA sequencing (Longo *et al.* 2011). Identifying the sources of contamination in various steps of genome sequencing should provide a better way to prevent future contaminations, implying the utility of our algorithm for routine genomics facilities.

The web-based service of our algorithm is also available at www.ezbiocloud.net/tools/contest16s in which the results of the ContEst16S algorithm, as well as the phylogenetic tree, are provided upon upload of a FASTA format genome assembly.

There is no perfect way of detecting all contamination events from whole genome assemblies unless the genome sequence is completely determined. For example, a contamination event by a taxonomically closely related organism cannot be differentiated, with high confidence, from micro-sequence heterogeneity of rRNA operons, sequencing errors or the presence of pseudogenes. Furthermore, given that a bacterial genome represents a mosaic genetic composition and has great potential to obtain a gene or gene clusters from other organisms, it is difficult to differentiate contaminated DNA from recently transferred DNA. However, some genes, including rRNA genes, are known to be rarely mobile and can be used to detect clear cases of contamination during genome sequencing projects. Finally, it is noteworthy that the proposed algorithm is dependent on the taxonomic coverage of a reference 16S rRNA database as contaminants are confirmed only if they showed >97 % sequence similarity to reference sequences.

The algorithm presented here should provide a robust and efficient way of detecting possible biological contaminations, which is demonstrated by the identification of a significant number of contaminated genome assemblies in the NCBI Assembly Database. Along with the software tools based on single-copy protein-coding genes (Parks *et al.* 2015, Tennessen *et al.* 2016), ContEst16S will improve our means of quality assurance, which is of primary importance in laboratories for routine genomics, clinical microbiology and public health.

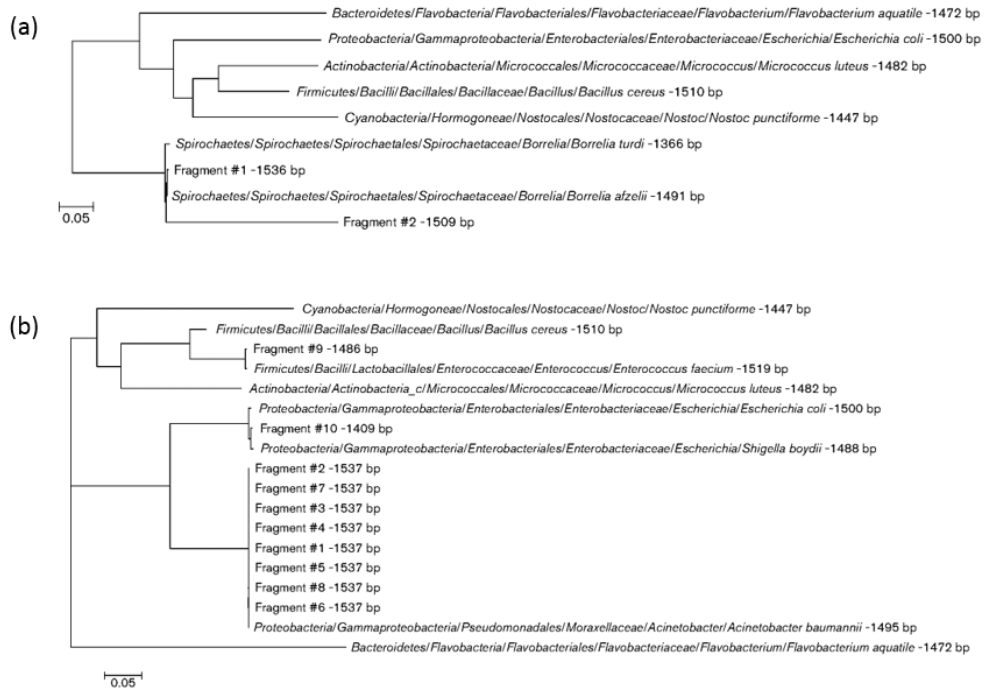

Figure 5. Maximum-likelihood phylogenetic tree of 16S rRNA gene fragments.

Figure 5-legend.

(a) Maximum-likelihood phylogenetic tree of two 16S rRNA gene fragments extracted from the *Borrelia afzelii* PKo genome (NCBI accession: GCF_000165595.2), one of which is a pseudogene. Bar, 0.05 changes per position.

(b) Maximum-likelihood phylogenetic tree of ten 16S rRNA gene fragments extracted from the *Acinetobacter baumannii* 45057_1 genome (Accession GCF_000682075.1), two of which were predicted as contaminants. Two probable sources of contaminations are strains belonging to *Enterococcus faecium* and *Escherichia coli* group that may include *Shigella* species. Bar, 0.05 changes per position.

Table 2. Top ten frequent contaminants.

Organisms	Count	Percentage (%)
Bacillus cereus group	140	13.61
Streptococcus pneumoniae	32	3.11
Escherichia coli group	26	2.53
Homo sapiens	24	2.33
Staphylococcus hominis	18	1.75
Staphylococcus aureus	14	1.36
Acinetobacter baumannii	13	1.26
Glycine max	13	1.26
Staphylococcus epidermidis	12	1.17
Bacillus subtilis	11	1.07
Others	726	70.55
Total	1029	100

Table2-legend. A total of 1029 contaminating organisms were identified from 594 whole genome assemblies that were predicted by the ContEst16S algorithm.

Chapter 4. Developing prediction tools for *Vibrio cholerae* phenotypes

4.1. Introduction

V. cholerae is a Gram-negative, spiral comma-shaped facultative pathogen, and is the etiologic agent of cholera a life-threatening diarrheal disease associated with the dehydration and hypovolemia (Farmer III *et al.* 2015). An estimated three million cases of diarrheal disease and approximately 100,000 deaths are caused by CT-producing strains of *V. cholerae* annually in regions where safe drinking water and sanitation are lacking (Zuckerman *et al.* 2007).

Despite some controversy as to whether an ancient disease was known as cholera, there is much evidence in ancient Asia and Europe about the existence of diseases, such as cholera at the time. However, relatively recently researchers have begun to analyze the cause of cholera. The first cholera epidemic of O1 type cholera was known to have occurred in India in 1817 and then appeared in Europe (Barua 1992). It caused many victims, but researchers did not know that *V. cholerae* was the causative agent of cholera. Several decades later, in 1884, a German microbiologist Robert Koch first reported that *Vibrio* was found in stools and the guts of patients with cholera living in Calcutta, India. That report first revealed that *V. cholerae* was the causative agent of cholera (Howard-Jones 1984), and the strain became the most widespread pathogen in the world. Since the first report in 1817, almost all continents have participated in seven global cholera epidemics (Faruque *et al.* 2003). The current pandemic is the seventh global pandemic that began in 1961 in the Philippines caused by the *V. cholerae* El Tor biotype as a new type from classic Asiatic cholera (Wallace *et al.* 1964). In 1968, this new type of cholera began in south-east Asia and traveled to western Asia, the USSR, and Mediterranean countries, including Egypt. After raging in Egypt severe diarrhea disease infiltrated the Africa continent. The disease breaking out in Guinea, Senegal, Mali, Sierra Leone, Ghana, Niger, Liberia, and Nigeria. About 6,300 people were killed by El Tor *Vibrio* during that period (Parnis 1971). In the early 1990s, the pandemic broke out again concurrently in South America and Southern Asia. In 1991, the life-threatening epidemic cholera caused by the O1 type had continued to spread to South America where tens of thousands of deaths occurred. However, the Asian cholera that emerged in Bengal during 1992-1993 was found to be a newly recognized strain called *V. cholerae* O139 (Ramamurthy *et al.* 1993). The O139 type affected at least ten countries in Southern Asia with symptoms including severe watery diarrhea and harsh dehydration (Tauxe and Barrett 1998). The two types of *V. cholerae* (O1 El Tor and O139) are causing

global epidemics with the current seventh cholera pandemic wave. The cholera epidemic of Haiti from 2010 to 2017 was the most large-scale epidemic of cholera. It was the most recent cholera epidemics that killed 4,672 people and hospitalized thousands more over ten years (Chowdhury *et al.* 2011).

4.1.1. O antigen serotypes

V. cholerae is classified serologically based on variations in the O antigen. O antigen specificity is determined by the O-specific polysaccharide, which is located on the external region of lipopolysaccharide (LPS). While more than 200 serogroups of *V. cholerae* have been identified, only serogroups O1 and O139 have been associated with cholera epidemics. *V. cholerae* strains that do not belong to serogroup O1 or O139 are frequently collectively referred to as non-O1/O139 *V. cholerae*. Non-toxigenic, non-O1/O139 *V. cholerae* may cause rare cases of gastroenteritis and sepsis, but does not cause cholera-like disease.

The *V. cholerae* O serogroups are detected by the O antigen on bacterial LPS. LPS is present in most Gram-negative bacteria and consist of three components: lipid A, an oligosaccharide core, and the O antigen at the surface of the outer membrane. The lipid A portion is composed of fatty acid and sugar chains that link the phospholipids in the outer membrane. The oligosaccharide core consists of sugars and sugar-derived substances. The O antigen is a polysaccharide that extends from the outer membrane and is formed of repeating oligosaccharide units. The O antigen has been found to be highly polymorphic, and even closely related species have few or no O antigen serogroups in common (Chatterjee and Chaudhuri 2003). O antigen genes involved in the biosynthesis of the polysaccharide are arranged in a cluster. These gene clusters are generally divided into three classes: genes for enzymes involved polysaccharide synthesis, genes for glycosyltransferases, and genes for polysaccharide processing. Because of the structural complexity of the gene cluster, identifying serotyping O antigen is not easy and tends to be inaccurate (Pengsuk *et al.* 2010).

Traditional biochemical methods used to detect the cholera agent in infected animals, the environment, and in clinical samples are very slow and error-prone. Moreover, identification of the serogroup of *V. cholerae* is implemented based on type-specific antisera analysis. This method requires many materials to cover over 200 serogroups of *V. cholerae*. The serogroup O1 strains are further classified into three serotypes of Ogawa, Inaba, and Hikojima by the serotype-specific antigens A, B, and C, respectively (Shimada and Sakazaki 1988). This means that three antigens are required for detecting only one serogroup. Current serotyping analyses of *V. cholerae* usually use standard polyvalent antisera produced by immunizing rabbits with heat-

killed bacteria. As the antigens are part of the outer-membrane LPS, the antigen-antibody reaction response works well within the host compared to other antigens. However, it remains challenging to produce mono-targeted specific antisera, and it is easy to lose specificity to a specific antigen. As polyvalent antisera analyses have low sensitivity, various immunoassays using monoclonal antibodies have been developed, such as enzyme-linked immunosorbent assays, to identify the serotype and the immunochromatography stripping test (Gustafsson and Holme 1985, Pengsuk *et al.* 2010). However, the labor-intensive and time-consuming problems of identification still remain.

V. cholerae strains are known to have more than 200 antigen groups (Bernardy *et al.* 2016). Most all of the serotypes are non-pathogenic strains and only two types, O1 and O139, are pandemic strains that cause cholera outbreaks. *V. cholerae* O1 strains classified two biotypes: 'El Tor' and 'Classical.' One of the strains causing the seventh pandemic is *V. cholerae* biovar. El Tor named after Tur Sinai (in Romanian Al-Tur) Egypt where the strain was first isolated (Aydanian *et al.* 2011). While El Tor was named after the first region from which it was isolated, the other serogroup got the name "Classical" because the type was first identified as the causative agent of cholera (Cvjetanovic and Barua 1972). The Classical biotype was regarded as being responsible for cholera from the second pandemic to the sixth pandemic. No strains are available to study the fifth pandemic from the first pandemics wave, except *V. cholerae* PA1849, which was the second cholera pandemic strain isolated in 2014 from the preserved intestine of a victim of the 1849 cholera outbreak in Philadelphia PA, USA. As strain PA1849 was revealed as a Classical strain by phylogenetic analysis, the Classical strain was supposed to be the causative agent for the sixth pandemic (Hu *et al.* 2016). However, only O1 El Tor and O139 are pandemic agents of the continuous seventh wave. Since the O139 serogroup was the first non-O1 serogroup associated with the causative pathogen of the epidemics, this serogroup has attracted worldwide attention (Faruque *et al.* 2003).

V. cholerae O1 is further divided by sub-serotype into 'Ogawa,' 'Inaba,' and 'Hikojima.' The two dominant serotypes are *V. cholerae* O1 Ogawa and *V. cholerae* O1 Inaba. Hikojima has been suggested to represent strains that undergo a high frequency of conversion (Sakazaki and TAMURA 1971, Stroeher *et al.* 1992). The Ogawa and Inaba serotypes differ by the addition of a single 2-O-methyl group in the non-reducing terminal saccharide of the Ogawa-specific polysaccharide. There are many reports of interconversion between the serotypes (Sack and Miller 1969, Stroeher *et al.* 1992, Ito *et al.* 1993, Colwell *et al.* 1995). The switch from Ogawa to Inaba arises from mutations in WbeT methyltransferase. A strain that possesses the wild-type *wbeT* gene is the Ogawa subtype, whereas a strain with a gene expressing a malfunctioning WbeT methyltransferase is the Inaba type.

O antigens are key toxic agents and are targets of the immune system. Because these antigens play a particular role in the host-pathogen interaction, understanding how the O antigen gene clusters facilitates the development of reliable and rapid molecular diagnostic platforms that can replace conventional serotyping. Thanks to current biotechnology, sequencing techniques facilitate analyzing microbial features, such as bacteria serotyping, and are rapid, accurate and reliable.

4.1.2. Cholera Toxin

The CT is the primary virulence agent in pathogenic strains of *V. cholerae*, and the principal cause of watery diarrhea in patients with cholera. The CT is an ADP-ribosylating toxin that increases cAMP and chloride secretion by the apical cystic fibrosis transmembrane conductance regulator. The CT is encoded by *ctxAB* and is classified in the super-family of AB toxins. The CT consists of two oligo-dimeric subunits, such as an enzymatic A subunit and a receptor binding B subunit. The A subunit is 27.2 kDa and consists of an enzymatic chain and alpha helix site for linkage, and the 11.6 kDa homo-pentameric B subunits interact with the host as binding-receptors. These two distinct parts participate to activating site-specific recombination carrying CTX elements into the chromosomes of *V. cholerae* strains infected by CTXϕ (Pearson *et al.* 1993).

While CT was discovered in 1951 (De *et al.* 1951), CTXϕ was first identified in *V. cholerae* in 1996 (Waldor *et al.* 1996). CTXϕ is a filamentous, lysogenic *V. cholerae*-specific bacteriophage that converts non-toxigenic strains to toxigenic strains and harbors the cholera toxin gene (*ctxAB*) (Waldor *et al.* 1996). The CTX contains three types of toxins. Two open reading frames (ORFs) (*ctxAB*) encode the A and B subunits of CT. Toxigenic *V. cholerae* also make a putative toxin known as zonula occludens toxin (Zot), which grows the permeability of the small intestinal mucosa by changing the composition of the intercellular tight junctions (Baudry *et al.* 1992). A third toxin is called the accessory cholera enterotoxin (Ace), which induces fluid accumulation in rabbit ligated ileal loops (Trucksis *et al.* 1993).

The CTXϕ genome is composed of a CTX- core and one or more copies of a repetitive sequence called the RS elements. The CTX-core encodes toxins that contain *ace*, *zot*, *ctxAB*, a core-encoded pilin (*cep*), and *rstRAB*, formerly known as the RS2 element. The *ctxAB* genes encode the A and B subunits of the CT, and the rest of the core region is involved in transcriptional regulation (*rstR*), replication (*rstA*), integration (*rstB*), packaging and secretion (*cep*, *orfU* (*pIII*), *ace,* and *zot*) of phage DNA. However, the RS1 element contains three core genes (*rstRAB*) and one peculiar gene, *rstC* which encodes an anti-repressor that facilitates CTXϕ gene expression (Waldor *et al.* 1996,

McLeod *et al.* 2005).

Under the proper conditions, toxigenic *V. cholerae* strains can be induced to produce extra-cellular CTX phage particles (Waldor *et al.* 1996, Faruque *et al.* 1998). Phages are propagated at a specific site on the chromosome of recipient *V. cholerae* strains, forming stable lysogens or a replicative form of the phage DNA in the extra-chromosome (Waldor *et al.* 1996). As the bacteriophage uses TCP as a receptor, the expression of TCP by the strains is a prerequisite for susceptibility to the phage. TCPs are found in human bacteria, and the bacteria also serve as a receptor for the CTX phage, revealing coevolution of genetic elements that mediate the transfer of infected pathogenic bacterial species and toxic genes (Faruque *et al.* 1998).

Changes to the CTX properties can be classified into two main features: (i) Variant CTX phages generated by recombination of CTX (CTX^{cla} and CTX^{ElTor}), RS1 (repetitive sequences of the CTX satellite phage), and a *ctxB* point mutation. (ii) Replacing CTX phages with *V. cholerae* itself (Kim *et al.* 2015). Because the mechanisms for these characteristic changes are relatively simple, the population of *V. cholerae* carrying the CTX phage is continuously changing.

The classical biotype *V. cholerae* is the causative agent of the sixth cholera pandemic that was prevalent until the 1960s. However, after the emergence of the new El Tor biotype strains in 1961, the classical strains declined during the next 30 years. The last reported classical biotype was *V. cholerae* str. 95412 (Mexico, 1987) O1 Classical Inaba (Choi *et al.* 2016) isolated in 1997. The composition of CTX within two biotype strains differs in the CTX phage array by containing several single nucleotide polymorphisms (SNPs) in the genes.

CTX phages are mainly classified by *rstR* and *ctxB* genotypes and further subdivided by SNPs throughout the phage genomes. Depending on the variant of the CTX phage, there are 11 types of CTX phage (CTX^{cla}, $CTX^{US\ Gulf}$, CTX^{AUS}, CTX-1, CTX-2, CTX-3, CTX-3b, CTX-4, CTX-5, CTX-6, and CTX-6b). The phage genome which harbors classical biotype strains and the phage genome of El Tor strains has similar genomic structure and sequences. The classical strains contain CTX^{cla}, whereas the El Tor strains have $CTX^{El\ Tor}$ (CTX-1 − 6b). Furthermore, within the El Tor types, the CTX types are separated into several subtypes by each gene and by their arrays. Because the combination and structures of the ORFs within CTX phages vary from host to host, the CTX types are also classified as the Wave1, Wave2, and Wave3 types following their host.

The El Tor type strains, the El Tor types are classified into three epidemic waves based on the comparative genomics analysis of SNPs in the genomes of El Tor type strains (Mutreja *et al.* 2011). The prototype El Tor strains that contain the CTX-1 type

are designated Wave 1 strains of which a representative strain is N16961 (Bangladesh, 1975). Wave 1 strains are defined as El Tor strains that produce El Tor-specific CT encoded by *ctxB3*. Wave 1 strains were prevalent in West Asia and India until early 1996 (Lee *et al.* 2009). Wave 2 strains are designated as El Tor strains as they have a tandem CTX-2 repeat which harbors *ctxB1* on chromosome II, and various arrays including TLC, RS1, and CTX are on chromosome I. Wave 2 strains were prevalent in East Asia and African countries in the mid-1990s. Wave 3 strains are defined as El Tor strains harboring an atypical CTX including CTX-3, 3b, 4, 5, 6 and 6b. The atypical CTX are on chromosome I, and no CTX elements are found on the small chromosome. Wave 3 strains are similar to Wave 1 strains except they possess *ctxB1* or *ctxB7* instead of *ctxB3* of Wave 1. Wave 3 strains were also prevalent in India and Haiti until the early 2000s.

Comparative genomics studies over the past ten years have revealed the dynamics of *V. cholerae* which harbors CTX phages, and well organize the types of CTX phages (Waldor *et al.* 1996, Faruque *et al.* 1998, Chun *et al.* 2009, Lee *et al.* 2009, Cho *et al.* 2010, Kim *et al.* 2014, Kim *et al.* 2014, Kim *et al.* 2015, Choi *et al.* 2016). The massive data set of CTX and *V. cholerae* which harbor CTX phage elements are enriched our knowledge of the variation in strains and the evolutionary history of cholera.

