Comparative computational analysis of *Vibrio cholera* strains from different geographical location

Submitted By:

Mandeep Singh

151501 [4rd Semester]

Jaypee University of Information Technology

Waknaghat, Solan, Himachal Pradesh - 173234

Comparative computational analysis of *Vibrio cholera* strains from diverse geographical location

Introduction

Vibrio cholera is a bacterial organism of cholera, an extreme diarrheal disease that happens most of the time in global shape. This bacterium contains a wide collection of biotypes and strains, tolerating and trading qualities for toxic substances, colonization factors, and antidote poison resistance, capsular polysaccharides that offer imperviousness to chlorine and new surface antigens; for instance, the lipopolysaccharide and O antigen compartment. The parallel or flat exchange of these harmfulness qualities by phage3, pathogenicity islands and other embellishment hereditary components gives bits of knowledge into how bacterial pathogens develop and advance to wind up distinctly new strains. *V. cholerae*, amid inter epidemic periods, is a tenant of harsh and estuarine waters, and in these situations, is related with the gram negative ton and another oceanic widely varied vegetation. The living being additionally enters a reasonable yet non-social state under specific conditions.

In the mid-2000s, many nations inside Africa for example, Mozambique, Democratic Republic of the Congo, and Tanzania, experienced flare-ups that frequently included more than 20,000 cases and a few 100 passing. Among that time, the difference in the frequency of cholera in Africa with respect to different parts of the world kept on developing. In last few years, the number of cases reported to world health organization in 2015 there were 172454 cases was reported from 42 countries in which there was 1304 deaths. There are many cases are not recorded due to limitations in surveillance systems and fear of impact on trade and tourism.

Methodology

Five different strains N16961**,** M66-2, 2010EL-1786, M-1293, **L-3226** from different geographical locations Bangladesh, Indonesia-Makssar, Haiti, Russia-Sulina-village-Dagestan., Moscow respectively. Workflow for comparative analysis is given below.

