ABSTRACT

Title of Dissertation: PHYLOGENOMIC AND STRUCTURAL ANALYSES OF *VIBRIO CHOLERAE* POPULATIONS AND ENDEMIC CHOLERA

Young-Gun Zo, Doctor of Philosophy, 2005

Dissertation directed by: Professor Rita R. Colwell
Marine Estuarine Environmental Sciences
University of Maryland, College Park

Cholera is a serious public health problem because of the high burden of morbidity. Recurrent pandemic cholera is sustained by an endemic epicenter in the Bay of Bengal region but the mechanism of endemism is not clearly understood. Recent information showing that the dynamics and seasonality of endemic cholera are linked with environmental parameters led to the hypothesis that the population dynamics of *V. cholerae*, the causative agent of cholera indigenous in natural aquatic environments, is the link causing variation in endemic cholera. To substantiate this hypothesis, the structure and dynamics of *V. cholerae* populations in the aquatic environments were investigated, employing three approaches.

First, the phylogeny of the family *Vibrionaceae* was analyzed to determine the phylogenetic boundary of *V. cholerae*. Phylogeny analysis using comparative genomics revealed that the species, *V. cholerae*, is a direct descendant of a common ancestor of the genus, with at least 25% of its genome subject to horizontal gene transfer from other vibrios.

The second approach was analysis of the population structure of *V. cholerae* using genomic fingerprinting, with the conclusion that there is a multilayered clonality and paraphyla within the species, with a subvar branch, *V. mimicus*. It was also concluded that all of the epidemic lineages of *V. cholerae* are highly clonal, forming a tight phylogenetic compartment. The nonpathogenic clones were found to be highly diverse and some showed significant association with fluctuations observed in the potential-host crustacean zooplankton compositions.

Finally, analyses of both the dynamics and compartmentalization of *V. cholerae* populations during endemic cholera outbreaks yielded a compartmentalized understanding of the mechanism of endemic cholera, namely that there are bodies of water in a cholera endemic area that serve as a reservoir of the bacterium and, therefore, a point source for the seasonal spread of cholera bacteria. The nature of a universal seasonal forcing that repeats the spread of the cholera bacterium from the point source each cholera season is not clear. Further study is recommended to identify those factors that determine both the point source reservoir and the mode of transportation resulting in spread of contaminated water from the reservoir.

PHYLOGENOMIC AND STRUCTURAL ANALYSES OF *VIBRIO CHOLERAE* POPULATIONS AND ENDEMIC CHOLERA

By

Young-Gun Zo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
Professor Rita R. Colwell, Chair
Associate Professor Anwar Huq
Professor Sam W. Joseph
Professor Michael R. Roman
Professor Estelle Russek-Cohen

Dedication

To my mother, my father, and Soyoung

Acknowledgements

Fishing this thesis, I find myself in debt to so many people who contributed various parts of the thesis.

I owe multitudes of appreciation to Prof. Rita R. Colwell. I appreciate her generosity of sharing her visions in the thesis and other researches, which is a great respect that this humble student can hardly earn from such an highly accomplished scientist like her, her patience of taking times and pains of reading, making senses out and editing my draft writing that was merely a woven of inarticulate ambiguity, and her extraordinarily inverted, steadfast commitment to her students, including academic, financial, and scientific supports, which made me keep indulging into the biology of *V. cholerae.*

I also present my sincere appreciation and thankfulness to Prof. Anwar Huq who has been the greatest, cordial supporter of me, my laboratory activities, and even my family. Particularly, supply of research topics and materials that involved his contribution had stretched half way across the globe and was essential to most part of this thesis. As an educator, he also taught me how to successfully manage scientific communications and cooperations in a networking environment. As an experienced researcher, he showed me how to maintain the integrity of one's research despite impediments coming from a variety of directions.

Completing this thesis, I would like to mark my expression of gratitude to all members of my advisory committee for reading the poorly readable drafts twice and giving valuable feedback on concepts, analyses, and interpretations. Ideas on the concepts and significances of researches in this thesis were improved thanks to Prof. Sam Joseph, ecological interpretation on Chesapeake Bay ecosystem was much motivated by Prof. Mike Roman's discoveries and comments, and various mathematical and statistical methods and data presentations were refined thanks to challenging questions and expertise of Prof. Estelle Russek-Cohen.

In the production of data, I thank to my colleagues who used work with me side by side in the laboratory. Dr. Nipa Chokesajjawatee, who authored several, related publications (Choopun et al.) that provided valuable information and materials to this thesis. She also provided much help in labor and in discussion in Chapters 2 and 3. Prof. Dong-Hun Lee of Chungbuk University, Korea has contributed in collection of sequence data, used in Chapter 2, while he was visiting the lab for a year. The large part of my indulgence into quantitative methods and computing skills was also motivated by working with him before I joined the lab at the Center of Marine Biotechnology. I thank Christopher Grim for being such a good company in the NIH project and sharing his luminescence data and etc. with Dr. Nipa Chokesajjawatee and me. I also thank Dr. Valerie Louis for providing *in silico* form of Chesapeake Bay data appeared in her thesis and papers and used in my Chapter 3. For other valuable influences and friendship, I thank Dr. Irma Rivera, Dr. Erin Lipp, Dr. H. and G. Matte's, Dr. Jongsik Chun, Dr. Sunny Jiang, and Dr. Paul Dunlap.

For producing data and generously allowing me to use them, I thank people involved in the NIH-granted project for the molecular epidemiology and ecology of cholera (1R01A139129-01): Prof. R. Bradley Sack in Johns Hopkins Univ., Prof. J. Glenn Morris Jr. and Prof. Afsar Ali in Univ. of Maryland, Baltimore, Prof. Ira M. Longini, Jr. and Prof. Azhar Nizam in Emory Univ., and Dr. Sirajul Islam, Dr. A. K. Siddique, and Dr. M. Yunus and numerous people in ICDDR, Bangladesh.

For the unlimited patience and faith on me, I appreciate my parents, wife Soyoung and her parents. Completion of this thesis being one of theirs small accomplishments is my wish. Though last, not least in joy, I thank my daughter Hannah for being born and giving endless joys at the difficult times of my family.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Primary problem addressed by this study

Cholera is an enteric disease with the key symptom being large volumes of watery stool, followed by dehydration and hypovolemic shock and eventually death in the most severe cases. These symptoms are the result of the disruptive effect on intestinal cells by cholera toxin (CT), a highly potent and irreversible stimulator of cellular adenylate cyclase. The toxin is produced by the bacterium *Vibrio cholerae* after infecting the intestine via contaminated drinking water or food (Sack *et al.*, 2004). While oral or intravenous rehydration therapy is effective in preventing death, the case fatality rate still reaches about 50% where proper treatment is not readily available, most notably in refugee camps (Siddique *et al.*, 1995). The burden of morbidity is also costly (Guerrant *et al.*, 2003; World Health Organization, 1992) because, in general, victims are in the morbid state for up to half a week, even with administration of antibiotics (Ryan *et al.*, 2000; Sack *et al.*, 2004).

The immense epidemic capacity of cholera is manifested by the seven historical pandemics of cholera (Barua, 1992). The worst among them is the recent and continuing seventh pandemic which began in Indonesia in 1961 after a hiatus of 33 years. Cholera affects more than 100 countries and every continent (Communicable Disease Surveillance and Response, World Health Organization, http://www.who.org). Unlike earlier pandemics, the seventh has been persistent, spanning more than 40 years. Recurrent cholera epidemics in Africa and Latin America raise, not only the question of geographical expansion of cholera-afflicted areas, but also the suspicion of establishment of new endemic foci of cholera in those continents (Kaper *et al.*, 1995; Lan & Reeves, 2002; Naidoo & Patric, 2002).

The persistence and geographical expansion of cholera outbreaks are concordant with the suspicion that spread of a well-known infectious disease may intensify with climate change, i.e., global warming and El Niño (Haines & Patz, 2004). Observation of a direct proportional response in the incidence of cholera cases associated with the El Niño events in the Equatorial Pacific (Checkley *et al.*, 2000; Speelmon *et al.*, 2000) and an 11-month latent link of cholera prevalence in the Ganges Delta with the dynamics of the Pacific El Niño index (Pascual *et al.*, 2000; Rodo *et al.*, 2002) strongly suggests that cholera will intensify in the future (certainly with more cases and possibly with wider and more frequent outbreaks), since global climate warming has been predicted. According to United Nations Intergovernmental Panel on Climate Change (IPCC), the global average temperature has risen about 0.6ºC in the last century, but the projection for the next century is between 3 to 10 times that value (Houghton *et al.*, 2001). By applying the projected trend proportionally, which was observed between cholera outbreaks in Latin America and El Niño events in the Equatorial Pacific (Checkley *et al.*, 2000; Speelmon *et al.*, 2000), human populations in tropical areas are predicted to be exposed to at least three times more cholera cases. The level of risk for populations in tropical areas being exposed to cholera-causing bacteria is predicted to increase via warmer temperature, based on enhanced survival and growth of cholera-causing bacteria in warmer temperatures (Louis *et al.*, 2003), and shifts in the availability and use of water resources with sea-level rise and increased brackish water. Predicted global warming also suggests that human populations in subtropical or template regions will be newly exposed to the risk of cholera by transformation of their environment to one that will be favorable for proliferation of *V. cholerae*. Therefore, the increasing threat of cholera for human populations globally is a genuine human health risk and a problem that must be solved, with respect to prediction and prevention, based on improved understanding of the mechanisms of disease outbreaks.

1.2 Rationale for the solution

Conventional public health prevention measures, such as current vaccination or use of antibiotics are not completely effective in preventing cholera. In spite of intense microbiological research over the last century, measures to prolong and enhance immunity of the intestine against colonization by the cholera bacterium have not been successful. Commercially available cholera vaccines produce transient immunity that lasts only four to six months. Application of the currently available vaccines is neither enforced nor recommended by public health administrations [CDC Cholera General Information (http://www.cdc.gov), WHO Cholera Fact Sheet (http://www.who.org)]. Emergence and spread of multiple antibiotic-resistant *V. cholerae* (MARV) also demonstrates the negative effects caused by pharmaceutical approaches to treat cholera (Sack *et al.*, 2004). Prophylaxis by administration of antibiotics during epidemic outbreaks is not effective, although it can reduce the severity of symptoms (Khan, 1982).

Understanding the mechanism of cholera outbreaks is more valuable for prediction as well as prevention of cholera. Because cholera is caused by consumption of contaminated food and water, the mechanism of epidemic cholera has been regarded as simply fecal contamination of water resources, like many other diarrheal diseases (Mintz *et al.*, 1994). Pandemics of cholera are also understood as being caused by geographical spread via fecal contamination as a result of intercontinental travel of victims or contaminated transmissive vehicles, such as water or food (Mintz *et al.*, 1994). Lack of outbreaks in developed countries, where proper sanitary measures are followed in treatment of municipal water supplies and food production systems, hold up with the current understanding of the mechanism of the pandemic as the "chain of fecal contamination", and, at the same time, it proves that prevention of epidemic cholera is possible by such efforts. The recent finding of reduction in cholera cases in villages of Bangladesh by using simple filtration of household water also supports the notion of a fundamental solution for prevention of cholera by supplying uncontaminated water and food (Colwell *et al.*, 2003).

In certain geographical areas, such as the region around Bay of Bengal, and in Southeast Asia and Africa, cholera is also endemic because the incidence of cholera is persistent throughout the year, with a clearly seasonal fluctuation, and choleragenic *V. cholerae* is easily detected in the aquatic environment throughout the year (Colombo *et al.*, 1997; Lan & Reeves, 2002; Longini *et al.*, 2002; Sack *et al.*, 2004). Especially, the Ganges delta area, known as the home of cholera pandemics, it is the very epicenter of the current pandemic. From an epidemiological perspective, curbing a cholera outbreak from its epicenter by proper management of the water and food supply is the effective approach. However, the co-occurrence of socioeconomic conditions of underdevelopment and a cholera-prone tropical climate makes this solution economically unaffordable by the people in cholera endemic areas (Pauw, 2003). Therefore, invention of a more affordable solution or more efficient ways of applying existing cholera-prevention measures is needed. Investigating the mechanisms of endemic cholera in those geographical areas will provide a foundation from which to meet the need.

1.3 Background

1.3.1 Epidemiology of endemic cholera

In reviewing the epidemiology of cholera, Faruque *et al.* (1998) summarized the characteristics of cholera in the Bay of Bengal areas as: (1) a high degree of clustering of cases by location and season; (2) highest rates of infection in children 1 to 5 years of age in areas of endemic infection; (3) antibiotic resistance patterns that frequently change from year to year; (4) clonal diversity of epidemic strains, and (5) protection against the disease by improved sanitation and hygiene and preexisting immunity.

In highly populated areas of Bangladesh, cholera outbreaks occur seasonally twice a year (Colwell, 1996; Longini *et al.*, 2002). From March to May, small peaks in the number of cases are observed. The larger peak occurs during September to December, after the monsoon. The beginning of the large peak of cholera coincides with high temperature, low water depth, and the beginning of the low precipitation season. When an epidemic starts, cholera cases occur simultaneously in multiple locations (Glass *et al.*, 1982). The senescence of the large peak coincides with the beginning of cold dry weather. Also notable is the difference in the peak of the cholera season between Bangladesh and West Bengal, India. In Calcutta, India, the highest peak season of cholera is April to June (Barua & Greenough III, 1992).

With respect to the inter-annual variability of endemic cholera, an important finding was made recently (Pascual *et al.*, 2000; Rodo *et al.*, 2002). Nonlinear regression along the time series data of cholera prevalence for the last two decades showed strong latent correlation with El Niño-Southern Oscillation (ENSO), explaining the variation in cholera dynamics by 70%. An additional intriguing finding is that the strength of the correlation dissipates as the time frame goes to the past (1893-1920 and 1920-1940). Therefore, modulation of endemic cholera dynamics by global environmental change is evident, while the role of local and basin wide links and modulator remains vague.

1.3.2 Pathogenicity in cholera

Pathogenicity of *V. cholerae* is associated with the toxigenic effect of cholera toxin (CT), which disrupts ion transport of human intestinal epithelial cells. When *V. cholerae* enters the small intestine, it colonizes the epithelium by means of vigorous flagella activity, toxin-coregulated pili (TCP), and other virulence factors that aid in the attachment and colonization of cells in the mucus-protected habitat. After successful colonization, *V. cholerae* cells secrete CT, which is composed of two subunits (A and B). The B subunit is a pentamer of a 11.6 kDa polypeptide and serves to bind the holotoxin to the eukaryotic cell receptor, the ganglioside G_{M1}. The A subunit is further divided into A_1 and A_2 subunits after cleavage by protease activity. The A_2 subunit is a 5.4 kDa polypeptide and its function is to link A_1 subunit to B subunit. The A_1 subunit is a 21.8 kDa polypeptide with an enzymatic activity for toxigenicity. It has structural homology with the catalytic region of *Pseudomonas aeruginosa* exotoxin A and diphtheria toxin. It irreversibly activates the adenylate cyclase in the eukaryotic cell by transferring the ADP-ribose moiety to $G_{S\alpha}$ protein, the adenylate cyclase activator in the eukaryotic cell membrane. The increase of cyclic AMP (cAMP) eventually leads to increased Cl^- secretion by intestinal cryptic cells to cause a trans-epithelial osmotic gradient.

The genes related to CT and TCP were found to be highly clustered to form two separate pathogenicity islands on the *V. cholerae* chromosome. Recent studies revealed that both pathogenicity islands were actually lysogenic phages, CTXΦ (Waldor & Mekalanos, 1996) and VPIΦ (Karaolis *et al.*, 1999), which can transduce non-toxigenic strains of *V. cholerae* and *V. mimicus* to harbor cholera toxin (Faruque *et al.*, 1999).

1.3.3 Taxonomy and diversity of the choleragenic organism

The bacterium causing cholera was originally discovered in 1854 by Pacini who applied the name *Vibrio* but described again by Koch in 1884 (Howard-Jones, 1984). Now named as *V. cholerae*, it is an eubacterial member of the genus *Vibrio* in the gamma-proteobacterium group which includes most of the enteric bacteria, such as *Escherichia coli* (Tison, 1999). Within the taxonomic family *Vibrionaceae*, which now includes the genera *Vibrio*, *Photobacterium*, *Salinovibrio*, *Enterovibrio*, and *Grimontia* (Garrity *et al.*, 2004), *V. cholerae* was the sole member of the family showing phenotypic characteristics of moderate growth in Nutrient Broth (Difco, 1984), which contains NaCl, but requiring supplementation of NaCl, and moderate growth at 42ºC in any conventional culture medium. In 1981, Davis *et al.* (1981) designated *V. cholerae* strains with an atypical phenotype (negative sucrose utilization) as *V. mimicus* and indicated it was a sister species of *V. cholerae*, based on the observation that the genomic DNA-DNA hybridization level fell below 70%, although the 16S rRNA nucleotide sequence, a key molecular criterion for species-level divergence of *V. cholerae*, was not at a divergent level of similarity (Chun *et al.*, 1999). Notable in the pathogenic perspective was the fact that the two species often share antigenic properties (Ansaruzzaman *et al.*, 1999) and possess cholera toxin genes (Faruque *et al.*, 1999; Shinoda *et al.*, 2004). Because of the similarity between *V. cholerae* and *V. mimicus*, efforts to design molecular criteria to distinguish the two species were made (Chun *et al.*, 1999) and failed, as shown by Vieira *et al.* (2001). Therefore, the taxonomic standing of *V. cholerae* and *V. mimicus* is uncertain, namely whether both are a single species or not.

V. cholerae is conventionally classified according to biotype and serogroup. Up to the present time, 210 serogroups have been identified (E. Arakawa, National Insitute of Infectious Disease, Japan, personal communication) according to the antigenic property of cell surface polysaccharides. Two serogroups, designated O1 and O139, respectively, are considered to be responsible for most of the epidemic and endemic cholera. *V. cholerae* of other serogroups are collectively called non-O1/non-O139 and diverse environmental isolates of *V. cholerae* belong to this group. However; other serogroups may also cause localized epidemics of cholera. For instance, *V. cholerae* O37 was found to be the etiologic agent of localized cholera in Czechoslovakia and Sudan. Possession of the cholera toxin gene (CT) was found to occur in *V. cholerae* of other serogroups in the aquatic environment of West Bengal (Chakraborty *et al.*, 2000). The O1 serogroup is further divided into three sub-serovars: Ogawa; Inaba; and Hikojima (Kay *et al.*, 1994).

V. cholerae O1 strains are also divided into the two biotypes, classical and El Tor (Kay *et al.*, 1994). Keys for differentiation of the two biotypes are bacteriophage susceptibility, hemagglutination of chicken erythrocytes, and patterns of sugar fermentation. Isolates from the sixth pandemic were classical *V. cholerae* while outbreaks in the seventh pandemic were caused, in most cases, by *V. cholerae* of El Tor biotype. The classical biotype was not found throughout the seventh pandemic, except in Bangladesh. The El Tor biotype is known for its less severe symptoms of cholera and better survival in the natural environment.

1.3.4 Ecology of *V. cholerae*

As *V. cholerae* is an indigenous inhabitant of the aquatic environment, the modes of life style it demonstrates in the aquatic environment versus the human intestine can be different. In its physiology, *V. cholerae* is a saprophytic organism when free-living or attached to detritus (Tison, 1999) and this aspect is applicable in both the aquatic environment and the human intestine.

The majority of *V. cholerae* isolated from the aquatic environment belong to the non-O1/non-O139 serogroup, while stool samples of cholera patients usually yield *V. cholerae* of the O1 or O139 serotype (Sack *et al.*, 2004). When *V. cholerae* cells are located in stressed environments, such as the oligotrophic water column, there are at least three known modes of response to enhance survival. First, *V. cholerae* cells attach to nutrient-providing and protective particulate matter. The cells also attach to zooplankton, including marine and freshwater copepods, of which they comprise the dominant microbial flora (Tamplin *et al.*, 1990), and in rarer cases to phytoplankton (Islam *et al.*, 1999). Second, *V. cholerae* cells undergo physiological changes when they enter a dormant state, called the viable-but-nonculturable (VBNC) state (Colwell & Grimes, 2000; Roszak & Colwell, 1987). In the VBNC state, *V. cholerae* cells are small and oval in shape. They cannot be cultivated in conventional media which support multiplication of normal vegetative cells. Third, the cells form aggregates, notably by a variant of *V. cholerae*, the rugose positive variant. These cells occur with a frequency of one such cell per thousand cells in clinical strains and small cells of the variant are typically embedded in a carbohydrate-based matrix. It is known that cells of rugose aggregates are resistant to disinfectants, such as chlorine (Ali *et al.*, 2002).

While *V. cholerae* can cause either symptomatic or asymptomatic cholera, it can also densely colonize the epithelial layer of the human intestine. Therefore, the life style of *V. cholerae* can be represented as three different types: parasitic; free-living (planktonic); and epibiotic. In fact, the parasitic life cycle can be regarded as the cycle of epibiotic proliferation.

1.4 Objectives and approaches

To enhance our capabilities to predict and prevent cholera outbreaks, the objective of this study was to elucidate the underlying mechanism of endemic cholera in the Ganges delta areas.

