Chapter 6

Genetic diversity of *Vibrio parahaemolyticus*

Genetic diversity of *Vibrio parahaemolyticus*

6.1 Introduction

The population dynamics of pathogenic bacteria are very important for diagnosis, treatment and epidemiological surveillance. Homologous recombination, gene rearrangements and mutation play a major role for generation of genetic diversity among bacterial population (Virdi and Sachdeva, 2005). Various methods was used previously to study the genetic diversity of bacteria like virulence gene profile, serotype, ribotype, Pulsed-Field Gel Electrophoresis (PFGE), Restriction Fragment Length Polymorphism (RFLP) analysis of rDNA, and Randomly Amplified Polymorphic DNA (RAPD) (Chowdhury et al., 2000; McLaughlin et al., 2005; Fuenzalida et al., 2006; Silvester et al., 2016; Wang et al., 2017). 16S rRNA gene is highly conserved among bacterial species and the comparison of 16S rRNA gene sequences of unknown strain is widely used for identification of bacteria and for understanding the phylogenetic relationships. However, the presence of multiple copies of the 16S rRNA gene in the genome is another important limiting factor that needs to be considered (Liu et al., 2017). In spite of using a single locus within the genome, use of multiple loci within genome increase the discriminatory power as well as cover the genome as proposed by Gevers et al. (2005). Multiple housekeeping genes which are present as a single copy within the genome and are not subjected to selective pressure was used for Multilocus Sequence Analysis (MLSA). MLSA has been used to understand the genetic variation of highly recombinogenic *Helicobacter pylori* (Osaki et al., 2013), *Salmonella* sp. (Achtman et al., 2012), *Listeria monocytogenes* (Salcedo et al., 2003), *Vibrio cholera* (Lee et al., 2006), *Streptococcus suis* (King et al., 2002) and *Bartonella henselae* (Iredell et al., 2003) *V. parahaemolyticus*

is a highly diverse species, therefore MLSA will provide a clear representation of the genetic diversity of *V. parahaemolyticus* in India.

6.2 Material and Methods

6.2.1 Phylogenetic analysis using 16S rRNA gene

The details of PCR amplification and sequencing of 16S rRNA gene is described in Chapter 1. All the 16S rRNA gene sequences were corrected using BioEdit (version 7.0.1). The corrected sequences were aligned in MEGA 7 (Kumar et al., 2016) using the ClustalW algorithm (Thompson et al., 1994). The nucleotide diversity (p), number of biotypes, polymorphic sites and biotype diversity (h) was analyzed using DnaSP 4.0 software (Rozas et al., 2003). Neighbor-Joining (NJ) (Saitou and Nei, 1987) algorithm was used to understand the evolutionary history of *V. parahaemolyticus* in MEGA7 software (Kumar et al., 2016). Tamura-Nei method (Tamura and Nei, 1993) was used to analyze the evolutionary relationship. The bootstrap values are denoted next to the branches. This shows the percentage at which associated taxa clustered together (Felsenstein, 1985). The final phylogenetic tree shows only the nodes with bootstrap support of more than 50 %. Pairwise deletion for indels was done to eliminate positions containing alignment gap and missing data MEGA7 was used for evolutionary analyses (Kumar et al., 2016).

Inferences were drawn from the demographic history of the isolated strains which were analyzed with the 16S rRNA gene sequences by using two different approaches discussed below, one of the approaches is that the null hypothesis of neutrality may be rejected when a population has undergone population expansion (Tajima, 1989). Tajima's *D* test (Tajima, 1989) was carried out using DnaSP 4.0 (Rozas et al., 2003) for

examining whether the population from different regions is at genetic equilibrium or not. The second approach was Fu's *Fs* test which is preferred over other tests because of its suitability for analyzing large samples (Ramos-Onsins and Rozas, 2002). The Fu's *Fs* test was used to tests deviations from neutrality which could be found under population expansion (Fu, 1997). For evaluation of possible historical events of population growth and decline, DnaSP 4.0 (Rozas et al., 2003) mismatch distribution analysis was done with the assumption of selective neutrality. A smooth wave-like mismatch distribution was displayed for a population that had experienced a rapid expansion in recent past.