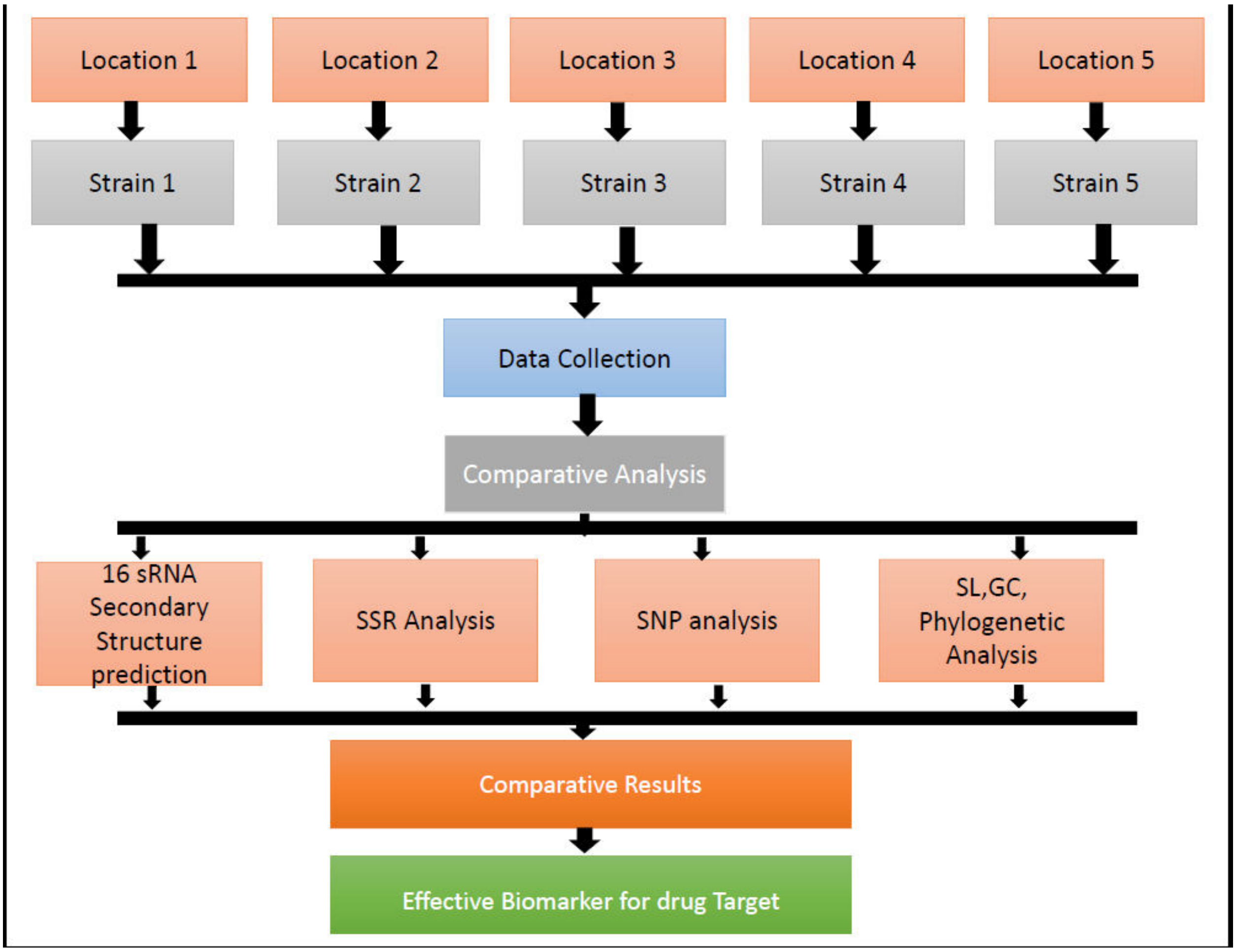


Figure 1. Workflow for the comparative analysis and screening of novel drug targets

Expected Results

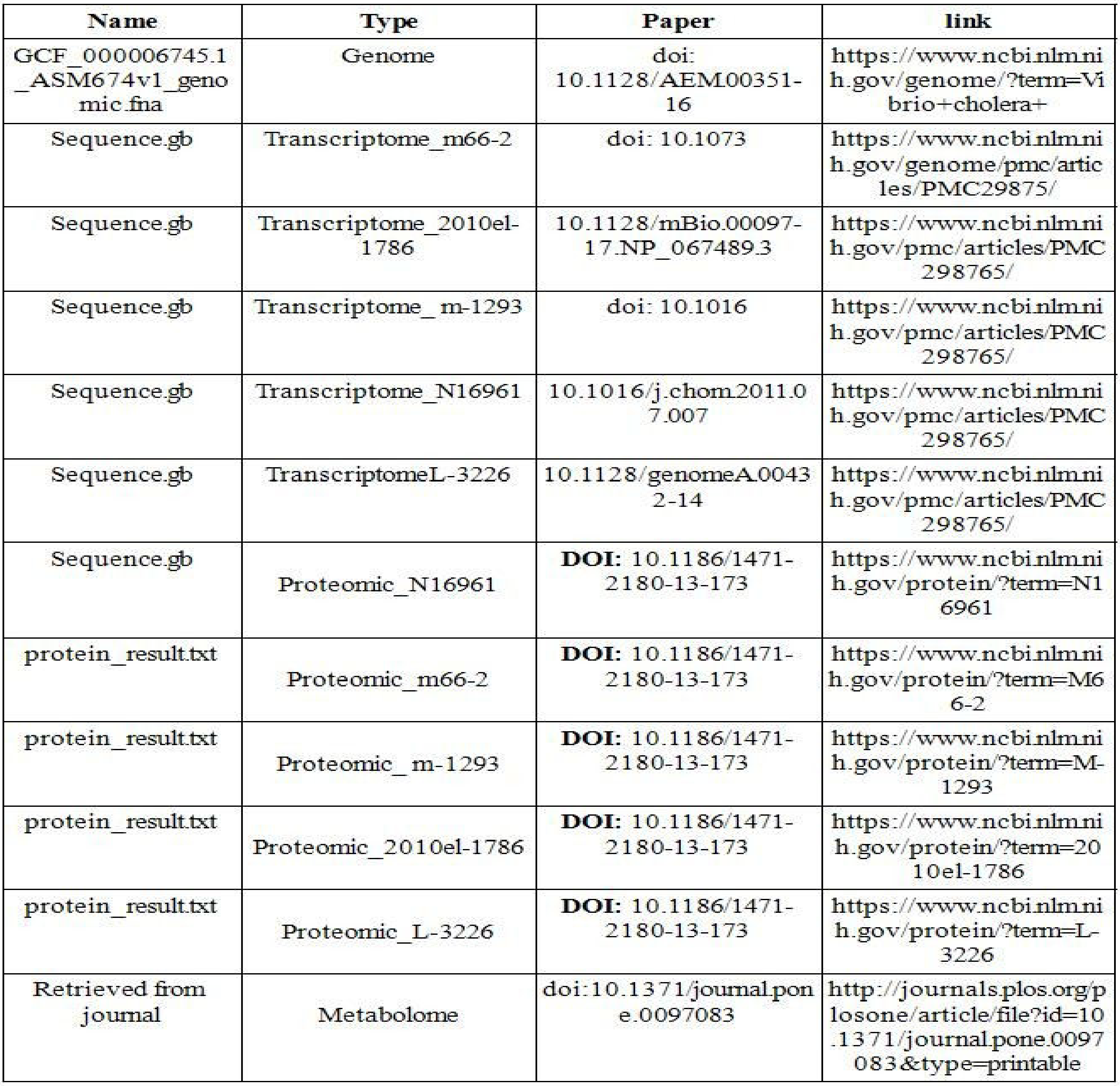
- Comparative analysis helps to understand the similarity and indentifying conserved regions across different strains.
- SSR and SNP screening help in identification of novel biomarkers for therapeutic studies.
- Geographical location basis variation and respective sequence length, codon percentage values, GC percentage will help in prioritization of novel variation factors.