Notable endemic characteristics of cholera in that geographical area are: (1) strong seasonality of cholera related to the spring-summer peak and a larger fall peak which develops after monsoon and lasts until the weather turns cold and dry (Barua & Greenough III, 1992); (2) environmental reservoirs of cholera, which is the general aquatic environment (Colwell & Huq, 1994); and (3) environmental modulation of seasonal and inter-annual dynamics of cholera prevalence (Rodo *et al.*, 2002). Therefore, approaches taken to define the mechanism of endemic cholera should include aquatic ecosystems as well as the clinical condition of human populations. It is also a sound postulation that the dynamics of *V. cholerae* populations in different water bodies of an area can be the link to environmental changes and cholera prevalence, which implies that an environmental reservoir of cholera is not a single entity, but is subject to dynamic changes related to seasonal and long-term climate change. The multiplicity and dynamic nature of the disease reservoir led us to analyze the various components of endemic cholera as many separate and dynamically-interacting entities. These entities comprise a single physical entity (e.g., an individual human or a body of water) or a collection of entities with discretely identifiable boundaries, so that all which is enclosed by a boundary carries attributes that are homogeneous in terms of ecological and epidemiological functions, but different from surrounding entities. Moreover, the system in which endemic cholera occurs can be defined as a collection of these entities and the dynamics of the system derives from flow and processes among entities with time. Therefore, they can be termed compartments of the system and the approach of a compartmental model can be used to determine the mechanism(s) of endemic cholera. This approach is further justified in that host populations and *V. cholerae* populations associated with the disease are also structurally multi-compartmental, exemplified by the differential age structure of victims of *V. cholerae* O1 and O139 (Cholera Working Group, 1993; Sack *et al.*, 2003).

The dimensions of compartments for aquatic environments and host populations include spatial and temporal compartments, such as location and season. In the case of the host population, age and socioeconomic strata can put more dimensions into the compartmental structures. To account for the diversity in bacterial populations causing cholera, *V. cholerae* populations need to be resolved into phylogenetic lineages and ecological niches, as compartments. Several publications (Beltran *et al.*, 1999; Farfán *et al.*, 2000; Stine *et al.*, 2000) have addressed this issue, but a uniform and consistent structure for bacterial species has not yet been resolved. The main obstacle is the limitation in methodology, especially in achieving significant sampling size. Another limitation is the uncertainty of the differentiation between *V. cholerae* and *V. mimicus*. Therefore, the objectives in the opening chapters of this study were to determine a phylogenetic definition of *Vibrio* species, notably to determine the boundary of the species *V. cholerae*, and to analyze the phylogenetic and ecological structure of the totality of the *V. cholerae* population.

Based on the approaches discussed above, a new practical goal of this study was to understand the compartmentalization and dynamics of the cholera bacteria and their habitat in cholera-endemic areas. The thesis comprises three sections: phylogeny; population structure; and population dynamics. In the first section, the phylogeny of the family *Vibrionaceae* is presented from the viewpoint of taxonomy and taxonomic boundaries of *V. cholerae*, the target population. The phylogeny was analyzed using comparative genomics and the molecular clock method. In the second section, the population structure of *V. cholera* is defined, using a genomic fingerprinting method that provides results reflecting the anatomical similarity of the genome of strains of *V. cholerae*, hence phylogenetic relatedness among the strains. The results are interpreted as an organization of phylogenetic compartments, each with correspondence to ecological niches challenged by ordination with environmental parameters of their source habitats. As the last section, the population dynamics and compartmentalization of *V. cholerae* during endemic cholera outbreaks was analyzed, providing a compartment model for a newly described mechanism for endemic cholera.

1.5 Significance of the research

Understanding the dynamics and mechanisms of infectious disease outbreaks is useful, not only for obtaining a quantitative description of past and on-going epidemics, but also for developing appropriate preventive measures for the disease itself, as well as a prediction model of the impact of environmental changes on the natural course of epidemics.

From public health and epidemiological perspectives, this research contributes to discovery and description of underlying mechanisms of endemism and insight into the variability of cholera, as well as the means for developing proper countermeasures against future outbreaks. Specifically, a dynamic model based on the compartment model can provide the ability to evaluate outcomes arising from changes in public health policy (drug administration, vaccination, or disinfection programs), as well as global or local environmental changes.

The approach proposed here, which included analysis of the structure and dynamics of *V. cholerae* populations, allows accountability of the role of *V. cholerae* populations in their natural aquatic environment. The dynamics of cholera epidemics is presented in terms of the population biology of the bacterium, *V. cholerae*. It is interpreted via proliferation and decline of diverse *V. cholerae* populations in the aquatic environment, as well as by the interaction of *V. cholerae* populations separated by habitat, e.g., the aquatic environment and the human intestine, and geographic location. The association of *V. cholerae* with plankton communities and the role of the microbial foodweb of aquatic ecosystems in controlling infectious disease comprise an additional, novel contribution of this approach. Heterotrophic bacteria, including *V. cholerae*, are active components of the microbial food web, together with zooplankton and phytoplankton. The dynamics of each component of the foodweb is determined by the ecological relationships that exist among the components that are, in turn, influenced by various abiotic factors; hence the dynamics of the diverse *V. cholerae* clones in aquatic environment can be best understood in the context of their ecological relationship with plankton communities and climate parameters. Therefore, the approach to ascertaining cholera dynamics via the population biology of *V. cholerae* in given aquatic environments extends our understanding of infectious disease in human populations as a function of the microbial foodweb. Ecological and population genetic approaches taken in this study to understand a diarrheal disease also will be applicable to other bacterial infectious diseases affected by environmental parameters.

Chapter 2. Phylogeny of *Vibrionaceae* and Speciation of *Vibrio cholerae*

2.1 Introduction

Prokaryotic organisms undergo binary fission, without the normal life stages typical of eukaryotic cells; therefore, the definition of species developed from studies of eukaryotic organisms does not apply. Instead, physiological, biochemical, and molecular characteristics have been employed to classify related organisms within systematic lineages equivalent to those of eukaryotic organisms(Cohan, 2001). The most recent consensus definition of bacterial species proposed by microbial systematists is >70% genome relatedness measured by the relative binding ratio (RBR) of DNA-DNA hybridization, and >97% similarity in 16S rRNA sequence (Stackebrandt *et al.*, 2002).

The prokaryote, *V. cholerae*, is a Gram-negative, curved rod belonging to the genus *Vibrio*, family *Vibrionaceae*. It is motile by means of a single sheathed polar flagellum and oxidase positive. It is a chemoorganotroph that can utilize D-glucose as a sole carbon and energy source. It is also capable of facultative anaerobe growth by respiratory as well as fermentative metabolism. The main characteristics distinguishing *V. cholerae* from other species in the genus *Vibrio* are ability to grow in nutrient broth without supplementation of extra NaCl, negative for esculin hydrolysis, arginine dihydrolase, lysine decarboxylase, ornithine decarboxylase, and positive for acid production from sucrose (Baumann & Schubert, 1984). It is autochthonous to the aquatic environment and has been found in water with a wide salinity range, especially in estuarine systems (Colwell & Spira, 1992; Colwell *et al.*, 1977; Kaper *et al.*, 1979).

As genomes of a large number of bacterial species are sequenced, the traditional systematics of bacteria are being challenged by the newly available genomic information (Stackebrandt *et al.*, 2002). Notably, the discovery of extensive lateral gene transfer (LGT) between species provides a picture of a network of organisms influencing the evolution of one organism, instead of a tree-like pedigree (Feil *et al.*, 2001; Ochman *et al.*, 2000). In the case of *V. cholerae* and other vibrios, the plasticity of their genomic makeup is pronounced, as evidenced by the presence of a super-integron structure in their genome (Rowe-Magnus *et al.*, 2003). The integrons are genetic elements functioning as a quenching mechanism, in which foreign genes are integrated into chromosomes of the *Vibrionaceae* as expressible open reading frames (ORFs). Therefore, to understand the evolutionary processes which led to the current composition of *V. cholerae* genomes, one must consider horizontal phylogeny (network), as well as vertical phylogeny, i.e., vertical flow of genetic information from the chain of direct ancestors.

The current gold standard in bacterial phylogeny is similarity of 16S rRNA sequence. Even with extensive LGT, theoretically, at least, the RNA molecule is believed to be resistant to horizontal influence due to its complexity in interaction with more than 50 other macromolecules and the presence of multiple copies in the genome (Stackebrandt *et al.*, 2002). Because the latter property contributes to "concerted evolution" of multicopy genes, the LGT effect is diluted out by the predominant copy of the molecule in a genome (Liao, 2000).

With the assumption of a complete lack of a LGT effect in the 16S rRNA genes, it can be stated that the gene phylogeny of 16S rRNA sequences represents the evolutionary path of organisms driven mainly by random mutation and genetic drift. In this case, genetic materials are inherited from the progenitor cell, and they are modified by mutation and genetic drift in the offspring population. Thus, the genealogy of 16S rRNA genes among related organisms can represent a “vertical evolutionary path” of the organisms (i.e., an organism phylogeny). In the case of single copy genes which lack the diluting effect by concerted evolution among multiple conserved copies of the same genes, the process of LGT (uptake and homologous recombination of foreign genes), which is wide-spread among bacteria, imposes an additional force of gene evolution, namely “horizontal gene flow”. The addition of new genes to the genomic content of an organism is one case and the generation of new alleles of genes by domain swapping between existing genes and their homologous foreign genes is the other case of “horizontal evolution”, the evolution caused by the horizontal forcing of LGT. When the assumption of a complete lack of LGT effect in 16S rRNA genes is not met, 16S rRNA genes may also be subject to this horizontal evolution. The known cases of horizontal transfer of an entire 16S rRNA operon among extreme halophiles (Boucher *et al.*, 2004) advocate that this assumption can be violated, at least in some taxons.

In fact, whether a 16S rRNA phylogeny represents the vertical evolutionary path of bacteria without significant bias caused by the effect of LGT can be challenged by examining the congruence of 16S rRNA gene phylogeny to organismic phylogeny (i.e., genome-wide phylogeny) among remotely related bacterial species. As the full genome sequences of diverse bacterial strains have been determined, the genome-wide phylogeny of bacterial strains has been analyzed, in contrast to the phylogeny of the 16S rRNA gene. Analysis of gene content in 13 complete genomes by Snel et al (1999) revealed that the phylogeny by similarity in gene content among the genomes was correlated with the 16S rRNA phylogeny. When the target organisms were expanded to 20 genomes and included partial information for *Schizosaccharomyces pombe*, *Homo sapiens*, and *Mus musculus*, the same congruence was observed (Tekaia *et al.*, 1999). Clarke *et al.* (2002) also found the topology in the genome tree built by mean pairwise sequence similarity among complete genomes of 28 bacteria, eight archaea, and one eukaryote (*Saccharomyces cerevisiae*) was consistent with the phylogeny based on 16S rRNA sequences. Therefore, the capacity of the 16S rRNA gene sequence to serve as an indicator of the vertical evolutionary path of the entire genome content appears to be robust against the bias of the LGT effect.

In spite of the benefits of concerted evolution and the high level of sequence conservation among 16S rRNA genes, there are drawbacks. The presence of multiple alleles of genes in a single cell hampers the resolution in studying local phylogeny, such as subspecies, and the small difference in sequences among species of the family makes 16S rRNA sequence analysis less useful in resolving supra-genus phylogeny. Therefore, other molecular clock sequences were used to fill this gap. These are collectively called "housekeeping genes" to indicate that they are functionally essential for the life of bacteria; therefore, they are believed to be highly conserved along the evolutionary path. Single copy small RNA sequences such as RNase P RNA (*rnpB*) and tmRNA (*ssr*) have also been suggested to have such a function (Haas & Brown, 1998; Schonhuber *et al.*, 2001).

Results of several studies, however, indicate a lack of congruence of 16S rRNA genealogy with that of other molecular clock sequences. For the *Vibrionaceae*, Hsp60 and RecA protein have received the most intensive study (Kwok *et al.*, 2002; Stine *et al.*, 2000; Thompson *et al.*, 2004a), only to conclude a lack of congruence with the phylogeny established by 16S rRNA. For *rnpB*, Maeda *et al.* (Maeda *et al.*, 2001) found a general congruence among nine *Vibrio* species but not with *V. pelagius*. The phylogeny of *Vibrionaceae* by tmRNA sequences have not been reported and this study is the first to do so. The tmRNA is a tRNA-like molecule that functions in the recovery of a ribosome stalled in the process of translation (Withey & Friedman, 2003). It is known to be present in all eubacterial species as a single copy gene and good clock-like behaviors have been reported in other families of bacteria (Felden *et al.*, 2001). It is also well established that the evolution of tmRNA gene is vulnerable to horizontal gene transfer because it often is the integration and excision target of mobile genetic cassettes among various bacteria (Williams, 2002), including *V. cholerae* (Rajanna *et al.*, 2003).

As more complete genomes of microorganisms are sequenced, it will be possible to use full genome information to construct a phylogeny of the bacteria, at least in some cases (Bansal & Meyer, 2002) ,and to test the congruence of 16S rRNA phylogeny to genome phylogeny, which must be constructed including horizontal as well as vertical gene flow. Several studies which compared the genomic sequences of “distantly-related” bacteria with 16S rRNA phylogeny concluded that 16S rRNA phylogeny is concordant with phylogenies drawn from genome-wide information (Bansal & Meyer, 2002; Snel *et al.*, 1999). Recently, such congruence among the enteric or symbiotic bacteria was demonstrated by Daubin *et al.* (2003). However, this congruence has never been tested at the “local level” of the *Vibrionaceae*, which comprise the natural inhabitants of the aquatic environment. Considering the uncertainties associated with a supra-genus level phylogeny of 16S rRNA sequences for the vibrio-like organisms (Ivanova *et al.*, 2004), a test for congruence must be done to delineate species evolution among members of the family.

Using current molecular information, the *Vibrionaceae* appear to be a monophyletic cluster of organisms. The 16S rRNA sequence phylogeny suggests that members of the family originated from a single common ancestral organism (an hypothesis revisited in this study, see Results). Another very convincing molecular aspect supporting a monophylum for the *Vibrionaceae* is the structure of the genomes of these organisms. Unlike the majority of bacterial genomes which comprise a single circular chromosome, the genomes of all known *Vibrionaceae* genomes include two chromosomes (Egan & Waldor, 2003). Their second, small chromosomes share the same scheme of replication origin, which initiates chromosome replication in a manner tightly synchronized with that of the large chromosome, to keep the mole ratio of one-to-one (Egan *et al.*, 2004). The presence of a super-integron island in all tested *Vibrionaceae* species examined to date (Rowe-Magnus *et al.*, 2003) is another predominant anatomical marker of the genomes of the *Vibrionaceae*. These two anatomical features of the genomes of *Vibrionaceae*, which are not found in other bacteria but are present in all members of the *Vibrionaceae*, provide exceptionally strong confidence for a monophyly of this group within the eubacterial kingdom. For this reason, species in the *Vibrionaceae* are ideal for the study of speciation and supra-species evolutionary processes in bacteria.

In this chapter, the phylogeny of *Vibrionaceae* is analyzed to determine speciation of the target organism of this study, *V. cholerae*. The primary interests are; (1) to define the species boundary of *V. cholerae* through determination of its sister species; (2) to quantify horizontal and vertical forcings, separately, in the composition of the *V. cholerae* genome. To accomplish these objectives, two approaches were taken. The 16S rRNA was used to resolve phylogeny and single copy genes of tmRNA were employed as a new clock molecule to reflect horizontal gene transfer across species. Using the published genomic sequences of *V. cholerae*, *V. vulnificus* and *V. parahaemolyticus* (hereafter, the "*Vibrio* Triad"), phylogeny was calibrated with the genome-wide information.

2.2 Materials and methods

2.2.1 Strains

A *Vibrionaceae* strains

The type strains of 49 *Vibrionaceae* species were obtained directly from the American Type Culture Collection (ATCC), Collection de l'Institut Pasteur (CIP), and the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ). Strains used in this study are referred to by their source collection identification number, employing the acronym of the source culture collection. *Vibrio rumoensis* S-1, type strain of the species, was provided by Dr. Yumoto of the Hokkaido National Industrial Research Institute, Sapporo, Japan. The type strain of *Vibrio diabolicus*, a purified DNA preparation was provided by Dr. Raguenes at the Institut Francais de Recherche pour l'Exploitation de la Mer, Centre de Brest, France as a gift, and was used to obtain DNA sequences. Two deep sea *Vibrio* strains (RC95 and RC96), provided by Dr. Reysenbach, Portland State University, Portland, Oregon, had been isolated from enrichment samples collected from hydrothermal vent areas of the Eastern Pacific Rise, off the coast of Oregon.

V. rumoiensis S-1 was grown in PYS-2 medium (8.0 g polypeptone, 3.0 g yeast extract, 5.0 g NaCl in one liter of distilled water; pH 7.5) at 25ºC. All other *Vibrionaceae* strains were grown in Marine Broth 2216e broth (Difco Laboratories) at 15ºC, 25ºC, or 35ºC depending on the optimum temperature of the strain.

Throughout the presentations of this study, identical type strains of species were used and they were designated with strain identification number of the source culture collection (e.g., ATCC or CIP). Unless strains were *V. cholerae* or *V. mimicus*, the same type strain was used for a species to generate and/or acquire molecular sequences of the species. As such, molecular sequences designated by the sequence accession number to reveal their sources were also from identical type strains, which can be verified by accessing the record of the sequence in public databases.

B *V. cholerae* and *V. mimicus* strains

To assess the polymorphisms of 16S rRNA and tmRNA sequences among strains of *V. cholerae*, representative *V. cholerae* and *Vibrio mimicus* strains were selected from our culture collection, Center of Marine Biotechnology, which contains environmental and clinical isolates from previous studies (Choopun, 2004; Choopun *et al.*, 2002; Kaper *et al.*, 1979; Kaper *et al.*, 1982; Kaper *et al.*, 1986; Rivera *et al.*, 2003). Representative strains of arbitrary groups, distinguished by differences in serotype, phenotype (i.e., luminescence or biotype), and geographic origin, were included, if the 16S rRNA sequence or tmRNA sequence was known.

Strains are listed in Table 2.3. All *V. cholerae* strains designated with *CT* in the table are clinical isolates. RC2 is the type strain of *V. cholerae* (ATCC 14035T = CECT 514T). The genome of RC145, *V. cholerae* N16961, has been sequenced (Heidelberg *et al.*, 2000) and RC215 is classical strain 569B, for which extensive pathogenicity and genetic data are available. ATCC 14547 (AT3) is the type strain of the species *Vibrio albensis*, a junior synonym of *V. cholerae* for luminescent *V. cholerae* isolates. RC44 (ATCC 25874), RC45 (Y334), RC47 (ATCC 25872) and RC48 (Y1= NRT 36S) are non-

O1 clinical isolates. The serogroups for RC44 and RC48 are not known, whereas RC45 and RC48 are serogroup O22 and O31, respectively; all other *V. cholerae* strains are environmental isolates from Chesapeake Bay (RC360 – RC549), coastal waters of Peru (P5 – P78), and ballast water of a cargo ship originated from Egypt (BLSTVC); RC5 is the type strain of *V. mimicus* (ATCC 33653T); RC54 – RC59 were isolated from environmental samples collected in West Bay, Louisiana; and RC217 – RC219 were environmental samples from Matlab, Bangladesh. Detailed characteristics of the strains are provided by Choopun (2004). The *V. cholerae* and *V. mimicus* strains were grown in Luria-Bertani (LB) broth (Difco Laboratories) at 37ºC.

2.2.2 Acquisition of molecular sequences

When sequences of genes and proteins were known and available in the public databases, they were downloaded from the GenBank database at the web site of the National Center for Biotechnology Information (NCBI), Bethesda, MD (http://www.ncbi.nlm.nih.gov). When sequences of 16S rRNA or tmRNA genes were unknown for a selected strain, sequencing was performed on the PCR products of the genes. In the case of the 16S rRNA gene, universal primers, p16SF1 and p16SR1 (Appendix A), were used and the PCR products were cloned into the pCR4TOPO sequencing vector, using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA). For the tmRNA gene, automatic sequencing was performed on PCR products using the primer pair, pTmVF1 and pTmR1 (Appendix A). BigDye termination sequencing of the PCR products was done using the ABI Prism 373 or 377 autosequencer (Applied Biosystems, Foster City, CA), according to the manufacturer’s instructions. The quality of base calling was assessed by GENESCAN (Applied Biosystems) and PHRED (Ewing &

Green, 1998; Ewing *et al.*, 1998). PQ value 20 (i.e., 99% accuracy) was used as an acceptable criterion for base calling and the results were also manually examined to confirm single base mutations. When base calling did not meet these criteria, sequencing was repeated until the required quality was achieved.

To sequence the flanking regions of tmRNA genes, the inverse PCR (IPCR) was performed. *Hae*II or *Mfe*I (New England Biolabs, Beverly, MA) digestion of genomic DNA of strains was done, according to instructions of the manufacturer. The digested DNA was subjected to ligation reaction at 15ºC overnight with T4 DNA ligase (Promega, Madison, WI). Using a set of primers (pIVPF and pIVPR; Appendix A) complementary to the internal sequences of *V. cholerae* tmRNA, the flanking region was amplified, and the products were sequenced, as described above.