4.1.3. Antibiotic resistance of the *V. cholerae* strains

Patterns of antibiotic resistance in *V. cholerae* have varied over time and by area, mainly due to the rapid acquisition of antibiotic resistance phenotypes through horizontal gene transfer of mobile elements that shifted among *Vibrio* species or other Gram-negative organisms. Self-transmissible mobile elements are critical for *V. cholerae* antibiotic resistance.

In 1979, the *V. cholerae* strains isolated from a Bangladesh outbreak showed that 16.7 % of the isolates had resistance to the five antibiotics, including ampicillin, kanamycin, streptomycin, tetracycline, and trimethoprim-sulfamethoxazole, and 10 % of the isolates were resistant to tetracycline and any four of these antibiotics (Glass *et al.* 1980, Glass *et al.* 1983). The epidemiological evaluation suggested that the onset of *V. cholerae* O1 was initiated by the introduction of a single multidrug-resistant strain (Glass *et al.* 1983). In 1986, a study of *V. cholerae* drug resistance patterns and screening of patients with cholera in Dhaka showed that none of these isolates was resistant to amoxicillin, chloramphenicol, nalidixic acid, streptomycin, and tetracycline (Nakasone *et al.* 1987). However, in late the 1980s, a study on the Bangladesh outbreak showed that nearly all Classical isolates were resistant to

tetracycline whereas the El Tor biotype strains were sensitive to tetracycline (Siddique *et al.* 1989). In 1991, tetracycline resistant El Tor strains re-emerged in the Bangladesh epidemic, and 70 % of the isolates were resistant to tetracycline, as well as other antibiotics (Siddique *et al.* 1992). In 1995, a Southern Indian epidemic study showed that several *V. cholerae* O1 strains were resistant to nalidixic acid which is a class of nitro-quinolone antibiotics (Jesudason and Saaya 1997). In the mid-1990s in the African epidemic study showed that all isolates from Tanzania and Rwanda were resistant to tetracycline, 80 − 100 % of *V. cholerae* O1 strains in Kenya and South Sudan and 65 − 90 % of isolates in Somalia were susceptible (Materu *et al.* 1997). The percentage of isolates from Somalia and Kenya resistant to chloramphenicol and co-trimoxazole, which is sulfone antibiotic remarkably increased from 15 % in 1994 to more than 90 % in 1996 (Materu *et al.* 1997). The O139 serogroup of *V. cholerae* newly emerged in 1992−1993. The new serogroup strains were more sensitive to ampicillin and tetracycline than the O1 strains (Albert *et al.* 1993, Sciortino *et al.* 1996).

Waldor et al. reported that the presence of self-transmissible transposon-like (SXT) element encoded resistance to sulfamethoxazole, streptomycin, and trimethoprim in *V. cholerae* O139 (Waldor *et al.* 1996). The SXT element of the O139 strains could be transferred to *V. cholerae* O1 in a *recA*-independent integrating manner into a recipient chromosome in a site-specific region (Waldor *et al.* 1996). A comparison study showed that O139 strains have increased susceptibility to chloramphenicol and streptomycin, but are increasingly resistant to ampicillin and neomycin (Mukhopadhyay *et al.* 1998). A study of the SXT element harbored in *V. cholerae* El Tor isolates in China reported that inducing the elements in the host strain lead to tetracycline resistance (Wang *et al.* 2016).

V. cholerae strains show a rapidly changing pattern of antibiotic resistance suggesting that mobile genetic elements are encoding antibiotic resistance in *V. cholerae* (Faruque *et al.* 1998).

Although cholera is a significant pandemic contagious diarrheal disease with a complex genetic history, there is no bioinformatics tool for cholera research. Conventional biochemical methods used to detect the cholera agent are labor intensive and require too much time. Three tools for detecting *V. cholerae* were developed in this study, including the O serogroup prediction tool, prediction of CTX phage elements within a genome of *V. cholerae*, and prediction of antibiotic resistance. These tools generate faster, more reliable and more precise results for researchers of cholera-related studies.

4.2. Methods

4.2.1. O Antigen serotyping

Genome data collection: Total 800 *V. cholerae* genome data are used for this study retrieved from EzBioCloud (Yoon *et al.* 2017) and are classified with OrthoANI (Lee *et al.* 2016) by the taxonomical approach. Within 800 genome data, and we select 796 genome data (Appendix-Table 1) with filtering by genome quality using ContEst16S (Lee *et al.* 2017) and CheckM (Parks *et al.* 2015).

Algorithm: Firstly the program extracts O antigen gene cluster within all genome data set assigned as *V. cholerae*, then gathers all CDSs within the cluster. Then, all CDSs from gathered O antigen related sequences are clustered with optimal options (90 – 95 % id values are applied depending on the variation of each CDS). After clustering, representative CDS sequences are picked up from each cluster. With representative CDS sequences, binary data (CDS present:1, absent:0) are generated from all genome data by searching representative sequences on all *V. cholerae* genome sequences. With the comparison of all binary data (representative - all genome data), the representative genomes are sorted by the same Jaccard-Index. If there is a new query genome data that need to be determined its O serotypes, by comparing query genome to representative CDS sequence data and representative Jaccard-Index profile, the program determines O antigen type of the query genome. The predicted serogroups results show the name of serogroups and the name of genes contained in the genome data. The gene names of predicted O antigen gene cluster shown by BPGN (Bacterial Polysaccharide Gene Nomenclature) gene naming (Reeves *et al.* 1996) or standard gene name instead of various ORF names.

Extraction of full O antigen gene cluster: For serogroup prediction, the extraction process of full O antigen gene clusters for comparing to other serogroups is needed. For searching O antigen gene cluster, *gmhD* (synonyms: *rfaD, hldD, waaD, nbsB, htmM, ECK3609, b3619, JW3594*) of *V. cholerae* str. N16961 is used as start site of O antigen gene cluster (Bik *et al.* 1996), and *ysh1* gene of *V. cholerae* known as *rjg* (right junction gene of O antigen gene cluster) is selected as end sequence of O antigen gene cluster (Sozhamannan *et al.* 1999). Using USEARCH (ver. 8) tool (Edgar 2010), the program extracts start position of O antigen gene cluster and end position within genome sequence and then makes the temporary file of extracted O antigen gene cluster sequences.

CDS assign within O antigen gene cluster: For prediction of a protein-coding gene in the gene cluster, dynamic programming of prokaryotic gene finding algorithm as Prodigal v2.6.3 (Hyatt *et al.* 2010) is used. For getting a gene name of predicted CDSs, UniProt (Apweiler *et al.* 2004), KEGG (Kanehisa and Goto 2000), and NCBI-nr

database (Pruitt *et al.* 2005) were used. The gene names were assigned official gene name by a nomenclature committee (Povey *et al.* 2001). However, numerous genes had no specific name because most research does not need to report all of their gene features. In such cases, gene names derived from an orthologous gene were used in a public database.

Making representative genes: In this study, total 13,561 CDSs were obtained from 538 full O antigen gene clusters extracted from the genome data set. It meant that only 538 of *V. cholerae* genomes had full O antigen gene clusters among the 796 genomes. (It is not because the *V. cholerae* genomes do not have O antigen, but because the sequencing process could not interpret the nucleotide sequences of O antigen cluster location). Then the genes were clustered with 0.90 minimum identity with USEARCH v8.0.1517 option. Because the two genes determining the location of O antigen gene cluster were covering all genome dataset with 90% minimum identity, the value of 90 also applied to clustering option. In the result of clustering, total 745 clusters were obtained, and then 745 of representative CDS sequences were picked up from each cluster. The representative sequences were determined by the criteria that were long enough to represent each cluster and had the least ambiguous sequences (Figure 7-b.).

Prediction of O antigen serotypes: To predict serotypes, a genome is firstly processed sequence-searching by USEARCH program with two query gene sequences, *gmhD* as start and *rjg* as end position genes. This step determines the full O antigen gene cluster within the query genome sequence. If the gene cluster is extracted successfully, the program finds out whether all of the representative genes are in the target query cluster or not. If specific representative sequence is in the cluster, the program records 1, otherwise records 0 with the following cut-off: > 90 % identity, > 90 % length coverage, < 1e-5 e-value and > 500 bit-score (e-value and bit-score calculated using Karlin-Altschul statistics (Karlin and Altschul 1990)). From the produced binary data, the program compares the data with the pre-calculated set of binary data. After that, the program predicts the nearest serogroup showing the highest Jaccard-Index (Jaccard 1912) with calculated reference-set. The predicted result shows the name of serogroup and the name of genes that are contained in the genome sequence data.

Prediction of O1 Ogawa, O1 Inaba, and O1 Hikojima: For predicting O1 serogroup to sub-serogroups (Ogawa, Inaba, and Hikojima), the representative sequences of *wbeT* (*rfbT*) were gathered from NCBI-nucleotide database or were directly extracted from genome data of specific strains. Total 36 distinctive sequences of serogroups that include two Classical Ogawa (O395, and M29), two Classical Inaba (A60, A68), three El Tor Ogawa (M66-2, MG116226, and A152), 26 El Tor

Inaba, and 1 Hikojima (FJ619106.1) are selected as reference sequences. With 36 representative *wbeT* sequences, query genome which predicted as O1 is classified to Ogawa or Inaba or Hikojima using USEARCH tool as the followed option: > 99 % identity, > 95 % length coverage, < 1e-10 e-value.

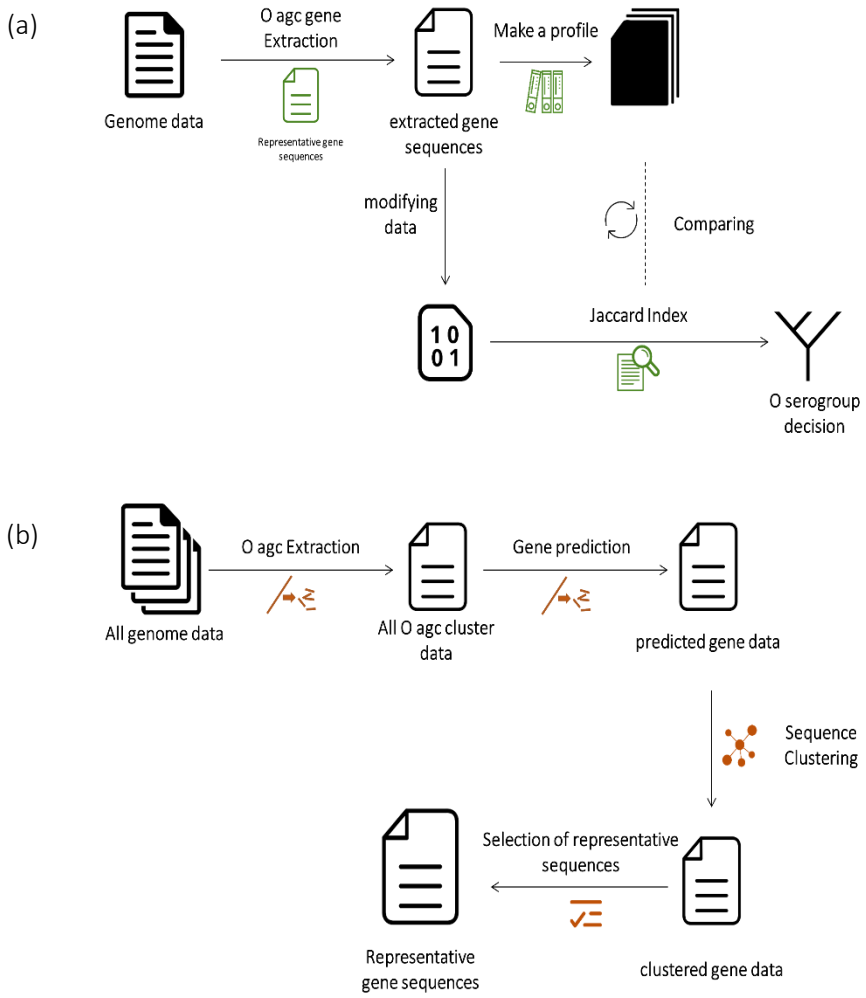

Figure 6. Algorithm of O antigen serotyping and process of making representative sequences**.**

Figure 6-legend. (a) Algorithm of O antigen serotyping. (b) Process of representative genes collection.

4.1.2. Prediction of Cholera Toxin genes

Algorithm: The program for prediction of CT extracts CTXϕ elements from query genome data by searching with representative genes. The extraction process of the CT elements from *V. cholerae* genome uses USEARCH (ver. 8) program as followed options: -ublast, -evalue 1e-5, -strand both, -id 0.35 –blast6out. And then, the program shows the gene name (*cep*, *orfU (pIII)*, *ace*, *zot*, *ctxA*, *ctxB*, *rstR*, *rstA*, *rstB* and *rstC*) types of biotypes of host strains, the location of the ORFs, and located chromosome of the ORFs. The program developed in this study, also shows the presence of TCP (toxin-coregulated pili) gene, and TLC region flanked by CTX elements.

Selection of representative genes: The representative genes were made by manually with references (Kim *et al.* 2014, Kim *et al.* 2014, Kim *et al.* 2015). The representative sequences cover most of CTX phage elements that include RS1, CTX^{cla}, CTX-1, CTX-2, CTX-3, CTX-3b, CTX-4, CTX-5, CTX-6, CTX-6b, CTX^{AUS}, $CTX^{US\ Gulf}$, and CTX^{O139} (Kim *et al.* 2015). X64098.1 (accession number) was used for reference sequence of TCP region, and *VC1466*, *VC1467*, *VC1468*, *VC1469*, and *VC1470* of str. N16961 (NC_002505.1) were used for reference sequence of TLC elements (Hassan *et al.* 2010). For prediction of CT genes, one reference sequence file which contains all CTX elements of *V. cholerae* N16961 strain, TLC sequence, and TCP sequences is used for gene extraction at first step and ten sub-sequence files (*ace*, *cep*, *orfU*, *zot*, *ctxA*, *ctxB*, *rstR*, *rstA*, *rstB*, and *rstC*) for the downstream classifying process were used.

Prediction for chromosome location of genome contigs: To designate chromosome as large-chromosome (chromosome I) or small-chromosome (chromosome II) for non-assigned contigs, the genome of str. N16961 was used as a reference genome. By searching the location of contigs in the large chromosome or small chromosome of reference genome using USEARCH tool (Edgar 2010), the location of contigs was predicted as Chromosome_I or Chromosome_II, or NA (Not Available).

Detection of TCP: For detecting presence of TCP, the sequences (X64098.1) of *V. cholerae* Z17561 were used that contains *tcpA*, *tcpB*, *tcpC*, *tcpD*, *tcpE*, *tcpF*, *tcpH*, *tcpI*, *tcpJ*, *tcpP*, *tcpQ*, *tcpR*, *tcpS*, *tcpT*, and *toxT* using USEARCH tool (Edgar 2010).

4.1.3. Prediction of antibiotic resistance

Method: For predicting antimicrobial resistance of *V. cholerae*, the RGI (Resistance Gene Identifier) program of the CARD database was used (Jia *et al.* 2016). The RGI program uses a manually curated data based on the molecular experiments, and automatically predicts antimicrobial resistance traits for the query genome using

DIAMOND (Buchfink *et al.* 2014) and Prodigal programs (Hyatt *et al.* 2010).

If the query genome predicted as 'PERFECT' or 'STRICT' by RGI program, the program developed in this study predicts as resistance. The 'PERFECT' sign means prefect-matching to the curated reference sequences and mutation in CARD database, and the 'STRICT' means functionally similar with known AMR (anti-microbial resistance) genes with curated similarity cut-offs by CARD database. From the results of RGI, the program developed in this study shows filtered information that includes the location of resistance-related gene, criteria of RGI, drug class, and resistance mechanism.

SXT elements analysis: To see the relation between antibiotic resistance of *V. cholerae* predicted by RGI and drug resistance derived by integration of SXT elements to the strains, six genes within the SXT gene cluster; *floR* (AY034138.1: 12159..13373), *strA* (AY034138.1: 15165..15968), *strB* (AY034138.1: 14329..15165), *sulII* (AY034138.1: 16029..16844), *tetR* (KT151664.1: 80500..81102), and *tetA* (KT151664.1: 81183..82385) were used. Because SXT elements are reported that integration of SXT induces antibiotic resistance to host bacteria, the correlation study between specific antibiotic resistance predicted by RGI and integration of specific gene in SXT elements was analyzed. Using USEARCH tools, the presence of the six genes were analyzed. The *floR* was used for detecting chloramphenicol resistance induced by SXT gene cluster, and *strAB* was used as the indicator of streptomycin resistance induced by SXT gene elements, and *sulII* was used for identifying SXT cluster inducing sulfonamide resistance, and *tetAR* was used to predict resistance to tetracycline by SXT elements.

To test interrater reliability between results of RGI and results of SXT elements prediction, the Cohen's kappa (κ) coefficient was used.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

The Cohen's kappa ranges from −1 to +1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents a perfect agreement between the raters (Marston 2010, McHugh 2012).

4.3. Results and Discussion

4.3.1. Prediction of O Antigen serotypes

Total 745 representative genes were extracted from 13,561 CDSs in 538 different genomes that have full O antigen gene cluster. As a result of analysis by searching with those representative gene sequences, species of *V. cholerae* had genes within O antigen gene cluster ranging from 12 to 41. In the disease-causing serogroups, the number of genes of O1 serogroup is 14 to 22, and O139 strains have 25 to 26 with entirely different structures of cluster between O1 and O139 (Figure 7).

O1, O139 Serogroup: Because most of the genome data available were O1 serogroups, the most of the results were assigned as O1 type (Table 3). Most of the O1 groups had almost the same structures of gene composition. The dominant type of O1 serogroup had 19 ORFs within O antigen gene cluster (about 24 kbp-25 kbp). However, some minor variations were found at the position from 12th (*wbeO*) to 15th (*wbeT*) ORFs. The position of variants contained transposon elements (*dde_yhhI*: *yhhI* containing transposase DDE domain), and transposase IS family (*insO*, *insN*). It can explain the reason of truncated *wbeT* gene in Inaba serotypes.

However, there were some deletions of ORFs at O antigen gene cluster in some case. The deleted ORFs were not because they were excluded from their original location, but because the genes just had a low identity with reference sequences. Moreover, the 'lost' sequences could not be identified as an ORFs by searching to a public database. It is possible that the 'lost' sequences are point mutation variant or in/del mutation by an unknown source such as mobile elements.

The number of genes of O1 serogroup was 14-22, but most of the genomes of O1 serogroup (> 96 %) have 19-20 boundary. The dominant type of O1 serogroup contained 19 ORFs (95.3 %: 428 out of 449), and second largest groups had 20 ORFs (< 1 %: 4 out of 449), and the number of other variants of O1 serogroups was 1-2. The largest group of O1 serogroup contained El Tor, Classical, El Tor Ogawa, El Tor Inaba, and just O1 serogroups.

Total 171 genome data of *V. cholerae* strains which do not have their property were newly predicted as O1, and three O1 strains (str. 2012HC-25 2012 Haiti; GCA_000788775.1, str. 87395 1983 Mexico; GCA_000348085.2, str. EM-1676A 2011 Bangladesh; GCA_000348345.2) were newly predicted as non-O1/O139 serotypes.

There were two types of O139 strains, but most of the serotypes had the same

structure with the structure of O antigen gene cluster of *V. cholerae* MO10 (India, 1992). Sixteen out of seventeen (94 %) were grouped with MO10, and one variant was *V. cholerae* A1330 (India, 1993) with one ORF different to major type.

Assuming that the genomic backbones of O139 are highly related with genomes of the 7th pandemic group, O1 antigen coding gene cluster can be readily transferable between environmental and clinical clones (Cho *et al.* 2010). In other words, the O antigen gene cluster can be mobile within *V. cholerae* species.

Ogawa, Inaba, and Hikojima serogroups: Representative sequences of *wbeT* gene contain 36 types. The *wbeT* genes of Inaba strains possess total six types of mutation; frameshift (deletion) - 6, frameshift (insertion) - 6, missense mutation (non-synonymous single amino acid change) - 3, nonsense mutation - 7, synonymous mutation - 1, and truncated ORF - 6.