Materials and methods

SSR is a tract of tandemly rehashed DNA themes that range long from two to five nucleotides, and are normally rehashes 5-50 times. Straightforward succession rehashes happen at a huge number of areas inside a living being's genome; furthermore, they have a higher transformation rate than different territories of DNA prompting to high hereditary differences. Straightforward grouping rehashes (SSRs) are regularly alludes to as Microsatellite. For instance, the succession TATATATATA is a dinucleotide microsatellite, and GTCGTCGTCGTCGTC is a trinucleotide microsatellite. Rehash units of four and five nucleotides are alludes to as tetra-and penta-nucleotide themes, individually. Microsatellites are circulates all through the genome [1] [2] [3]. Many are situated in non-coding parts of the human genome and are consequently do not create proteins, notwithstanding they can likewise be situated in administrative locales and inside the coding district. Here, we have utilized Microsatellite ID apparatus (MISA) calculation for limitation and recognizable proof of microsatellite(s) sort notwithstanding the event of a specific microsatellite sort as per the individual themes or unit measure. The parameters utilized while executing the Perl script utilizing MISA calculation to define microsatellite (unit estimate/least number of repeats) are, (1/10); (2/6); (3/5); (4/5); (5/5); (6/5); and most extreme number of bases interfering with two SSRs in a compound microsatellite was set to 100 base sets.

Results: -

The SSR analysis is on selected dataset on the *vibriocholera.* There were two chromosomes in selected dataset. Size of one sequence is 2961149 and size of another sequence is 107215 .the total number of SSR repeats in chr1 is 7 and in CHR2 is 3.the SSR types is seq1 is p6,p1,p3and p2 . The SSR type are in seq2 is p6, p6 and p1. SSR (TGA) 5 size is 15 starts from 1515993 to 1516007. SSR (G) 11 size is 11 starts from 1234653 to 1234663. SSR (A) 10 size is 10 starts from 361768 to 361777.   SSR (AACAGA) 54 size is 54 starts from 137106 to 361777. SSR (ACCAGA) 14 size is 84 starts from 303939 to 304022. SSR (G) 11 size is 11 starts from 1003188 to 1003198.SSR (CGC) 15 size is 15 starts from 1519461 to 1519475. SSR (C) 10 size is 10 starts from 2553344 to 2553353. SSR (CA) 6 size is 12 starts from 2658669 to 2658680.   SSR (TGCTGT) 23 size is 138 starts from 187759 to 187896. SSR (ACCAGA) 14 size is 84 starts from 303939 to 304022. SSR (G) 11 size is 11 starts from 1003188 to 1003198. Unit size of SSR is 1 and 6.unit size of seq1 is 1,2,3 and 6