2.2.3 Phylogenetic analyses

A Estimation of genetic distances and tree-building

RNA and protein sequences were aligned with the aid of the multiple alignment software CLUSTALX (Thompson *et al.*, 1997). Genetic distances were calculated using the Jukes-Cantor (JC) model (Jukes & Cantor, 1969) for RNA sequences and Dayhoff matrix model (Dayhoff *et al.*, 1978) for protein sequences, respectively. The Neighbor-Joining (NJ) method (Saitou & Nei, 1987) was used to build trees, employing the distance matrices. In the case of 16S rRNA sequences, it was anticipated that sequencing errors might be contained in sequences obtained from public databases. Therefore, the JC model was selected to minimize the effect of those errors on the interpretation of phylogeny. The model assigns an equal amount of evolutionary distance to all types of base transitions and transversions. Therefore, this model was expected to be most robust to errors from inaccurate base callings during sequencing. It is also known that this model does not require correction for multiple changes of one base locus, when the tree is interpreted as a molecular clock, because the order of nodes was always conserved in the clock-like behavior of the molecule (Felsenstein, 2004b). The significance of branching was tested by the criteria of either > 70% bootstrap support or > 50% bootstrap support. The former criterion was based on the observation of the Hillis and Bull (1993) simulation, in which branches with >70% bootstrap support had a 95% probability of being valid. According to Felsenstein's review (2004b) of the simulation, the criterion is not applicable in general, but it can be appropriate when the amount of information in a data set is large. In the case of the 16S rRNA sequence sets used in this study, the condition of data set size was met by having an aligned length of 1,356 bp, which is at the high end of size ranges among typical genes. Even though the tmRNA sequences were shorter in length, they were analyzed in the same way, because the results indicated that the amount of phylogenetic information in tmRNA genes was comparable to that of the 16S rRNA genes.

The criterion of >50% cut-off was based on the argument of Berry and Gascuel (1996) in which both Type I and Type II errors were treated equally. In addition, this condition deals with the phenomenon of overall reduction in bootstrap support values by violation of site-independence assumptions. The molecules of 16S rRNA and tmRNA have complex secondary structures, which leads to coevolution of a pair of nucleotide sites via complementary mutation. Therefore, some of the sites in these RNA genes show non-independence among sites. Recently Galtier (2004) demonstrated the influence of non-independence by simulation in a ribosomal RNA data set (*c.f.*, another example of the effect of non-independence is the PCR-based data set, such as RAPD, AFLP and rep-PCR, as employed in Chapters 3 and 4 of the present study. By the nature of the PCR reaction, two loci of a genome have to participate in order to produce a single bit of data. Therefore, the independence assumption is always violated. When the linear dependence among PCR product loci is removed by ordination, bootstrap support increases significantly (Appendix C)). Because the low support threshold, however, can result in a greater probability for misinterpretation arising from increased Type I error, a branch was considered significant only when the branch was supported by additional information. For example, a tree based on the Hasegawa-Kishino-Yano (HKY) substitution model (Hasegawa *et al.*, 1985), where greater weight is assigned to transversions than to transitions of nucleotide bases, was used to support branches in the NJ trees, when target sequences were considered free of sequencing error (i.e., tmRNA sequences were obtained with the intense quality control used in this study or from genomic sequencing projects). When using the HKY substitution model, quartet puzzling (Strimmer & von Haeseler, 1996) was employed for tree building to create a new tree completely independent from the tree obtained using JC model – NJ clustering. In simulation experiments, Strimmer and von Haeseler (1996) demonstrated that the quartet puzzling (QP) algorithm performs slightly better than the NJ method in depicting true phylogenetic relationships. JC distance estimation, Dayhoff distance estimation, NJ tree-building, and bootstrapping were done using MEGA2 software (Kumar *et al.*, 2001). Quartet puzzling based on the HKY model was done using TREEPUZZLE version 5.2 (Schmidt *et al.*, 2002), which also provided bootstrap-like support values for internal branches. Branches showing QP support values from 90% to 100% can be considered very strongly supported. Branches with lower reliability (> 70%) can, in principle, also be trusted but they were interpreted by comparison with other branches in the tree, i.e. by using relative support values.

B Comparison of tmRNA phylogeny and 16S phylogeny

Because the tmRNA used as a molecular clock is a novel approach for bacterial phylogeny, its properties as a molecular clock in making phylogenetic inferences of *Vibrionaceae* were analyzed in comparison to the standard molecular clock, namely the 16S rRNA sequences. The comparison was made with distance matrices and NJ trees from the two kinds of molecular sequences for the type strains of *Vibrionaceae*.

In the case of overall comparison with distance matrices, Mantel's test (Manly, 1997) was used to examine the existence of correlation between the matrix of JC distances from pairs of 16S rRNA sequences and the JC distance matrix from corresponding pairs of tmRNA sequences. Matrices were randomized 50,000 times to calculate probabilities for random occurrence of the observed correlation, using the DMAUSE module implemented in the ADE-4 package (Thioulouse *et al.*, 1997).

To estimate the relative evolutionary rate of the tmRNA sequences, the slope of the linear regression between the two JC distance matrices was calculated with distance matrices standardized to zero means and unit variance. Standardization was required because the unit of distance measurement for 16S rRNA and tmRNA differed according to the difference in sequence length. The 95% confidence interval (CI) for the estimated slope was determined following Manly's method (1997), where lack of correlation between the residual matrix and one of the standardized matrices was treated as the criterion for inclusion of the true slope of CI. Arbitrary slope values close to the slope estimate were assigned as ‘presumptive’ true slopes, used to produce the residual matrix. The significance of the correlation between the residual matrix and a standardized matrix for each presumptive slope value was estimated by Mantel’s test, with the critical value of 2.5% (i.e., two-tailed test). Practically, the lower bound and upper bound of the CI was determined by interpolating a value between the two presumptive values on a plot of significances across presumptive values, one of which produced insignificant correlation, while the other yielded significant correlation. If the slope was significantly different from one, the evolutionary rate by tmRNA is interpreted to be different from the rate by 16S rRNA.

To detect differences in the distance matrices of the two clock molecule sequences, the distance-matrix rate (DMR) test developed by Syvanen (2002) was used. This test is a graphical method, where distance value in one matrix is treated as the independent variable while distance value in the other matrix is the dependent variable only for the purpose of plotting (i.e., variables are not interpreted as independent-dependent relationship because there is no causality relationship between the two molecular clocks). An ordinary linear regression was performed and 95% CI of predicted values were determined as a function of slope and the independent variable. This method can reveal behaviors of molecular sequences deviating from the molecular clock hypothesis, based on the neutral evolution model that imposes a Poisson distribution of neutral and random substitution events as the mechanism of clock behavior (i.e., constant mutation rate) of molecular clocks. The results are presented on, and interpreted from, a graph where each pair of length-normalized genetic distances is plotted.

2.2.4 Phylogenomics of the *Vibrio* triad

A Determination of orthologous proteins and orthologous quartets

To compare the contents of two different genomes, one should identify genes of the equivalent function in both genomes and obtain phylogenetic distances from only those equivalent pairs. However, the typical protein sequence comparison identifies pairs of homologous genes without knowledge of their share of origin or functional equivalence. Homology of sequences can be of two types: orthology or paralogy. Homologous sequences are orthologous if they were separated by a speciation process. If a gene exists in a species and that species diverges into two species, then the copies of this gene in the resulting species are orthologous. Homologous sequences are paralogous if they were separated by gene duplication and diverge into genes of different functions. A pair of sequences, each of which is orthologous to the other is called an ortholog, whereas a pair that is paralogous is called a paralog. For genome-wide phylogeny analysis, phylogenetic distances only among orthologs should be considered. Therefore, a criterion that ensures exclusion of paralogs from the phylogenetic analysis was required.

Protein sequences of *V. cholerae* N16961, *V. vulnificus* CMCP6, *V. parahaemolyticus* RIMD2210633 (named here as the *Vibrio* Triad) and *E. coli* K12 were downloaded from GenBank, and a BLAST database (Altschul *et al.*, 1997) was constructed. The best hit pairs (BeT) of query protein and subject protein were catalogued from the BLASTP program output, which was commanded to search all proteins of the *Vibrio* Triad from the entire genomes of the *Vibrio* Triad. Because the presence of paralogs (i.e., genes sharing ancestry, but diverged with different functions) can mislead phylogenetic inferences, a stringent criterion that discards a set of genes with any possibility of being a paralog was used. BeT pairs with BLAST Expect value of less than 10^{-10} were collected (Clarke *et al.*, 2002; Tatusov *et al.*, 2000), and proteins were clustered to form a single cluster by pooling any protein in the BeT pairs that had a common protein (e.g., Expect value of one assigned to a hit can be interpreted as meaning that one might expect to see one match with a similar score simply by chance). The total of 13,204 proteins from the three genomes was pooled into 2878 clusters, comprising 1 – 200 proteins. Among them, those clusters which included three proteins, each of which was from one species, were selected. Further screening was done to determine whether the three BeT were symmetric (i.e., proteins in a BeT identify each other as unique matches). Thus, clusters with three proteins, one from each species, identified as a unique BeT on a genome, was selected as orthologs of the *Vibrio* Triad.

B Analysis of the topology of quartets

A quartet of orthologous proteins was decided by determining the BeT of orthologs of the *Vibrio* Triad on the *E. coli* K12 genome. Each of three proteins in the ortholog set of the *Vibrio* Triad was queried on the *E. coli* protein database using BLASTP. When all three queries returned a unique protein from the *E. coli* genome, with the BLAST Expect value of less than 10^{-10}, the four proteins were accepted as a quartet of orthologs protein in the four species. The four protein sequences were aligned by CLUSTALW. The tree topology of the quartets was evaluated by the maximum likelihood (ML) method implemented in the TREEPUZZLE version 5.2 (Schmidt *et al.*, 2002) based on HKY model. It weights the three possible tree topologies by their posterior probabilities (P_i). The probabilities were determined as $P_i = L_i / (L_1 + L_2 + L_3)$ by Bayes' theorem, where i is 1, 2, or 3 representing one of the three possible trees and L_i is the maximum likelihood of a tree (Strimmer & von Haeseler, 1997). When the weight support is larger than 95%, the topology is accepted as fully resolved.

To analyze uniform or differential distribution of genes with different quartet topologies on the chromosomes of *V. cholerae*, Rao's spacing test for uniformity in circular space, implemented in the S-PLUS library CIRCSTAT (Jammalamadaka & SenGupta, 2001), was used. The Fisher-Freeman-Halton exact test on contingency tables was employed using the STATXACT version 6 (Cytel Software, Cambridge, MA), either as an exact method or Monte Carlo approximation, when the former was not possible because of limited resources.

C Distribution of repeated sequences on complete genomes

Repeated DNA motifs, such as BIME, IRU box, Rep, and chi (Versalovic & Lupski, 1998), were searched on the *Vibrio* Triad genomes using the BLASTN program, with the minimal word size set as W=7 and expect cutoff value at 1000. Locations of matches were recorded and fragments of DNA sequence covering 100-bp forward and 100-bp backward were catalogued. The collected sequences were aligned with the query motif sequence, using CLUSTALW version 1.82 (Chenna *et al.*, 2003) to identify the boundary of the corresponding motifs. To analyze the secondary structure of the transcribed versions of the repeated motifs, the MFOLD program (Zuker, 2003) was used, with conditions of 37ºC and 1 M NaCl.

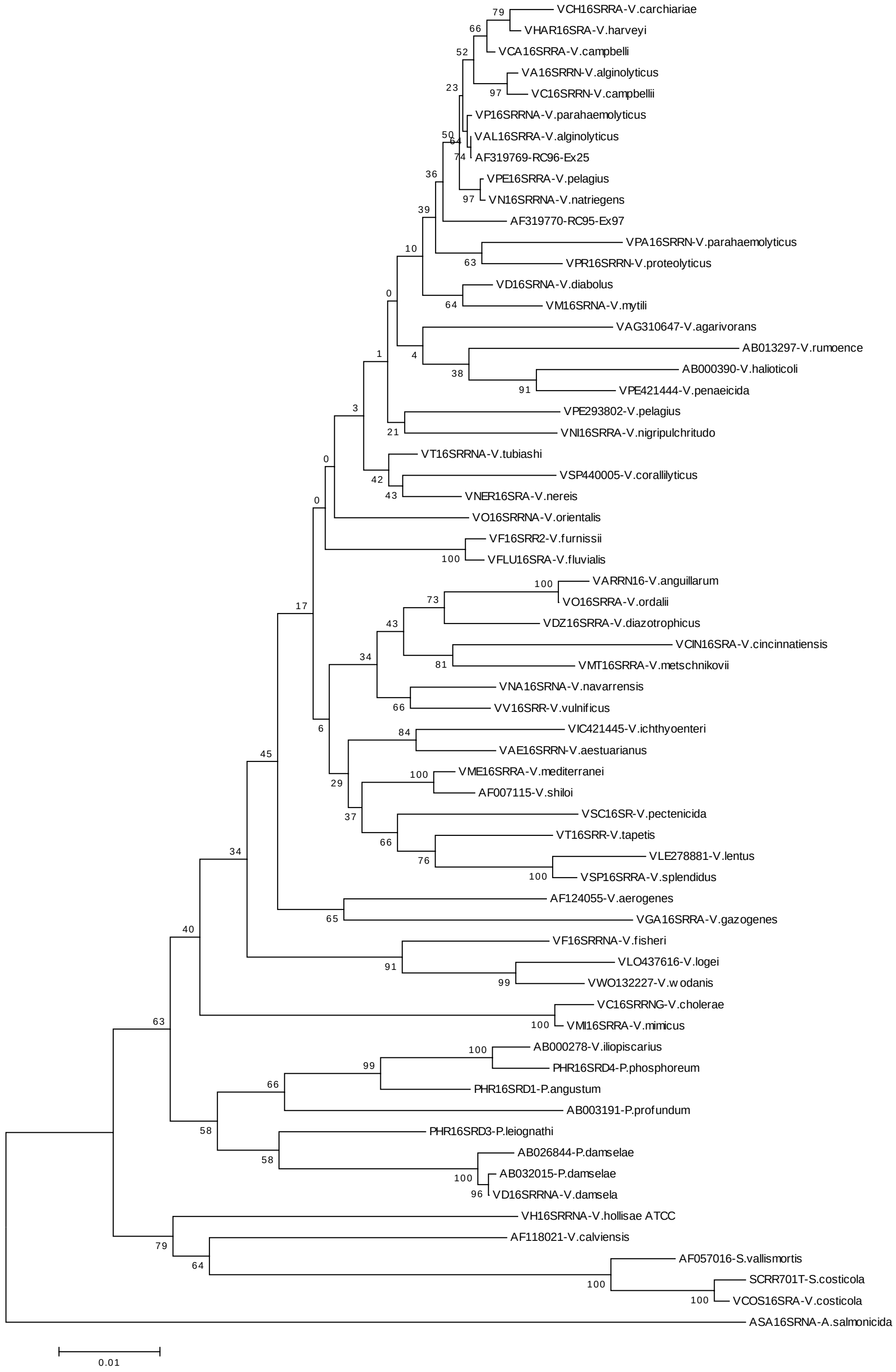


Figure 2.1. Neighbor-Joining tree produced by applying the Jukes-Cantor model to 16S rRNA sequences, with % branch support (shown next to branches) from 1000 bootstrapping (Label: GenBank Accession No. – species name).

2.3 Results and discussion

2.3.1 Phylogeny by 16S rRNA

The 16S rRNA sequences of *Vibrionaceae* clearly showed a monophyletic and multi-generic structure (Figure 2.1). The root branch from the outgroup, *Aeromonas salmonicida*, bifurcated to form *Vibrio-Photobacterium* and *Salinovibrio-V. hollisae-V. calviensis* clusters with significant bootstrap support of 63% and 79%, respectively. According to the results, the assignment of *V. hollisae* and *V. calviensis* to the genus *Vibrio* is not concordant with the phylogeny based on 16S rRNA sequences. When this study was underway, Thompson *et al.* (2003) reclassified *V. hollisae* as a sole member of the new genus *Grimontia*, based on 16S rRNA sequence and phenotypic data. *V. calviensis,* assigned to the genus *Vibrio* mainly on the basis of 16S rRNA sequence similarity to *V. hollisae*, has been reported to have been isolated only once, from 0.2 µm-filtered pelagic seawater (Denner *et al.*, 2002). Reclassification has not been published. It is concluded that *Vibrio* and *Photobacterium* share an ancestry which has diverged from other genera of the family, namely *Grimontia*, *Salinovibrio* and *Enterovibrio*, sister genus to *G. hollisae* (Thompson *et al.*, 2002). The separation of *Photobacterium* and *Vibrio* was also clear, based on the observation that the *Photobacterium* species formed a single cluster with significant bootstrap support (58%).

Regarding the phylogeny of *V. cholerae*, *V. mimicus* was the only species of the *Vibrionaceae* to link with it and they formed a two-species complex (Vcm complex). Based on the interpretation from bootstrap support values, the 16S rRNA tree, however, did not resolve the relationship of Vcm complex to other *Vibrionaceae* species. In fact, this observation could be applied to other *Vibrio* species, in general. Unlike many other eubacterial genera, the extensive diversity of *Vibrio* species appears to limit manifestness of supra-species phylogeny in the genus. Consequently, reevaluation of existing taxonomy of *Vibrio* species is inevitable as more complete information on the species becomes available.

The genetic distance between *V. cholerae* and *V. mimicus* was 0.0045, smaller than distances between the subspecies of *P. damselae* (0.006) and the subspecies of *S. costicola* (0.013). A similar distance (0.0055) was observed between *V. mediterranei* and *V. shiloi*. The latter was recently reclassified as *V. mediterranei*, based on DNA-DNA hybridization, fatty acid profile, genomic fingerprinting, and biochemical tests (Thompson *et al.*, 2001). A similar reclassification was reported for *V. carchariae*, which turned out to be a branch of *V. harveyi* (Gauger & Gomez-Chiarri, 2002). These patterns of reclassification indicate that strains, previously recognized as separate species by traditional methods relying on a few dozen biochemical tests, are now combined into single species based on results of molecular analysis, e.g., 16S rRNA sequencing, genomic fingerprinting, and DNA-DNA hybridization. In fact, there are three other cases of two traditionally separated *Vibrio* species with small genetic distances in their 16S rRNA sequence: *V. pelagius*-*V. natriegens*, *V. fluvialis*-*V. furnissii* and *V. anguillarum*-*V. ordalii*. Together with these pairs of species, the low level of divergence between *V. cholerae* and *V. mimicus* by 16S rRNA sequence also requires reevaluation of separate species status for *V. mimicus*.

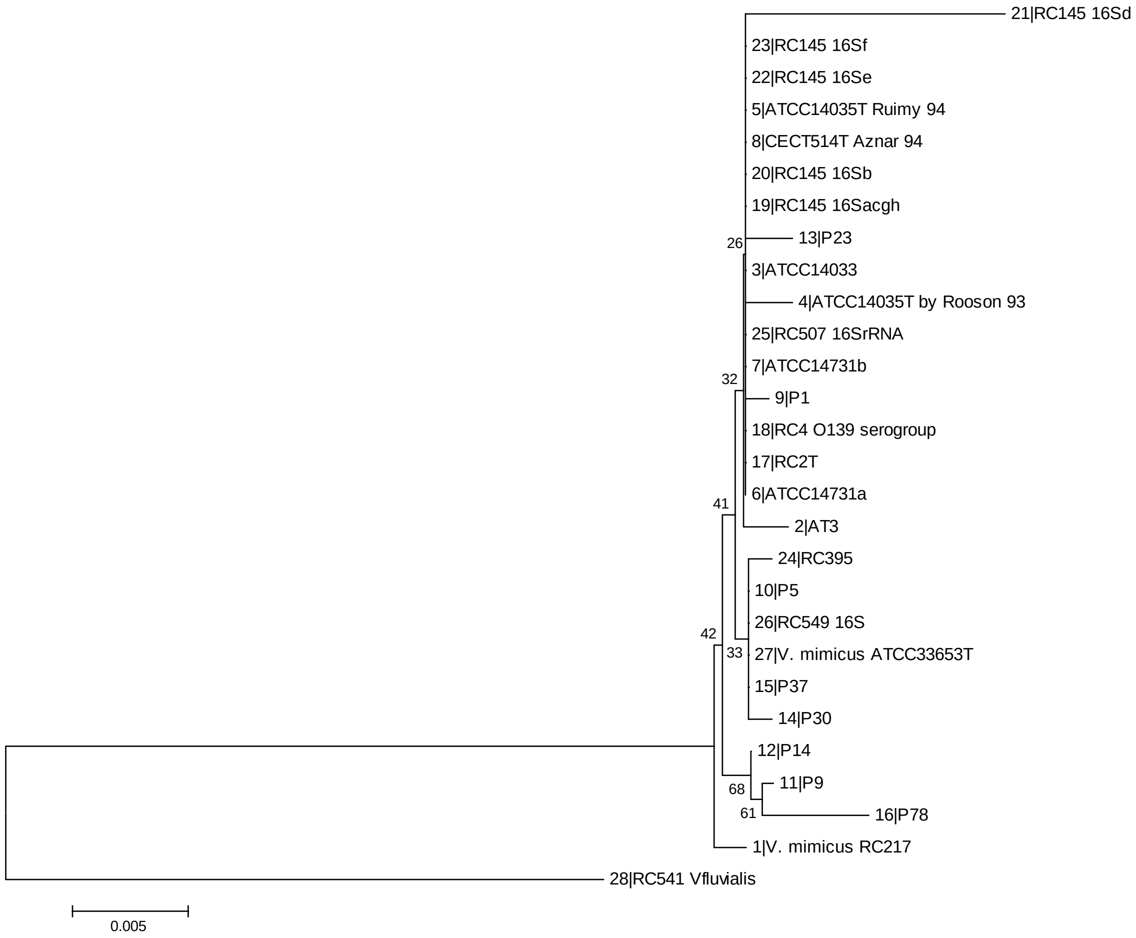


Figure 2.2. Neighbor-Joining tree generated using the Jukes-Cantor model on 16S rRNA sequences of *V. cholerae* and *V. mimicus*, with the % branch support from 1000 bootstrapping shown next to the branch (Label: serial number | strain name and other information). Characteristics of selected strains, such as toxigenicity can be found in Table 2.1. For strain RC145 (N16961), the genome of which has been fully sequenced, eight copies of the 16S rRNA genes were found and are labeled as copy ‘a’ to ‘h’. *V. fluvialis* strain RC541 served as outgroup marker. All sequences shown without species names are *V. cholerae*.