6.2.2 PCR amplification and sequencing for MLSA

Bacterial culture and genomic DNA isolation methodologies are described in Chapter 1. Four loci were selected for Multilocus Sequence Analysis. The housekeeping genes, *dnaE* (DNA polymerase III, alpha subunit), and *recA* (RecA protein) were used for chromosome I and *dtdS* (Threonine 3-dehydrogenase), *pyrC* (Dihydro-oratase) were used as a housekeeping genes for chromosome II. The PCR amplification were performed using the primers described at the *V. parahaemolyticus* MLST website (http://pubmlst.org/vparahaemolyticus). The PCR reaction volume was 25 µl, consist of 2.5 µl 10X PCR buffer, 1 µl of 25 mM $MgCl_2$, 1 µl of 5 pmol forward and reverse primer, 0.5 µl of 25 mM dNTPs and 0.2 µl of Taq DNA Polymerase. The thermal profile consisted of initial denaturation for 2 min. at 95 °C, 35 cycles of denaturation at 94 °C for 30 s, annealing temperature for 45 s and extension at 72 °C for 30 s, final extension for 10 min. at 72 °C. The PCR products were visualized by electrophoresis in a 1.8 % (w/v) agarose gel. The PCR samples were sequenced in both directions using an ABI 3730xl capillary sequencer (Applied Biosystems, Foster City, CA) to check the validity of the

sequence data. The forward and reverse sequences were aligned using the software DNA Baser. The sequence of forward stand was proofread using the sequence of complementary strand.

6.2.3 Genetic diversity study using MLST

The nucleotide diversity (p), number of biotypes, polymorphic sites and biotype diversity (h) was analyzed using DnaSP 4.0 software (Rozas et al., 2003). The ratio between synonymous (*dS*) and nonsynonymous (*dN*) substitutions was calculated following the method of the Tamura-Nei model using Mega 7 (Kumar et al., 2016). The synonymous (*dS*) and nonsynonymous (*dN*) substitutions ratio was used to test the hypothesis for neutrality (*dS/dN*); if $dS/dN<1$, then nonsynonymous sites are under selective constraint or purifying pressure (negative selection); $dS/dN>1$ indicates positive selection, and $dS/dN=1$ indicates neutrality.

6.2.4 Sequence types and clonal complexes determination

Sequence types (ST) and Clonal complexes (CC) were identified using Public databases for molecular typing and microbial genome diversity (https://pubmlst.org/vparahaemolyticus/). During the search for different sequence types, clonal complexes match option (exact or nearest) was selected in sequence definition database of *V. parahaemolyticus*. The sequence type profile of the entire four housekeeping genes of each strain was used for identification of clonal complex.

6.2.5 The phylogenetic analysis

Neighbour-Joining (NJ) trees were constructed using concatenated sequences of each STs. About 2226 bp concatenated sequences of each locus were analyzed with CLC genomics workbench using the Kimura 2-parameter model to understand the genetic distance. The statistical support of the nodes in the phylogenetic tree was assessed through 1,000 bootstrap resampling.

6.3 Results

6.3.1 Genetic diversity study using 16S rRNA gene

From 183 numbers of 16S rRNA gene sequences, 214 variable sites were spotted (126 parsimony informative) which comprised 138 biotypes. Most of the biotypes revealed uniqueness to particular individual strain. The nucleotide composition of 16S rRNA gene was similar to GC bases reported in the 16S rRNA gene region. The mean base-pair composition was 25.2 % A, 20.7 % T, 22.1 % C and 32.1 % G. For all the three populations, the biotype diversity (h) indicated higher values and it ranged from 0.9759 to 0.9943 (Table 6.1). The sequence data revealed a wide range of nucleotide diversity (p) and the values ranged from 0.0056 to 0.0140 (Table 6.1).