With the program developed in this study, total 315 genomes are newly predicted as Ogawa or Inaba from 444 O1 strains. 273 non-assigned genomes got new serotype names as Ogawa or Inaba. Among the 273 genomes, most of the strains were designated as El Tor Ogawa (85 %), and rest were two Classical Ogawa, 39 El Tor Inaba, and 1 unknown type (Figure 8-a). In the strains predicted as Inaba which harboring mutations in *wbeT*, all six types of mutation were detected; 30 % of frame shift-del, 12 % of frame shift-ins, 26 % of missense, 21 % of nonsense, 3 % of synonymous, and 8 % of truncated ORF (Figure 8-b). Among 171 of re-predicted strains, 75 % of strains were predicted as same types of original ones. Six strains were newly predicted as Inaba from Ogawa (Figure 8-c). In this case, because all of the mutation of the *wbeT* gene in the isolates were nonsense mutation, the isolates could have been incorrectly serotyped derived from experimental error. However, 36 cases were predicted as Ogawa instead of original Inaba types. It is possibly explained by the discrepancy between genotypes and phenotypes. Whereas Inaba isolates harbor a wild type *wbeT* gene, the strains cannot express B determinants which encoded by wild type *wbeT* gene. Because WbeT is the methyltransferase, it is possible that the mutation occurs at downstream of methyl transferring to the sugar. However, the reason is unclear.

If the case of 21 % of Ogawa prediction (from Inaba isolates) were regarded as mis-prediction, the accuracy of prediction tool for subserotyping program is about 80 % (Figure 8-c). The *wbeT* gene is a reasonably good marker for classification of O1 subserotypes. However, there were some exceptions with a discrepancy between genotype and phenotype such as prediction of Inaba strain to Ogawa strain.

Figure 7. Genetic structures of O antigen gene cluster of *V. cholerae* strains

Table 3. Representative strains and all serogroups of *V. cholerae*

Serogroups	Representative Strain	Accession	Country	Num.of strains in Serogroup	Year of Isolation	Num.of strains in Serogroup
O1	O395	GCA_000021625.1	India	428	1965	428
O1	A131	GCA_001259315.1	India	1	1989	1
O1	A185	GCA_001253295.1	Colombia	1	1992	1
O1	A241	GCA_001256675.1	Vietnam	1	1989	1
O1	A76	GCA_001259495.1	Bangladesh	1	1982	1
O1	4679	GCA_001247245.1	Bangladesh	1	1999	1
O1	7685	GCA_001255915.1	Kenya	1	2009	1
O1	7686	GCA_001258535.1	Kenya	1	2009	1
O1	GP143	GCA_001260075.1	Bahrain	1	1978	1
O1	6191	GCA_001249515.1	Kenya	1	2005	1
O1	A103	GCA_001254575.1	ND	1	1990	1
O1	12129	GCA_000174115.1	Australia	1	1985	1
O1	V109	GCA_001257255.1	India	1	1990	1
O1	LMA3984-4	GCA_000195065.1	Brazil	4	2007	4
O1	M2140	GCA_001887635.1	Australia	2	1977	2
O1	I-1471	GCA_000818865.1	Russia	1	2011	1
O1	I-1300	GCA_000967785.1	Russia	1	1999	1
O1	A325	GCA_001254095.1	Argentina	1	1993	1
O139	MO10	GCA_000152425.1	India	16	1992	16
O139	A1330	GCA_001257215.1	India	1	1993	1
O14	MZO-2	GCA_000153985.3	Bangladesh	1	2001	1
O144	254-93	GCA_000737025.1	India	1	1993	1
O16	877-163	GCA_001402745.1	Bangladesh	1	2002	1
O27	10432-62	GCA_000969265.1	Philippines	1	1962	1
O37	MZO-3	GCA_000168935.3	Bangladesh	3	2001	3
O39	AM-19226	GCA_000153785.3	Bangladesh	1	2006	1
O49	1154-74	GCA_000969235.1	India	1	1974	1
O65	981-75	GCA_000736925.1	India	1	1975	1
O77	8-76	GCA_000736935.1	India	1	1976	1
O80	1421-77	GCA_000736785.1	India	1	1977	1
O89	984-81	GCA_000736775.1	India	1	1981	1
non-O1/O139	1587	GCA_000168895.2	Peru	1	1587	1
non- O1/O139	2012EL-1759	GCA_000710155.1	Haiti	1	2012	1
non- O1/O139	2012Env-2	GCA_000788495.1	Haiti	1	2012	1
non- O1/O139	2012Env-32	GCA_000788675.1	Haiti	1	2012	1
non- O1/O139	2012Env-92	GCA_000788755.1	Haiti	1	2012	1
non- O1/O139	2012HC-25	GCA_000788775.1	Haiti	1	2012	1
non- O1/O139	2521-89	GCA_002216685.1	United States	1	1989	1
non- O1/O139	623-39	GCA_000154005.2	ND	1	ND	1
non- O1/O139	CISM_1163068.5	GCA_002097815.1	Mozambique	1	2012	1
non- O1/O139	DL4211	GCA_001953365.1	United States	1	2008	1
non- O1/O139	DL4215	GCA_001953375.1	United States	1	2008	1
non- O1/O139	Drakes2013	GCA_001543505.1	United States	1	2013	1
non- O1/O139	FDAARGOS_103	GCA_001471585.2	Germany	2	ND	2
non- O1/O139	FORC_055	GCA_002313025.1	South Korea	1	2014	1
non- O1/O139	HE-39	GCA_000220765.3	Haiti	4	2010	4
non- O1/O139	HC-43B1	GCA_000279435.1	Haiti	4	2010	4
non- O1/O139	HC-1A2	GCA_000304775.1	Haiti	5	2010	5
non- O1/O139	HE-25	GCA_000279265.1	Haiti	1	2010	1
non- O1/O139	HE-45	GCA_000279285.1	Haiti	1	2010	1
non- O1/O139	L15	GCA_001718095.1	Sweden	1	2006	1
non- O1/O139	EM-1676A	GCA_000348345.2	Bangladesh	1	2011	1
non- O1/O139	87395	GCA_000348085.2	Mexico	1	1983	1
non- O1/O139	OYP2A12	GCA_002284395.1	United States	1	ND	1
non- O1/O139	OYP2D07	GCA_002284425.1	United States	1	2009	1
non- O1/O139	OYP3F10	GCA_002284355.1	United States	1	ND	1
non- O1/O139	OYP5F10	GCA_002284235.1	United States	1	2009	1
non- O1/O139	OYP6D06	GCA_002284205.1	United States	1	2009	1
non- O1/O139	OYP6E07	GCA_002284185.1	United States	1	2009	1
non- O1/O139	OYP6F10	GCA_002284265.1	United States	1	2009	1
non- O1/O139	S12	GCA_001735565.1	Australia	1	2009	1
non- O1/O139	TMA 21	GCA_000174295.1	Brazil	1	1982	1
non- O1/O139	TP	GCA_001857485.1	United States	1	2000	1
non- O1/O139	YB1A01	GCA_001402185.1	United States	4	2009	4
non- O1/O139	YB1C07	GCA_001402285.1	United States	3	2009	3
non- O1/O139	YB2G01	GCA_001411585.1	United States	7	2009	7
non- O1/O139	YB4B03	GCA_001402605.1	United States	2	2009	2

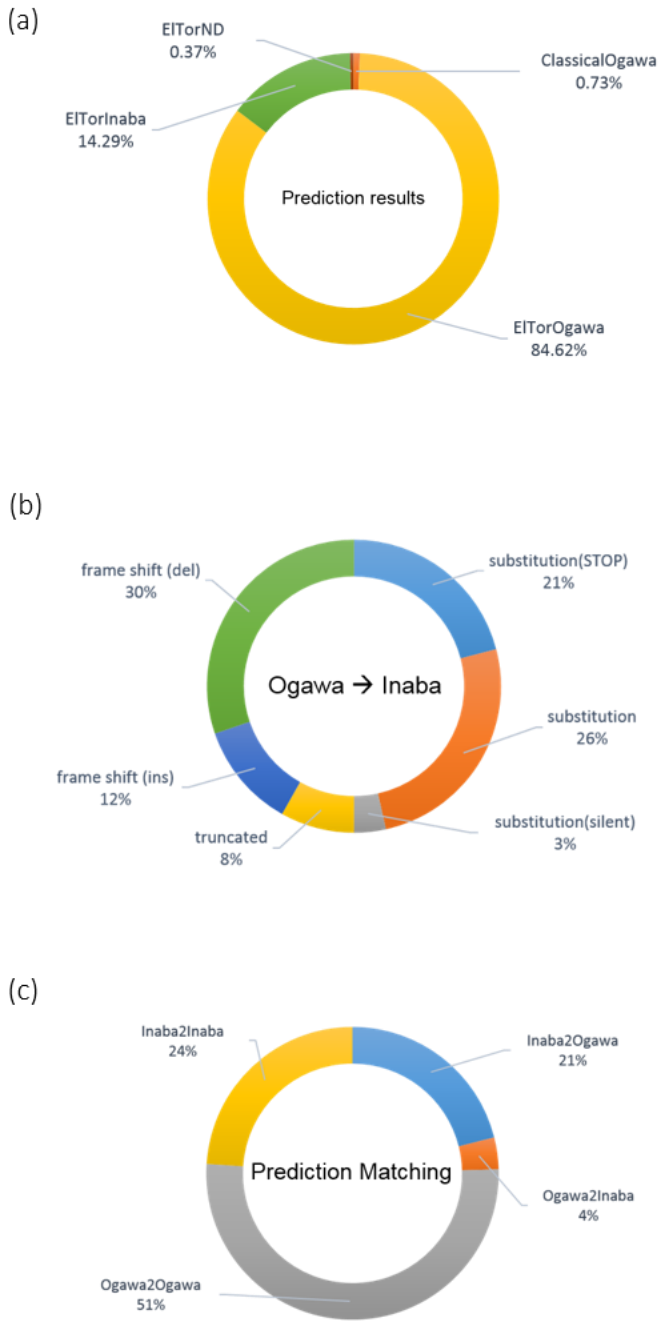

Figure 8. Statistics of O1 sub types prediction (Ogawa and Inaba)

Figure 8-legend.

(a) All newly predicted stains

(b) Assortment of mutation of *wbeT* gene in predicted as Inaba strains

(c) Prediction types.

Ogawa2Inaba: Predicted as Ogawa from Inaba strain,

Inaba2Ogawa: Predicted as Inaba from Ogawa strain,

Ogawa2Ogawa: correctly predicted,

Inaba2Inaba: correctly predicted.

4.3.2. Prediction of Cholera Toxin genes

By using 12 reference gene sequences (*ace*, *cep*, *orfU*, *zot*, *ctxAB*, *rstRABC*, TLC, and TCP), the prediction tool successfully extracted CTX phage elements, RS1 elements, TLC, and TCP. Moreover, with ten reference gene sequence files which contain all variant of each gene, the program successfully classified types of ORFs. All classified results are in Figure 9 and Figure 10.

The strain O395 as classical biotype was predicted as harboring two CTX^{cla}-core on chromosome I and chromosome II respectively, one RS2 (*rstRAB*) on chromosome I, and three TLC elements in tandem repeat manner on chromosome I. The representative strain of Wave-1 N16961 was predicted as harboring one CTX-1 core, one RS1, two TLC with RS1: CTX:TLC:TLC array on chromosome I. Strain MJ-1236 as representation of Wave-2 was predicted as having one CTX-2 core on the chromosome II and no elements on chromosome I. Two strains of Wave-3 were also well predicted. Strain 330073_B (2013, Bangladesh) was predicted as having CTX-3, and str. 7Mo (2015, Tanzania) was predicted as harboring CTX-3b on chromosome I (Figure 11).

There is no information about CTX-core elements of CTX^{AUS} type except *rstR* and *ctxB* yet (Kim *et al.* 2015). Although there is no chemo-taxonomical experimental reference, the program successfully predicted information of CTX elements harbored in M2140 (1977, Australia) (Figure 12). The CTX^{AUS} type phage is known to be possessed $rstR^{cla}$ and *ctxB*2. The predicted results showed that CTX harbored in M2140 had $rstR^{cla}$, *ctxB*2 types, and *rstA* (CTX-2), *rstB* (CTX-1), *cep* (CTX-1), *ace* (US Gulf), *zot* (CTX-1), and $ctxA^{cla}$ (Table 4). Because the combination of $rstR^{cla}$ and *ctxB*2 can be only in CTX^{AUS} type, it is a reasonable decision to predict that the strain M2140 harbors CTX^{AUS} type phage.

Almost all structure and class about CTX harbored in *V. cholerae* information is known (Lee *et al.* 2009, Kim *et al.* 2014, Kim *et al.* 2014, Kim *et al.* 2015). Because this program refers almost all information from previous CTX researches, this program can be regarded as the gold standard for prediction tool for CTX study.

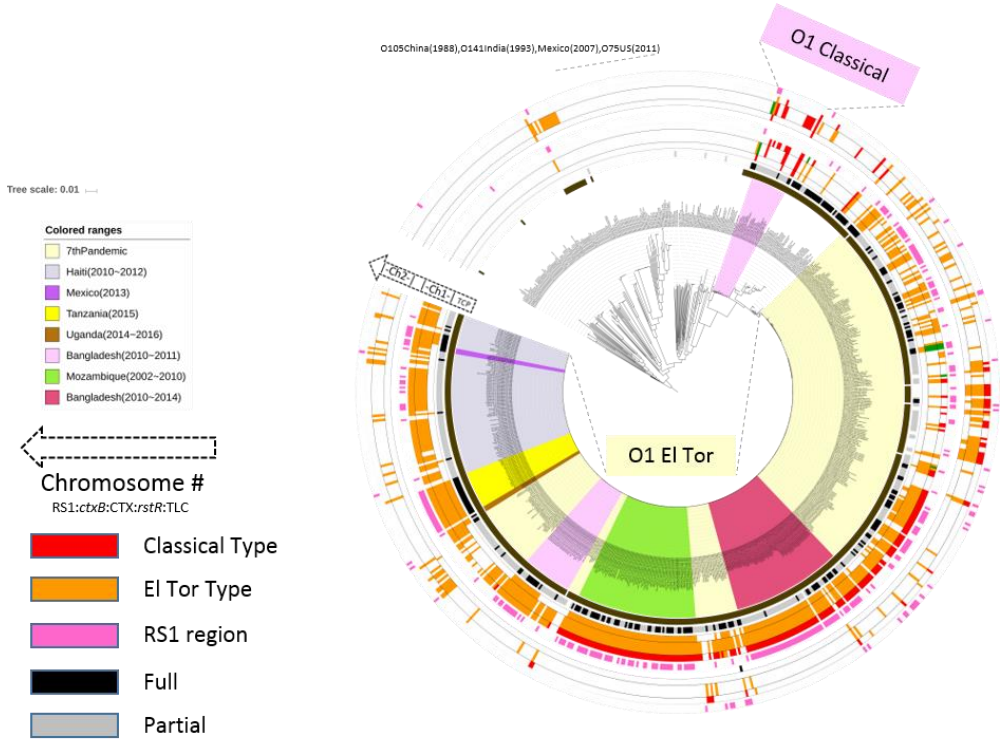

Figure 9. Distribution of *V. cholerae* strains harboring CTX (All strains)

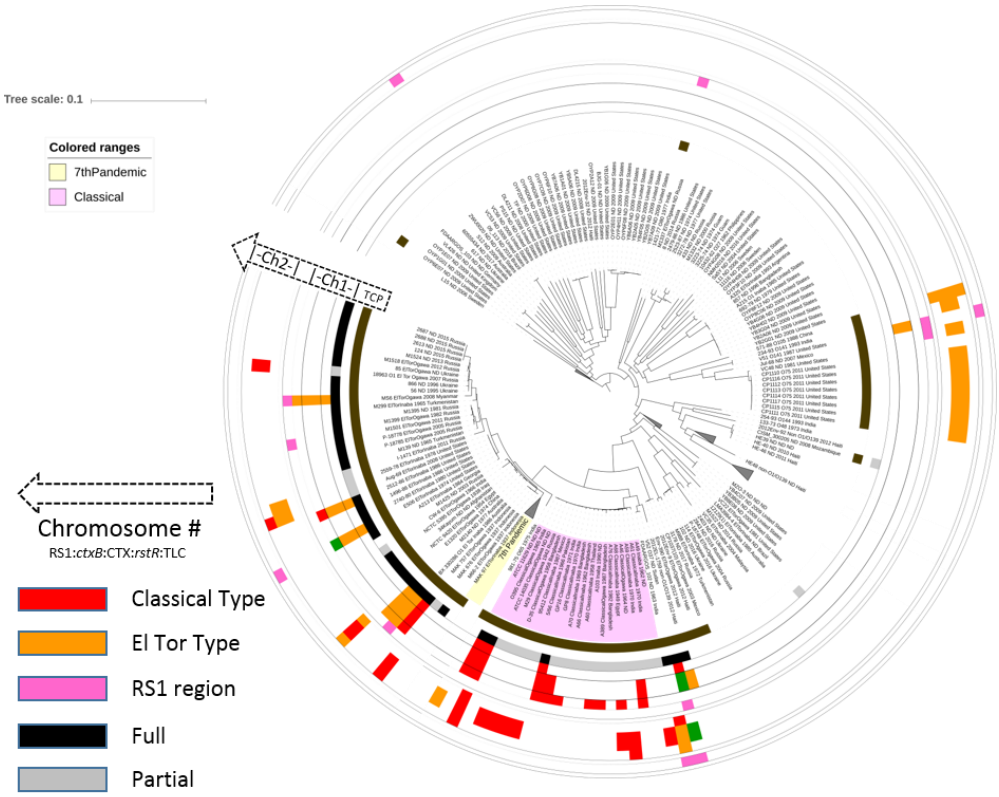

Figure 10. Distribution of *V. cholerae* strains harboring CTX (except 7th pandemic strains)