Figure 2.1 also shows the distribution of 16S rRNA genes for multiple strains of two species: *Vibrio parahaemolyticus* and *Vibrio alginolyticus*. Classification of the two species and their sister species, in most cases had been determined using conventional phenotypic identification schemes. The 16S rRNA-based phylogeny revealed some ambiguity in the definition of species established using conventional identification methods. Thus, it is hypothesized that the genomes of strains within the *V. parahaemolyticus*, *V. alginolyticus*, *V. harveyi*, *V. proteolyticus* and *V. campbellii* complex are so plastic, with extensive intermingling by LGT, that the results are chimerical genomes.

Speciation of *V. cholerae* determined by 16S rRNA sequences is characterized as (1) early deep branching from a common ancestor of the genus *Vibrio* and (2) containing only *V. mimicus* as a significant sister species. To pursue this finding in greater detail, sequences differentiating *V. mimicus* from *V. cholerae* strains were examined (Figure 2.2). Divergence among the sequences arising from artifacts, e.g., sequencing error, was visualized from the four sequences (serial number 4, 5, 8 and 17) of the type strain of *V. cholerae* (RC2^T = ATCC14035^T = CECT514^T). These four sequences were independently sequenced for the same strain by different authors. Three sequences clustered into one complex, while the forth sequence diverged within the cluster. When the 16S rRNA sequence divergence of the type strain of *V. mimicus* (ATCC33653^T = RC5^T) was compared with the *V. cholerae* sequences (Figure 2.2), it was found to be identical to that of three *V. cholerae* strains (P5, RC549 and P37). The only significant divergence was observed in the complex of three *V. cholerae* strains (P9, P14 and P78) isolated from the coastal waters of Peru. Therefore, divergence of the 16S rRNA sequence of *V. mimicus* is concluded to be negligible, compared to sequences of the *V. cholerae* population in general.

2.3.2 Phylogeny employing tmRNA sequence

In general, estimates of genetic distance by 16S rRNA and tmRNA sequences between pairs of strains were congruent by having a strong linear proportional relationship (r = 0.78, P < 0.001, Mantel's test). The slope between the standardized distance matrices for 16S rRNA and tmRNA was, however, significantly different from that with 95% CI between 0.71 and 0.88, indicating that the clock rate (i.e., mutation rate) of the tmRNA sequence is faster than the16S rRNA sequence, the implication of which is that tmRNA is under a less stringent functional constraint. A faster clock rate also suggests that tmRNA better serves as a molecular clock to resolve local evolution patterns, provided that the clock rate is also slow enough to avoid saturation of informative sites by recurrent substitutions. An important property of a molecular sequence that is selected to serve as a molecular clock is that it will not be saturated with mutations on informative sites (Page & Holmes, 1998). If saturation occurs, the resulting homoplasy will be misleading in interpreting phylogenetic relationships.

To determine whether there was saturation of tmRNA sequences by substitutions and to evaluate compliance of the molecular clock hypothesis in the evolution of tmRNA sequences among the *Vibrionaceae*, the distance matrix rate test (DMR) was performed (Syvanen, 2002). The regression for the distribution of 1,596 distance pairs was linear between tmRNA and 16S rRNA distances (Figure 2.3). Especially noteworthy is that the variation of distances at the high ends of the distance ranges could be fully explained by the linear relationship between the two distance variables and their 95% CI, based on the neutral evolution model of the molecular clock hypothesis. Visual inspection of the DMA plot, which showed little indication of saturation, together with the high significance in Mantel's test, the conclusion that there was no indication of mutation saturation of tmRNA sequences along the time span inferred by the range of 16S rRNA distances (i.e., along the evolutionary time of *Vibrionaceae*) was substantiated. Therefore, the tmRNA sequence can serve usefully as a molecular clock, without misleading homoplasy.

As determined by slope estimation, using Mantel's test (see above), the number of mutations (i.e., JC distance) per site was higher for tmRNA sequences than for 16S rRNA sequences, i.e., a slope value of 0.3. Therefore, the faster rate of tmRNA mutation per site was confirmed by the DMR test. Considering the total number of mutations along the full length of the aligned sequences, the slope between 16S rRNA and tmRNA mutations was 1.2, implying that 16S rRNA carried 20% more mutations than tmRNA, in spite of the slower clock rate per site, most probably caused by the longer length (hence, more informative sites) of 16S rRNA. The conclusion, then, is that 16S rRNA is, indeed, a more accurate clock (i.e., less vulnerable to bias by a few mutation sites), and tmRNA carries less evolutionary information (about 80%).

Compliance of the molecular clock hypothesis for tmRNA mutations was examined by counting the number of strain pairs whose genetic distances were located out of 95% CI in the DMR test plot (Figure 2.3) (hereafter, referred to as an outlier). Out of 1,596 pairs of strains, 1,422 (89%) fell within the 95% CI while 174 (11%) were outliers, indicating that the majority of base differences in a pair of tmRNA sequences can be explained by neutral and random base substitution events in each of the sequences.

However, a significant number of pairs (11%) showed sequence differences between tmRNA or 16S rRNA sequences, significantly deviating from the molecular clock hypothesis. Because the mutation rate of the 16S rRNA sequence conforms to the molecular clock hypothesis when only a local level of evolution is considered, such as evolution within a taxonomic group below the level of kingdom (Syvanen, 2002), the source sequence deviating from the hypothesis must be tmRNA rather than 16S rRNA. From this finding, the source of variation in the tmRNA sequence among outliers most likely includes other molecular mechanisms, in addition to neutral and random base substitutions.

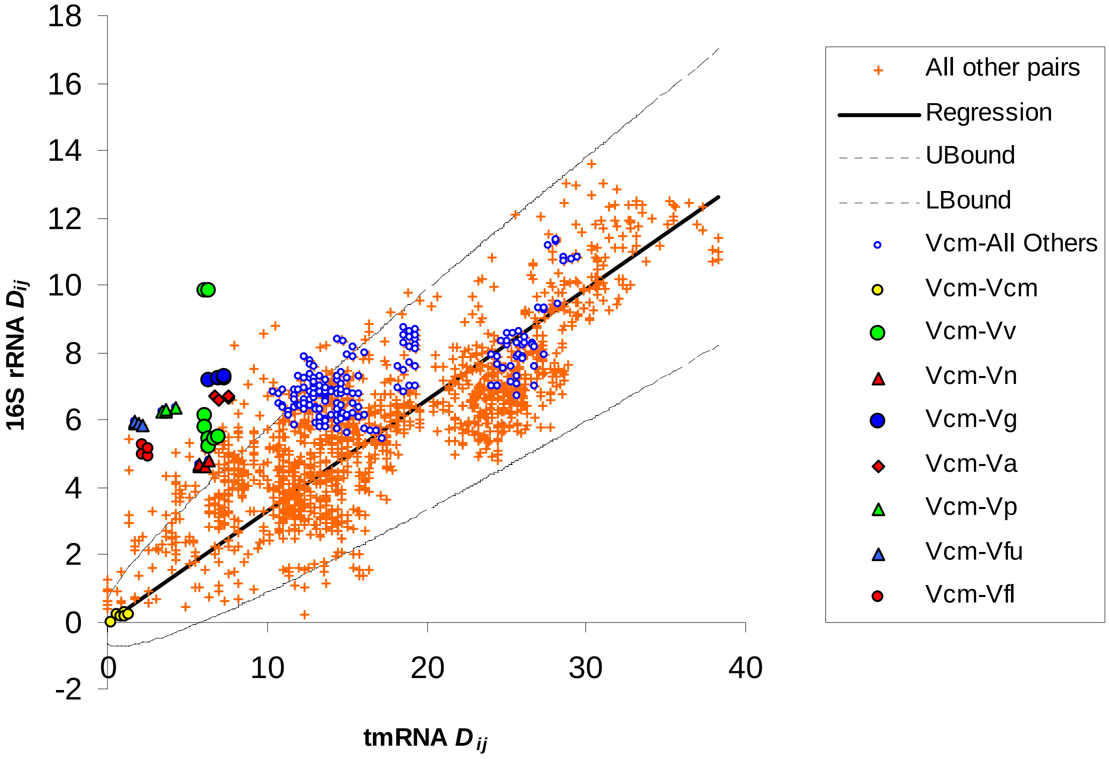


Figure 2.3. Distance matrix rate (DMR) test plot between 16S rRNA and tmRNA sequences using Jukes-Canter distance per 100 aligned bases of *i*th strain and *j*th strain (D_{ij}). The regression line (solid line) was derived from the total number of replacements along the full length of the sequences and transformed to be consistent with the D_{ij} values. The same transformation was done for 95% CI values (dashed lines). D_{ij} pairs were plotted with crosses, except for pairs including *V. cholerae* or *V. mimicus* (Vcm). Pairs including Vcm were further separated by species of their counterpart strains: *V. cholerae* or *V. mimicus* (Vcm), *V. aerogenes* (Va), *V. gazogene* (Vg), *V. fluvialis* (Vfl), *V. furnissii* (Vfu), *V. vulnificus* (Vv), *V. proteolyticus* (Vp) and *V. navarrensis* (Vn).

To understand outlying deviations, distribution of the outliers was further examined using the DMA test (Figure 2.3). Interestingly, outliers occurred only when the genetic distances were relatively small, i.e., $D_{ij} < 20$ for the tmRNA and $D_{ij} < 10$ for 16S rRNA. Knowing that the error level determining 95% CI is proportional to phylogenetic distance between two strains, confinement of outliers within a relatively short range of distance occurred in two cases. The first is when the extent of outlier deviation is constant, regardless of genetic distance. As the error level increases with genetic distance between strain pairs, the relative magnitude of outlying deviation becomes smaller, so that the outlying deviation can be detected only when the distances are small. The second case occurs when the outlying deviation is reversely proportional to phylogenetic distance, i.e., it increases in proportion to the relatedness of strains. From results of this study, it was difficult to determine which case prevailed. However, it can be concluded that the force causing outlying deviation in tmRNA sequences among closely related species predominates over neutral mutations of the molecular clock hypothesis.

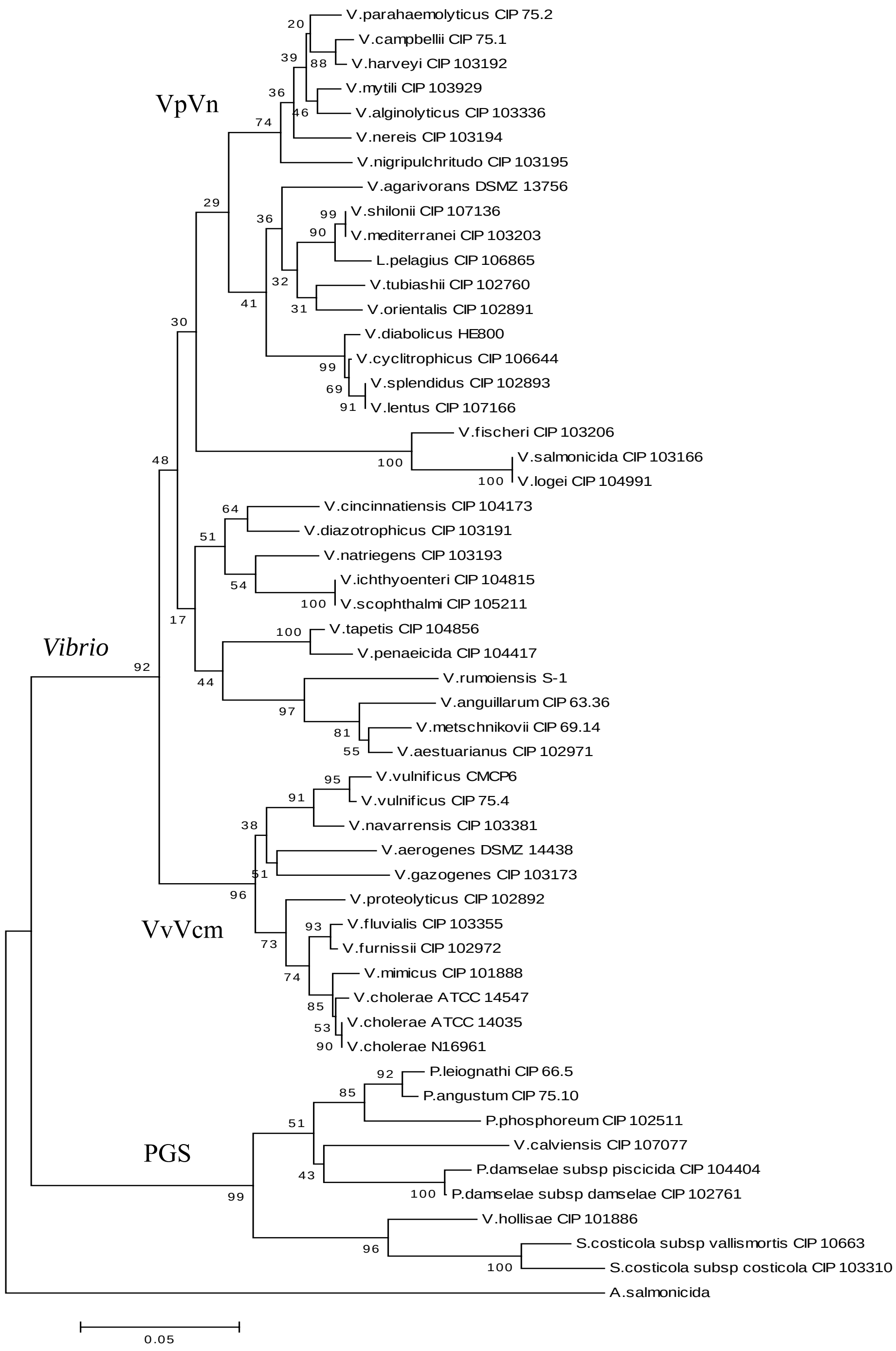


Figure 2.4. Neighbor-Joining tree developed using the Jukes-Cantor model for tmRNA sequences with % branch support from 1000 bootstrapping shown next to branches (Branches producing significant clusters are labeled by cluster names).

With an interest in the distribution of outlying deviations across hierarchies of taxonomic units, i.e., species, supra-species, and genus, clusters of species were identified from the tmRNA tree (Figure 2.4), and their distribution was analyzed using the results of the DMR test. Species were divided into two supra-genus clusters (*Vibrio* and PGS). The genus *Vibrio* displayed two supra-species clusters (VpVn and VvVcm). PGS was a supra-genus cluster whose distance range (from zero to maximum D_{ij} values in Table 2.1), based on tmRNA and 16S rRNA sequences, was slightly larger than the range for outliers, i.e., 0-20 by tmRNA or 0-10 by 16S rRNA. The other two clusters comprised intra-genus clusters, with a distance range less than that of the outliers. Therefore, the threshold of relatedness among strains at which outlying deviation is the predominant driver for the tmRNA sequence variation, is located around the genus differentiation. This estimation is additionally supported by proximity of maximum D_{ij} values for the *Vibrio* cluster (Table 2.1) with respect to the range of outliers.

Table 2.1. Characteristics of the clusters derived from the tmRNA tree.

			Total	Outliers		Maximum D_{ij}	
		No. of	strain		% to total		16S
Cluster	Level of cluster	species	pairs	No	strain pairs	tmRNA	rRNA
PGS	Supra-genus	7	55	18	33%	23.26	10.12
Vibrio	Genus	38	990	155	16%	20.42	9.84
VpVn	Supra-species	7	21	3	14%	5.16	3.05
VvVcm	Supra-species	9	66	44	67%	8.83	9.84

Regarding distribution of outliers among species within a cluster, two alternative hypotheses were explored: namely that outlying deviations occur uniformly across all species and, alternatively, that deviations arose only for a given species, rather than universal among related species. As shown in Table 2.2, the latter was supported by the highly skewed distributions of outliers for only a few species in each cluster. All three outliers in cluster VpVn contained *V. mytili* as the common species. The 32 (73%) of 44 outlying distance estimates in the VvVcm cluster had strains of the Vcm complex. Using the parsimonious counting method (Table 2.2), the three farthest outlier species of clusters PGS and *Vibrio* involved 94% and 54% of the total outliers, respectively.

The most notable discrepancy between trees constructed from tmRNA and 16S rRNA sequences was clustering of *V. vulnificus*, *V. fluvialis* and *V. cholerae* based on tmRNA sequences. In contrast to the tmRNA tree topology, the Vcm complex did not have any sister species around the lineage, based on 16S rRNA sequences (Figure 2.1). This difference in inference from 16S rRNA and tmRNA sequences on relationships of *V. cholerae* with other vibrios was reflected in both the frequency of outliers and the magnitude of outlying deviations in the Vcm complex. In the *Vibrio* and VvVcm clusters, Vcm species were most frequently outliers. Extreme outliers in the DMR plot also occurred in the distance estimates for Vcm species paired with *V. vulnificus*, V. *fluvialis*, *V. proteolyticus*, *V. navarrensis* or *V. aerogenes* (Figure 2.3). Since the DMR plot produced a band of outlying Vcm points at the high end of 16S rRNA distance, the discrepancy is caused by various levels of similarity for tmRNA sequence of Vcm species with other VvVcm species.

Table 2.2. Distribution of outliers by species in each cluster

Cluster	Species	No. of strains	No. of strain	Direct outlier count [a]		Parsimonious outlier count [a]	
				No.	%[b]	No.	%[b]
PGS	*V. calviensis*	1	10	7	70%	7	70%
	P. angustum	1	10	6	60%	4	40%
	P. leiognathi	1	10	4	40%		
	P. damselae	2	20	6	30%		
	S. costicola	2	20	6	30%	6	30%
	G. hollisae	1	10	3	30%		
	P. profundum	1	10	2	20%		
	P. iliopiscarium	1	10	1	10%		
	P. phosphoreum	1	10	1	10%		
Vibrio	*V. cholerae* and *V. mimicus*	4	176	60	37%	60	34%
	V. diazotrophicus	1	44	15	34%	13	30%
	V. wodanis	1	44	13	30%	10	23%
	V. cincinnatiensis	1	44	11	25%	8	18%
	V. proteolyticus	1	44	11	25%	6	14%
	V. harveyi	1	44	10	23%	5	11%
	V. mytili	1	44	10	23%	3	7%
	V. natriegens	1	44	10	23%	4	9%
	V. vulnificus	2	88	19	22%	8	9%
	V. aerogenes	1	44	9	20%	2	5%
	V. gazogenes	1	44	9	20%	1	2%
	V. lentus	1	44	8	18%	8	18%
	V. agarivorans	1	44	7	16%	5	11%
	V. tubiashii	1	44	7	16%	3	7%
	V. anguillarum	1	44	7	16%		
	V. furnissii	1	44	7	16%		
	V. parahaemolyticus	1	44	7	16%		
	V. alginolyticus	1	44	6	14%		
	V. fluvialis	1	44	6	14%		
	V. mediterranei	2	88	11	13%	2	2%
	V. diabolicus	1	44	5	11%		
	V. ichthyoenteri	1	44	5	11%		
	V. navarrensis	1	44	5	11%		
	V. pelagius	1	44	5	11%		
	V. salmonicida	1	44	5	11%		
	V. scophthalmi	1	44	5	11%		
	V. halioticoli	1	44	4	9%		
	V. logei	2	88	8	9%		
	V. tapetis	1	44	4	9%		
	V. fischeri	2	88	7	8%		
	V. aestuarianus	1	44	3	7%		

	V. penaeicida	1	44	3	7%		
	V. nereis	1	44	2	5%		
	V. orientalis	1	44	2	5%		
	V. splendidus	1	44	2	5%		
	V. campbellii	1	44	1	2%		
	V. metschnikovii	1	44	1	2%		
	V. nigripulchritudo	1	44	0	0%		
VpVn	*V. mytili*	1	6	3	50%	3	50%
	V. alginolyticus	1	6	1	17%		
	V. harveyi	1	6	1	17%		
	V. parahaemolyticus	1	6	1	17%		
	V. campbellii	1	6	0	0%		
	V. nereis	1	6	0	0%		
	V. nigripulchritudo	1	6	0	0%		
VvVcm	*V. cholerae* and *V. mimicus*	4	44	32	73%	32	73%
	V. vulnificus	2	22	15	68%	6	27%
	V. proteolyticus	1	11	7	64%	3	27%
	V. aerogenes	1	11	9	82%	2	18%
	V. gazogenes	1	11	9	82%	1	9%
	V. furnissii	1	11	6	55%		
	V. fluvialis	1	11	5	45%		
	V. navarrensis	1	11	5	45%		

a: Direct outlier count was made by counting outlying distance estimates for any pair of strains. In this case, an outlier was counted twice: once for each of the strains in a pair. In parsimonious counting, it is assumed that outlying deviation is caused by only one of the strains in a pair. The percent value of outliers for a species was considered to indicate the likelihood of a species to cause the outlier deviation. Therefore, the species with greater outlier frequency, i.e., greater % of outliers by direct count, was considered as the strain causing outlying deviation in a pair of strains. By this method, an outlier was counted only for a species causing outlier deviation. In cluster VpVn, *V. mytili* is concluded to be the outlier species because all three outliers in the cluster had this species in common. In the VvVcm cluster, the % value using the parsimonious outlier count for the *Vibrio* cluster, was considered as the likelihood indicator because the *Vibrio* cluster had at least four times more pairs and outliers. Thus, the indicator was more precise than the % direct outlier count from only 11 pairs.

b: % to number of strain pairs

Therefore, properties of the outlying deviations are as follows. The outlying deviation is the predominant source of sequence variation of tmRNA among species when they belong to the same genus. It occurs only for certain species rather than as a general phenomenon of all species of a genus. The tmRNA phylogeny within a genus can be concordant with that of 16S rRNA, if outlying species are eliminated from the phylogeny construction. Finally, outlying species are not closely related by 16S rRNA but by tmRNA.