Name	Type	Paper	link
GCF_000006745.1_ASM674v1_genomic.fna	Genome	doi: 10.1128/AEM.00351-16	https://www.ncbi.nlm.nih.gov/genome/?term=Vibrio+cholera+
Sequence.gb	Transcriptome_m66-2	doi: 10.1073	https://www.ncbi.nlm.nih.gov/genome/pmc/articles/PMC29875/
Sequence.gb	Transcriptome_2010el-1786	10.1128/mBio.00097-17.NP_067489.3	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298765/
Sequence.gb	Transcriptome_ m-1293	doi: 10.1016	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298765/
Sequence.gb	Transcriptome_N16961	10.1016/j.chom.2011.07.007	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298765/
Sequence.gb	TranscriptomeL-3226	10.1128/genomeA.00432-14	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298765/
Sequence.gb	Proteomic_N16961	**DOI:** 10.1186/1471-2180-13-173	https://www.ncbi.nlm.nih.gov/protein/?term=N16961
protein_result.txt	Proteomic_m66-2	**DOI:** 10.1186/1471-2180-13-173	https://www.ncbi.nlm.nih.gov/protein/?term=M66-2
protein_result.txt	Proteomic_ m-1293	**DOI:** 10.1186/1471-2180-13-173	https://www.ncbi.nlm.nih.gov/protein/?term=M-1293
protein_result.txt	Proteomic_2010el-1786	**DOI:** 10.1186/1471-2180-13-173	https://www.ncbi.nlm.nih.gov/protein/?term=2010el-1786
protein_result.txt	Proteomic_L-3226	**DOI:** 10.1186/1471-2180-13-173	https://www.ncbi.nlm.nih.gov/protein/?term=L-3226
Retrieved from journal	Metabolome	doi:10.1371/journal.pone.0097083	http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0097083&type=printable

Table 1: Omics data Collection

Data Collection for Vibrio cholera

These are five strains .which i selected on the basis of pathogenicity.Their information are retrived from NCBI ,Pubmed and many other resources.

1.**NCBI Retrieval**

Strain1:- Vibrio cholerae O1 biovar El Tor str. N16961 (g-proteobacteria)Table 1. Information retrieved from Genome Assembly **ASM674v1** out of 578 assemblies of NCBI.

ORGANISM NAME	VIBRIO CHOLERAE O1 BIOVAR EL TOR STR. N16961 (G-PROTEOBACTERIA)
INFRASPECIFIC NAME	STRAIN: N16961
BIOSAMPLE	SAMN02603969
SUBMITTER	TIGR
DATE	2001/01/09
ASSEMBLY LEVEL	COMPLETE GENOME
GENOME REPRESENTATION	FULL
REFSEQ CATEGORY	REFERENCE GENOME
GENBANK ASSEMBLY ACCESSION	GCA_000006745.1 (LATEST)
REFSEQ ASSEMBLY ACCESSION	GCF_000006745.1 (LATEST)
REFSEQ ASSEMBLY AND GENBANK ASSEMBLY IDENTICAL	YES

Table 2. Information retrieved from Taxonomy Browser of NCBI for Vibrio Cholerae.

Taxonomy ID	243277
Inherited blast name	g-proteobacteria
Rank	no rank
Genetic code	Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names	
Synonym	Vibrio cholerae O1 biovar eltor str. N16961
Equivalent name	Vibrio cholerae serotype O1 biotype ElTor

	strain N16961
Equivalent name	Vibrio cholerae serotype O1 biotype El Tor strain N16961
Equivalent name	Vibrio cholerae El Tor N16961

Table 3. Information retrieved from Bio sample

Identifiers	BioSample: SAMN02603969; Sample name: AE003852
Organism	Vibrio cholerae O1 biovar eltor str.
Attributes	**Strain** N169161 **Serotype** O1 **Biotype** E1 O1 **Geographical location** Bangladesh

STRAIN 2.VIBRIO CHOLERAE M66-2 (G-PROTEOBACTERIA)

Table 3. Information retrieved from Genome Assembly ASM2160v1 out of 578 assemblies of NCBI.

ORGANISM NAME	VIBRIO CHOLERAE M66-2 (G-PROTEOBACTERIA)
INFRASPECIFIC NAME	Strain: M66-2
BIOSAMPLE	SAMN02603897
SUBMITTER	TEDA School of Biological Sciences and Biotechnology
DATE	2009/04/20
ASSEMBLY LEVEL	COMPLETE GENOME
GENOME REPRESENTATION	FULL
REFSEQ CATEGORY	REFERENCE GENOME
GENBANK ASSEMBLY ACCESSION	GCA_000021605.1 (LATEST)
REFSEQ ASSEMBLY ACCESSION	GCA_000021605.1 (LATEST)

REFSEQ ASSEMBLY AND GENBANK ASSEMBLY IDENTICAL	YES

Table 4. Information retrieved from Taxonomy Browser of NCBI for Vibrio Cholerae.