Table 6.1 Gene flow and genetic differentiation based on 16S rRNA gene sequences of *V. parahaemolyticus*

Population	No. of Sequences	No. of segregating sites	No. of Biotypes (b)	Biotype diversity (bd)	Average no. of differences	Nucleotide diversity
West Bengal	71	188	62	0.9943	19.8503	0.0140
Andhra Pradesh	61	66	43	0.9759	8.2530	0.0056
Gujarat	51	201	43	0.9906	13.2902	0.0093
Overall	183	340	138	0.9901	14.4898	0.0102

Phylogenetic analysis of 16S rRNA gene sequences showed that exclusive clades were not formed by the biotypes from distinct geographical locations and individual sites. The mix of biotypes was observed in the phylogenetic tree which was indicated by the spread of biotypes across all the regions. The biotypes pairwise comparison between nucleotide indicated mismatch distribution for all the samples was unimodal (Figure 6.1). The values of Tajima's D (-2.59470; $p < 0.001$) test of selective neutrality and Fu's *Fs* (-32.422; $p < 0.001$) test indicated highly negative and significant. This shows that, population expansion occurred suddenly and population differentiation took place in a short period of time (Figure 6.2).

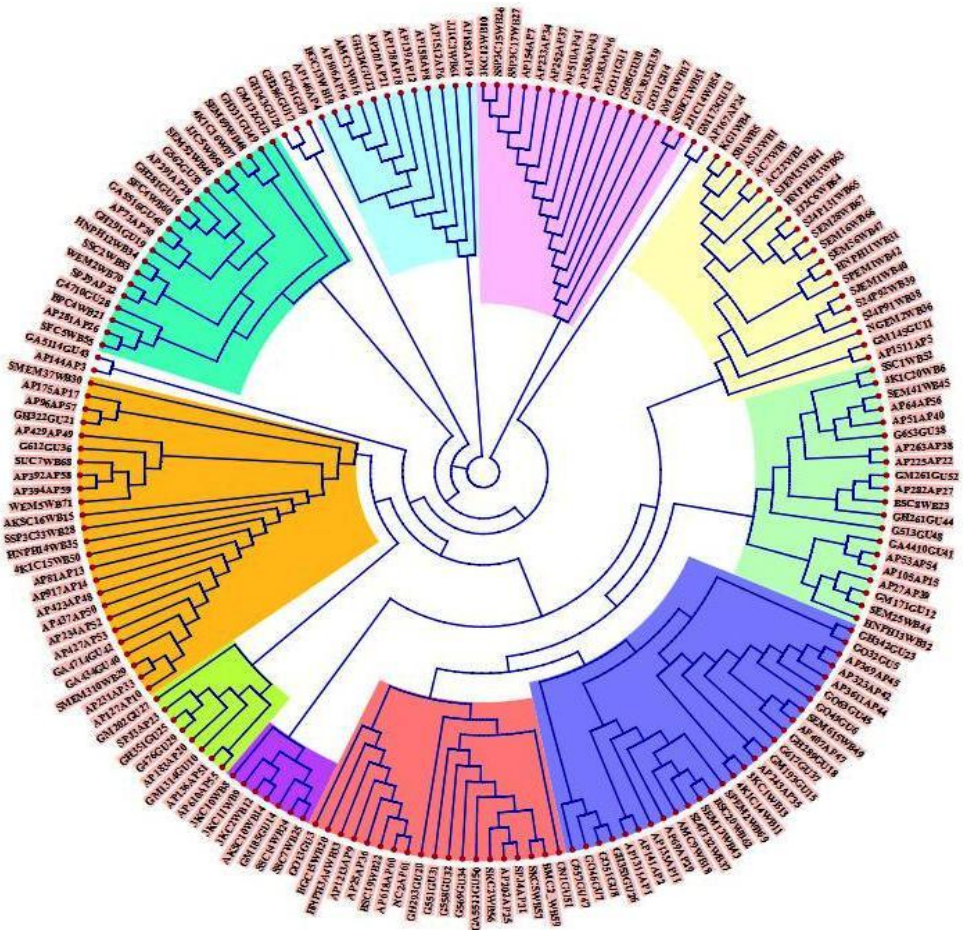

Figure 6.1 Evolutionary relationships among biotypes of *V. parahaemolyticus* collected from three states of India.