From these properties, the best explanation for the uncoupling of evolutionary processes based on 16S rRNA and tmRNA for certain species, as in the case of *V. cholerae*, is lateral gene transfer (LGT) of tmRNA genes from species within a cluster to outlying species of another cluster. LGT would be less effective in the case of 16S rRNA because of functional constraint and dilution effect, if multiple copies of the gene are present. However, functional constraint for tmRNA is relatively less, so that sequences can be exchanged between species within the same genus, perhaps with minor modifications. A species phylogenetically remote from a cluster of closely related species would receive tmRNA from strains within the cluster by LGT. Therefore, the observed relationship between 16S rRNA and tmRNA sequences for the *Vibrionaceae* can be explained by events of LGT during evolution of the strain cluster.

The implication of LGT-driven sequence variation that occurs in essential bacterial genes is a forcing toward homogenization of the genomic makeup of strains of related bacterial species. The tmRNA sequences of *V. cholerae*, *V. vulnificus* and *V. fluvialis* provide one such example. That is, they are not sister species, but their tmRNA sequences have converged to a close similarity, sufficiently high to comprise the VvVcm cluster. If the same kind of convergence occurs for many of housekeeping genes, the genomic makeup of the species carrying those genes will become more similar.

2.3.3 Evidence for lateral transfer of the tmRNA gene

To substantiate the LGT hypothesis, the portability of tmRNA genes among closely related strains was assessed for the Vcm complex by analysis of allelic diversity and sequence variability among flanking regions of the tmRNA gene.

Allelic diversity of the tmRNA sequences for *V. cholerae* and *V. mimicus* is presented in Table 2.3. Eight alleles (VC1 – VC8) were found in *V. cholerae* and five (VM1 – VM4, and VC1) in *V. mimicus*. Significant phylogenetic separation of the VM from VC alleles was detected (Figure 2.5). Using the JC-NJ tree method of analysis, clusters of VC and VM alleles were observed with >50% bootstrap support. These clusters were strongly supported by results of analysis using the KHY-QP tree method at > 95% branch support. Therefore, VC alleles, in general, are distributed among *V. cholerae* and VM alleles among *V. mimicus* strains. However, one *V. mimicus* strain, RC217, carried the VC1 allele, the predominant *V. cholerae* allele. Divergence among the VC alleles was calculated according to one to three pairwise base differences. Sequence differences within the cluster of four VM alleles also ranged from 1 to 3 nucleotides. The pairwise difference between VC1 and VM1 – VM4 ranged from 4 to 6 nucleotides. Therefore, divergence in the tmRNA sequences of the respective strain clusters cannot explain the appearance of the VC1 allele in *V. mimicus* RC217. Instead, LGT of the tmRNA gene between *V. cholerae* and *V. mimicus* is a more logical explanation than an accumulation of point mutations.

Table 2.3. Distribution of tmRNA alleles among 30 strains of *V. cholerae* and *V. mimicus*.

Allele	*V. cholerae* strains	*V. mimicus* strains	Number of alleles
VC1	RC3^{aCT}, RC4^{cCT}, RC33 ^{cCT}, RC44, RC45, RC47, RC48, RC145 ^{bCT}, RC215^{aCT}, P9, P30, BLSTVC	RC217^{CT}	13
VC2	RC2 ^{aCT}		1
VC3	ATCC14547		1
VC4	RC521, P5		2
VC5	RC360, RC549		2
VC6	RC507		1
VC7	P78		1
VC8	P23		1
VM1		RC5	1
VM2		RC59	1
VM3		RC54, RC57, RC218^{CT}, RC219^{CT}	4
VM4		RC55^{CT}, RC56^{CT}	2

a: O1 classical biotype

b: O1 El Tor biotype

c: O139 serogroup

CT: cholera toxin (CT) positive and toxin-coregulated pili (TCP) positive

To seek further evidence for LGT among strains of *V.* cholerae and *V. mimicus*, the chromosomal regions flanking the tmRNA genes of *V. cholerae* and *V. mimicus* (Table 2.3) were sequenced using inverse PCR. All of the downstream region sequences were conserved, including the gene coding for the small protein B (SmpB), one of the macromolecules that participate in the ribosome rescue function of tmRNA. However, variability was found upstream of the tmRNA genes. In the case of CT positive strains carrying the *Vibrio* pathogenicity island (VPI), a phage-like integrase gene (*int*) was observed always to flank upstream of the tmRNA genes. At the same position, a strain that did not carry the VPI had VC0816, an hypothetical protein shown to be located at the upstream junction of the VPI on the *V. cholerae* N16961 chromosome. It was recently reported that the VPI uses the chromosomal tmRNA location as the target of insertion and excision via its integrase and a transposase (VpiT) (Rajanna *et al.*, 2003). Results from that study also showed that the excision process caused functionally deleterious mutations in the tmRNA gene. Therefore, an intact, or at least a partially intact tmRNA sequence from a VPI-positive strain has to be transferred to the chromosome of recipient strains when a VPI-negative strain is transformed to VPI-positive via excision-transformation. In the case of the VPI-positive *V. mimicus* RC217, the VC1 allele from a *V. cholerae* VPI-donor strain most likely was transferred to RC217 when it acquired the VPI. This example supports the hypothesis that at least a part of the tmRNA gene can be horizontally transferred by piggy-backing on mobile transposable elements using tmRNA as the target of chromosomal insertion.

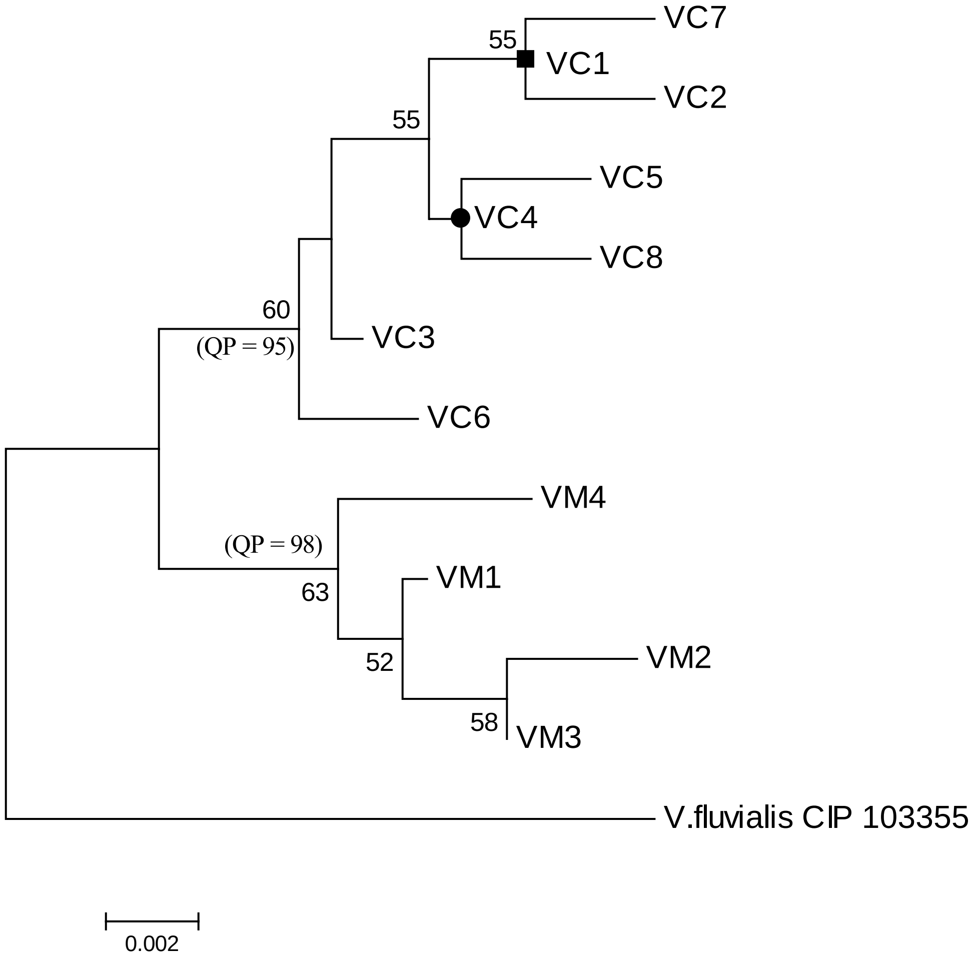


Figure 2.5. Neighbor-Joining tree produced using the Jukes-Cantor model with sequences of 12 different tmRNA alleles from 30 *V. cholerae* and *V. mimicus* strains. Percent branch support shown next to the branches from 1000 bootstrapping. Percent QP support values from 1000 puzzling steps of QP on HKY model are in parentheses. The nodes marked with closed square and closed circle are the locations of the VC1 and VC4 alleles, respectively.

2.3.4 Phylogenetic analysis employing other housekeeping genes

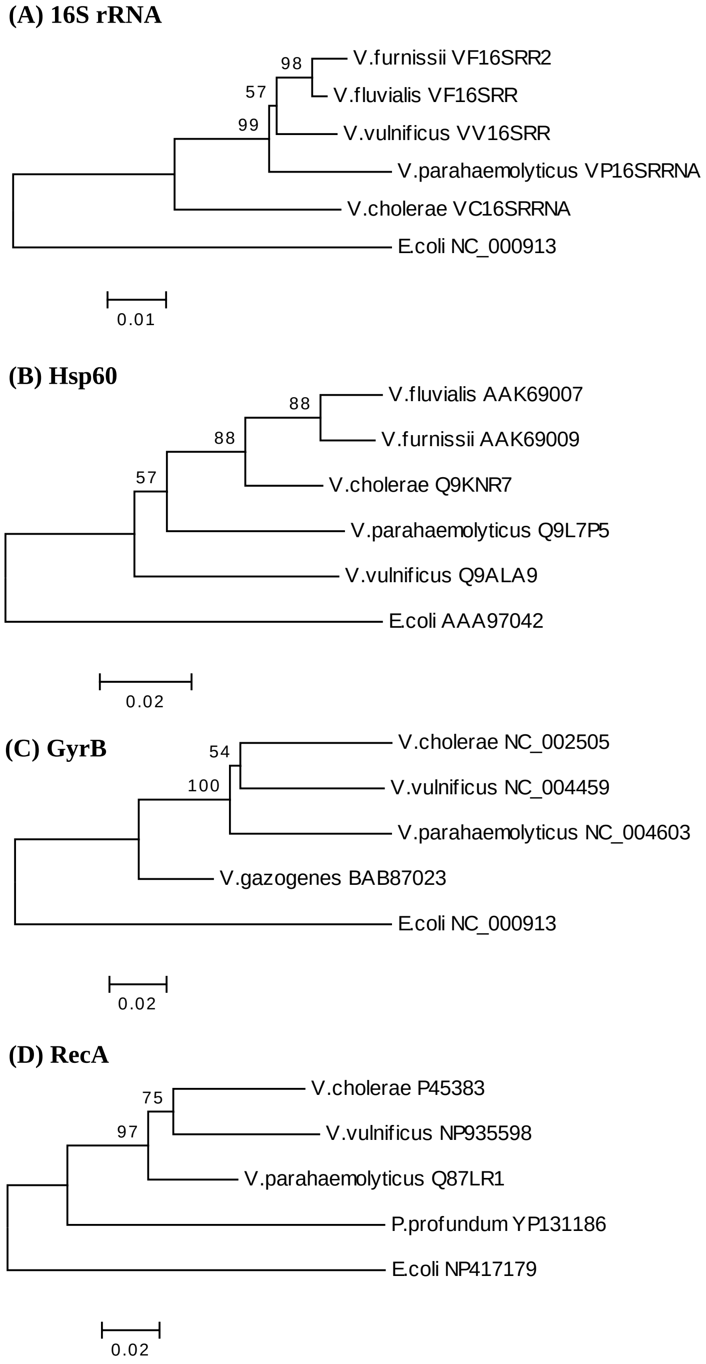
According to the observations described above, tmRNA genes are affected more by LGT than 16S rRNA because of the relatively loose functional constraints. Although strong resistance against LGT sustains the accuracy of the 16S rRNA phylogeny, the genome-wide phylogeny for bacterial species, nevertheless, will be affected by LGT, especially if a significant portion of the genome is affected by LGT. Therefore, the question of how general the LGT effect is for the bacterial genome must be answered and prevalence of LGT among bacterial genes are considered in this section. A set of housekeeping genes were selected and their sequences collected from public databases. Because the information available for these genes is not as extensive as for tmRNA, *V. cholerae*, where significant uncoupling between the 16S rRNA and tmRNA phylogenies has been observed, was selected for analysis, rather than the complete range of the *Vibrionaceae*.

Using tmRNA phylogeny, sister species of *V. cholerae* can be defined as those species within the VvVcm complex along with *V. fluvialis*, *V. furnissii*, *V. vulnificus*, and *V. gazogenes*. Unlike typical sister species, which share direct ancestry, these are sister species because of similar genetic information arising from homogenization via LGT. Housekeeping genes available for these species were *hsp60*, *gyrB*, and *recA*. Because a partial DNA sequence for a protein coding gene can be misleading in a phylogenetic analysis, only the most complete protein sequences were used. When the topologies of the Hsp60, GyrB, and RecA protein sequences were compared with 16S rRNA (Figure

2.6), they were found to be congruent with the tree topology of tmRNA, rather than that of 16S rRNA.

The Hsp60 gene provided a unique opportunity to compare two species of greatest interest, namely *V. fluvialis* and *V. furnissii*, with *V. cholerae*. By having strong bootstrap support for the *V. fluvialis* and *V. furnissii* clustering with *V. cholerae*, the Hsp60 of *V. cholerae* was most similar to those of *V. fluvialis* and *V. furnissii*. Their topology was identical to tmRNA (Figure 2.4). Recently, the nucleotide-based phylogeny for the *recA* gene was reported by Thompson *et al.* (2004a), in which the partial DNA sequences of *recA* for *V. fluvialis* and *V. furnissii* clustered very closely with *V. cholerae*. Therefore, a sister species relationship between *V. cholerae* and *V. fluvialis* and *V. furnissii* appeared to be more broadly spread among diverse housekeeping genes. In the case of the GyrB and RecA genes, *V. vulnificus* was a sister species suitable for comparison with *V. cholerae*.

The *V. gazogenes* and *V. cholerae* relationship based on the GyrB gene, however, appeared to be dissimilar to the case from the tmRNA gene, in that *V. gazogenes* was more distantly related than *V. parahaemolyticus* to *V. cholerae*. Thus, tree topology based on tmRNA, shown in Figure 2.4, cannot be generalized to all housekeeping genes. Considering that LGT is stochastic, i.e., involving random encountering processes between donor and recipient organisms, the observation implies that the outcome of LGT may vary from organism to organism and from gene to gene. Therefore, LGT on housekeeping genes causes not only an uncoupling of evolution between 16S rRNA and other genes, but it also results in an uncoupling in the direction of sequence variation among the housekeeping genes themselves. In fact, the more general nature of variability in LGT direction of different genes is presented in the next section based on the phylogenomic analyses.



(A) 16S rRNA

(B) Hsp60

(C) GyrB

(D) RecA

Figure 2.6. Neighbor-Joining trees prepared using the Jukes-Cantor model for 16S rRNA and the Dayhoff model for amino acid sequences of the Hsp60, GyrB, and RecA genes (percent bootstrap support from 1000 bootstrapping is shown above branches; taxon labels are species names, followed by accession number).

2.3.5 Phylogenomics of the *Vibrio* triad

The complete genomic sequences for *V. parahaemolyticus* and *V. vulnificus* are available in the NCBI GenBank, making it possible to evaluate the contradiction in the phylogenetic relationships established by 16S rRNA and tmRNA using genome-wide comparisons of gene similarity for *V. parahaemolyticus*, *V. vulnificus*, and *V. cholerae*. According to the tmRNA topology, *V. vulnificus* is a member of the sister cluster of *V. cholerae* while *V. parahaemolyticus* is only remotely related. In contrast, according to the 16S rRNA phylogeny, *V. vulnificus* is more related to *V. parahaemolyticus* than to *V. cholerae*. Therefore, the question whether *V. vulnificus* (and, therefore, *V. furnissii* and *V. fluvialis*) is a sister species of *V. cholerae* can be answered by quantitatively contrasting the relatedness of the *V. vulnificus – V. cholerae* pairing to the *V. vulnificus- V. parahaemolyticus* pairing.

Relatedness among genomes can be assessed from two separate aspects: content and structure. The former corresponds to information contained in the genome and the latter is physical organization of information contained in the genome. Relatedness according to content can be expressed by presence/absence of homologous genes and gene phylogeny when homologous genes are present. Relatedness by genome structure can be determined by shape and number, and size of chromosomes, synteny (i.e., identity in physical ordering of genes along a chromosome), and distribution of chromosome "landscaping" motifs, such as intergenic repeated sequences and super-integron islands. In this section, structural relatedness of genomes based on the features cited here and the cohesion of evolutionary paths of genetic content were examined.

A Chromosome synteny

While all species of the *Vibrionaceae* contain two circular chromosomes, the size of the chromosomes of *V. vulnificus* and *V. parahaemolyticus* are similar, e.g., 3.8 and 1.8 Mbp, respectively, in their large and small chromosomes. *V. cholerae*, however, has smaller chromosomes than *V. vulnificus* and *V. parahaemolyticus*, namely 3.0 and 1.1 Mbp, respectively.

When 1,535 orthologs were identified and plotted along linearized nucleotide coordinates from the replication origin (Figure 2.7), a more compact and linear distribution was found between the genomes of *V. vulnificus* and *V. parahaemolyticus*, indicating strong synteny. The starkest contrast in synteny was in chromosome 2 of *V. cholerae*, which does not have significant synteny with *V. vulnificus*. Another contrast is the frequent reciprocal synteny, i.e., symmetrically located in the negative strand across the replication origin, obvious when compared with *V. cholerae* but not *V. parahaemolyticus*. Therefore, in both quantity of genes and their alignment in the genome, *V. vulnificus* is more closely related to *V. parahaemolyticus* than to *V. cholerae*.

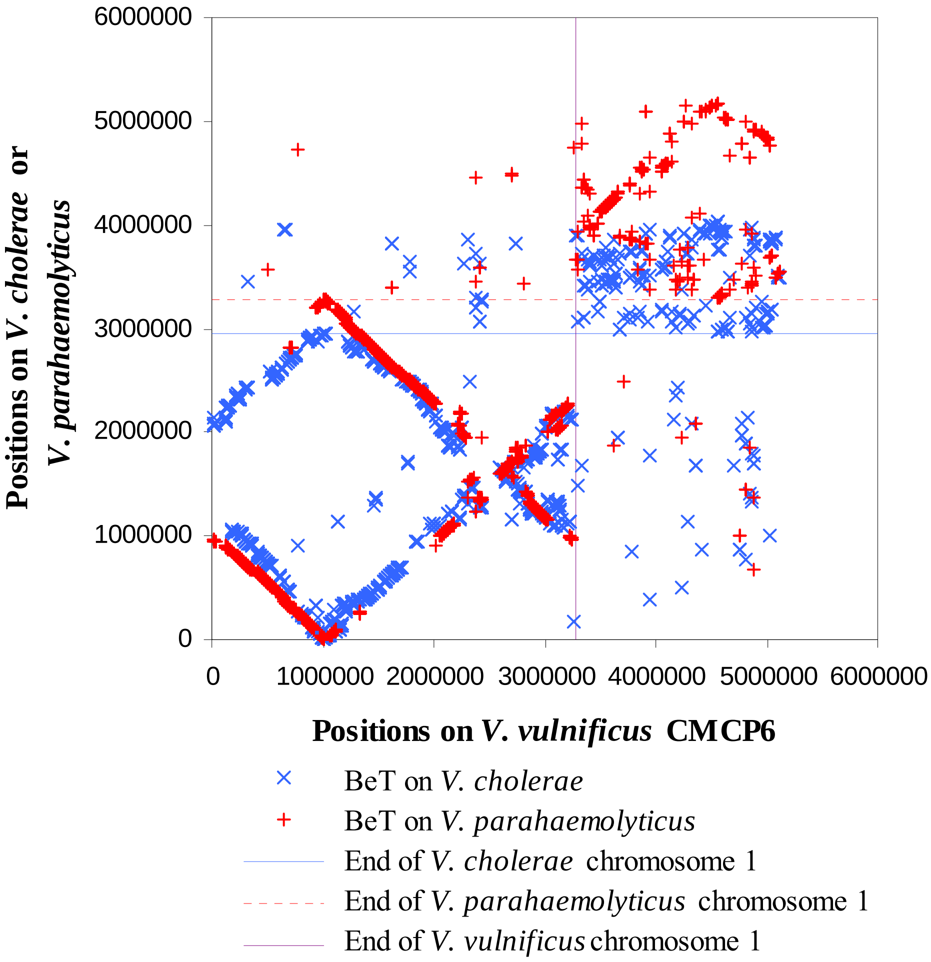


Figure 2.7. Distribution of *Vibrio* Triad orthologs along genomes from replication origin. The origin of the small chromosome is concatenated at the end of the larger chromosome. Symbols: Blue X is *V. cholerae* coordinates in the Y axis, and Red + *V. parahaemolyticus* coordinates.