Taxonomy ID	579112
Inherited blast name	g-proteobacteria
Rank	no rank
Genetic code	Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names	
Synonym	Vibrio cholerae O1 biovar eltor str. M66-2
Equivalent name	Vibrio cholerae strain M622-2
Equivalent name	Vibrio cholerae str M66-2
Equivalent name	Vibrio cholerae str M66-2

Table 5. Information retrieved from Bio sample

Biosample	
Strain	M66-2
Geographical location	Indonesia,Makssar
Collection date	1937
Sero type	O1

STRAINS 3: - VIBRIO CHOLERAE O1 STR. 2010EL-1786 (G-PROTEOBACTERIA)

Table 6. Information retrieved from Genome Assembly ASM131818v1 out of 578 assemblies of NCBI.

ORGANISM NAME	VIBRIO CHOLERAE O1 STR. 2010EL-1786 (G-PROTEOBACTERIA)
INFRASPECIFIC NAME	Strain 2010EL-1786

BIOSAMPLE	SAMN065529
SUBMITTER	Centers for Disease Control and Prevention
DATE	2011\11\10
ASSEMBLY LEVEL	COMPLETE GENOME
GENOME REPRESENTATION	FULL
REFSEQ CATEGORY	REFERENCE GENOME
GENBANK ASSEMBLY ACCESSION	GCA__000166455.2 (latest)
REFSEQ ASSEMBLY ACCESSION	GCA__000166455.2 (latest)
REFSEQ ASSEMBLY AND GENBANK ASSEMBLY IDENTICAL	YES

Table 7. Information retrieved from Taxonomy Browser of NCBI for Vibrio Cholerae.

Taxonomy ID	914149
Inherited blast name	g-proteobacteria
Rank	no rank
Genetic code	Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names	
Synonym	Vibrio cholerae O1 strain 2010EL-1786
Equivalent name	Vibrio cholerae O1 strain 2010EL-1786
Equivalent name	Vibrio cholerae O1 strain 2010EL-1786
Equivalent name	Vibrio cholerae O1 strain 2010EL-1786

Table 8. Information retrieved from Bio-sample of NCBI for Vibrio Cholerae.

Biosample	
Strain	2010EL-1786

Geographical location	Haiti
Collection date	2010
Sero type	O1
Host	Homosapiens
Isolation source	stool sample from patient with cholera
Isolation source	stool sample from patient with cholera

STRAIN 4:- V.CHOLERAE O1 BIOVAR EL TOR STR. M-1293

Table 9. Information retrieved from Genome Assembly ASM131818v1 out of 578 assemblies of NCBI.

ORGANISM NAME	VIBRIO CHOLERAE O1 BIOVAR EL TOR (G-PROTEOBACTERIA)
INFRASPECIFIC NAME	Strain: M-1293
BIOSAMPLE	SAMN02666511
SUBMITTER	RARI
DATE	2014\06\12
ASSEMBLY LEVEL	CONTIG
GENOME REPRESENTATION	FULL
REFSEQ CATEGORY	REFERENCE GENOME
GENBANK ASSEMBLY ACCESSION	GCA__000705295.1 (latest)
REFSEQ ASSEMBLY ACCESSION	GCA__000705295.1 (latest)
REFSEQ ASSEMBLY AND GENBANK ASSEMBLY IDENTICAL	YES

Table 10. Information retrieved from Taxonomy Browser of NCBI for Vibrio Cholerae.

Taxonomy ID	696
Inherited blast name	g-proteobacteria
Rank	no rank
Genetic code	Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names	
Synonym	Vibrio cholerae biovar eltor
Equivalent name	Vibrio cholerae O1 biovar eltor
Equivalent name	Vibrio cholerae O1biovar eltor

Table 11. Information retrieved from Bio-sample of NCBI for Vibrio Cholerae.

Bio sample	
Strain	M-1293
Geographical location	Russia , Sulina village, Dagestan
Collection date	1994
Sero type	O1
Host	Homo sapiens
Isolation source	Missing
Isolation source	Missing

Strain 5:- Vibrio cholerae O1 biovar El Tor str. L-3226 (g-proteobacteria)

Introduction about the strain:-Draft entire genome sequencing of the Vibrio cholerae O1 El Tor clinical strain L3226, disengaged in Moscow in 2010, was completed. Different changes in the destructiveness related portable components were resolved in its genome that separated this strain from the reference V. cholerae O1 El Tor strain N16961.