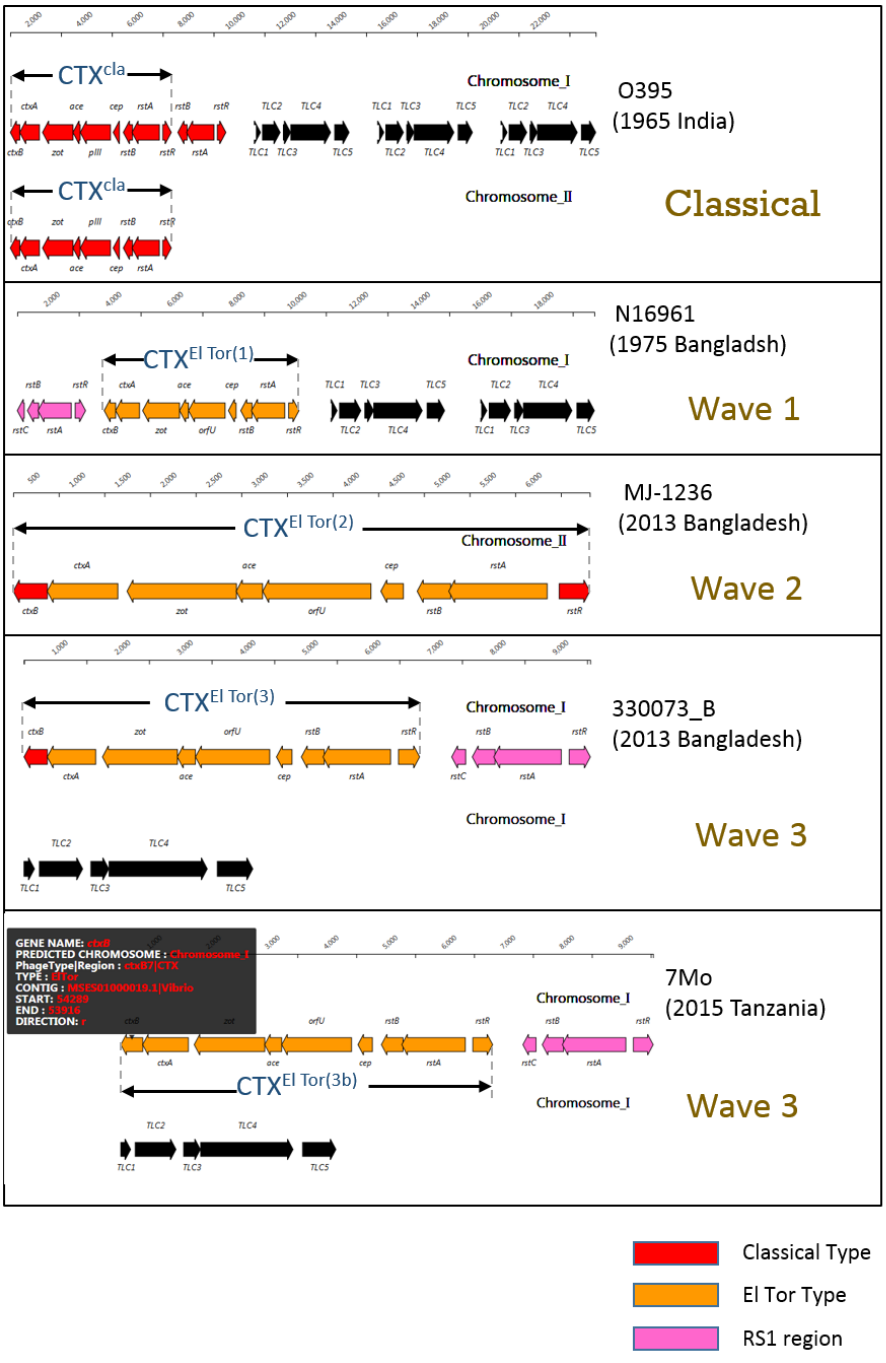

Figure 11. Predicted structures of various CTX harbored in five strains

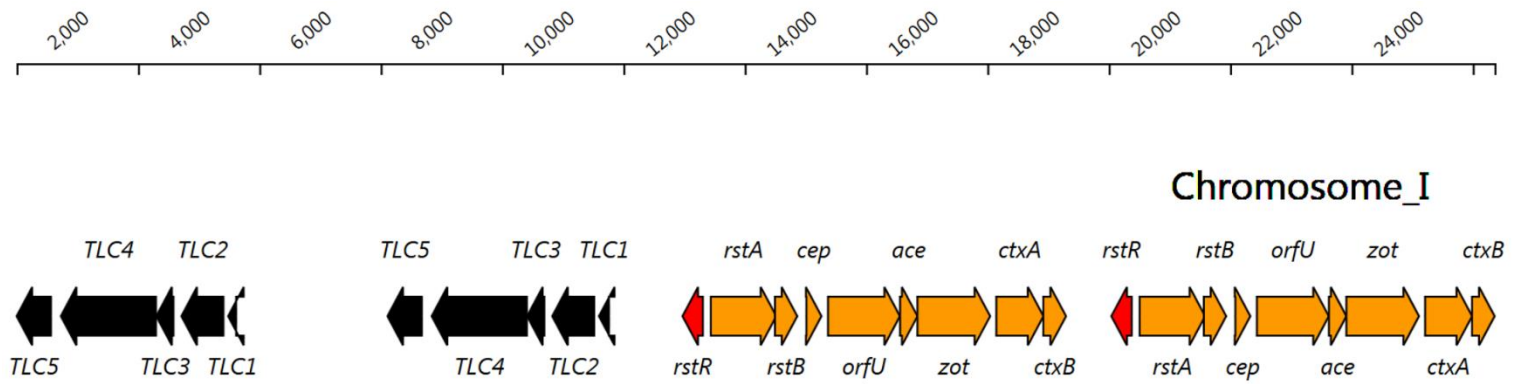

Figure 12. Predicted genetic structure of CTX harbored in str. M2140

Table 4. Information of predicted CTX elements in M2140

gene	Class	gene type
rstR	Classical	CTX-cla
rstA	ElTor	CTX-2
rstB	ElTor	CTX-1
cep	ElTor	CTX-1
orfU	ElTor	CTX-1
ace	ElTor	CTX-US Gulf
zot	ElTor	CTX-1
ctxA	ElTor	CTX-cla
ctxB	ElTor	ctxB2

4.3.3. Prediction of antibiotic resistance

All of the *V. cholerae* strains were resistant to at least one drug. In the case of Haiti strain (*V. cholerae* 2012EL-2176. 2012 Haiti), the isolate was resisting antibiotics through a combination of 39 different ways ('PERFECT': 9, 'STRICT': 30) which was the maximum number of resistance among the all 796 *V. cholerae* strains (Figure 13). With 'PERFECT or STRICT' criterion, *V. cholerae* had a minimum resistance count of 2 and the maximum value was 39, the average of counting for resistance was 9.3, and the median value was 10.5 (Figure 13).

Total 17 types of antibiotic resistances were predicted, and five types of abundant antibiotics mechanisms were estimated. While *V. cholerae* species were mostly resistant with 'PERFECT' against carbapenem (70 %), sulfone antibiotic (52.7 %), phenicol antibiotic (52.5 %), and sulfonamide (52.7 %), followed by penem (1.3 %) (Figure 14). Eight types of antibiotics (nucleoside (99.7 %), cephamycin antibiotic (99.3 %), rifamycin antibiotic (99.3 %), streptogramin antibiotic (99.2 %), penem antibiotic (96.5 %), monobactam antibiotic (96.2 %), cephalosporin (95 %) and tetracycline antibiotics (84.3 %)) were abundantly susceptible for the species (Figure 14).

With considering that most of *V. cholerae* are resistant to chloramphenicol (Kitaoka *et al.* 2011) and, the 'PERFECT' term is maybe reasonable to decide whether the drug is efficient or not. However, considering that resistance to fluoroquinolones in the Africa and Asia is growing (Saha *et al.* 2005, Saha *et al.* 2006, Islam *et al.* 2009), and study in Bangladesh over a ten-year period reports that the minimum inhibitory concentration (MIC) of ciprofloxacin had been increased an 83 fold (Kim *et al.* 2010), it is more reasonable to consider 'PERFECT or STRICT' term as a decision of resistance to drugs. As using 'PERFECT or STRICT' cut-off, *V. cholerae* strains are mostly resistant to are a macrolide, fluoroquinolone, and penam (99.87 % of strains had resistance to those drugs), followed by carbapenem (88.94 %), phenicol (71.61 %) (Figure 14). Macrolides have shown efficacy in adults and children with cholera (Khan *et al.* 2002, Bhattacharya *et al.* 2003, Kaushik *et al.* 2010, Das *et al.* 2014). However, resistance strains to erythromycin have been reported in recent years in the South Asia (Faruque *et al.* 2003). Whereas rare cases of resistance to azithromycin have been reported, and azithromycin is regarded as the last line of treatment for cholera patients. Whereas macrolide antibiotic has a second highest unsusceptible drug in 'PERFECT or STRICT' term, but also macrolides are most efficacy drug to treatment of cholera on 'PERFECT' criterion. Considering *V. cholerae* resistance to macrolides, the term 'STRICT' can be possibly interpreted as having a high possibility of resistance to drugs.

Tetracycline is one of the most efficient drugs for cholera treatment. The two drugs of tetracycline class are tetracycline and doxycycline. Due to the broad spectrum of activity, tetracyclines are widely used for other indications. In the result of this study, the *V. cholerae* strains having resistance to tetracyclines are about 15 % ('PERFECT': 0.12 %, 'STRICT': 15.57 %) among the total of 796 genomes. 15.57 % of *strictly* resistant strains have a high possibility of having resistance to tetracyclines. In fact, recently, resistance to tetracycline and doxycycline has been reported and there is cross-resistance between the two antibiotics, although *V. cholerae* strains circulating in recent years have been relatively sensitive to doxycycline than tetracycline (Sack *et al.* 1978, Siddique *et al.* 1989, Mwansa *et al.* 2007, Talkington *et al.* 2011, Tran *et al.* 2012, Díaz-Quiñonez *et al.* 2014). The result of this study shows that there were no resistant strains to tetracycline before 1970s, but after the cholera emergence in the 1970s, the strains having resistance were increasing continuously (Figure 15-a).

The almost 100 % of *V. cholerae* strains were resistant against fluoroquinolone antibiotic, macrolide antibiotic, and penam all period with 'PERFECT or STRICT' criterion (most of results are predicted as 'STRICT') (Figure 15-a.), and the percentile of *V. cholerae* resistant strains to carbapenem is always over 80 %. Notably, since the 1970s the rate of resistant strains to the phenicol antibiotic, diaminopyrimidine, aminoglycoside, sulfonamide, and tetracycline has been steadily increasing, but before the 1970s there were no resistant strains (Figure 15-a). It is reasonably interpreted that the usage of antibiotics evoked resistance genes, and the genes spread to other strains by in tandem mobile elements.

To evaluate correlation between results of RGI prediction and resistance triggered by SXT elements, the trend of presence in *V. cholerae* genomes were analyzed (Figure 15-b). The presence patterns of four drug resistance SXT element were similar with resistant strains trend predicted by RGI. However, there was no evidence that the SXT element triggers the resistance of the strains from predicted results by RGI. So, from the both data, which include results of antibiotic resistance by RGI and results of prediction of SXT elements presence, the Cohen's kappa coefficient (κ) was used to evaluate their correlation. The κ of the tetracycline data is 0.940293. It strongly supports that the SXT element triggers the resistance of *V. cholerae* strains to tetracycline. The κ of the streptomycin which is drug class of aminoglycoside is 0.739879. The value means that correlation between two results is substantially supported. The κ values of the sulfonamide and chloramphenicol are 0.59942 and 0.40793 respectively. Considering that 0.41-60 of κ is regarded as moderate correlation, it is possibly interpreted that the SXT elements contributes to the increasing resistance to the drugs (Table 5).

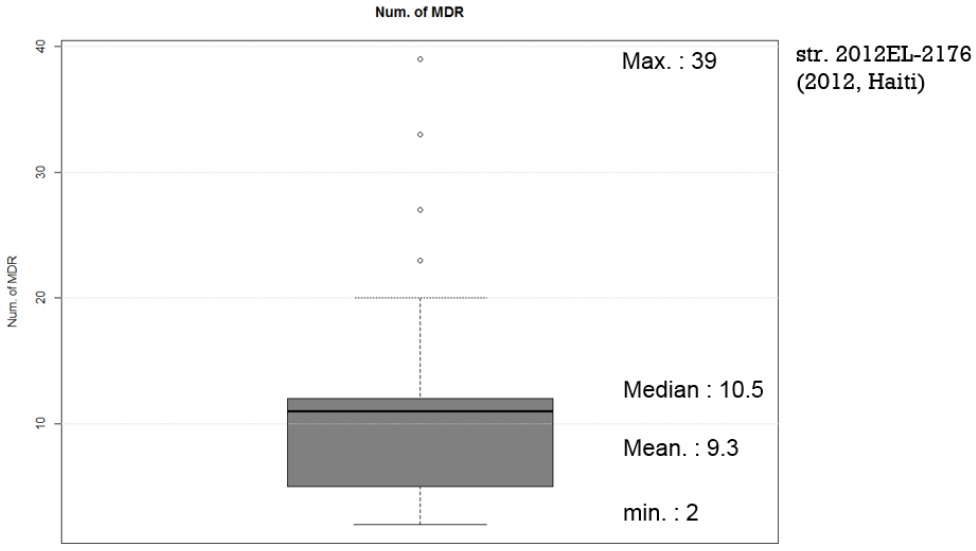

Figure 13. Boxplot of predicted antibiotic resistance number by each strain

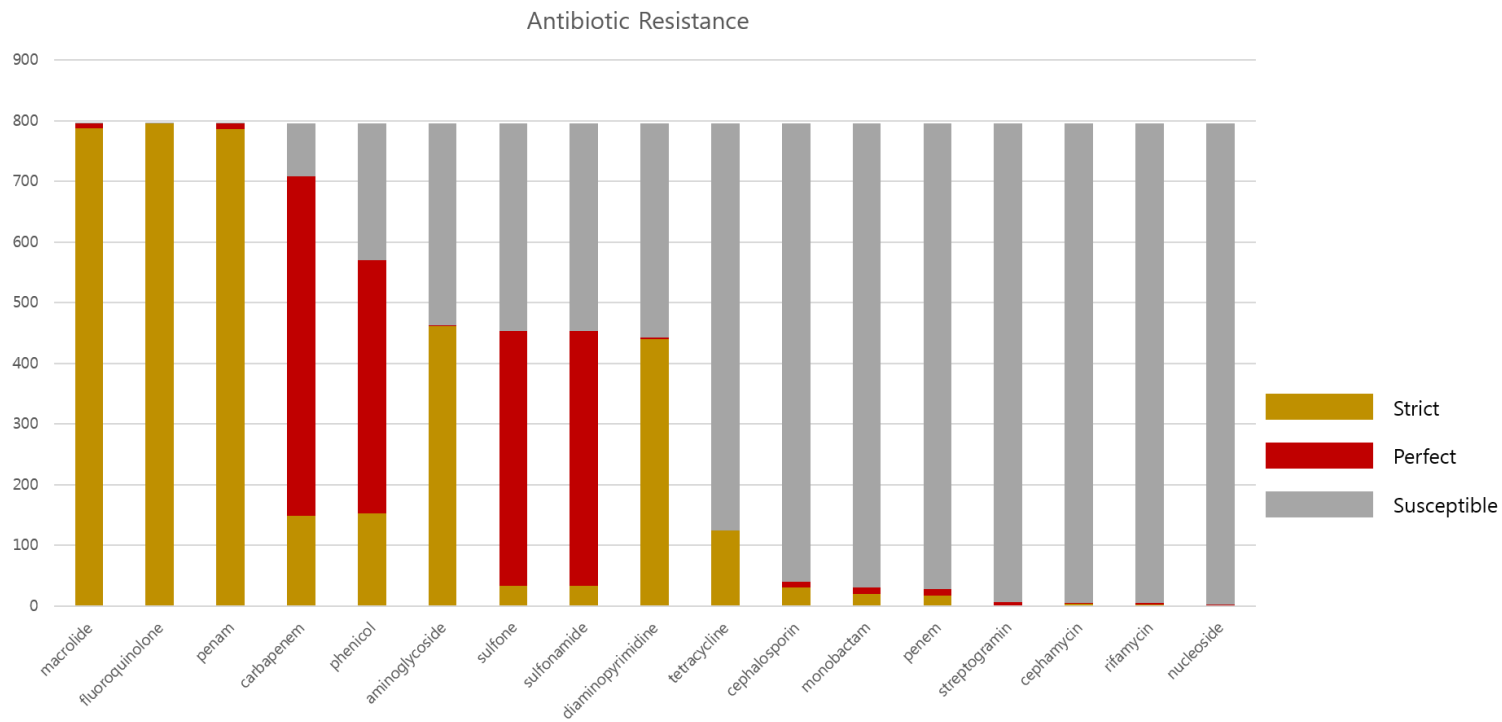

Figure 14. Antibiotic resistance of *V. cholerae* by drug types

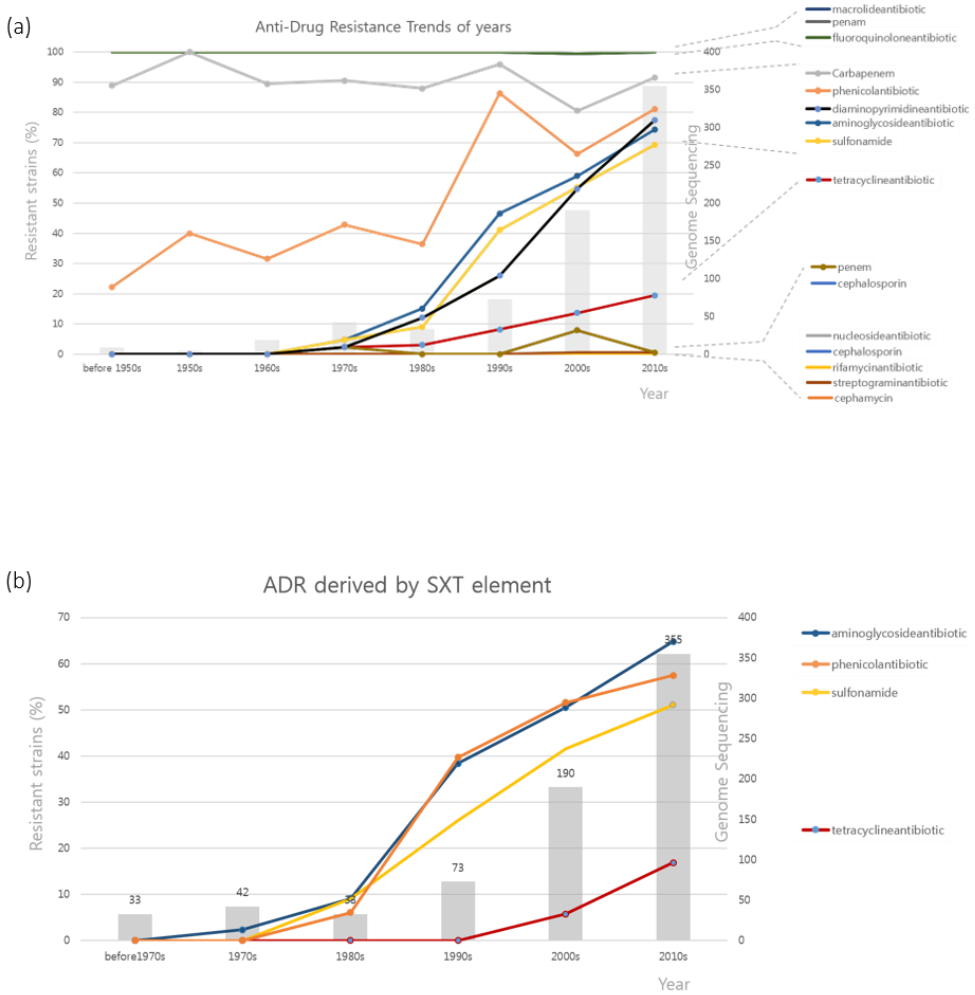

Figure 15. *V. cholerae* anti-drug resistance trend of years

Figure 15 – legend (a) Anti-drug resistance trend of years predicted by RGI. (b) ADR trend of years derived by SXT element.

Table 5. Cohen's Kappa between prediction by RGI and prediction of SXT derived antibiotic resistance of *V. cholerae*

Predicted Resistance	Cohen's Kappa (κ)
Tetracycline R	0.940293
Aminoglycoside R	0.739879
Sulfonamide R	0.59942
Chloramphenicol R	0.40793

Table 5-legend

Cohen's Kappa

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

		SXT elements drived	
		R	S
RGI	R	a	b
	S	c	d

R: resistance
S: susceptible

$$Pr(a) = \frac{a+d}{a+b+c+d}$$

$$Pr(e) = \left(\frac{(a+c)}{a+b+c+d} \times \frac{(a+b)}{a+b+c+d}\right) + \left(\frac{(b+d)}{a+b+c+d} \times \frac{(c+d)}{a+b+c+d}\right)$$

κ	agreement
≤ 0	No agreement
0.01 – 0.20	None or slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect agreement

Chapter 5. Conclusion

Microbial studies have reached a new era over the past decade. The study of microbes is an essential academic field for the public health and welfare of humankind. The microbiology field has made tremendous progress with a powerful combination of genome sequencing and bioinformatics-driven analyses of sequencing data. Bioinformatics helps us understand how bacteria evolve and function and interact with each other. Developing new tools for biology is indispensable for the coevolution of informatics and biology.

Here, the robust, objective, and readily usable tools, including OrthoANI, ContEst16S, and tools to predict *V. cholerae* phenotypes have developed in this study. The OrthoANI program successfully classifies microbial species, and the ContEst16S program improves the quality of genome data. Many researchers around the world are already using these two programs. The phenotype prediction programs for *V. cholerae* can predict the O antigen serotype, cholera toxin, and antibiotic resistance, which will help microbiologists contribute to our understanding of the microbial world. These programs will be the basis for the development of pathology and general microbiology.