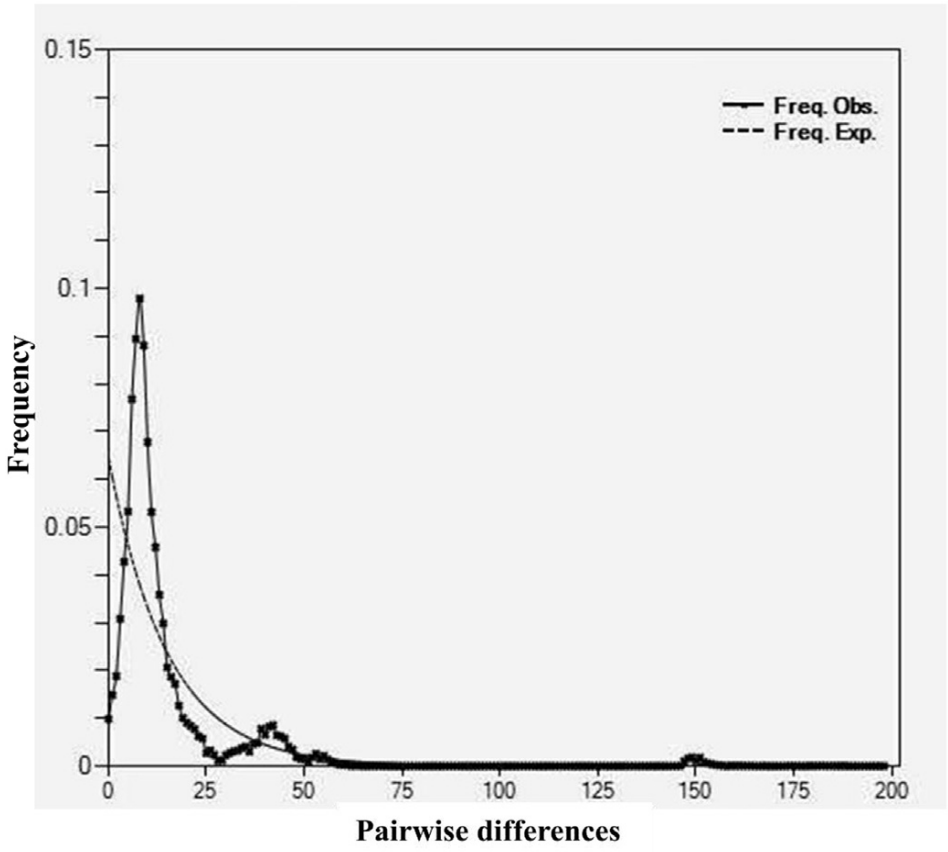

Figure 6.2 Mismatch distribution of 183 numbers of 16S rRNA gene sequences of *V. parahaemolyticus* from three populations (West Bengal, Andhra Pradesh, Gujarat) representing the observed and expected pairwise differences under the sudden population expansion model

6.3.2 Sequence diversity

The genetic diversity of 90 different isolates of *V. parahaemolyticus* was analyzed by using four different housekeeping genes under MLST scheme. The detail nucleotide diversity, polymorphic sites, biotype of each locus is summarized in Table 6.2. The number of polymorphic sites observed in each locus varied *recA* (59), *dnaE* (55), *pyrC* (45) and *dtdS* (81). The nucleotide diversity ranges from 0.0112 to 0.027. *dnaE* (0.0112) and *pyrC* (0.027) showed the lowest diversity. 90 sequence of each locus comprised of different biotypes *recA* (44), *dnaE* (55), *pyrC* (52) and *dtdS* (69). The biotype diversity

ranges from 0.9737 to 0.9928. Synonymous with nonsynonymous (*dS/dN*) substitutions ratio in the entire four loci was higher than 1 indicating positive selection.

Table 6.2 Nucleotide and allelic diversity of housekeeping genes used in MLST of *V. parahaemolyticus* isolates

Gene	Fragment size	No. of Sequences	No. of segregating sites	No. of Biotypes (b)	Biotype diversity (bd)	Nucleotide diversity	dS/dN ratio
recA	750	85	59	44	0.9737	0.0215	2.685
dnaE	582	86	55	55	0.9855	0.0112	2.599
pyrC	512	89	45	52	0.9826	0.0115	1.69
dtdS	481	90	81	69	0.9928	0.027	3.07

6.3.3 Sequence Types (ST) and Clonal Complexes (CC)

Thirty-eight sequence types were identified using individual sequence definition of four housekeeping genes of 90 stains. Out of 36 STs, 23 STs were found within single isolates whereas rest of the STs found within the isolates ranges from 2 to 4. ST-428 is the most frequent, found within four strains. Four strain belongs to the ST-3 and another 2 strains belong to ST-281 were identified, which is the pandemic strain of *V. parahaemolyticus*. Individual sequence definition of 4 housekeeping genes of 38 strains did not match with the existing database present in the PubMed. Out of 38 sequence types two clonal complex CC3 and CC281 were identified which comprised 6 isolates (Table 6.3).