B Chromosomal distribution of intergenic repeated sequences

Since the prime source of reciprocal synteny is replication inversion (Rocha, 2004), finding prevalent reciprocal synteny in the *V. cholerae* chromosomes raises the possibility of a wide distribution of interspersed repeated sequences (IRS) in the *V. cholerae* genome. Because high occurrence of IRS affects the structure and stability of genomes by increasing the possibility of genome inversion via homologous recombination among the IRS (Achaz *et al.*, 2003), the presence of IRS in only one species can result in its divergence from the evolutionary paths of closely related species.

When all known bacterial repeated sequences cited by Versalovic and Lupski (1998) were searched in three *Vibrio* genomes, only two kinds of inverted repeat sequences were found (Figure 2.8 and Figure 2.9), both related to the enterobacterial repetitive intergenic consensus (ERIC) sequences but with size and sequence divergences. The presence of these sequences in *V. cholerae* has been reported previously (Rivera *et al.*, 1995), but their frequency and the significance of their distribution were not analyzed. In the section below, the distribution of these two ERIC-like intergenic motifs (referred to here as VCRIC motifs) across the three *Vibrio* genomes was examined from the perspective of their effect on the genome structure and phylogeny of *Vibrio* spp.

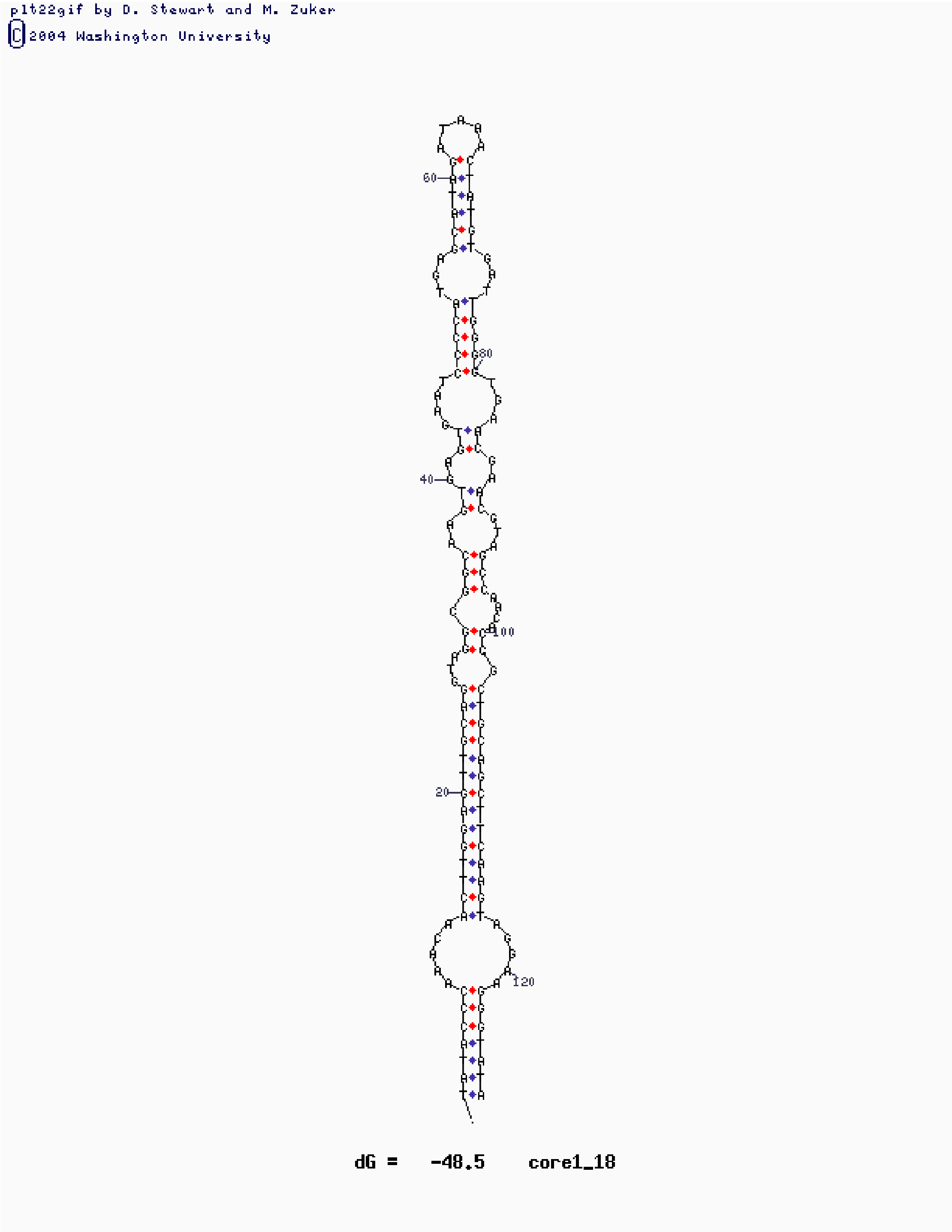

plt22gif by D. Stewart and M. Zuker
©2004 Washington University
60
80
40
100
20
120
dG = -48.5 core1_18

Figure 2.8. Secondary structure of a typical 128-bp VCRIC RNA sequence.

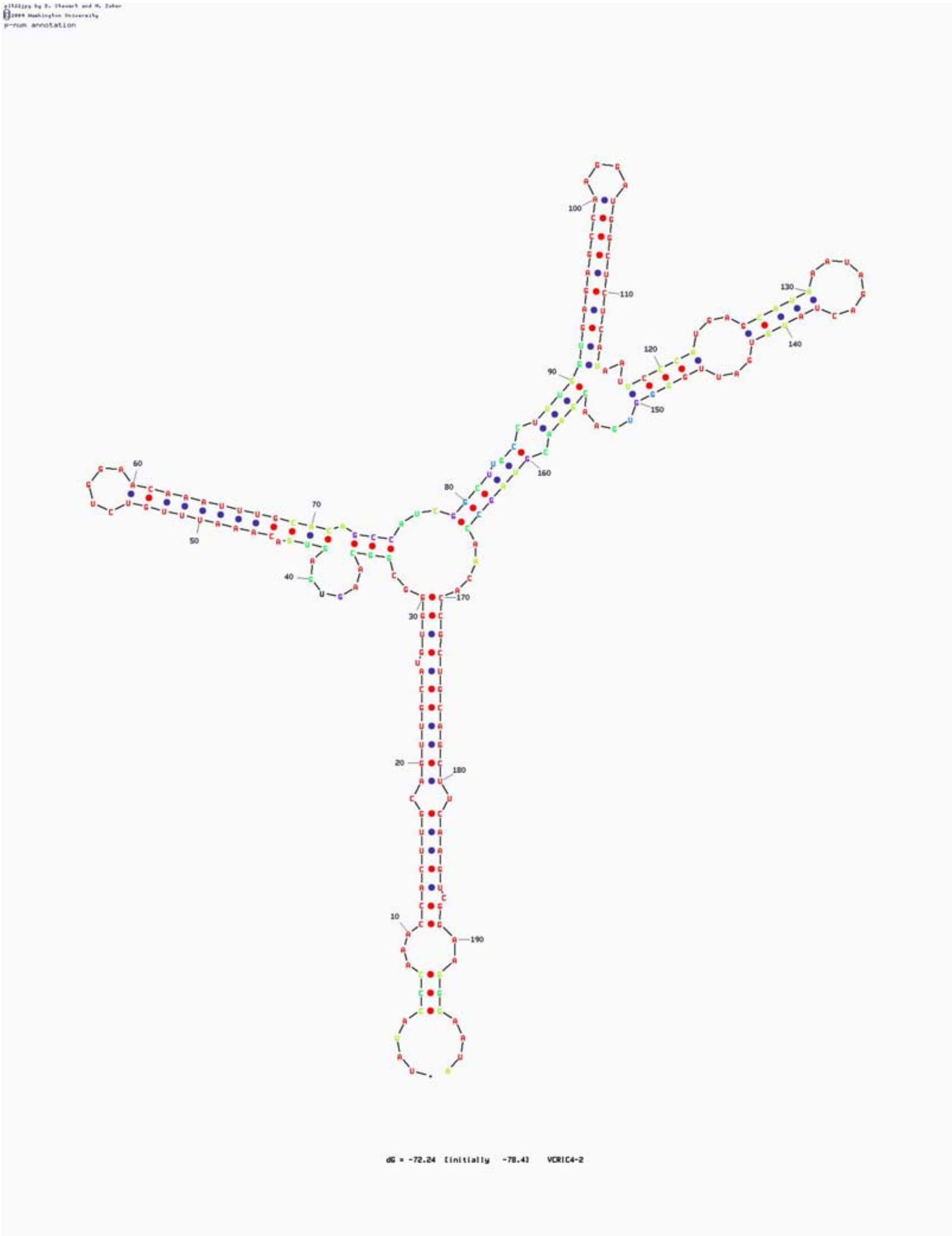

plt22jpg by D. Stewart and M. Zuker
© 2004 Washington University
p-num annotation
100
110
130
140
120
90
150
60
70
160
80
50
40
30
170
20
180
10
190
dG = -72.24 [initially -78.4] VCRIC4-2

Figure 2.9. Secondary structure of a typical 198-bp VCRIC RNA sequence

According to results of a BLAST search (Figure 2.10), the 128-bp motif (labeled "motif" because the sequences have variability within a conserved structure, as shown in Figure 2.11) was found 108 times along both chromosomes of *V. cholerae*, while the 198-bp motif was found to occur only 10 times in the same genome. Besides the 208 loci, more than one hundred instances of significant partial matches of the 128-bp motif (i.e., >25 bp long and < 10^{-5} BLAST expect value) were found along both chromosomes of *V. cholerae*. They occurred mainly in the intergenic loci, suggesting that they represent remnants from numerous recombination events. From the distribution of the intact 118 loci, it is noteworthy that the motifs are not located within highly-conserved gene clusters, e.g., rRNA operons or ribosomal protein operons, or within large gene cassettes, such as VPI, the mannose sensitive haemoagglutin (MSHA) operon, the VCR super-integron island, and lipopolysaccharide (LPS) cell wall antigen synthesis genes (Figure 2.10). When variability of the 128-bp VCRIC motif was examined by alignment of the 32 most frequent sequences, 11 nucleotide locations were found to vary significantly (Figure 2.11) and were all located at the loop regions of the simulated secondary structure (Figure 2.8). A BLAST search using the consensus VCRIC sequence (Figure 2.11), found only six equivalent copies in the *V. vulnificus* genome and none in *V. parahaemolyticus*. The 128-bp *V. vulnificus* VCRIC motifs differed from the VCRIC by two bases.

Therefore, distribution of the VCRIC among *Vibrio* species can be summarized as being widespread and active in recombining VCRIC motifs in *V. cholerae*. Extending these findings to relatedness among the three *Vibrio* species, *V. vulnificus* can be considered chimerical between *V. cholerae* and *V. parahaemolyticus*. The high genome synteny observed between *V. vulnificus* and *V. parahaemolyticus* supports a close relatedness between them; however, the lack of synteny in *V. cholerae* could arise from the widely distributed VCRIC. While species-specific distributions can explain the prevalence of reciprocal synteny in the *V. cholerae* genome and the weakness of synteny between *V. cholerae* and other vibrios, the reason for low copy number and resulting weakness of the effects of VCRIC in *V. vulnificus* is not clear, requiring further investigation of the function of VCRIC motifs in *Vibrio* species in general.

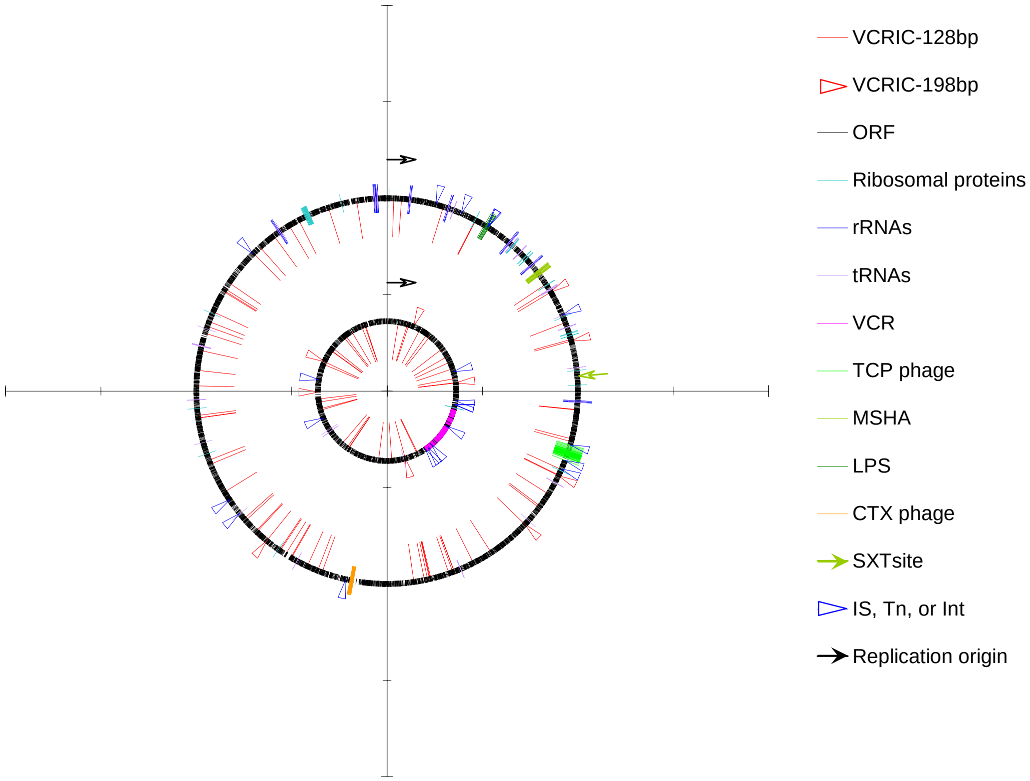


Figure 2.10. Distribution of ERIC-like sequences on the large chromosome (outer circle) and small chromosome (inner circle) of *V. cholerae* O1 El Tor N16961. Black circles are formed by the protein coding ORF loci. One hundred and eight red inward spines are 128-bp VCRIC sequence loci and ten red triangles on the outside of circles are 198-bp VCRIC loci.

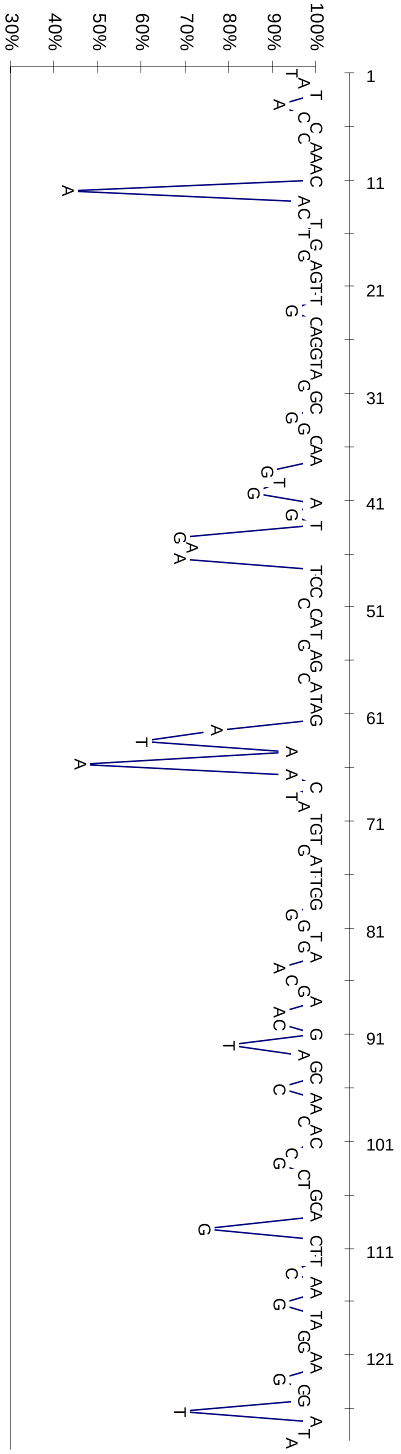
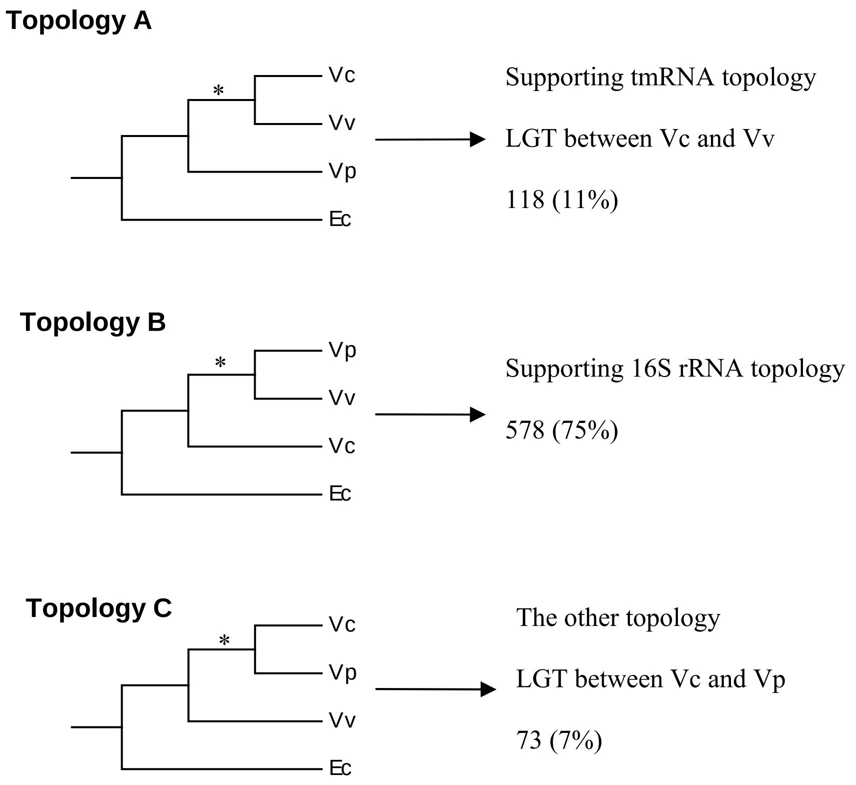


Figure 2.11. Conservation of nucleotides along consensus 128-bp VCRIC DNA sequences. The consensus sequence was derived from 35 intact 128-bp fragments.



Supporting tmRNA topology

LGT between Vc and Vv

118 (11%)

Supporting 16S rRNA topology

578 (75%)

The other topology

LGT between Vc and Vp

73 (7%)

Figure 2.12. Possible outcomes from a quartet comprised of *V. cholerae* (Vc), *V. vulnificus* (Vv), *V. parahaemolyticus* (Vp) and the outgroup *E. coli* (Ec). The quartet is regarded as resolved when the branch marked with * is significant by Bayesian posterior probability ($P > 0.95$). *E. coli* is regarded as the truthful outgroup, based on the difference in genome anatomy (i.e., two chromosomes and super-integrons in *Vibrionaceae*). The numbers indicate the number and fraction of resolved quartets (769 orthologous quartets) with the indicated topology.

C Phylogenetic analysis of quartets of common orthologs

Besides differences in the structure of *Vibrio* genomes, the sequence information contained in the genomes can provide a more complete understanding of *V. cholerae* speciation. By supplementing the 16S rRNA-based phylogeny, analysis of results of a genome-wide search of the effect of LGT, as seen in the case of tmRNA, can help determine whether evolution of the *Vibrio* species is driven primarily by a genome-wide cohesion of vertical gene transfer (Daubin *et al.*, 2003) or is significantly reflected in the genome-wide LGT phenomenon (Gogarten *et al.*, 2002).

As shown in Figure 2.12, three kinds of topologies can be constructed from a quartet of homologous genes. The topology of a gene tree comprising three *Vibrio* genes and a truthful outgroup representative can produce one of the three outcomes when phylogenetic information is significantly divergent to be detectible. Otherwise, the tree is non-resolving.