REFERENCES

Albert, M. J., M. Ansaruzzaman, P. K. Bardhan, A. Faruque, S. M. Faruque, M. Islam, D. Mahalanabis, R. B. Sack, M. A. Salam and A. K. Siddique (1993). Large epidemic of cholera-like disease in Bangladesh caused by *Vibrio cholerae* 0139 synonym Bengal. *The Lancet* **342**(8868): 387-390.

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson and P. H. Nielsen (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* **31**(6): 533.

Alkan, C., S. Sajjadian and E. E. Eichler (2010). Limitations of next-generation genome sequence assembly. *Nature methods* **8**(1): 61.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**(17): 3389-3402.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez and M. Magrane (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research* **32**(suppl_1): D115-D119.

Aydanian, A., L. Tang, J. G. Morris, J. A. Johnson and O. C. Stine (2011). Genetic Diversity of O-Antigen Biosynthetic Regions in *Vibrio cholerae*. *Applied and environmental microbiology*.

Barua, D. (1992). History of cholera. Cholera, Springer**:** 1-36.

Baudry, B., A. Fasano, J. Ketley and J. Kaper (1992). Cloning of a gene (zot) encoding a new toxin produced by *Vibrio cholerae*. *Infection and immunity* **60**(2): 428-434.

Beaz-Hidalgo, R., M. J. Hossain, M. R. Liles and M.-J. Figueras (2015). Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for Aeromonas genomes in the GenBank database. *PLoS One* **10**(1): e0115813.

Bernardy, E. E., M. A. Turnsek, S. K. Wilson, C. L. Tarr and B. K. Hammer (2016). Diversity of clinical and environmental isolates of *Vibrio cholerae* in natural transformation and contact-dependent bacterial killing indicative of type VI secretion system activity. *Applied and environmental microbiology*: AEM. 00351-00316.

Bhattacharya, M., D. Dutta, T. Ramamurthy, D. Sarkar, A. Singharoy and S. Bhattacharya (2003). Azithromycin in the treatment of cholera in children. *Acta Paediatrica* **92**(6): 676-678.

Bik, E. M., A. E. Bunschoten, R. J. Willems, A. C. Chang and F. R. Mooi (1996). Genetic organization and functional analysis of the otn DNA essential for cell wall polysaccharide synthesis in *Vibrio cholerae* O139. *Molecular microbiology* **20**(4): 799-811.

Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall and M. J. Stanhope (2001). Universal trees based on large combined protein sequence data sets. *Nature genetics* **28**(3): 281.

Buchfink, B., C. Xie and D. H. Huson (2014). Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**(1): 59.

Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. Stoutland, W. Regala, A. Georgescu, L. Vergez, M. Land and V. Motin (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences* **101**(38): 13826-13831.

Chatterjee, S. and K. Chaudhuri (2003). Lipopolysaccharides of *Vibrio cholerae*: I. Physical and chemical characterization. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1639**(2): 65-79.

Cho, Y.-J., H. Yi, J. H. Lee, D. W. Kim and J. Chun (2010). Genomic evolution of *Vibrio cholerae*. *Current opinion in microbiology* **13**(5): 646-651.

Choi, S. Y., S. M. Rashed, N. A. Hasan, M. Alam, T. Islam, A. Sadique, F.-T. Johura, M. Eppinger, J. Ravel and A. Huq (2016). Phylogenetic diversity of *Vibrio cholerae* associated with endemic cholera in Mexico from 1991 to 2008. *MBio* **7**(2): e02160-02115.

Chowdhury, F., M. A. Rahman, Y. A. Begum, A. I. Khan, A. S. Faruque, N. C. Saha, N. I. Baby, M. Malek, A. R. Kumar and A.-M. Svennerholm (2011). Impact of rapid urbanization on the rates of infection by *Vibrio cholerae* O1 and enterotoxigenic *Escherichia coli* in Dhaka, Bangladesh. *PLoS neglected tropical diseases* **5**(4): e999.

Chun, J., C. J. Grim, N. A. Hasan, J. H. Lee, S. Y. Choi, B. J. Haley, E. Taviani, Y.-S. Jeon, D. W. Kim and J.-H. Lee (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences* **106**(36): 15442-15447.

Chun, J. and F. A. Rainey (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *International journal of systematic and evolutionary microbiology* **64**(2): 316-324.

Colwell, R., A. Huq, M. Chowdhury, B. Xu and P. Brayton (1995). Serogroup conversion of *Vibrio cholerae*. *Canadian journal of microbiology* **41**(10): 946-950.

Cvjetanovic, B. and D. Barua (1972). The seventh pandemic of cholera. *Nature* **239**(5368): 137.

Díaz-Quiñonez, A., I. Hernández-Monroy, N. Montes-Colima, A. Moreno-Pérez, A. Galicia-Nicolás, H. Martínez-Rojano, C. Carmona-Ramos, M. Sánchez-Mendoza, J. C. Rodríguez-Martínez and L. Suárez-Idueta (2014). Outbreak of *Vibrio cholerae* Serogroup O1, Serotype Ogawa, Biotype El Tor Strain—La Huasteca Region, Mexico, 2013. *MMWR. Morbidity and mortality weekly report* **63**(25): 552.

Das, S., A. Rahman, M. Chisti, S. Ahmed, M. Malek, M. Salam, P. Bardhan and A. Faruque (2014). Changing patient population in D haka H ospital and M atlab H ospital of icddr, b. *Tropical Medicine & International Health* **19**(2): 240-243.

Daubin, V. and M. Gouy (2001). Bacterial molecular phylogeny using supertree approach. *Genome Informatics* **12**: 155-164.

De, S. N., J. Sarkar and B. Tribedi (1951). An experimental study of the action of cholera toxin. *The Journal of pathology and bacteriology* **63**(4): 707-717.

Dobrindt, U., B. Hochhut, U. Hentschel and J. Hacker (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* **2**(5): 414.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5): 1792-1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19): 2460-2461.

Farmer III, J., J. Michael Janda, F. W. Brenner, D. N. Cameron and K. M. Birkhead (2015). *Vibrio*. *Bergey's Manual of Systematics of Archaea and Bacteria*: 1-79.

Faruque, S. M., M. J. Albert and J. J. Mekalanos (1998). Epidemiology, Genetics, and Ecology of Toxigenic*Vibrio cholerae*. *Microbiology and molecular biology reviews* **62**(4): 1301-1314.

Faruque, S. M., A. A. Alim, M. J. Albert, K. N. Islam and J. J. Mekalanos (1998). Induction of the lysogenic phage encoding cholera toxin in naturally occurring strains of toxigenic *Vibrio cholerae* O1 and O139. *Infection and immunity* **66**(8): 3752-3757.

Faruque, S. M., D. A. Sack, R. B. Sack, R. R. Colwell, Y. Takeda and G. B. Nair (2003). Emergence and evolution of *Vibrio cholerae* O139. *Proceedings of the National Academy of Sciences* **100**(3): 1304-1309.

Gargis, A. S., L. Kalman, M. W. Berry, D. P. Bick, D. P. Dimmock, T. Hambuch, F. Lu, E. Lyon, K. V. Voelkerding and B. A. Zehnbauer (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature biotechnology* **30**(11): 1033.

Glass, R., M. Huq, J. Lee, E. Threlfall, M. Khan, A. Alim, B. Rowe and R. Gross (1983). Plasmid-borne multiple drug resistance in *Vibrio cholerae* serogroup O1, biotype El Tor: evidence for a point-source outbreak in Bangladesh. *Journal of Infectious Diseases* **147**(2): 204-209.

Glass, R. I., I. Huq, A. Alim and M. Yunus (1980). Emergence of multiply antibiotic-resistant *Vibrio cholerae* in Bangladesh. *Journal of Infectious Diseases* **142**(6): 939-942.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme and J. M. Tiedje (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology* **57**(1): 81-91.

Gustafsson, B. and T. Holme (1985). Rapid detection of*Vibrio cholerae* O1 by motility inhibition and immunofluorescence with monoclonal antibodies. *European journal of clinical microbiology* **4**(3): 291-294.

Hassan, F., M. Kamruzzaman, J. J. Mekalanos and S. M. Faruque (2010). Satellite phage TLCφ enables toxigenic conversion by CTX phage through dif site alteration. *Nature* **467**(7318): 982.

Hay, C. W. and K. Docherty (2006). Comparative analysis of insulin gene promoters: implications for diabetes research. *Diabetes* **55**(12): 3201-3213.

Howard-Jones, N. (1984). Robert Koch and the cholera vibrio: a centenary. *British medical journal (Clinical research ed.)* **288**(6414): 379.

Hu, D., B. Liu, L. Feng, P. Ding, X. Guo, M. Wang, B. Cao, P. R. Reeves and L. Wang (2016). Origins of the current seventh cholera pandemic. *Proceedings of the National Academy of Sciences* **113**(48): E7730-E7739.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**(1): 119.

Islam, M. S., S. M. Midzi, L. Charimari, A. Cravioto and H. P. Endtz (2009). Susceptibility to fluoroquinolones of *Vibrio cholerae* O1 isolated from diarrheal patients in Zimbabwe. *JAMA* **302**(21): 2321-2322.

Ito, T., K. Hiramatsu, Y. Ohshita and T. Yokota (1993). Mutations in the *rfbT* gene are responsible for the Ogawa to Inaba serotype conversion in *Vibrio cholerae* O1. *Microbiology and immunology* **37**(4): 281-288.

Iyer, L. M., V. Anantharaman, M. Y. Wolf and L. Aravind (2008). Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *International journal for parasitology* **38**(1): 1-31.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist* **11**(2): 37-50.

Jesudason, M. and R. Saaya (1997). Resistance of *Vibrio cholerae* 01 to nalidixic acid. *The Indian Journal of Medical Research* **105**: 153-154.

Jia, B., A. R. Raphenya, B. Alcock, N. Waglechner, P. Guo, K. K. Tsang, B. A. Lago, B. M. Dave, S. Pereira and A. N. Sharma (2016). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*: gkw1004.

Johnson, J. L. and W. B. Whitman (2007). Similarity analysis of DNAs. Methods for General and Molecular Microbiology, Third Edition, American Society of Microbiology**:** 624-652.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**(1): 27-30.

Karlin, S. and S. F. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences* **87**(6): 2264-2268.

Kaushik, J. S., P. Gupta, M. Faridi and S. Das (2010). Single dose azithromycin versus ciprofloxacin for cholera in children: a randomized controlled trial. *Indian pediatrics* **47**(4): 309-315.

Khan, W. A., D. Saha, A. Rahman, M. A. Salam, J. Bogaerts and M. L. Bennish (2002). Comparison of single-dose azithromycin and 12-dose, 3-day erythromycin for childhood cholera: a randomised, double-blind trial. *The Lancet* **360**(9347): 1722-

1727.

Kim, E. J., D. Lee, S. H. Moon, C. H. Lee and D. W. Kim (2014). CTX prophages in *Vibrio cholerae* O1 strains. *J Microbiol Biotechnol* **24**(6): 725-731.

Kim, E. J., D. Lee, S. H. Moon, C. H. Lee, S. J. Kim, J. H. Lee, J. O. Kim, M. Song, B. Das and J. D. Clemens (2014). Molecular insights into the evolutionary pathway of *Vibrio cholerae* O1 atypical El Tor variants. *PLoS pathogens* **10**(9): e1004384.

Kim, E. J., C. H. Lee, G. B. Nair and D. W. Kim (2015). Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. *Trends in microbiology* **23**(8): 479-489.

Kim, H. B., M. Wang, S. Ahmed, C. H. Park, R. C. LaRocque, A. S. Faruque, M. A. Salam, W. A. Khan, F. Qadri and S. B. Calderwood (2010). Transferable quinolone resistance in *Vibrio cholerae*. *Antimicrobial agents and chemotherapy* **54**(2): 799-803.

Kim, M., H.-S. Oh, S.-C. Park and J. Chun (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of systematic and evolutionary microbiology* **64**(2): 346-351.

Kitahara, K., Y. Yasutake and K. Miyazaki (2012). Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **109**(47): 19220-19225.

Kitaoka, M., S. T. Miyata, D. Unterweger and S. Pukatzki (2011). Antibiotic resistance mechanisms of *Vibrio cholerae*. *Journal of medical microbiology* **60**(4): 397-407.

Konstantinidis, K. T. and J. M. Tiedje (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* **102**(7): 2567-2572.

Konstantinidis, K. T., A. Ramette and J. M. Tiedje (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **361**(1475): 1929-1940.

Krichevsky, M., L. Moore, W. Moore, R. Murray, E. Stackebrandt, M. Starr and H. Trper (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**: 463464.

Lalley, P., S. J. O'Brien, N. Creau-Goldberg, M. Davisson, T. Roderick, G. Echard, J. Womack, J. Graves, D. Doolittle and J. Guidi (1987). Report of the committee on comparative mapping. *Cytogenetic and Genome Research* **46**(1-4): 367-389.

Lee, I., Y. O. Kim, S.-C. Park and J. Chun (2016). OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *International journal of systematic and evolutionary microbiology* **66**(2): 1100-1103.

Lee, I., M. Chalita, S.-M. Ha, S.-I. Na, S.-H. Yoon and J. Chun (2017). ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S rRNA gene sequences. *International journal of systematic and evolutionary microbiology* **67**(6): 2053-2057.

Lee, J. H., S. Y. Choi, Y.-S. Jeon, H. R. Lee, E. J. Kim, B. M. Nguyen, N. T. Hien, M. Ansaruzzaman, M. S. Islam and N. A. Bhuiyan (2009). Classification of hybrid and altered *Vibrio cholerae* strains by CTX prophage and RS1 element structure. *The Journal of Microbiology* **47**(6): 783-788.

Li, X., Y. Huang and W. B. Whitman (2015). The relationship of the whole genome sequence identity to DNA hybridization varies between genera of prokaryotes. *Antonie van Leeuwenhoek* **107**(1): 241-249.

Logan, N. A. and P. De Vos (2009). Bacillus. Bergey's manual of systematic bacteriology, Springer. **3:** 21-128.

Longo, M. S., M. J. O'Neill and R. J. O'Neill (2011). Abundant human DNA contamination identified in non-primate genome databases. *PLoS One* **6**(2): e16410.

Marston, L. (2010). Introductory statistics for health and nursing using SPSS, Sage Publications.

Materu, S., O. Lema, H. Mukunza, C. Adhiambo and J. Carter (1997). Antibiotic resistance pattern of *Vibrio cholerae* and *Shigella* causing diarrhoea outbreaks in the eastern Africa region: 1994-1996. *East African medical journal* **74**(3): 193-197.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* **22**(3): 276-282.

McKusick, V. A. and F. H. Ruddle (1987). A new discipline, a new name, a new journal,*Elsevier*.

McLeod, S. M., H. H. Kimsey, B. M. Davis and M. K. Waldor (2005). CTXφ and *Vibrio cholerae*: exploring a newly recognized type of phage–host cell relationship. *Molecular microbiology* **57**(2): 347-356.

Mukherjee, S., M. Huntemann, N. Ivanova, N. C. Kyrpides and A. Pati (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in genomic sciences* **10**(1): 18.

Mukhopadhyay, A. K., A. Basu, P. Garg, P. K. Bag, A. Ghosh, S. Bhattacharya, Y. Takeda and G. B. Nair (1998). Molecular epidemiology of reemergent *Vibrio cholerae* O139 Bengal in India. *Journal of clinical microbiology* **36**(7): 2149-2152.

Mutreja, A., D. W. Kim, N. R. Thomson, T. R. Connor, J. H. Lee, S. Kariuki, N. J. Croucher, S. Y. Choi, S. R. Harris and M. Lebens (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**(7365): 462.

Mwansa, J., J. Mwaba, C. Lukwesa, N. Bhuiyan, M. Ansaruzzaman, T. Ramamurthy, M. Alam and G. B. Nair (2007). Multiply antibiotic-resistant *Vibrio cholerae* O1 biotype El Tor strains emerge during cholera outbreaks in Zambia. *Epidemiology & Infection* **135**(5): 847-853.

Myers, E. W. and W. Miller (1988). Optimal alignments in linear space. *Bioinformatics* **4**(1): 11-17.

Mylvaganam, S. and P. P. Dennis (1992). Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. *Genetics* **130**(3): 399-410.

Nakasone, N., M. Iwanaga and R. Eeckels (1987). Characterization of *Vibrio cholerae* 01 recently isolated in Bangladesh. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **81**(5): 876-878.

Nawrocki, E. P. and S. R. Eddy (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22): 2933-2935.

Nawrocki, E. P., S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones and J. Tate (2014). Rfam 12.0: updates to the RNA families database. *Nucleic acids research* **43**(D1): D130-D137.

Pak, T. R. and A. Kasarskis (2015). How next-generation sequencing and multiscale data analysis will transform infectious disease management. *Clinical Infectious Diseases* **61**(11): 1695-1702.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*: gr. 186072.186114.

Parnis, R. (1971). Cholera in West Africa.

Pearson, G. D., A. Woods, S. L. Chiang and J. J. Mekalanos (1993). CTX genetic element encodes a site-specific recombination system and an intestinal colonization factor. *Proceedings of the National Academy of Sciences* **90**(8): 3750-3754.

Pengsuk, C., S. Longyant, S. Rukpratanporn, P. Chaivisuthangkura, P. Sridulyakul and P. Sithigorngul (2010). Development of monoclonal antibodies for simple detection and differentiation of *Vibrio mimicus* from *V. cholerae* and *Vibrio* spp. by dot blotting. *Aquaculture* **300**(1-4): 17-24.

Perna, N. T., G. Plunkett III, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor and H. A. Kirkpatrick (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. *Nature* **409**(6819): 529.

Povey, S., R. Lovering, E. Bruford, M. Wright, M. Lush and H. Wain (2001). The HUGO gene nomenclature committee (HGNC). *Human genetics* **109**(6): 678-680.

Prentis, P. J., A. Vesey, N. Meyers and P. Mather (2004). Genetic structuring of the stream lily Helmholtzia glaberrima (Philydraceae) within Toolona Creek, south-eastern Queensland. *Australian Journal of Botany* **52**(2): 201-207.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **33**(suppl_1): D501-D504.

Ramamurthy, T., S. Garg, R. Sharma, S. Bhattacharya, G. B. Nair, T. Shimada, T. Takeda, T. Karasawa, H. Kurazano and A. Pal (1993). Emergence of novel strain of *Vibrio cholerae* with epidemic potential in southern and eastern India. *The Lancet* **341**(8846): 703-704.

Reeves, P. R., M. Hobbs, M. A. Valvano, M. Skurnik, C. Whitfield, D. Coplin, N. Kido, J. Klena, D. Maskell and C. R. Raetz (1996). Bacterial polysaccharide synthesis and gene nomenclature. *Trends in microbiology* **4**(12): 495-503.

Richter, M. and R. Rosselló-Móra (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences* **106**(45): 19126-19131.

Rosselló-Móra, R. and R. Amann (2015). Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology* **38**(4): 209-216.

Sack, D. A., S. Islam, H. Rabbani and A. Islam (1978). Single-dose doxycycline for cholera. *Antimicrobial agents and chemotherapy* **14**(3): 462-464.

Sack, R. B. and C. E. Miller (1969). Progressive changes of *Vibrio* serotypes in germ-free mice infected with *Vibrio cholerae*. *Journal of bacteriology* **99**(3): 688-695.

Saha, D., W. A. Khan, M. M. Karim, H. R. Chowdhury, M. A. Salam and M. L. Bennish (2005). Single-dose ciprofloxacin versus 12-dose erythromycin for childhood

cholera: a randomised controlled trial. *The Lancet* **366**(9491): 1085-1093.

Saha, D., M. M. Karim, W. A. Khan, S. Ahmed, M. A. Salam and M. L. Bennish (2006). Single-dose azithromycin for the treatment of cholera in adults. *New England Journal of Medicine* **354**(23): 2452-2462.

Sakazaki, R. and K. TAMURA (1971). Somatic antigen variation in *Vibrio cholerae*. *Japanese Journal of Medical Science and Biology* **24**(2): 93-100.

Schmieder, R. and R. Edwards (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one* **6**(3): e17288.

Sciortino, C., J. A. Johnson and A. Hamad (1996). Vitek system antimicrobial susceptibility testing of O1, O139, and non-O1 *Vibrio cholerae*. *Journal of clinical microbiology* **34**(4): 897-900.