Table 6.3 Geographic distribution of different STs identified from MLST analysis

Sequence Types	No. of strains	Clonal Complex	Country (strain identified)
424	2		India, Ecuador
281	2	281	China, Thailand
225	2		China
150	2		China
1823	2		China
1390	1		Sri Lanka, Turkey
1551	2		China
673	1		Sri Lanka
954	1		China
1308	1		China
2165	1		China
150	2		China
810	1		USA
1838	2		China
1835	1		China
1060	2		China, Canada
470	1		China, Italy
831	2		USA
2090	1		China
888	1		China
646	1		China
428	4		Vietnam, China
396	1		Sri Lanka, China, Italy
1689	1		China
1310	1		China
3	2	3	Chile, India, Korea, Japan, Peru, Bangladesh, and Thailand
459	2		China
1321	1		China
1479	1		China
1969	1		China
2134	1		China
250	1		Thailand
393	1		China
1991	4		China
978	2		Thailand
684	1		China
355	1		China, Sri Lanka, Philippines

6.3.4 Geographical distribution of the STs.

The geographical distributions of different STs identified in the present study were shown in Table 6.3. The most pandemic strain of *V. parahaemolyticus* ST-3, was isolated in four continents which was also identified in the present study. CC3 strains were isolated primarily in Chile, India, Korea, Japan, Peru, Bangladesh, and Thailand (González-Escalona et al., 2008). Most of the isolated strain of *V. parahaemolyticus* belongs to non-pandemic and were isolated in China (*n*-16). Other pandemic strains of *V. parahaemolyticus* (CC281) isolated in the present study were reported from Thailand and China.

6.3.5 Phylogenetic analysis

Neighbor-joining (NE) tree was constructed using concatenated sequences of four housekeeping genes of thirty-six different STs as shown in Figure 6.3. All the thirty-eight STs divided into two different lineages (lineages A and B). Lineages A was a major part constructed with 22 different STs whereas lineages B consist of 14 different STs. Lineages A was again subdivided into five different clades, I, II, III, IV and V. Lineages B was subdivided into two clades, VI and VII. The pandemic strain of *V. parahaemolyticus* ST3 form clades V with other three STs, ST470, ST978, ST2090 and ST1991 whereas other pandemic strain ST281 form clades I with four different STs, ST646, ST1479, ST1835 and ST2165.