By matching all 1,535 orthologous proteins of the *Vibrio* Triad to the *E. coli* K12 genomic protein database, 1,090 orthologous protein quartets were determined, corresponding to 23% - 28% of the total ORFs in the three *Vibrio* genomes. The trees of 769 quartets (71%) were resolved, and 578 quartets (75% of those resolved) produced topologies identical to the 16S rRNA tree (Topology B in Figure 2.12). Topology A, which supports the tmRNA tree, was found among 118 (11%) of the resolved quartets. The rest of the resolved orthologs (73 quartets or 7% of the resolved) were found to comprise Topology C, in which the *V. cholerae* proteins are more related to *V. parahaemolyticus* orthologs than to *V. vulnificus* orthologs. The distribution indicates that the majority of the genes on genomes of *Vibrio* followed the same path of evolution as with 16S rRNA, while some proportion was influenced by LGT with neighboring species. Therefore, the genomic makeup of *V. cholerae* inferred from the 20% of the total protein is understood as follows: three quarters of the genes evolved in cohesion with each other and to 16S rRNA, but about one quarter of the genes in the genome are significantly influenced by neighboring species via LGT.

To determine the significance of such levels of cohesion versus LGT among *Vibrio* spp., the results of quartet analysis were compared with other taxonomic groups. Recently, Daubin *et al.* (2003) published quartet analysis results for various taxa levels: four species, four genera, and two families. The *Vibrio* data reported here were compared with the published data by plotting taxonomic groups according to proportion of topology outcome (Figure 2.13). The proportion of orthologous quartets with phylogeny congruent with the 16S rRNA phylogeny can be interpreted as the level of cohesion of the genes to a vertical evolutionary path. The overall results indicate that cohesion with the 16S rRNA phylogeny is the universally prevalent force in intra-family and intra-genus evolution (>50% of total quartets or >75% of the resolved), except for the case of *Streptococcus* where non-resolving cases were > 85%. When non-resolving quartet incidence is very high, comparisons between cohesion and LGT lack confidence because the fraction of quartets that is resolved decreases by an order of magnitude, yielding poor representation of the whole genome. In most cases, this occurred in intra-species comparisons. Therefore, a weak level of sequence divergence among strains of the same species is believed to be the cause of the low analytical confidence.

In comparing those cases with *Vibrio*, two interesting points characterizing the unique situation of the genus *Vibrio* can be noted. The first is that *Vibrio* showed the highest level of LGT and the second is that the proportion of non-resolving cases occurred at the level of family rather than genus. The latter can also derive from a strong forcing by LGT. Technically, repeated LGT of a gene among various species can increase ambiguity in delineating sequence phylogeny. Therefore, we can expect the presence of unique mechanisms among *Vibrio* species causing extensive LGT. One possibility, inferred from differences in the habitat of the genera (Figure 2.13) is the uniqueness of the *Vibrio* species habitat. Unlike other genera, the natural habitat of *Vibrio* spp. is the aquatic environment, a relatively more homogenizing (open) environment, perhaps conducive to LGT.

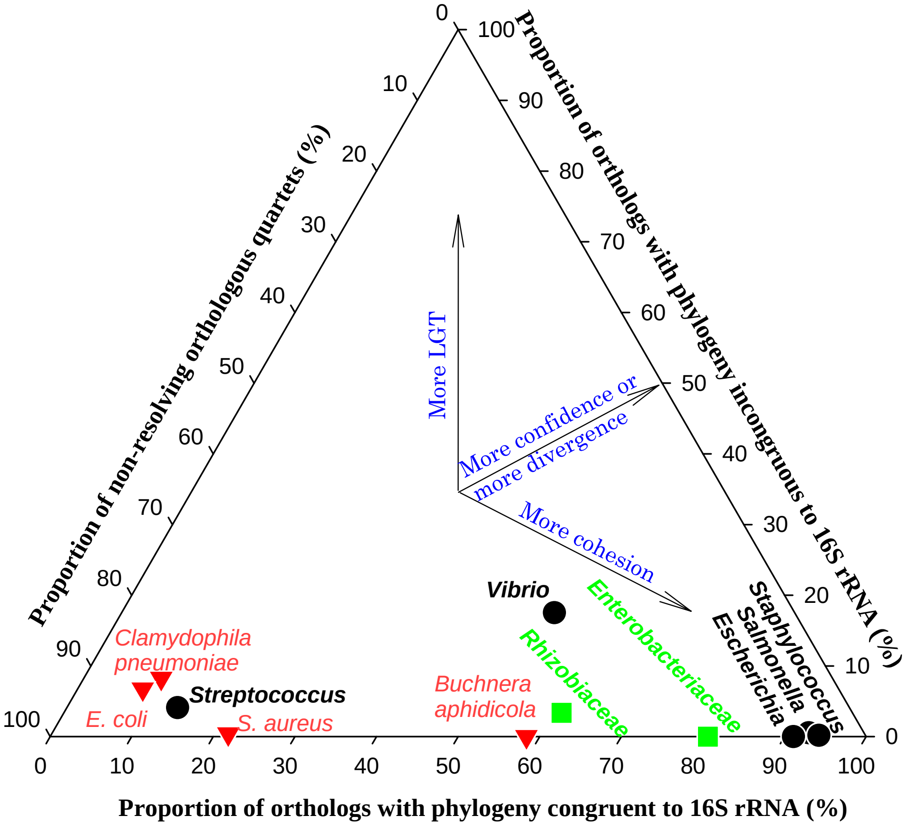
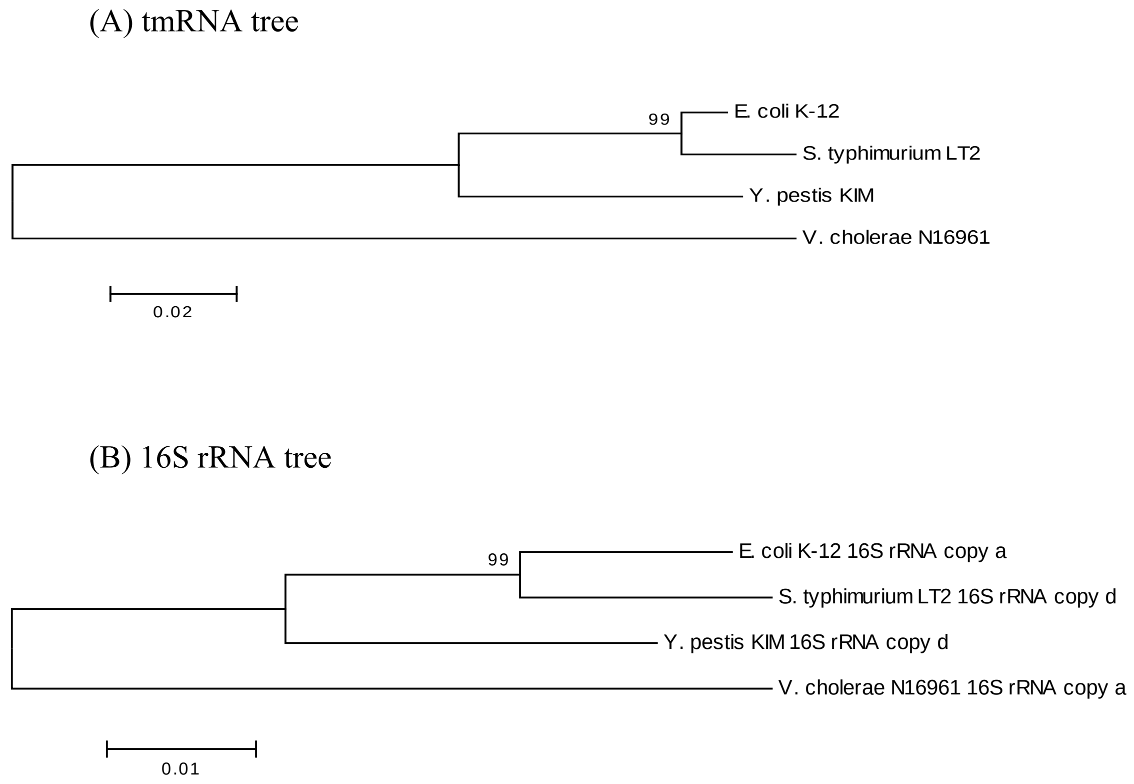


Figure 2.13. Distribution of *Vibrio* and other bacterial taxonomic groups by proportion of different quartet topologies derived from protein orthologs. *Vibrio* showed 53:18:29 ratio of orthologs congruent to 16S rRNA, incongruent to 16S rRNA, and non-revolving, respectively. Data for groups other than *Vibrio* are from Daubin *et al.* (2003). Genus level taxonomic groups are marked in bold letters and black circles. Family level taxonomic groups are marked in green letters and green squares. Species level taxonomic groups are marked in red letters and red triangles. The *Enterobacteriaceae* (*Salmonella typhimurium*, *E. coli*, and *Yersinia pestis*) employed *V. cholerae* as the outgroup. The alpha-proteobacteria family is represented by *Rhizobiaceae* (i.e., *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, and *Mesorhizobium lot* against the outgroup *Brucella melitensis*). *Buchnera aphidicola*, an aphid endosymbiont, is represented by three strains isolated from different host species. Thus, an exceptionally high level of divergence among genome content was reported. The proportion of orthologs with quartet topology incongruous to the 16S rRNA includes both topology A and C shown in Figure 2.12. Three arrows within the triangular chart indicate interpretation of the three axes, as described in the text.

From the results presented here, we conclude that the proteins of *Vibrio* are strongly influenced by LGT. Going back to the original hypothesis of interest, namely the question of whether tmRNA can represent LGT in many proteins, the tmRNA topology for *Vibrio* (Topology A in Figure 2.12) was the predominant of the two possible topologies of LGT (Topology A and C). Therefore, tmRNA evolution can be concluded to reflect the stronger trend of LGT among genes in the *Vibrio* spp. The other genera and families shown in Figure 2.13 appear to be only weakly affected by LGT, since < 10% of the total orthologs are LGT topologies. When 16S rRNA and tmRNA phylogeny were inferred from the trees (Figure 2.14), results for all genera and families included in the analysis indicate that the two phylogenies are completely congruent with each other, implying a lack of sufficient influence of LGT on the tmRNA phylogeny to create digression from the vertical evolutionary paths in those groups. In conclusion, the tmRNA phylogeny reported here for *Vibrio* and other supra-species taxonomic groups is consistent with the hypothesis that tmRNA phylogeny can serve as an indicator of a significant influence of LGT on bacterial genomes.


(A) tmRNA tree

99
E. coli K-12
S. typhimurium LT2
Y. pestis KIM
V. cholerae N16961
0.02

(B) 16S rRNA tree

99
E. coli K-12 16S rRNA copy a
S. typhimurium LT2 16S rRNA copy d
Y. pestis KIM 16S rRNA copy d
V. cholerae N16961 16S rRNA copy a
0.01

Figure 2.14. tmRNA and 16S rRNA trees for *Enterobacteriaceae*. NJ clustering using distances based on the Jukes-Cantor model was used to analyze sequences extracted from the genomic sequences. Among multiple copies of the 16S rRNA genes, the most frequent allele was selected. Bootstrap support from 1000 bootstrapping is shown above the branches.

D Distribution of different topologies of quartets

A follow up question that concerns the unusually high incidence of LGT among species of the genus *Vibrio* is whether the probability of a gene being subject to LGT is the same for all housekeeping genes or if it depends on certain properties of each gene. While evenness in the frequency of sequence of LGT can indicate a genome-wide generality of LGT, distributions skewed toward a particular collection of genes support the presence of selective forces for LGT during evolution of the *Vibrio* species. In this section, the common occurrence and selectivity of LGT in *Vibrio* were investigated by analyzing the distribution of LGT orthologs in *Vibrio* genomes.

The distribution of orthologs by quartet topology is shown in Figure 2.15 and the assumption of uniform distribution was tested by Rao's spacing test for uniformity in circular space (Table 2.4). Excluding the segment of VCR super-integrons from chromosome 2, distributions of total orthologs in both chromosomes were not significantly different from uniform distribution. The orthologs resolved from the three kinds of topologies showed significantly uneven distribution in both chromosomes. However, this may be an artifact of the occurrence of segments where orthologous genes are absent or very rare, e.g., from S1 to S7, and VCR island. Genes in these regions are considered as species-specific or strain-specific genes. To overcome this problem from the native physical coordinates of the chromosomes, circular "logical" coordinates of orthologs were created for each chromosome by serially listing orthologs in their order on the physical coordinates. There was no gap in the logical coordinates between orthologs and all orthologs were of the same size. Therefore, the orthologs demonstrated an ideal even distribution. Rao's test using the logical coordinates was performed on the ortholog groups differentiated by their quartet topologies and the result was that none of the two LGT topologies deviated significantly from uniform distribution of LGT among common orthologs of *Vibrio*. Therefore, it is concluded that LGT is spatially a generalized phenomenon in *V. cholerae*, i.e., all genes of the genome being subject to LGT rather than LGT being preferential to specific clusters of genes.

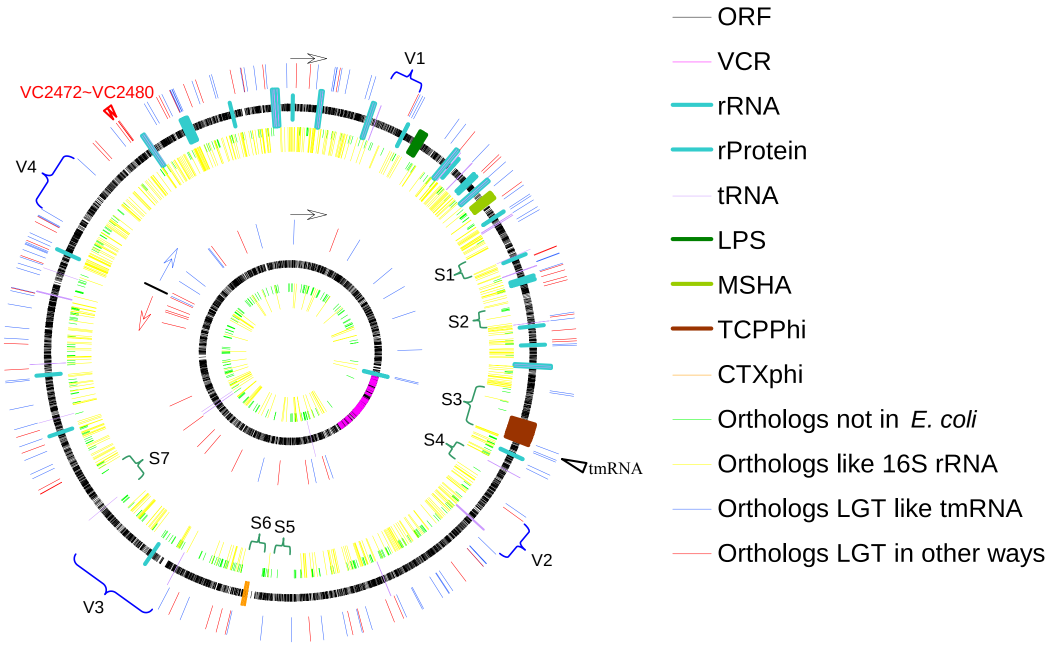


Figure 2.15. Distribution on the two chromosomes of *V. cholerae* N16961 of orthologous proteins common to the genomes of the *Vibrio* Triad and/or *E. coli* K12. The two circles comprising short black spines are the ORFs of all proteins. The super-integron island comprising *V. cholerae* repeated (VCR) sequences is shown by pink spines. Orthologs common to the *Vibrio* Triad, but not to *E. coli*, are plotted as short green spines. Orthologs common to the four genomes (i.e., quartets of orthologous proteins) are separated by their quartet phylogeny (yellow spines: topology congruent to 16S rRNA, blue spines: LGT topology congruent to tmRNA, red spines: other LGT topology). The black spine, from which the red arrow and the blue arrow are heading in opposite directions on the small chromosome, indicates the arbitrary split point between the two zones where only one of the LGT topologies is predominant among the LGT orthologs. The black triangle and the red triangles point out the locations of particular genes (labeled) on the large chromosome. The sections of the large chromosome indicated by brackets (S1-S7, and V1-V4) mark the chromosome sections devoid of orthologs or LGT orthologs.

Table 2.4. Significance (*P*) values from Rao's spacing test for uniformity in circular space (Jammalamadaka & SenGupta, 2001) on orthologs of the two chromosomes of *V. cholerae* N16961[a].

			Total	Quartet Topology[c]		
Coordinates[b]	*V. cholerae* chromosome	Total orthologs[a]	resolved quartets[c]	16S rRNA	tmRNA	Other
Physical	1	>0.1	<0.001	<0.001	<0.05	>0.1
	2[d]	>0.1	<0.01	<0.05	>0.1	>0.1
Logical	1	NA[e]	>0.1	>0.1	>0.1	>0.1
	2	NA	>0.1	>0.1	>0.1	>0.1

a: orthologs common to *Vibrio* Triad and *E. coli* K12 genomes

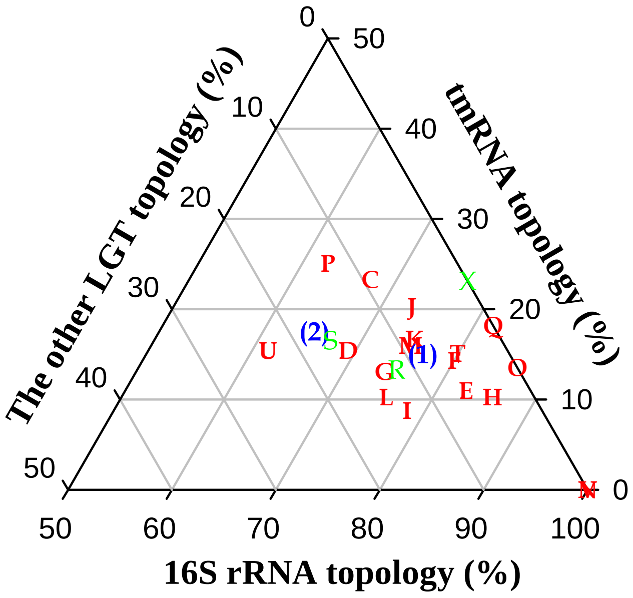
b: physical coordinates are base location from replication origins, while logical coordinates are artificially created by arranging orthologs according to the order of their occurrence. The former has base pair (bp) units, but one orthologous gene locus is the unit in the latter.

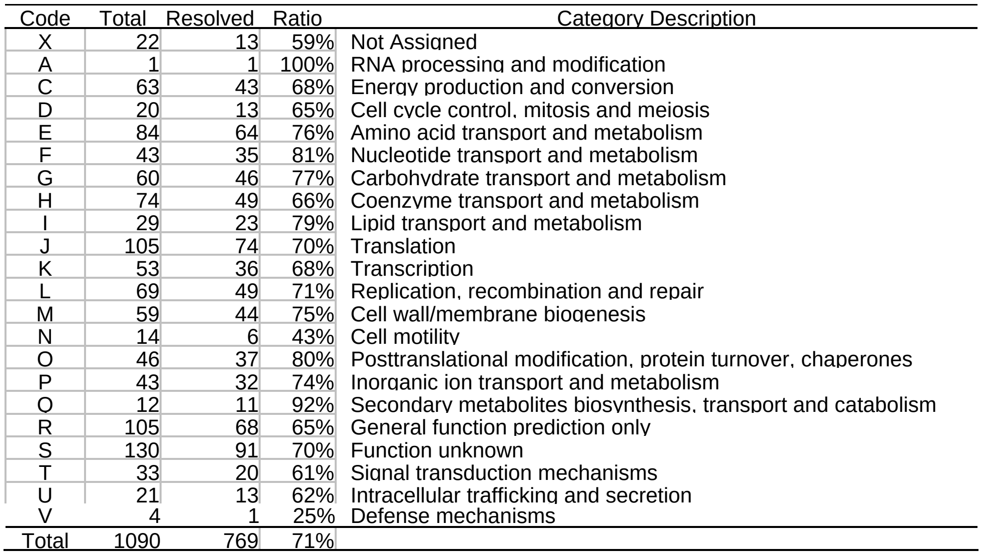
c: quartets and topologies are as in Figure 2.12.

d: chromosome 2 contained the VCR super-integron which did not carry orthologs of interest. This large segment caused a non-uniform distribution of any category of orthologs on the physical coordinate. To avoid this noise, the VCR segment was removed from chromosome 2 for this analysis.

e: not applicable

Noting the generalized occurrence of LGT throughout the entire genome, we can conclude that LGT can occur with various species serving as donor organism. However, LGT from a given donor can be preferential and fixed by a recipient via selective pressure. Firstly, functions of genes can be considered to be one of the most powerful factors in selective pressure for LGT. Events of LGT, followed by homologous recombination, will introduce alleles which are novel to the recipient. However, not all alleles will be fixed by a recipient, since stochastic chance, in the case of neutral alleles, or an advantage of a new allele to the recipient under selective pressure will play a role. The best example of the latter is when a variant of a species is adapting to a new environment. The variant will be under selective pressure to acquire new alleles of genes that support its adaptation more effectively. When a set of alleles already acquired by better adapted, existing inhabitants of the new environment are horizontally transferred to the variant, the set of alleles functionally related to the adaptation can become simultaneously fixed in the variant, resulting in an uneven distribution of the LGT among different functional classes of genes. Therefore, an observation of high incidence of LGT skewed toward particular functional classes may indicate the occurrence of such adaptive events in the recipient organism.