Shimada, T. and R. Sakazaki (1988). A serogroup of non-O1 *Vibrio cholerae* possessing the Inaba antigen of *Vibrio cholerae* O1. *Journal of applied bacteriology* **64**(2): 141-144.

Siddique, A., K. Zaman, A. Baqui, K. Akram, P. Mutsuddy, A. Eusof, K. Haider, S. Islam and R. Sack (1992). Cholera epidemics in Bangladesh: 1985-1991. *Journal of diarrhoeal diseases research*: 79-86.

Siddique, A. K., K. Zaman, Y. Majumder, Q. Islam, I. Bashir, P. Mutsuddy and A. Eusof (1989). Simultaneous outbreaks of contrasting drug resistant classic and El Tor *Vibrio cholerae* 01 in Bangladesh. *Lancet* **2**(8659): 396.

Sone, T., K. Kasahara, H. Kimura, K. Nishio, M. Mizuguchi, Y. Nakatsumi, K. Shibata, Y. Waseda, M. Fujimura and S. Nakao (2007). Comparative analysis of epidermal growth factor receptor mutations and gene amplification as predictors of gefitinib efficacy in Japanese patients with nonsmall cell lung cancer. *Cancer* **109**(9): 1836-1844.

Sozhamannan, S., Y. K. Deng, M. Li, A. Sulakvelidze, J. B. Kaper, J. A. Johnson, G. B. Nair and J. G. Morris (1999). Cloning and Sequencing of the Genes Downstream of thewbf Gene Cluster of *Vibrio cholerae* Serogroup O139 and Analysis of the Junction Genes in Other Serogroups. *Infection and immunity* **67**(10): 5033-5040.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21): 2688-2690.

Stroeher, U. H., L. E. Karageorgos, R. Morona and P. A. Manning (1992). Serotype conversion in *Vibrio cholerae* O1. *Proceedings of the National Academy of Sciences* **89**(7): 2566-2570.

Stropko, S. J., S. E. Pipes and J. D. Newman (2014). Genome-based reclassification of *Bacillus cibi* as a later heterotypic synonym of *Bacillusindicus* and emended description of *Bacillus indicus*. *International journal of systematic and evolutionary microbiology* **64**(11): 3804-3809.

Sun, S., R. Ke, D. Hughes, M. Nilsson and D. I. Andersson (2012). Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PloS one* **7**(8): e42639.

Talkington, D., C. Bopp, C. Tarr, M. B. Parsons, G. Dahourou, M. Freeman, K. Joyce, M. Turnsek, N. Garrett and M. Humphrys (2011). Characterization of toxigenic *Vibrio cholerae* from Haiti, 2010–2011. *Emerging infectious diseases* **17**(11): 2122.

Tauxe, R. V. and T. J. Barrett (1998). Cholera and *Vibrio cholerae*: new challenges from a once and future pathogen. Emerging Infections 2, American Society of Microbiology**:** 125-144.

Tennessen, K., E. Andersen, S. Clingenpeel, C. Rinke, D. S. Lundberg, J. Han, J. L. Dangl, N. Ivanova, T. Woyke and N. Kyrpides (2016). ProDeGe: a computational protocol for fully automated decontamination of genomes. *The ISME journal* **10**(1): 269.

Tindall, B. J., R. Rosselló-Mora, H.-J. Busse, W. Ludwig and P. Kämpfer (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology* **60**(1): 249-266.

Tran, H. D., M. Alam, N. V. Trung, N. Van Kinh, H. H. Nguyen, V. C. Pham, M. Ansaruzzaman, S. M. Rashed, N. A. Bhuiyan and T. T. Dao (2012). Multi-drug resistant *Vibrio cholerae* O1 variant El Tor isolated in northern Vietnam between 2007 and 2010. *Journal of medical microbiology* **61**(3): 431-437.

Trucksis, M., J. E. Galen, J. Michalski, A. Fasano and J. B. Kaper (1993). Accessory cholera enterotoxin (Ace), the third toxin of a *Vibrio cholerae* virulence cassette. *Proceedings of the National Academy of Sciences* **90**(11): 5267-5271.

Waldor, M. K., H. Tschäpe and J. J. Mekalanos (1996). A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. *Journal of bacteriology* **178**(14): 4157-4165.

Wallace, C. K., A. E. Fabie, O. Mangubat, E. Velasco, C. Juinio and R. A. Phillips (1964). The 1961 cholera epidemic in Manila, Republic of the Philippines. *Bulletin of the World Health Organization* **30**(6): 795.

Wang, R., D. Yu, J. Yue and B. Kan (2016). Variations in SXT elements in epidemic *Vibrio cholerae* O1 El Tor strains in China. *Scientific reports* **6**: 22733.

Ward, N. and C. M. Fraser (2005). How genomics has affected the concept of microbiology. *Current opinion in microbiology* **8**(5): 564-571.

Willenbrock, H., P. F. Hallin, T. M. Wassenaar and D. W. Ussery (2007). Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome biology* **8**(12): R267.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin and E. V. Koonin (2002). Genome trees and the tree of life. *TRENDS in Genetics* **18**(9): 472-479.

Yap, W. H., Z. Zhang and Y. Wang (1999). Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *Journal of bacteriology* **181**(17): 5201-5209.

Yi, H. and J. Chun (2015). *Neisseria weaveri* Andersen et al. 1993 is a later heterotypic synonym of *Neisseria weaveri* Holmes et al. 1993. *International journal of systematic and evolutionary microbiology* **65**(2): 463-464.

Yoon, S.-H., S.-M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo and J. Chun (2017). Introducing

EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International journal of systematic and evolutionary microbiology* **67**(5): 1613-1617.

Yuan-Hai, Y., W. Peng, W. Yan-Hua, W. Hai-Bin, Y. Dong-Zheng, H. Rong and J.-Z. Zhang (2010). Assessment of comparative genomic hybridization experiment by an in situ synthesized CombiMatrix microarray with Yersinia pestis vaccine strain EV76 DNA. *Biomedical and Environmental Sciences* **23**(5): 384-390.

Zuckerman, J. N., L. Rombo and A. Fisch (2007). The true burden and risk of cholera: implications for prevention and control. *The Lancet infectious diseases* **7**(8): 521-530.