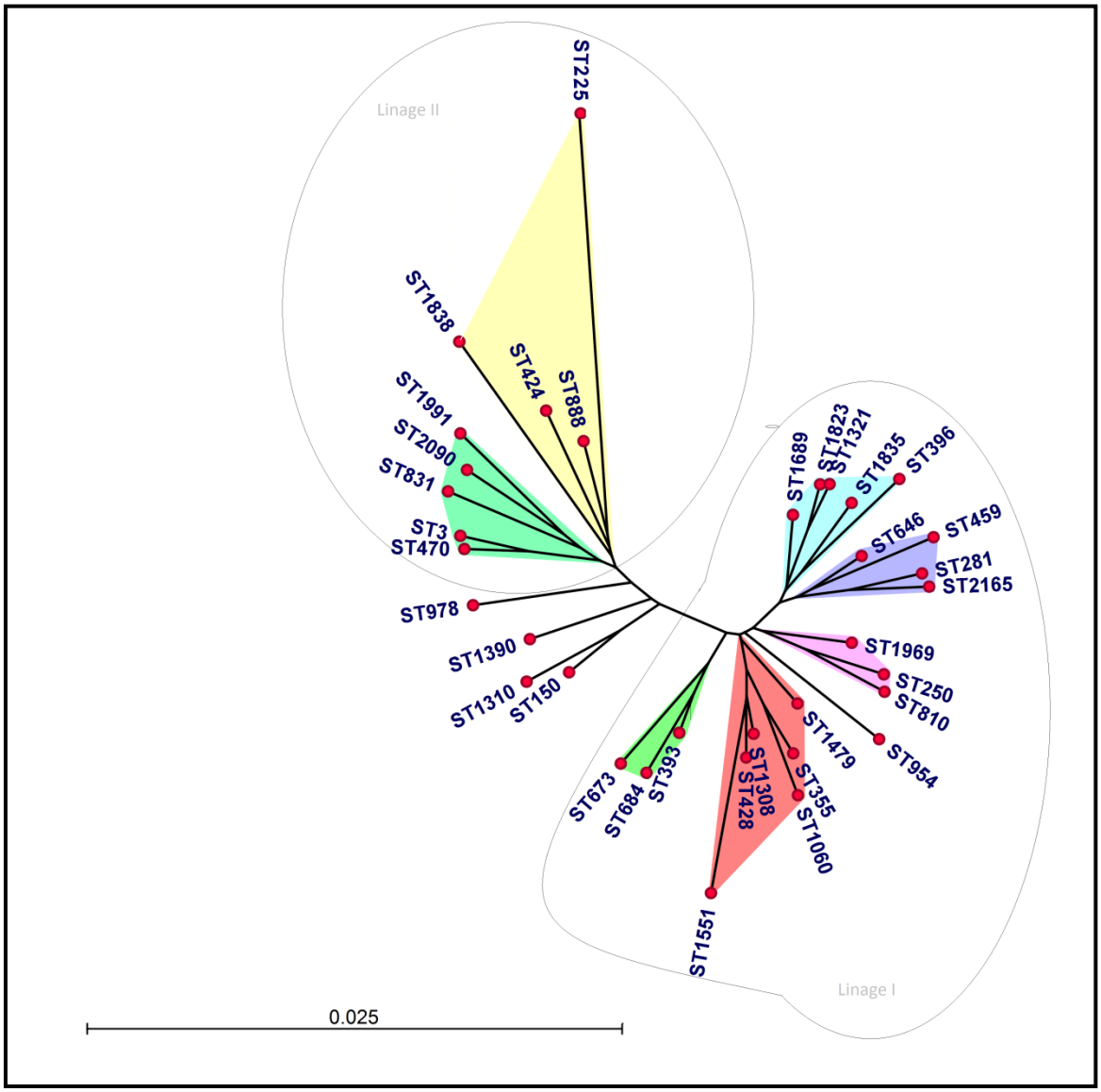

Figure 6.3 Neighbor-Joining tree of 36 concatenated sequences of four housekeeping genes under MLST of *V. parahaemolyticus* isolated from shrimp farms of India. Different colors representing different clusters

6.4 Discussion

The biological diversity in nature is an outcome of the evolutionary dynamics of populations and the geographical history of that region where the population arises (Lomolino et al., 2006). The proper method to assess current genetic variation to recreate

the demographic history of populations is very important for understanding the impact of different historical factors responsible for the evolution of a population (Hope et al., 2014). The study on genetics of microorganisms with historical demography emphasized mainly on host-associated bacteria to understand disease epidemics (Wirth et al., 2006; Comas et al., 2013; Holt et al., 2013). For understanding the genetic diversity of *V. parahaemolyticus,* population analysis was done using 16S rRNA gene from the isolates obtained from 3 different geographical locations. Bacterial populations should be demarcated cautiously, considering important factors like degree of genetic and ecological diversity (Avitia et al., 2014). Another important key factor to be considered is molecular resolution at which populations are defined (Kopac and Cohan, 2011). A study with more variable loci will help to identify newer population. The present study examined the conserved loci 16S rRNA gene region to know the phylogenetic pattern which may result in coarse population demarcation. However, individual populations for each lineage were demarcated by analyzing the genetic structure of *V. parahaemolyticus* across sampling sites. The demarcation of populations considered both the genetic and ecological aspects. The value of genetic diversity estimated within the sampled populations was in the range of natural bacterial populations as reported by Vos and Velicer, (2006). This indicated that the estimated diversity was in the range of normal polymorphism for bacterial populations.