Table 12. Information retrieved from Genome Assembly ASM131818v1 out of 578 assemblies of NCBI.

ORGANISM NAME	VIBRIO CHOLERAE O1 BIOVAR EL TOR STR. L-3226 (G-PROTEOBACTERIA)
INFRASPECIFIC NAME	Strain: L-3226
BIOSAMPLE	SAMN02630793
SUBMITTER	RARI
DATE	2014\03\25
ASSEMBLY LEVEL	CONTIG
GENOME REPRESENTATION	FULL
REFSEQ CATEGORY	REFERENCE GENOME
GENBANK ASSEMBLY ACCESSION	GCA__000705295.1 (latest)
REFSEQ ASSEMBLY ACCESSION	GCA__000705295.1 (latest)
REFSEQ ASSEMBLY AND GENBANK ASSEMBLY IDENTICAL	YES

Table 13. Information retrieved from Taxonomy Browser of NCBI for Vibrio Cholerae.

Taxonomy ID	1458274
Inherited blast name	g-proteobacteria
Rank	no rank
Genetic code	Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names	
Synonym	Vibrio cholerae L3226

Table 14. Information retrieved from Bio-sample of NCBI for Vibrio Cholerae.

Bio sample	
Strain	L-3226

Geographical location	Russia , Moscow
Collection date	Oct-2010
Sero type	O1
Host	Homo sapiens
Isolation source	stool from a Russian tourist who visited India in 2010
Isolation source	stool from a Russian tourist who visited India in 2010

Scripts used in project:

Calculating the length, total nucleotides, dinucleotide sequence GC and AT counts

```
#Calculating the length, total nucleotides, dinucleotide sequence GC and AT counts
$DNA="TACCGTGTAAGCTGCGTATGCGATCGTACGCGTGTGCGGT";
#length of DNA
($length=length$DNA);
print"the length of DNA $length\n";
$a=($DNA=~tr/A//);
$b=($DNA=~tr/C//);
$c=($DNA=~tr/G//);
$d=($DNA=~tr/T//);
$Total=$a+$b+$c+$d;
print"total bases in DNA $Total:\n";
#count of GC
$GC=($DNA=~s/GC/GC/g);
print"the total number of dinucleotide GC in DNA :$GC:\n";
#count of AT
$AT=($DNA=~s/AT/AT/g);
print"the total number of dinucleotide AT in DNA:$AT:\n";
#percentage of GC
$GCper=($GC/($Total)*100);
print"the percentage of GC: $GCper:\n";
exit;
```

Script for Phylogentic Analysis:

```
#Author Mandeep
#Date :27/04/2017
use strict;
use warnings;

@ARGV =  ('a,b|c', 'c,d|e', 'a,d|e') unless @ARGV;

my %HoA;
foreach ( @ARGV ) {
   m/^([a-z])[,]([a-z])[|]([a-z])$/ ;
   push @{$HoA{$1}}, $2;
}
print "\n===========\@HoA=====\n";
print "from->to\n";
while (my ($key, $values) = each %HoA) {
   print $key, "=>   [", join(',', @$values), "]\n";
}
```

References

1. Heidelberg, J. F., J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, *et al.* "DNA Sequence of Both Chromosomes of the Cholera Pathogen Vibrio Cholerae." [In eng]. *Nature* 406, no. 6795 (Aug 03 2000): 477-83.

2. Clark, K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. "Genbank." [In eng]. *Nucleic Acids Res* 44, no. D1 (Jan 04 2016): D67-72.

3. Retrieved from GenBank. Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I, complete sequence**.** NCBI Reference Sequence: NC 002505.1.

4. Retrieved from GenBank. Vibrio cholerae O1 biovar El Tor str. N16961 chromosome II, complete sequence. NCBI Reference Sequence:
5. Updated global burden of cholera in endemic countries. Ali M, Nelson AR, Lopez AL, Sack D. (2015). PLoS Negl Trop Dis 9(6): e0003832. doi:10.1371/journal.pntd.0003832.
6. Retrieved from GenBank. Vibrio cholerae O1 biovar El Tor str. M66-2chromosome I, complete sequence**.** NCBI Reference Sequence: NC 00891.1.

7. Retrieved from GenBank. Vibrio cholerae O1 biovar El Tor str. M66-2 chromosome II, complete sequence. NCBI Reference Sequence: 00891.1