APPENDIX

Table 1. Information of strains used in this study

Appendix table 1. Information of strains used in this study

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	07-2425	GCA_003311905.1	ND	ND	MDR
Vibrio cholerae	09_113	GCA_003312945.1	2018	Brazil	MDR
Vibrio cholerae	102	GCA_002196095.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	10432-62	GCA_000969265.1	1962	Philippines	MDR, CTX, Serotyping
Vibrio cholerae	1074-78	GCA_001857405.1	1978	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	11116	GCA_002890525.1	2006	Sweden	MDR
Vibrio cholerae	114	GCA_002196225.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	1154-74	GCA_000969235.1	1974	India	MDR, CTX, Serotyping
Vibrio cholerae	1157-74	GCA_000736875.1	1974	India	MDR, CTX, Serotyping
Vibrio cholerae	116059	GCA_000348045.2	1992	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	116063	GCA_000348065.2	1978	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	116-17b	GCA_001292745.1	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	11S	GCA_002076185.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	12129(1)	GCA_000174115.1	1985	Australia	MDR, CTX, Serotyping
Vibrio cholerae	124	GCA_003057085.1	2015	Russia	MDR
Vibrio cholerae	1270D	GCA_003130495.1	1994	Russia	MDR
Vibrio cholerae	1311-69	GCA_000736855.1	1969	India	MDR, CTX, Serotyping
Vibrio cholerae	133-73	GCA_000736765.1	1973	India	MDR, CTX, Serotyping
Vibrio cholerae	1346	GCA_001253035.1	2005	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	1362	GCA_001260295.1	2005	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	1421-77	GCA_000736785.1	1977	India	MDR, CTX, Serotyping
Vibrio cholerae	146N	GCA_002918345.1	1994	India	MDR
Vibrio cholerae	146P	GCA_002918335.1	1994	India	MDR
Vibrio cholerae	147	GCA_002196105.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	1496-86	GCA_001857325.1	1986	United States	MDR, CTX, Serotyping
Vibrio cholerae	153	GCA_002204095.1	2011	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	155	GCA_002196155.1	ND	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	1587	GCA_000168895.2	1587	Peru	MDR, CTX, Serotyping
Vibrio cholerae	16241D	GCA_003130465.1	1994	Russia	MDR
Vibrio cholerae	1627	GCA_001247835.1	2005	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	169D	GCA_003130485.1	1993	Russia	MDR
Vibrio cholerae	17609	GCA_002078825.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	186	GCA_003015005.1	2011	Ukraine	MDR
Vibrio cholerae	19886	GCA_002078715.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	1Mo	GCA_002076425.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	2009V-1046	GCA_000237405.2	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	2009V-1085	GCA_000237425.2	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	2009V-1096	GCA_000237445.2	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	2009V-1116	GCA_000237465.2	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	2009V-1131	GCA_000237485.2	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-142	GCA_000788425.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-143	GCA_000788415.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-144	GCA_000788435.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-145	GCA_000788535.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-146	GCA_000788555.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-147	GCA_000788575.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-148	GCA_000788595.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-150	GCA_000788615.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010AA-151	GCA_000788635.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-1749	GCA_000237505.2	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-1786	GCA_000166455.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-1792	GCA_000166495.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-1798	GCA_000166475.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-1961	GCA_000237525.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-2010H	GCA_000237545.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010EL-2010N	GCA_000237565.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2010V-1014	GCA_000237585.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2011EL-1089	GCA_000237605.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2011EL-1137	GCA_000237645.2	2009	Haiti	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	2011EL-301	GCA_000257415.2	2011	Russia	MDR, CTX, Serotyping
Vibrio cholerae	2011V-1021	GCA_000237665.2	2011	Dominican Republic	MDR, CTX, Serotyping
Vibrio cholerae	2012EL-1759	GCA_000710155.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012EL-2176	GCA_000765415.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-131	GCA_000788655.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-2	GCA_000788495.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-32	GCA_000788675.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-326	GCA_000788695.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-9	GCA_000788715.2	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-90	GCA_000788735.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-92	GCA_000788755.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012Env-94	GCA_000788855.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-07	GCA_000788875.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-08	GCA_000789115.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-10	GCA_000789135.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-11	GCA_000789035.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-12	GCA_000789155.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-15	GCA_000789075.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-16	GCA_000789055.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-17	GCA_000788915.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-18	GCA_000788895.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-19	GCA_000789095.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-21	GCA_000788995.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-22	GCA_000789015.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-24	GCA_000788795.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-25	GCA_000788775.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-31	GCA_000788835.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-32	GCA_000788935.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-33	GCA_000788975.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-34	GCA_000788815.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2012HC-35	GCA_000788955.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	2014V-1107	GCA_003311945.1	2014	United States	MDR
Vibrio cholerae	2015V-1076	GCA_003311815.1	2015	United States	MDR
Vibrio cholerae	2016V-1018	GCA_003312035.1	2016	United States	MDR
Vibrio cholerae	2016V-1062	GCA_003311825.1	2016	United States	MDR
Vibrio cholerae	2016V-1091	GCA_003312065.1	2016	United States	MDR
Vibrio cholerae	2016V-1111	GCA_003311965.1	2016	United States	MDR
Vibrio cholerae	2016V-1114	GCA_003312085.1	2016	United States	MDR
Vibrio cholerae	2017V-1038	GCA_003311805.1	2017	United States	MDR
Vibrio cholerae	2017V-1070	GCA_003311865.1	2017	United States	MDR
Vibrio cholerae	2017V-1085	GCA_003311895.1	2017	United States	MDR
Vibrio cholerae	2017V-1105	GCA_003311975.1	2017	United States	MDR
Vibrio cholerae	2017V-1110	GCA_003312005.1	2017	United States	MDR
Vibrio cholerae	2017V-1124	GCA_003311885.1	2017	United States	MDR
Vibrio cholerae	2017V-1144	GCA_003312015.1	2017	United States	MDR
Vibrio cholerae	2017V-1176	GCA_003312095.1	2017	United States	MDR
Vibrio cholerae	20390	GCA_002078815.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	2044	GCA_003013485.1	1966	Iraq	MDR
Vibrio cholerae	20478	GCA_002076785.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	20-a_11	GCA_003056705.1	1995	Ukraine	MDR
Vibrio cholerae	21027	GCA_002078705.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	21B	GCA_002076775.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	220075-6	GCA_002807865.1	2014	Bangladesh	MDR
Vibrio cholerae	220076-6	GCA_002807825.1	2014	Bangladesh	MDR
Vibrio cholerae	22043200_C1	GCA_002808465.1	2013	Bangladesh	MDR
Vibrio cholerae	22043202_C1	GCA_002808165.1	2013	Bangladesh	MDR
Vibrio cholerae	22043204_C1	GCA_002808435.1	2013	Bangladesh	MDR
Vibrio cholerae	22043300_C6	GCA_002808215.1	2013	Bangladesh	MDR
Vibrio cholerae	22044108_C3	GCA_002808275.1	2013	Bangladesh	MDR
Vibrio cholerae	22087102_C2	GCA_002808145.1	2013	Bangladesh	MDR
Vibrio cholerae	22087500_C9	GCA_002808345.1	2013	Bangladesh	MDR
Vibrio cholerae	234-93	GCA_000737005.1	1993	India	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	2403	GCA_002196295.1	ND	Ukraine	MDR
Vibrio cholerae	2479-86	GCA_001857305.1	1986	United States	MDR, CTX, Serotyping
Vibrio cholerae	2497-86	GCA_001857355.1	1987	United States	ContEst16S
Vibrio cholerae	2512-86	GCA_001857245.1	1986	United States	MDR, CTX, Serotyping
Vibrio cholerae	2521-89	GCA_002216685.1	1989	United States	MDR, CTX, Serotyping
Vibrio cholerae	2523-87	GCA_001857345.1	1974	United States	MDR, CTX, Serotyping
Vibrio cholerae	2523-88	GCA_003311755.1	ND	ND	MDR
Vibrio cholerae	254-93	GCA_000737025.1	1993	India	MDR, CTX, Serotyping
Vibrio cholerae	2559-78	GCA_001857145.1	1978	United States	MDR, CTX, Serotyping
Vibrio cholerae	2613	GCA_003057055.1	2015	Russia	MDR
Vibrio cholerae	2631-78	GCA_001857225.1	1978	United States	MDR, CTX, Serotyping
Vibrio cholerae	2633-78	GCA_001857425.1	1978	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	2687	GCA_003057075.1	2015	Russia	MDR
Vibrio cholerae	2688	GCA_003056975.1	2015	Russia	MDR
Vibrio cholerae	2740-80	GCA_001683415.1	1980	United States	MDR, CTX, Serotyping
Vibrio cholerae	28	GCA_002196175.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	2843	GCA_003057115.1	2016	Russia	MDR
Vibrio cholerae	2Mo	GCA_002076705.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	31	GCA_001281585.1	2011	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	3178	GCA_003057035.1	2017	Russia	MDR
Vibrio cholerae	31Ki	GCA_002076535.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	3223-74	GCA_001743085.1	1974	Guam	MDR, CTX, Serotyping
Vibrio cholerae	3225-74	GCA_001857365.1	1974	Guam	MDR, CTX, Serotyping
Vibrio cholerae	3265/80	GCA_000786345.1	2014	Russia	MDR, CTX, Serotyping
Vibrio cholerae	3272-78	GCA_001857265.1	1977	United States	MDR, CTX, Serotyping
Vibrio cholerae	330013_C1	GCA_002808485.1	2013	Bangladesh	MDR
Vibrio cholerae	330033_C1	GCA_002808265.1	2013	Bangladesh	MDR
Vibrio cholerae	330073_A	GCA_002807765.1	2013	Bangladesh	MDR
Vibrio cholerae	330073_B	GCA_002807965.1	2013	Bangladesh	MDR
Vibrio cholerae	330081	GCA_002808105.1	2014	Bangladesh	MDR
Vibrio cholerae	330110	GCA_002807895.1	2014	Bangladesh	MDR
Vibrio cholerae	330113	GCA_002807785.1	2014	Bangladesh	MDR
Vibrio cholerae	330440_C1	GCA_002808065.1	2013	Bangladesh	MDR
Vibrio cholerae	330590	GCA_002807705.1	2014	Bangladesh	MDR
Vibrio cholerae	330898_C2	GCA_002808225.1	2013	Bangladesh	MDR
Vibrio cholerae	330920_A	GCA_002807945.1	2013	Bangladesh	MDR
Vibrio cholerae	330920_B	GCA_002807875.1	2013	Bangladesh	MDR
Vibrio cholerae	331721_C1	GCA_002808415.1	2013	Bangladesh	MDR
Vibrio cholerae	34Kayum	GCA_002196395.1	ND	Afghanistan	MDR, CTX, Serotyping
Vibrio cholerae	May-00	GCA_000237685.2	2005	United States	MDR, CTX, Serotyping
Vibrio cholerae	Jun-46	GCA_000237705.2	2006	United States	MDR, CTX, Serotyping
Vibrio cholerae	Aug-54	GCA_000237725.2	2008	United States	MDR, CTX, Serotyping
Vibrio cholerae	Jul-68	GCA_001857505.1	2007	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	Aug-69	GCA_000237745.2	2008	United States	MDR, CTX, Serotyping
Vibrio cholerae	May-82	GCA_000237765.2	2005	United States	MDR, CTX, Serotyping
Vibrio cholerae	36KI	GCA_002078055.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	39	GCA_001281595.1	2011	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	39	GCA_002204075.1	2011	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	39361	GCA_002078635.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	39Ki	GCA_002076475.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	4110	GCA_001257035.1	1995	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	4111	GCA_001252855.1	2002	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	4113	GCA_001259055.1	2003	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	4121	GCA_001252075.1	2004	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	4122	GCA_001253455.1	2007	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	41D	GCA_003130475.1	1998	Russia	MDR
Vibrio cholerae	4260B	GCA_000330905.1	1993	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4295STDY6534216	GCA_900324445.1	ND	ND	MDR
Vibrio cholerae	4295STDY6534232	GCA_900324425.1	ND	ND	MDR
Vibrio cholerae	4295STDY6534248	GCA_900324455.1	ND	ND	MDR
Vibrio cholerae	43	GCA_001281615.1	1994	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	4322	GCA_001249315.1	2004	India	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	433	GCA_002196305.1	ND	Russia	MDR, CTX, Serotyping
Vibrio cholerae	4339	GCA_001260335.1	2004	India	MDR, CTX, Serotyping
Vibrio cholerae	43Ki	GCA_002078595.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	4488	GCA_001258555.1	2006	India	MDR, CTX, Serotyping
Vibrio cholerae	4519	GCA_001248505.1	2005	India	MDR, CTX, Serotyping
Vibrio cholerae	4536	GCA_001255835.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4538	GCA_001259715.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4551	GCA_001252055.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4552	GCA_001259235.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4585	GCA_001252895.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4593	GCA_001257895.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4600	GCA_001253055.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4605	GCA_001259135.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4623	GCA_001254635.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4642	GCA_001250435.1	2006	India	MDR, CTX, Serotyping
Vibrio cholerae	4646	GCA_001258995.1	2007	India	MDR, CTX, Serotyping
Vibrio cholerae	4656	GCA_001258495.1	2006	India	MDR, CTX, Serotyping
Vibrio cholerae	4661	GCA_001259875.1	2001	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4662	GCA_001260175.1	2001	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4663	GCA_001256015.1	2001	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4672	GCA_001249795.1	2000	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4675	GCA_001254815.1	2001	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	4679	GCA_001247245.1	1999	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	47610	GCA_002076645.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	47623	GCA_002076485.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	4784	GCA_001254335.1	2009	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	48055	GCA_002076315.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	490-93	GCA_000737015.1	1993	Thailand	MDR, CTX, Serotyping
Vibrio cholerae	5	GCA_002196165.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	5/66	GCA_000754625.1	1966	Pakistan	MDR, CTX, Serotyping
Vibrio cholerae	5473-62	GCA_000736795.1	1962	Philippines	MDR, CTX, Serotyping
Vibrio cholerae	56	GCA_002204105.1	1995	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	56	GCA_001281665.1	1994	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	571-88	GCA_000736945.1	1988	China	MDR, CTX, Serotyping
Vibrio cholerae	5879	GCA_002911455.1	1972	Russia	MDR
Vibrio cholerae	5Mo	GCA_002076665.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	6/67	GCA_001641745.1	1967	India	MDR, CTX, Serotyping
Vibrio cholerae	60555434	GCA_003260135.1	2017	Australia	MDR
Vibrio cholerae	617	GCA_002114205.1	ND	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	6191	GCA_001249515.1	2005	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6193	GCA_001257835.1	2005	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6194	GCA_001251935.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6197	GCA_001254955.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6201	GCA_001261555.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6210	GCA_001259475.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6212	GCA_001251975.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6214	GCA_001248645.1	2007	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	6215	GCA_001252775.1	2005	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	623-39	GCA_000154005.2	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	63-93 (MO45)	GCA_000736845.1	1992	India	MDR, CTX, Serotyping
Vibrio cholerae	692-79	GCA_001857285.1	1979	United States	MDR, CTX, Serotyping
Vibrio cholerae	76	GCA_001899465.1	2011	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	7685	GCA_001255915.1	2009	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	7686	GCA_001258535.1	2009	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	7687	GCA_001251435.1	2009	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	7714	GCA_002076415.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	7Mo	GCA_002076615.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	8	GCA_003057015.1	2014	Russia	MDR
Vibrio cholerae	81	GCA_000786335.1	2014	Russia	MDR, CTX, Serotyping
Vibrio cholerae	85	GCA_002196255.1	ND	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	857	GCA_001729125.1	1996	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	866	GCA_002204085.1	1996	Ukraine	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	87395	GCA_000348085.2	1983	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	8-76	GCA_000736935.1	1976	India	MDR, CTX, Serotyping
Vibrio cholerae	877-163	GCA_001402745.1	2002	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	89	GCA_002196135.1	2016	Ukraine	MDR, CTX, Serotyping
Vibrio cholerae	8Mo	GCA_002076695.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	9507	GCA_003096135.1	1974	Russia	MDR
Vibrio cholerae	95412	GCA_000348105.2	1987	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	981-75	GCA_000736925.1	1975	India	MDR, CTX, Serotyping
Vibrio cholerae	984-81	GCA_000736775.1	1981	India	MDR, CTX, Serotyping
Vibrio cholerae	9Mo	GCA_002076635.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	A10	GCA_001254655.1	1979	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A103	GCA_001254575.1	1990	ND	MDR, CTX, Serotyping
Vibrio cholerae	A131	GCA_001259315.1	1989	India	MDR, CTX, Serotyping
Vibrio cholerae	A152	GCA_001252875.1	1991	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	A154	GCA_001253155.1	1991	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	A1552	GCA_002997215.1	1992	Peru	MDR
Vibrio cholerae	A1552	GCA_002892855.1	1992	United States	MDR
Vibrio cholerae	A1552	GCA_003097695.1	1992	United States	MDR
Vibrio cholerae	A177	GCA_001249995.1	1992	Colombia	MDR, CTX, Serotyping
Vibrio cholerae	A18	GCA_001252495.1	1977	India	MDR, CTX, Serotyping
Vibrio cholerae	A185	GCA_001253295.1	1992	Colombia	MDR, CTX, Serotyping
Vibrio cholerae	A186	GCA_001248135.1	1992	Argentina	MDR, CTX, Serotyping
Vibrio cholerae	A19	GCA_001250235.2	1971	ND	MDR, CTX, Serotyping
Vibrio cholerae	A193	GCA_001248865.1	1992	Bolivia	MDR, CTX, Serotyping
Vibrio cholerae	A200	GCA_001255295.1	1992	Argentina	MDR, CTX, Serotyping
Vibrio cholerae	A201	GCA_001261515.1	1992	Argentina	MDR, CTX, Serotyping
Vibrio cholerae	A213	GCA_001248945.1	1984	Georgia	MDR, CTX, Serotyping
Vibrio cholerae	A215	GCA_001259995.1	1985	United States	MDR, CTX, Serotyping
Vibrio cholerae	A22	GCA_001255155.1	1979	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A241	GCA_001256675.1	1989	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	A245	GCA_001261135.1	1989	Vietnam	MDR, CTX, Serotyping
Vibrio cholerae	A27	GCA_001260995.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	A29	GCA_001253235.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	A3_296	GCA_003057775.1	2017	Brazil	MDR
Vibrio cholerae	A31	GCA_001253695.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	A32	GCA_001250455.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	A325	GCA_001254095.1	1993	Argentina	MDR, CTX, Serotyping
Vibrio cholerae	A330	GCA_001257215.1	1993	India	MDR, CTX, Serotyping
Vibrio cholerae	A346(1)	GCA_001247525.1	1994	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A383	GCA_001257975.1	2002	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A389	GCA_001259795.1	1987	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A4	GCA_001254055.1	1973	ND	MDR, CTX, Serotyping
Vibrio cholerae	A46	GCA_001259555.1	1964	ND	MDR, CTX, Serotyping
Vibrio cholerae	A487(1)	GCA_001261535.1	2007	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A488(1)	GCA_001257075.1	2006	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A488(2)	GCA_001250615.1	2006	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A49	GCA_001253835.1	1962	ND	MDR, CTX, Serotyping
Vibrio cholerae	A5	GCA_001254675.1	1989	Angola	MDR, CTX, Serotyping
Vibrio cholerae	A59	GCA_001254535.1	1970	India	MDR, CTX, Serotyping
Vibrio cholerae	A6	GCA_001255575.1	1957	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	A60	GCA_001248195.1	1958	Thailand	MDR, CTX, Serotyping
Vibrio cholerae	A61	GCA_001250935.1	1970	India	MDR, CTX, Serotyping
Vibrio cholerae	A66	GCA_001260915.1	1962	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A68	GCA_001259635.1	1949	Egypt	MDR, CTX, Serotyping
Vibrio cholerae	A70	GCA_001248905.1	1969	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	A76	GCA_001259495.1	1982	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	AG-7404	GCA_000348125.2	1991	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	AG-8040	GCA_000348145.2	1991	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	AM-19226	GCA_000153785.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	Amazonia	GCA_000223095.2	1991	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	ATCC 11629	GCA_001471455.2	ND	ND	MDR
Vibrio cholerae	ATCC 14035 (T)	GCA_000621645.1	ND	ND	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	B33	GCA_000174315.1	2004	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	BJG-01	GCA_000221465.1	ND	United States	MDR, CTX, Serotyping
Vibrio cholerae	BRV8	GCA_001292785.1	ND	United Kingdom	MDR, CTX, Serotyping
Vibrio cholerae	BX 330286	GCA_000174335.1	1986	Australia	MDR, CTX, Serotyping
Vibrio cholerae	C5	GCA_001887395.1	1957	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	C6706	GCA_001857435.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	CIRS 101	GCA_000175695.1	2002	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0005	GCA_002099125.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0008	GCA_002099115.1	2005	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0010	GCA_002099095.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0014	GCA_002099065.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0015	GCA_002099055.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0016	GCA_002099035.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0017	GCA_002099015.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0018	GCA_002098995.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0019	GCA_002098965.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0034	GCA_002098955.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0035	GCA_002098935.1	2002	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0074	GCA_002098915.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0079	GCA_002098875.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_0091	GCA_002098885.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_091	GCA_002098845.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_100	GCA_002098835.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_101	GCA_002098765.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1019828.5	GCA_002098805.1	2010	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1019829.2	GCA_002098795.1	2010	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1020229.6	GCA_002098755.1	2010	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1020231.9	GCA_002098715.1	2010	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1020234.0	GCA_002098705.1	2010	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_105	GCA_002098695.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_1163068.5	GCA_002097815.1	2012	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_120	GCA_002098675.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_121	GCA_002098655.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_122	GCA_002098605.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_134	GCA_002098625.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_146	GCA_002098595.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_147	GCA_002098555.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_151	GCA_002098525.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_152	GCA_002098495.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_153	GCA_002098535.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_154	GCA_002098515.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_178	GCA_002098445.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_179	GCA_002098435.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_188	GCA_002098415.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_189	GCA_002098425.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_191	GCA_002098365.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_196	GCA_002098355.1	2003	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_296	GCA_002098305.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300043	GCA_002098295.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300055	GCA_002097735.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300205	GCA_002097745.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300208	GCA_002098345.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300209	GCA_002098335.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300215	GCA_002098255.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_300506	GCA_002097755.1	2008	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_302015	GCA_002098225.1	2009	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_302029	GCA_002098235.1	2009	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_326	GCA_002098215.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_347	GCA_002098195.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_374	GCA_002098155.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_375	GCA_002098145.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_382	GCA_002098135.1	ND	Mozambique	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	CISM_398	GCA_002098085.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_399	GCA_002098075.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_420	GCA_002098055.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_505	GCA_002098065.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_510	GCA_002098005.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_511	GCA_002097985.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_655630.3	GCA_002097975.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_655664.3	GCA_002097995.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_655665.0	GCA_002097925.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_710180.8	GCA_002097895.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_740115.4	GCA_002097905.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_769845.7	GCA_002097915.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_770067.4	GCA_002097845.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_770180.8	GCA_002097835.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_780298.0	GCA_002097825.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CISM_S/Nida	GCA_002097765.1	ND	Mozambique	MDR, CTX, Serotyping
Vibrio cholerae	CMR001	GCA_001860225.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR004	GCA_001858585.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR007	GCA_001860265.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR008	GCA_001860285.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR009	GCA_001860295.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR010	GCA_001860315.1	2010	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR011	GCA_001860345.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR012	GCA_001860365.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR013	GCA_001860385.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR014	GCA_001860395.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR015	GCA_001860425.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR016	GCA_001860445.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR017	GCA_001860465.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR018	GCA_001860485.1	ND	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR019	GCA_001858475.1	ND	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR020	GCA_001858445.1	ND	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR021	GCA_001858455.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CMR022	GCA_001858465.1	2011	Cameroon	MDR, CTX, Serotyping
Vibrio cholerae	CP1030(3)	GCA_000279555.1	2008	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	CP1032(5)	GCA_000279305.1	1991	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	CP1033(6)	GCA_000304755.1	2000	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	CP1035(8)	GCA_000304915.2	2004	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	CP1037(10)	GCA_000302965.1	2003	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	CP1038(11)	GCA_000279325.1	2003	Zimbabwe	MDR, CTX, Serotyping
Vibrio cholerae	CP1040(13)	GCA_000302985.1	2004	Zambia	MDR, CTX, Serotyping
Vibrio cholerae	CP1041(14)	GCA_000279245.1	2004	Zambia	MDR, CTX, Serotyping
Vibrio cholerae	CP1042(15)	GCA_000279345.1	2010	Thailand	MDR, CTX, Serotyping
Vibrio cholerae	CP1044(17)	GCA_000303045.1	1991	Peru	MDR, CTX, Serotyping
Vibrio cholerae	CP1046(19)	GCA_000281655.1	1995	Peru	MDR, CTX, Serotyping
Vibrio cholerae	CP1047(20)	GCA_000279785.1	1995	Peru	MDR, CTX, Serotyping
Vibrio cholerae	CP1048(21)	GCA_000279395.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	CP1050(23)	GCA_000303065.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	CP1110	GCA_000387585.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1111	GCA_000387625.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1112	GCA_000387645.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1113	GCA_000387665.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1114	GCA_000387685.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1115	GCA_000387605.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1116	GCA_000387725.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CP1117	GCA_000387705.1	2011	United States	MDR, CTX, Serotyping
Vibrio cholerae	CRC1106	GCA_001887455.1	1962	India	MDR, CTX, Serotyping
Vibrio cholerae	CRC711	GCA_001887435.1	1964	India	MDR, CTX, Serotyping
Vibrio cholerae	CW-6	GCA_001617665.1	1966	India	MDR, CTX, Serotyping
Vibrio cholerae	D-35	GCA_000961975.1	1958	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	DL4211	GCA_001953365.1	2008	United States	MDR, CTX, Serotyping
Vibrio cholerae	DL4215	GCA_001953375.1	2008	United States	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	Drakes2013	GCA_001543505.1	2013	United States	MDR, CTX, Serotyping
Vibrio cholerae	E1162	GCA_001887495.1	1962	China	MDR, CTX, Serotyping
Vibrio cholerae	E1320	GCA_001887415.1	1974	China	MDR, CTX, Serotyping
Vibrio cholerae	E306	GCA_000487955.1	2013	China	MDR, CTX, Serotyping
Vibrio cholerae	E506	GCA_001887475.1	1974	United States	MDR, CTX, Serotyping
Vibrio cholerae	E7946	GCA_002749635.1	1978	Bahrain	MDR
Vibrio cholerae	E9120	GCA_001887655.1	1961	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	EC-0009	GCA_000348165.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EC-0012	GCA_000348185.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EC-0027	GCA_000348205.2	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EC-0051	GCA_000348225.2	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EC-051	GCA_001282605.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EDC-020	GCA_000348245.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EDC-022	GCA_000348265.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1536	GCA_000348285.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1542	GCA_001187255.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1543	GCA_001186515.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1546	GCA_000348305.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1626	GCA_001186485.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1652A	GCA_001186505.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1654	GCA_001186575.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1676A	GCA_000348345.2	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1688	GCA_001186565.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1690	GCA_001186595.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1690A	GCA_001186585.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1706	GCA_001186645.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	EM-1727	GCA_000348365.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	Env-390	GCA_001854425.1	2012	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	FC1105	GCA_002194295.1	2003	India	MDR, CTX, Serotyping
Vibrio cholerae	FC1225	GCA_002194335.1	2001	India	MDR, CTX, Serotyping
Vibrio cholerae	FC1341	GCA_002194265.1	2002	India	MDR, CTX, Serotyping
Vibrio cholerae	FC1384	GCA_002194245.1	2000	India	MDR, CTX, Serotyping
Vibrio cholerae	FC1817	GCA_002194305.1	1994	India	MDR, CTX, Serotyping
Vibrio cholerae	FC1877	GCA_002194155.1	1995	India	MDR, CTX, Serotyping
Vibrio cholerae	FC2271	GCA_002194235.1	1997	India	MDR, CTX, Serotyping
Vibrio cholerae	FC2273	GCA_002194215.1	1998	India	MDR, CTX, Serotyping
Vibrio cholerae	FC3611a	GCA_002194165.1	1999	India	MDR, CTX, Serotyping
Vibrio cholerae	FC3611b	GCA_002194185.1	1997	India	MDR, CTX, Serotyping
Vibrio cholerae	FDAARGOS_102	GCA_001525525.2	1963	India	MDR, CTX, Serotyping
Vibrio cholerae	FDAARGOS_103	GCA_001471585.2	ND	Germany	MDR, CTX, Serotyping
Vibrio cholerae	FDAARGOS_223	GCA_002073335.2	ND	United States	MDR
Vibrio cholerae	FJ147	GCA_000963555.1	2005	China	MDR, CTX, Serotyping
Vibrio cholerae	FORC_055	GCA_002313025.1	2014	South Korea	MDR, CTX, Serotyping
Vibrio cholerae	G_33	GCA_002102575.1	1986	Guinea	MDR, CTX, Serotyping
Vibrio cholerae	G4222	GCA_000338075.1	2001	South Africa	MDR, CTX, Serotyping
Vibrio cholerae	GP140	GCA_001253315.1	1978	Malaysia	MDR, CTX, Serotyping
Vibrio cholerae	GP143	GCA_001260075.1	1978	Bahrain	MDR, CTX, Serotyping
Vibrio cholerae	GP145	GCA_001250035.1	1979	India	MDR, CTX, Serotyping
Vibrio cholerae	GP152	GCA_001249715.1	1979	India	MDR, CTX, Serotyping
Vibrio cholerae	GP16	GCA_001251495.1	1971	India	MDR, CTX, Serotyping
Vibrio cholerae	GP160	GCA_001254435.1	1980	India	MDR, CTX, Serotyping
Vibrio cholerae	GP60	GCA_001261335.1	1973	India	MDR, CTX, Serotyping
Vibrio cholerae	GP8	GCA_001253575.1	1970	India	MDR, CTX, Serotyping
Vibrio cholerae	H1	GCA_000275645.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-02A1	GCA_000221445.1	ND	Haiti	ContEst16S
Vibrio cholerae	HC-02C1	GCA_000305525.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-06A1	GCA_000234375.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC1037	GCA_002946655.1	2014	Haiti	MDR
Vibrio cholerae	HC-17A1	GCA_000304935.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-17A2	GCA_000305675.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-19A1	GCA_000234965.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-1A2	GCA_000304775.1	2010	Haiti	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	HC-20A2	GCA_000279415.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-21A1	GCA_000234945.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-22A1	GCA_000234925.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-23A1	GCA_000234395.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-28A1	GCA_000234415.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-32A1	GCA_000234905.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-33A2	GCA_000234885.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-36A1	GCA_000474965.1	2010	ND	MDR, CTX, Serotyping
Vibrio cholerae	HC-37A1	GCA_000305585.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-38A1	GCA_000221485.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-39A1	GCA_000302775.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-40A1	GCA_000221345.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-41A1	GCA_000302755.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-41B1	GCA_000304955.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-42A1	GCA_000279185.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-43A1	GCA_000234435.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-43B1	GCA_000279435.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-44C1	GCA_000305565.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-46A1	GCA_000279455.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-46B1	GCA_000305605.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-47A1	GCA_000279955.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-48A1	GCA_000221365.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-48B2	GCA_000234865.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-49A2	GCA_000220725.2	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-50A1	GCA_000302835.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-50A2	GCA_000304995.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-51A1	GCA_000303105.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-52A1	GCA_000302855.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-55A1	GCA_000302875.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-55B2	GCA_000305645.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-55C2	GCA_000305015.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-56A1	GCA_000302895.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-56A2	GCA_000279205.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-57A1	GCA_000303005.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-57A2	GCA_000279375.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-59A1	GCA_000305195.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-59B1	GCA_000305545.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-60A1	GCA_000305055.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-61A1	GCA_000234455.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	HC-61A2	GCA_000304795.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-62A1	GCA_000305075.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-62B1	GCA_000305625.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-64A1	GCA_000327105.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-65A1	GCA_000327125.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-67A1	GCA_000327145.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-68A1	GCA_000327165.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-69A1	GCA_000305695.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-70A1	GCA_000221385.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-71A1	GCA_000327185.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-72A2	GCA_000327205.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-77A1	GCA_000305095.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-78A1	GCA_000327225.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-7A1	GCA_000318485.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-80A1	GCA_000327245.3	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-81A1	GCA_000318505.2	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HC-81A2	GCA_000303125.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HCUF01	GCA_000220745.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	HE-09	GCA_000221405.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE-16	GCA_000303085.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE-25	GCA_000279265.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE39	GCA_000220765.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	HE-40	GCA_000305115.2	2010	Haiti	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	HE-45	GCA_000279285.1	2010	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE46	GCA_001857515.1	2011	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE-46	GCA_000305135.2	2011	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HE48	GCA_000220785.2	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	HFU-02	GCA_000221425.1	ND	Haiti	MDR, CTX, Serotyping
Vibrio cholerae	I-1181	GCA_001597715.1	1994	Russia	MDR, CTX, Serotyping
Vibrio cholerae	I-1187	GCA_001661905.1	1994	Russia	MDR, CTX, Serotyping
Vibrio cholerae	I-1263	GCA_000735705.1	1997	Russia	MDR, CTX, Serotyping
Vibrio cholerae	I-1300	GCA_000967785.1	1999	Russia	MDR, CTX, Serotyping
Vibrio cholerae	I-1471	GCA_000818865.1	2011	Russia	MDR, CTX, Serotyping
Vibrio cholerae	ICDC-VC661	GCA_002313005.1	2006	China	MDR, CTX, Serotyping
Vibrio cholerae	IDH-06787	GCA_002899735.1	2014	India	MDR
Vibrio cholerae	IDHO1_726	GCA_001247885.1	2009	India	MDR, CTX, Serotyping
Vibrio cholerae	IEC224	GCA_000250855.1	1990s	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	InDRE 3140	GCA_000740515.2	2013	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	InDRE 4262	GCA_000953775.1	2013	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	InDRE 4354	GCA_000953755.1	2013	Mexico	MDR, CTX, Serotyping
Vibrio cholerae	J8YRS KAGUNGA	GCA_002078795.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	KW3	GCA_001318185.1	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	L11	GCA_001718105.1	2006	Sweden	MDR, CTX, Serotyping
Vibrio cholerae	L15	GCA_001718095.1	2006	Sweden	MDR, CTX, Serotyping
Vibrio cholerae	L-3226	GCA_000600255.1	2010	Russia	MDR, CTX, Serotyping
Vibrio cholerae	LMA3984-4	GCA_000195065.1	ND	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	M1030	GCA_002196335.1	ND	Turkmenistan	MDR, CTX, Serotyping
Vibrio cholerae	M1275	GCA_001517845.1	1993	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M-1293	GCA_000705295.1	1994	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1327	GCA_001641765.1	1998	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1332	GCA_003056995.1	2000	Russia	MDR
Vibrio cholerae	M1337	GCA_002196375.1	ND	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1344	GCA_002196275.1	ND	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M139	GCA_001637545.1	1965	Turkmenistan	MDR, CTX, Serotyping
Vibrio cholerae	M1395	GCA_001515105.1	1981	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1399	GCA_001515085.1	1982	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1425	GCA_003056955.1	2003	Russia	MDR
Vibrio cholerae	M1429	GCA_000960915.1	2004	Russia	ContEst16S
Vibrio cholerae	M1501	GCA_001637575.1	2011	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1518	GCA_001641685.1	2012	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1522	GCA_001515165.1	2014	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M1524	GCA_001641705.1	2013	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M2140	GCA_001887635.1	1977	Australia	MDR, CTX, Serotyping
Vibrio cholerae	M29	GCA_000709105.1	1942	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M299	GCA_001637555.1	1965	Turkmenistan	MDR, CTX, Serotyping
Vibrio cholerae	M66-2	GCA_000021605.1	1937	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	M818	GCA_000966385.1	1970	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M888	GCA_001521835.1	1970	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M888D	GCA_001617675.1	1970	Russia	MDR, CTX, Serotyping
Vibrio cholerae	M988	GCA_001515115.1	1972	Turkmenistan	MDR, CTX, Serotyping
Vibrio cholerae	MAK 676	GCA_000753725.1	1937	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	MAK 757	GCA_000153865.1	1937	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	MAK 97	GCA_000939665.1	1937	Indonesia	MDR, CTX, Serotyping
Vibrio cholerae	MBN17	GCA_001250795.1	2004	India	MDR, CTX, Serotyping
Vibrio cholerae	MBRN14	GCA_001249085.1	2004	India	MDR, CTX, Serotyping
Vibrio cholerae	ME-7	GCA_001515095.1	1966	India	ContEst16S
Vibrio cholerae	MG116025	GCA_001254895.1	1991	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	MG116226	GCA_001254355.1	1991	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	MJ-1236	GCA_000022585.1	1994	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	MJ1485	GCA_001250195.1	1994	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	MO10	GCA_000152425.1	1992	India	MDR, CTX, Serotyping
Vibrio cholerae	MS6	GCA_000829215.1	2008	Myanmar	MDR, CTX, Serotyping
Vibrio cholerae	MZO-2	GCA_000153985.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	MZO-3	GCA_000168935.3	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	N16961	GCA_000006745.1	1975	Bangladesh	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	N16961	GCA_003063785.1	1975	Bangladesh	MDR
Vibrio cholerae	NCTC 5395	GCA_001887515.1	1938	Iraq	MDR, CTX, Serotyping
Vibrio cholerae	NCTC 9420	GCA_001887615.1	1954	Egypt	MDR, CTX, Serotyping
Vibrio cholerae	Nep-21106	GCA_000348465.2	2003	Nepal	MDR, CTX, Serotyping
Vibrio cholerae	Nep-21113	GCA_000348485.2	2003	Nepal	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-004A	GCA_000348385.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-006C	GCA_000348405.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-008D	GCA_000348425.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-010F	GCA_000348445.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-011	GCA_001186655.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-019	GCA_001186755.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-021	GCA_001186725.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-04	GCA_001186665.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-042	GCA_001186735.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-048	GCA_001186675.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-05	GCA_001186785.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-068	GCA_001187185.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-078	GCA_001186495.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-079	GCA_001187225.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-080	GCA_001187245.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-081	GCA_001186805.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCC-083	GCA_001186835.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-01	GCA_001186825.1	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-012	GCA_001186915.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-013	GCA_001186925.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-016A	GCA_001186965.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-017	GCA_001186985.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-02	GCA_001186855.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-029	GCA_001186995.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-03	GCA_001187265.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-033	GCA_001187065.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-037	GCA_001187025.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-04	GCA_001186905.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-043	GCA_001187085.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-044	GCA_001187015.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-045	GCA_001187095.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-047	GCA_001187145.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-048	GCA_001187175.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-053	GCA_001187105.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-054	GCA_001187165.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NHCM-06	GCA_001186885.1	2011	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	NIH41	GCA_000736865.1	1941	India	MDR, CTX, Serotyping
Vibrio cholerae	NMH2016	GCA_002251495.1	2016	United States	MDR, CTX, Serotyping
Vibrio cholerae	O1S	GCA_002076155.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O2	GCA_002076255.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O395	GCA_000021625.1	1965	ND	MDR, CTX, Serotyping
Vibrio cholerae	O3MU	GCA_002076465.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O3S	GCA_002076575.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O5MU	GCA_002076585.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O6MU	GCA_002076175.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O7MU	GCA_002076545.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O7S	GCA_002076165.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	O9S	GCA_002076245.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	18963	GCA_003096115.1	2007	Russia	MDR
Vibrio cholerae	OO4	GCA_002076455.1	2015	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	OYP1E07	GCA_002284455.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP1G01	GCA_002284495.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP2A12	GCA_002284395.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP2C05	GCA_002284475.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP2D07	GCA_002284425.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP2E01	GCA_002284415.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP3F10	GCA_002284355.1	2009	United States	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	OYP4B01	GCA_002284365.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP4G08	GCA_002284325.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP4H06	GCA_002284315.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP4H08	GCA_002284255.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP4H11	GCA_002284245.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP5F10	GCA_002284235.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP6D06	GCA_002284205.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP6E07	GCA_002284185.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP6F08	GCA_002284175.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP6F10	GCA_002284265.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP6G08	GCA_002284155.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP7C09	GCA_002284075.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP8A01	GCA_002284125.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP8C06	GCA_002284095.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	OYP8F12	GCA_002284115.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	P13762	GCA_001639085.1	1988	Uzbekistan	MDR, CTX, Serotyping
Vibrio cholerae	P-18748	GCA_002196055.1	2004	Russia	MDR, CTX, Serotyping
Vibrio cholerae	P-18778	GCA_002196065.1	2005	Russia	MDR, CTX, Serotyping
Vibrio cholerae	P-18785	GCA_000338215.2	2005	Russia	MDR, CTX, Serotyping
Vibrio cholerae	P18899	GCA_000966395.1	2006	Russia	MDR, CTX, Serotyping
Vibrio cholerae	P18899-D	GCA_000966375.1	2006	Russia	MDR, CTX, Serotyping
Vibrio cholerae	PCS-022	GCA_000569115.2	ND	ND	MDR, CTX, Serotyping
Vibrio cholerae	PCS-023	GCA_000348505.2	2010	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	PhVC-311	GCA_001027505.1	2011	Philippines	MDR, CTX, Serotyping
Vibrio cholerae	PhVC-326	GCA_001027485.1	2011	Philippines	MDR, CTX, Serotyping
Vibrio cholerae	PhVE-5	GCA_001027495.1	2011	Philippines	MDR, CTX, Serotyping
Vibrio cholerae	PIC018	GCA_001543465.1	2007	Bangladesh	MDR, CTX, Serotyping
Vibrio cholerae	PRL5	GCA_001256355.1	1980	India	MDR, CTX, Serotyping
Vibrio cholerae	PRL64	GCA_001261075.1	1992	India	MDR, CTX, Serotyping
Vibrio cholerae	PS15	GCA_000318075.1	ND	United States	MDR, CTX, Serotyping
Vibrio cholerae	R17644	GCA_000965285.1	1997	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RC9	GCA_000174275.1	1985	Kenya	MDR, CTX, Serotyping
Vibrio cholerae	RND18826	GCA_000500735.1	2005	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND18899	GCA_000500695.1	2006	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND19187	GCA_000500675.1	2010	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND19188	GCA_000710445.1	2010	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND19191	GCA_000710455.1	2010	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND6878	GCA_000500715.1	2012	Russia	MDR, CTX, Serotyping
Vibrio cholerae	RND81	GCA_000763075.1	2014	Russia	MDR, CTX, Serotyping
Vibrio cholerae	S000100_C5	GCA_002808365.1	2013	Bangladesh	MDR
Vibrio cholerae	S000600_C10	GCA_002808075.1	2013	Bangladesh	MDR
Vibrio cholerae	S002300_B	GCA_002807835.1	2013	Bangladesh	MDR
Vibrio cholerae	S002300_E	GCA_002807985.1	2013	Bangladesh	MDR
Vibrio cholerae	S002502	GCA_002807725.1	2013	Bangladesh	MDR
Vibrio cholerae	S002506	GCA_002807735.1	2013	Bangladesh	MDR
Vibrio cholerae	S002604	GCA_002808155.1	2014	Bangladesh	MDR
Vibrio cholerae	S003008	GCA_002808205.1	2014	Bangladesh	MDR
Vibrio cholerae	S003202	GCA_002807975.1	2014	Bangladesh	MDR
Vibrio cholerae	S003806	GCA_002807805.1	2014	Bangladesh	MDR
Vibrio cholerae	S023202	GCA_002808125.1	2014	Bangladesh	MDR
Vibrio cholerae	S023208	GCA_002807925.1	2014	Bangladesh	MDR
Vibrio cholerae	S040602_C1	GCA_002808325.1	2013	Bangladesh	MDR
Vibrio cholerae	S042100	GCA_002808355.1	2013	Bangladesh	MDR
Vibrio cholerae	S042408	GCA_002808305.1	2014	Bangladesh	MDR
Vibrio cholerae	S081300_C2	GCA_002808405.1	2013	Bangladesh	MDR
Vibrio cholerae	S12	GCA_001735565.1	2009	Australia	MDR, CTX, Serotyping
Vibrio cholerae	Sa5Y	GCA_003063885.1	2004	United States	MDR
Vibrio cholerae	SIO	GCA_001857455.1	2000	United States	MDR, CTX, Serotyping
Vibrio cholerae	TEM/04/01-001	GCA_002076745.1	2012	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	TEM/10/01-002	GCA_002076265.1	2012	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	TEM/12/12-001	GCA_002076735.1	2011	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	TEM/15/01-005	GCA_002078695.1	2012	Tanzania	MDR, CTX, Serotyping