Tajima's D test is based on the assumption of mutation-drift equilibrium and it provides a test of neutrality (Tajima, 1989; Nei and Kumar, 2000). An example is that Tajima's D test value is negative for a recent population bottleneck with the absence of purifying selection. The present study has got highly negative and significant value of

Tajima's D test indicating a sudden population expansion with population differentiation which occurred in a short time period. Fu's *Fs* statistics is based on the infinite sites model of mutation. He suggests estimating the probability of observing a random sample with a number of alleles equal to or smaller than the observed value under given the observed level of diversity and the assumption that all of the alleles are selectively neutral. The value of F is negative for an excess number of alleles. This can be observed from a recent population expansion and the present study also showed a similar pattern. The phylogeny of *V. parahaemolyticus* using 16S rRNA gene sequences displayed an admixture of biotypes. This is a typical case of lack of genetic differentiation among different localities, which may be due to the short time period available for population differentiation.

16S rRNA gene is a noncoding and evolutionary very conserved among the species. There are nine hypervariable regions in the 16S rRNA gene that, allow mutation. Again the presence of multiple copies of 16S rRNA gene within the genome might be might be misleading. Therefore, to get a clear picture of genetic diversity and global epidemiology of *V. parahaemolyticus*, MLST was used. It has been extensively used around the world to understand the population structure and epidemiology of *V. parahaemolyticus* (Urmersbach et al., 2014; Gonza´lez-Escalona et al., 2008; Han et al., 2014). In the present study, 36 STs were identified which comprised 56 different isolates of *V. parahaemolyticus*. The biotype diversity of individual locus ranges from 0.9737 to 0.9928 which is quite high. The synonymous to nonsynonymous substitutions ratio (*dS/dN*) is an index of selection pressure of nature on proteins (Hanada et al., 2007). The *dS/dN* of all the four different loci used in the present study was higher than 1 clearly

indicating that the population is evolving. A separate study carried out by Gonza´lez-Escalona et al (2008) also showed *dS/dN >1* for the entire seven genes, indicating positive selection. In contradictory to our findings Urmersbach et al (2014) found *dS/dN* was zero or close to zero, indicating negative selection.

In the present study, ST3 and ST281 were identified with international distribution. ST3 was globally distributed and was reported from Asia, North America, South America, Europe and Africa (Han et al., 2014). ST3 was evolved from an ancestor of CC3. CC3 first time emerged in India in 1996 and become a global epidemic clone of *V. parahaemolyticus*. Another ST281 belongs to clonal complex 281 identified in the present study from shrimp sample was also reported from Thailand and China (Theethakaew et al., 2013). ST 281 serovar O1:K10 was identified from shrimp samples of Thailand. Four different STs (ST1551, ST1060, ST2090 and ST888) identified in the present study were also isolated from clinical samples in China. However, they did not belong to any clonal complex. The phylogenetic analysis was carried out to understand the evolutionary relationship among the different STs identified from pubMLST database. The pandemic strain ST3 and ST281 form a cluster with other non-pandemic stains which clearly indicates the evolutionary closeness. Phylogenetic analysis carried out by Han et al. (2014) showed that the pandemic and non- pandemic STs are clustered together. However, the phylogenetic analysis carried out by Theethakaew et al. (2013) showed that STs from clinical samples structure into distinct clusters from the STs isolated from environmental samples like shrimp and water samples.

6.5 Conclusion

The genetic diversity study using 16S rRNA gene sequences reveals high genetic diversity and a lack of genetic structure. This is a typical case of admixture population possible due to the limiting time for the population to differentiate. To reconfirm on genetic diversity, MLST was carried out. The study revealed a high genetic diversity of isolated strains of *V. parahaemolyticus* in India based on MLST. The small number of STs isolated from environmental samples related to clinical strain reported from China. The phylogenetic analysis revealed that, pandemic and non-pandemic STs are clustered together which clearly showed lack of genetic structure. This is attributed to short generation time, random mutations, rapid reproduction and genetic recombination of *V. parahaemolyticus*.