Species Name	Strain Name	Accession	Isolated Year	Country	Studies
Vibrio cholerae	TEM/25/01-004	GCA_002076235.1	2012	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	TEM/29/01-003	GCA_002078755.1	2012	Tanzania	MDR, CTX, Serotyping
Vibrio cholerae	TM 11079-80	GCA_000174255.1	1980	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	TMA 21	GCA_000174295.1	1982	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	TP	GCA_001857485.1	2000	United States	MDR, CTX, Serotyping
Vibrio cholerae	TSY216	GCA_001045415.1	2010	Thailand	MDR, CTX, Serotyping
Vibrio cholerae	UG010	GCA_003205765.1	2016	Uganda	MDR
Vibrio cholerae	UG020	GCA_003205635.1	2016	Uganda	MDR
Vibrio cholerae	UG026	GCA_003205685.1	2014	Uganda	MDR
Vibrio cholerae	UG040	GCA_003205735.1	2015	Uganda	MDR
Vibrio cholerae	UG042	GCA_003205655.1	2015	Uganda	MDR
Vibrio cholerae	UG046	GCA_003205565.1	2015	Uganda	MDR
Vibrio cholerae	UG054	GCA_003205555.1	2015	Uganda	MDR
Vibrio cholerae	UG060	GCA_003205755.1	2014	Uganda	MDR
Vibrio cholerae	UG071	GCA_003205705.1	2014	Uganda	MDR
Vibrio cholerae	UG086	GCA_003205675.1	2015	Uganda	MDR
Vibrio cholerae	V109	GCA_001257255.1	1990	India	MDR, CTX, Serotyping
Vibrio cholerae	V212-1	GCA_001248465.1	1991	India	MDR, CTX, Serotyping
Vibrio cholerae	V5	GCA_001252675.1	1989	India	MDR, CTX, Serotyping
Vibrio cholerae	V51	GCA_000152465.2	1987	United States	MDR, CTX, Serotyping
Vibrio cholerae	V52	GCA_000167935.2	ND	Sudan	MDR, CTX, Serotyping
Vibrio cholerae	VC0101557	GCA_002407455.1	2001	South Korea	MDR, CTX, Serotyping
Vibrio cholerae	VC1761	GCA_000299515.2	2009	Malaysia	MDR, CTX, Serotyping
Vibrio cholerae	VC22	GCA_001729195.1	1981	United States	MDR, CTX, Serotyping
Vibrio cholerae	VC35	GCA_000299495.2	2004	Malaysia	MDR, CTX, Serotyping
Vibrio cholerae	VC4370	GCA_000299535.2	2008	Malaysia	MDR, CTX, Serotyping
Vibrio cholerae	VC48	GCA_001857165.1	1981	United States	MDR, CTX, Serotyping
Vibrio cholerae	VC53	GCA_001857155.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	VC56	GCA_001857175.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	VCC19	GCA_000438805.2	1994	Brazil	MDR, CTX, Serotyping
Vibrio cholerae	VcN1	GCA_002843255.1	2017	Bangladesh	MDR
Vibrio cholerae	BC1071	GCA_900185995.1	ND	ND	MDR
Vibrio cholerae	VL426	GCA_000174235.1	ND	United Kingdom	MDR, CTX, Serotyping
Vibrio cholerae	W4-13	GCA_002217575.1	2013	India	MDR, CTX, Serotyping
Vibrio cholerae	YB1A01	GCA_001402185.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB1G06	GCA_001402365.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB2A05	GCA_001402535.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB2A06	GCA_001402375.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB2G01	GCA_001411585.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB2G05	GCA_001402415.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB2G07	GCA_001402425.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB3B05	GCA_001402545.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB3G04	GCA_001402275.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4B03	GCA_001402605.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4C07	GCA_001402285.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4F05	GCA_001402265.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4G05	GCA_001402575.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4G06	GCA_001402255.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB4H02	GCA_001402585.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB5A06	GCA_001402435.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB6A06	GCA_001402445.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB7A06	GCA_001402335.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB7A09	GCA_001402595.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YB8E08	GCA_001402655.1	2009	United States	MDR, CTX, Serotyping
Vibrio cholerae	YN2011004	GCA_001029975.1	2011	China	MDR, CTX, Serotyping
Vibrio cholerae	YN89004	GCA_001030035.1	1989	China	MDR, CTX, Serotyping
Vibrio cholerae	YN97083	GCA_001030015.1	1997	China	MDR, CTX, Serotyping
Vibrio cholerae	YN98296	GCA_001184775.1	1998	China	MDR, CTX, Serotyping
Vibrio cholerae	ZWU0020	GCA_000812045.1	ND	United States	MDR, CTX, Serotyping

국문 초록 (Abstract in Korean)

최근의 유전체 시퀀싱 기술의 발전으로 유전체학은 다양한 미생물학에서 중요한 역할을 담당해 왔다. 방대한 양의 유전체 데이터를 분석하기 위해서는 적당한 알고리즘의 이용과 적절한 생물정보학적 도구들의 개발이 절실하다. 세균의 유전체 분석 절차는 어셈블리, 유전자 탐사, 그리고 유전자 표지 순으로 진행 된다. 두 개 또는 그 이상의 유전체를 서로간 비교하는 것은 비교 유전체학이라고 한다. 비교유전체학의 목적은 다중 유전체를 비교하여 생물학적 함의와 생물학적 표지 등의 비교와 예측이다. 본 연구에서는 이러한 비교 유전체학의 목적에 맞는 세 가지 프로그램을 개발하였다.

최근의 세균의 종 개념은 이전에 사용된 표현형을 이용한 방법 보다 객관적인 유전체를 이용한 관련성 연구에 기반한다. 현재 OGRI (Overall Genomic Relatedness Index)라고 불리는 쌍방향 유전체 서열 유사성은 세균 분류학 및 식별에 사용되고 있다. OGRI를 계산하는 데 가장 널리 사용되는 알고리즘은 Average Nucleotide identity (ANI)이다. 그러나 BLAST를 사용하는 기존 ANI는 쿼리 시퀀스의 선택에 따라 상호 계산에서 서로 다른 값을 산출했다. 이러한 불일치를 해결하기 위해 본 연구에서는 orthology를 기반으로 한 새로운 알고리즘은 OrthoANI라는 개발 되었다. 기존 알고리즘에서 쿼리 유전체와 대상 유전체 간 ANI값은 쿼리 유전체 만을 조각 내었지만, 새로운 알고리즘에서는 쿼리와 대상 유전체 모두를 조각 낸다. 유사성은 오로지 양방향으로 orthology가 있을 때만 계산하는 것으로 한다. OrthoANI는 기존의 ANI와 상관 관계를 잘 이루며, 양 방향 값 또한 차이가 나지 않는다. OrthoANI는 유전자 표지나 유전자 탐사 등의 과정은 없이 분류학의 목적에 맞게 바로 사용할 수 있는 프로그램이다. 또한, 이 프로그램은 간편하고, 재 생산성이 있으며, 믿을 수 있는 분류학 프로그램이다.

NGS의 사용이 미생물학 연구에서 보다 일상화 됨에 따라 오염을 포함한 유전자 서열의 품질에 관한 우려가 커지고 있다. 오염은 잘 못된 진단이라는 문제로 이어질 수 있기 때문에 임상 실험실에서 특히 중요하다. 유전체의 품질을 관리하는 시스템 개발은 일반 미생물 실험실에서도 매우 중요하다. 이런 맥락에서 16S rRNA 유전자 서열을 이용한 원핵생물 유전체의 오염 탐지 알고리즘을 갖는 ContEst16S라는 새로운 프로그램이

개발 되었다.

또한, 본 연구에서는 콜레라균의 표현형 예측 프로그램이 새로이 개발되었다. 본 연구에서 개발 된 프로그램은 콜레라균의 O 항원형 타입과 콜레라 독소 파지의 존재, 그리고 항생제 내성을 띠는지에 대한 예측 정보를 제공한다. O 항원형 예측 프로그램은 유전자 클러스터를 시각화하여 사용자에게 보여준다. 콜레라 독소 파지의 존재 예측 프로그램은 콜레라 독소 파지들의 유전정보를 이용해 타입 별로 파지 요소의 정보를 보여준다. 항생제 내성 예측 프로그램은 RGI (CARD-The Comprehensive Antibiotic Resistance Database)이라는 외부 프로그램을 사용한다.

시퀀싱 데이터에서 나오는 문자열은 생물학적 문제에 대한 결정적인 대답을 제공하지 못할 수 있다. 생화학적 검증이 없다면 그것은 그냥 예측일 뿐이다. 그러나 생물정보학에 의한 예측은 과학자들에게 매우 강력한 영향을 미치기 때문에 본 연구는 생물학 분야에 충분히 가치가 있는 연구라고 할 수 있다. OrthoANI는 분류학에 대한 표준을 제공하며 ContEst16S는 연구자들이 오염된 미생물 유전체에 대한 정보를 확인할 수 있게 해 주며, 콜레라균 표현형 예측프로그램은 O 항원 및 독성 인자를 식별하고, 항생제 저항성을 예측하는 등 콜레라균 연구에 대한 통찰력을 제공한다.

주요어: 콜레라균, O 항원형, 콜레라 독소, 콜레라균 항생제 내성, 세균 유전체학, 비교 유전체학, OrthoANI, ContEst16