Code	Total	Resolved	Ratio	Category Description
X	22	13	59%	Not Assigned
A	1	1	100%	RNA processing and modification
C	63	43	68%	Energy production and conversion
D	20	13	65%	Cell cycle control, mitosis and meiosis
E	84	64	76%	Amino acid transport and metabolism
F	43	35	81%	Nucleotide transport and metabolism
G	60	46	77%	Carbohydrate transport and metabolism
H	74	49	66%	Coenzyme transport and metabolism
I	29	23	79%	Lipid transport and metabolism
J	105	74	70%	Translation
K	53	36	68%	Transcription
L	69	49	71%	Replication, recombination and repair
M	59	44	75%	Cell wall/membrane biogenesis
N	14	6	43%	Cell motility
O	46	37	80%	Posttranslational modification, protein turnover, chaperones
P	43	32	74%	Inorganic ion transport and metabolism
Q	12	11	92%	Secondary metabolites biosynthesis, transport and catabolism
R	105	68	65%	General function prediction only
S	130	91	70%	Function unknown
T	33	20	61%	Signal transduction mechanisms
U	21	13	62%	Intracellular trafficking and secretion
V	4	1	25%	Defense mechanisms
Total	1090	769	71%	

Figure 2.16. Distribution of orthologous protein quartets from the *Vibrio* Triad and *E. coli* by proportion of the three topologies in each COG functional category (A to V) and in each chromosome (1 and 2). A total of 769 resolved cases (672 in chromosome 1 and 97 in chromosome 2) were plotted. Note only a partial triangular chart is presented. The single protein in Category A, with tmRNA topology, is not shown. Category N and V are depicted by the overlapping letters at the bottom-right apex. Note Category R, S and X are not valid functional categories. The ratio of orthologs shown for chromosome 1 was 77:15:8 as the ratio of 16S rRNA topology, tmRNA topology, and other LGT topology.

This possibility was tested by classifying the *Vibrio* Triad orthologs according to functional categories of clusters of orthologous groups (COG) (Tatusov *et al.*, 2000). The table of frequencies of orthologs (i.e., a table of 3 columns and 19 rows comprising the three topologies of quartets in Figure 2.12 and COG categories in Figure 2.16, except for X, R, and S, which are not functional but arbitrary groups) were tested using the Fisher-Freeman-Halton exact test with Monte Carlo approximation. The 99% CI of significance was estimated as $P = 0.54 \sim 0.57$, indicating absence of dependence between COG categories and quartet topologies. Therefore, function specific preference in either case of LGT topologies was not significant. However, this result could also arise from the low power of the test statistic because frequencies of LGT for each category were low (maximum of 20 and mainly less than 12 orthologs).

The scatter graph (Figure 2.16) showed a pattern in agreement with the hypothesis that genes related to environmental adaptation will experience more LGT. In the case of the COG categories of central dogma (J, K and L), about 75% of the proteins were commonly coherent with the 16S rRNA phylogeny, but with a moderate level of variation (10% - 20%) in frequency of the two LGT topologies. In contrast, the C, P and U categories, related to environmental adaptation via energy conversion, ion transport and material secretion, receptively, showed higher LGT. The most extreme deviation was Category P, with higher LGT from *V. vulnificus* to *V. cholerae*. It is interesting to note that genes regulating ion content of cells show greater deviation, since salinity is a prominent factor in confining the habitat of an aquatic organism. The optimal range of salinity for different species has been reported for *Vibrio* species (Baumann *et al.*, 1984). Data correlated with LGT between *V. cholerae* and *V. vulnificus* in Category P includes those genes containing NaCl requirement for growth of *V. cholerae*, *V. vulnificus,* and *V. parahaemolyticus* (5 mM, 130 mM and 160 mM, respectively).

To overcome the difficulty in detecting selective pressure as a forcing for LGT in *Vibrio* species, distribution of linkage patterns of genes was examined. Because a set of genes coding for a given function often occurs in a cluster, i.e., an operon, a gene cassette or island, the recipient genome in the LGT may carry a set of adjacently-linked genes that share the same LGT phylogeny, e.g., Topology A or Topology C (Figure 2.12). The presence of such a LGT gene cluster in the *V. cholerae* genome would indicate an adaptive incorporation of the foreign alleles by the recipient. Such LGT clusters were sought by examining the genome of *V. cholerae* for chromosome segments containing more than two orthologous genes of the same functional category with the same LGT topology within a 10 kb length. The segments of chromosome that met these criteria included VC2472~VC2480 (Figure 2.15), the largest, containing Topology C. However, functional relatedness among the genes in the segment could not be measured because functions of those genes in the particular segment were unknown. The next large segment that was detected was located within a large operon coding for 24 ribosomal proteins. Among those, 5 ribosomal proteins were Topology A, similar to the tmRNA. Because ribosomal proteins are highly conserved and interacting directly with rRNA, this segment offers a clear example of selective pressure leading to incorporation of the *V. vulnificus* –related alleles via LGT.

Another method for classifying physical linkage among genes is to look at distribution chromosome by chromosome. Table 2.5 provides a contingency table constructed from results of the quartet analysis. Independence between the row and column variables was tested. Because Fisher's exact test yielded significance *P* < 0.05, the source of the interaction between quartet topology and chromosome was determined by comparing the observed frequencies to expected frequencies. While chromosome 1 carried ortholgos consistent with expected frequency, chromosome 2 showed an increased proportion of Topology C at the expense of reduced 16S rRNA topology (Table 2.5 and Figure 2.16). Therefore, chromosome 2 of *V. cholerae* can be concluded to have been influenced by *V. parahaemolyticus* more than chromosome 1. When the physical distribution of LGT orthologs on chromosome 2 was examined (Figure 2.15), a very interesting, uneven distribution was observed. The distribution of Topology A LGT and Topology C LGT were highly skewed to opposite sides of the chromosome across the split (split line between blue and red arrows at 880000 bp location as shown in Figure 2.15) and VCR islands. In the segment from the split line to the VCR island, moving in a clockwise direction, 13 orthologs were Topology A and four 4 orthologs Topology C. In the other segment, spanning clockwise from the split line to the VCR island, four orthologs were Topology A and 13 Topology C. When these frequencies were tested by Fisher's exact test, the hypothesis of independence between topologies and the segment was rejected (*P*< 0.01), with the conclusion being that physical linkages among LGT orthologs spanned nearly half of chromosome 2 (*ca.* 500 kb). This result also implies a unique and selective fixation process in the history of the chromosome, resulting in a differential distribution in the two kinds of LGT. The distribution is possible when the chromosome contains homologous recombination with the long fragment (~500 kb), half the size of the small chromosome and originating from chromosomes of *V. vulnificus* or

V. parahaemolyticus. This phenomenon merits further study since it implies a megabase transformation mechanism, such as conjugation, and selective fixation.

In summary, analysis of the distribution of LGT genes among orthologous proteins common to *Vibrio* and *E. coli,* corresponding to 28% of the whole genome ORFs, were performed. The results indicated that LGT is widespread throughout the genome of *V. cholerae*. Both chromosomes carry genes acquired by LGT along both chromosomes in a relatively uniform distribution. However, uneven distributions were observed when the direction (i.e., the counterpart species from which LGT alleles were imported) of the LGT was taken into considerations. On a scale of a 10 kb-sized polycistronic operon fragment, the operon of ribosomal proteins carried only one type of LGT topology. Physical linkage in one type of LGT was also found on a scale of 500 kb, half the size of the small chromosome of *V. cholerae*. The former provides an example of selective pressure in a particular direction during incorporation of the genes received via lateral gene transfer, whereas the latter suggests that large scale LGT is possible

Table 2.5. Frequencies of orthologous protein quartets on each chromosomes of *V. cholerae* N16961 determined from tree topologies of quartet analysis.

Chromosome	Congruent to 16S rRNA (Topology B)[a]	Congruent to tmRNA (Topology A)	Other LGT (Topology C)	Sum
Chromosome 1	515 (505)[b]	101 (103)	56 (64)	672 (87%) [c]
Chromosome 2	63 (73)	17 (15)	17 (9)	97(13%)
Sum	578 (75%)[c]	118 (15%)	73 (9%)	769

a: Topologies of trees in Figure 2.12

b: Expected frequencies, assuming independence between topology and the chromosome

c: Ratio to grand sum

2.4 Conclusion

The hypothesis that tmRNA phylogeny can depict the horizontal evolutionary force involved in extensive LGT was supported by at least two major pieces of evidence: (1) the phylogeny was consonant with a stronger influence of LGT from *V. vulnificus* than from *V. parahaemolyticus*; and (2) the congruence with 16S rRNA phylogeny when the influence of LGT was negligible, as in the case of *Enterobacteriaceae*. Because tmRNA phylogeny provides a stable and better resolution, both at the species and supra-species levels, it is a useful tool in phylogenetics that complements 16S rRNA phylogeny.

V. cholerae speciation can be better understood by the approach taken in this study. In agreement with 16S rRNA phylogeny, the genomic phylogenies of *Vibrio* indicate that the *V. cholerae – V. mimicus* complex is unique and an early branching within the genus *Vibrio*. However, 25% of the *V. cholerae* genome was strongly influenced by LGT from neighboring species, e.g., *V. fluvialis*, *V. furnissii*, and *V. vulnificus*. It is concluded that the *V. cholerae – V. mimicus* complex represents a singleton compartment of the genus *Vibrio*, without sister species, as determined by phylogenetic analysis. However, the compartment is not totally isolated but open to LGT from other related species sharing habitats with *V. cholerae*. When the nature of LGT was analyzed, all genes were found to be susceptible to LGT, along with selective incorporation of genes from a particular species. In the case of *V. cholerae*, greater incorporation of *V. vulnificus* than *V. parahaemolyticus* genes was observed on the large chromosome, but both were found to contribute equally in the makeup of the small composition of the small chromosome..

From the viewpoint of epidemiology, the outer boundary of the species, *V. cholerae*, is of particular interest, namely to determine the span of the target population under surveillance. According to the results of the phylogenomics analysis presented here, the species definition of *V. cholerae* by 16S rRNA addresses *ca*. 75% of its genome and includes *V. mimicus* within the species boundary of *V. cholerae* on the basis of the difference in 16S rRNA sequence being within sequencing error and variation accountable by different copies within a given strain.

Chapter 3. Diversity and Structure of *V. cholerae* Populations

3.1 Introduction

3.1.1 Evolution of epidemic cholera vibrios

One of the underscored features of cholera epidemiology is the rapid evolution of epidemic *V. cholerae* strains, described as “shifts” in the cholera vibrios. Biological properties of *V. cholerae* O1 serovar strains isolated during the 7th pandemic were different from those isolated during the 6th pandemic: the 7th pandemic isolates were resistant to the antibiotic polymyxin B and caused hamagglutination with chicken, goat, and sheep erythrocytes (Kay *et al.*, 1994). This shift in the biotype of cholera bacteria has been emphasized by the designation of the two different types as “classical” and “El Tor” (the latter taking the name of the quarantine station where the 7th pandemic biotype was first described).

In 1992, a wide-spread cholera outbreak caused by a strain of *V. cholerae* without the O1 antigen occurred in the Bay of Bengal area (Cholera Working Group, 1993). Because the primary tool used at the time for identifying the causative agent of cholera was an anti-O1 antibody, a shift in the serotypic characteristics of the cholera bacteria induced a great concern among public health authorities. This new serovar was named *Vibrio cholerae* O139, and its biotype characteristics were determined to be closer to El Tor than to the classical *V. cholerae*. Further molecular studies led to the observation that the strain had evolved from *V. cholerae* O1 El Tor by the insertion of new lipopolysaccharide (LPS) synthesis genes (Waldor *et al.*, 1994) and a multi-drug resistant mobile element, "costin SXT" (Waldor *et al.*, 1996). While it has not yet caused global pandemics of cholera, *V. cholerae* O139 has been shown to coexist with *V. cholerae* O1 El Tor in the aquatic environment and has been isolated from clinical cases of cholera in the Bay of Bengal (Faruque *et al.*, 2003a), Southeast Asia (Hoge *et al.*, 1996), and rural areas of China (Qu *et al.*, 2003).

This information, together with the knowledge that non-O1 and non-O139 *V. cholerae* are also autochthonous to the aquatic environment, on the observed shifts in composition of *V. cholerae* O1 and O139 led to the hypothesis that the interaction of epidemic *V. cholerae* strains with environmental strains is a major mechanism of the evolution of new types of the cholera *Vibrio* (Bik *et al.*, 1995; Faruque *et al.*, 2003a; Faruque *et al.*, 2003b; Faruque *et al.*, 2004; Lipp *et al.*, 2002). The differential rate of infection by *V. cholerae* O1 and O139 strains related to the age of victims in cholera outbreaks (Dalsgaard *et al.*, 1999; Sack *et al.*, 2003) illustrates the effect of shifts by microevolution. Therefore, investigation of the diversity and interaction between epidemic and environmental *V. cholerae* should provide new understanding of the type and scale of microevolution occurring in this pathogen. If substantiated, this hypothesis be the basis for development of useful tools in the weapon-and-armor race between pathogenic bacteria and their human host.

3.1.2 Diversity and clonality of bacterial populations

To define diversity and the structure of a species, the concept of what constitutes an individual must first be established, because the richness in number of individuals in a population is what diversity would represent. In the microbiology laboratory, a strain is the operational unit of the individual. Two bacterial isolates which prove to be the same strain have an identical genetic makeup. When strains are decedents from a common ancestor, they share a majority of their genetic makeup, but are not identical. That is, they are clonal (Spratt, 2004). Therefore, a clonal complex comprises a collection of directly related individuals. It is recognized that there is ambiguity in this terminology, because there is no clear-cut definition, or measure, of the extent of similarity or divergence among clonal strains. Technically, isolates of a bacterial species that are indistinguishable by genotype are labeled as clones, with the implication that they descended from the same recent ancestor. Because of frequent horizontal gene transfer and homologous recombination among closely related bacteria, similarity among strains that diverged further in the past would lose identity in genetic construction, so they would no longer be clonal. Clonality of a given population indicates the tendency to maintain a constant genetic makeup within the population and is determined by the rate of gene traffic among strains. That rate, in turn, is determined by many factors, such as the native tendency for gene exchange among species, the strength and direction of selective pressures of the habitat, and the level of habitat segmentation (Smith *et al.*, 1993). Therefore, by observing the clonality of a species, its phylogenetic or ecological compartmentalization can be inferred. For example, a recent epidemic growth of *Neisseria meningitidis* population, and the presence of two separate divisions within *Rhizobium meliloti* were inferred from the clonality of their populations (Smith *et al.*, 1993). Similarly, a weakness of compartmentalization (defined as a panmictic structure) was observed in *Pseudomonas stutzeri*, which has the highest mean genetic diversity (Rius *et al.*, 2001).

By using a battery of molecular methods, a complex subspecies-level of diversity among bacterial species is now recognized (Schloter *et al.*, 2000). Although various levels of clonality have been observed, depending on the species and the sampled populations, the predominant driving force leading to persistence of clonality in a complex is believed to be selective pressure from the ecological niche. Therefore, a stabilized clonal complex is considered an ecotype (Cohan, 2001), meaning a subvar distinguished from others of the same species by its ecological properties. In comparison with the systematics of macroorganisms, a bacterial ecotype corresponds to the conventional species when clonality imposed by selective pressure from the niche is interpreted as cohesion of the genetic makeup of the individuals. Cohan (2001) also suggested that the current classification of a bacterial species might better correspond to the genus of eukaryotic organisms.

While linking ecological selective pressures to the structure of a bacterial population is new to microbiology, an evolutionary perspective for habitat in animal ecology has been strongly promoted by Southwood (1977), who designed the theory on "habitat template", a classification based on defining habitats in terms of two main characteristics - scope for growth and disturbance of the environment. According to this theory, selective pressure, in combination with these two factors, determines the composition of animal community in an habitat template. Modeling the habitat preference of dinoflagellate bloom species by using nutrient concentration and the force of mixing, as the two axes of independent variables (Smayda & Reynolds, 2001) provides a good example of the successful application of the theory at the scale of the microorganism. Observing the relationship between clonality of a bacterial population and distribution of their habitat templates will provide a clear understanding of bacterial population structure.

3.1.3 Diversity and structure of *V. cholerae* derived from previous studies

By using the allelic distribution of multiple genetic loci, such as multilocus enzyme electrophoresis (MLEE) or multilocus sequence typing (MLST), strong clonality of the epidemic *V. cholerae* O1 classical, O1 El Tor, and O139 strains has been demonstrated (Beltran *et al.*, 1999; Farfán *et al.*, 2000; Farfán *et al.*, 2002; Garg *et al.*, 2003; Stine *et al.*, 2000), reflecting the fact of the epidemic explosion as the mode of population expansion. Expanded populations of the species *V. cholerae*, including representatives of the majority of serotype collections, also showed a monophyletic clonal structure; however, their diversity far exceeded that of epidemic clones. Most of the diversity of the species *V. cholerae* has been explained via study of non-O1 environmental strains (Beltran *et al.*, 1999; Stine *et al.*, 2000). Genetic diversity of the average genetic locus (*H*) was estimated to be 0.4 – 0.5 (Beltran *et al.*, 1999; Farfán *et al.*, 2000), a moderate level of allelic diversity for known bacterial species, e.g., $H = 0.3$ for *E. coli* (Selander *et al.*, 1987) and $H = 0.9$ for *P. stutzeri* (Rius *et al.*, 2001). The level of linkage disequilibrium, which indicates the non-random association of alleles occupying a given locus on a genome and is estimated from the index of association between loci (I_A), devised by Smith *et al.* (1993), ranged 1.3 to 1.8 and also indicated that the rate of recombination between different strains to be moderate, compared with other bacteria (Beltran *et al.*, 1999; Farfán *et al.*, 2000). However, these studies could not resolve the clonal structure and phylogeny among clonal complexes at the level between the outer boundary of the species and epidemic clones, mainly because of limitations of the method and the representation of the global population by an arbitrary collection of strains.

Linked to the focus of this chapter is the study by Choopun (2004) who investigated the phylogenetic structure of *V. cholerae* isolates from a single ecosystem, the Chesapeake Bay. ERIC-PCR genomic fingerprinting, devised to provide phylogenetic inference, was used to study a census type of collection of *V. cholerae* isolates. That study revealed that the population can be divided into multiple layers of phylogenetic clusters: three primary clusters, several intermediate clusters and numerous subclusters. The three primary clusters indicated separation of *V. cholerae*, *V. mimicus*, and a novel lineage, using the phenotypic characteristics of *V. cholerae* and the genotypic characteristics of *V. mimicus*. Consequently, the question was raised with respect to the species definition of *V. mimicus* and the results demanded better resolution of phylogenetic clusters.

3.1.4 Objectives

As mentioned previously, genetic shifts in the cholera bacterium can impact the dynamics of a cholera epidemic and could result in new pandemics of cholera. The sources for such a genetic shift would be microevolution occurring in the cholera-causing clonal complexes of *V. cholerae* and the interaction of cholera-causing clonal complexes with other bacteria sharing the habitat. The strongest candidate for the latter case would be the diverse environmental *V. cholerae* because they share a significant part of their genomic makeup with the cholera-causing *V. cholerae*. Kinship between cholera-causing *V. cholerae* and non-toxigenic environmental *V. cholerae* would allow lateral gene transfer (LGT) to occur more readily than with unrelated bacteria. Therefore, in any study of cholera dynamics, it is crucial to understand how diverse *V. cholerae* strains differ from each other and how they interact in their respective habitats, generating gene flow among the strains. Investigation of the intra-species phylogenetic structure of *V. cholerae* populations should be the first step in achieving this objective. There are other significant findings that can be gleaned from this investigation. Firstly, by obtaining detailed knowledge of the population structure of cholera-causing bacteria, it would be possible to develop a model of the cholera dynamics that is based on a sound mechanistic understanding by clearly identifying entities responsible for the processes and events in a cholera epidemic. Insights into population structure can reveal which phylogenetic compartments are directly responsible for cholera epidemics and which non-toxigenic compartments interact with them. Approaches to mechanistic modeling can produce results superior to conventional risk factor (or predictor) assessment since the former can provide fundamental aspects of the disease that are robust against changes in the ecosystem, clinical environment, and human immunity. The latter is based on regression analysis of cholera or *V. cholerae* dynamics against a given set of environmental and clinical factors collected during surveys. Results of such analysis would yield assessment of the quantitative relationships between components of the disease and the surveyed factors (Louis *et al.*, 2003; Sack *et al.*, 2003). Because the surveyed factors were selected without consideration of whether they are relevant properties of entities directly modulating the disease dynamics or proxy predictors incidentally correlated to one of the true predictors, the empirically deduced predictors will not be easily applicable to other systems or after changes in the ecosystem, clinical environment, and/or human immunity. Secondly, recognizing the range of phylogenetic compartments of cholera-causing bacteria is an important issue for public health decision-making, because adequate disease prevention or monitoring programs (e.g., vaccination or surveillance at the point of entry) are required to cover as much of the range of the cholera-causing bacteria as possible. Knowledge of the population structure of *V. cholerae* can also help in devising effective methods for monitoring the entire range of cholera-causing bacteria, besides specific serogroup such as O1 or O139.

Understanding the population structure of a species involves (1) determination of individuals or a finite collection of individuals as the unit of the phylogenetic compartment; (2) delineation of the relationships between those finite unit compartments; (3) estimations of the dynamics of the structure via ecological and evolutionary characterization of the unit compartments (e.g., estimation of the stability of structure and identification of factors shaping or modulating the structure). In the present study, these three aspects of population structure were analyzed with the limitation of the specific background of each component.

As described above, there have been several previous studies of the population structure of *V. cholerae*, with a variety of interests and methods employed. However, they failed to provide a species-wide structure with resolution fine enough to determine the finite phylogenetic compartments for building mechanistic compartment models. This is because the span of the strain collection was arbitrary or too narrow, there was bias in the genome-wide phylogeny from using only a single gene or a few genetic loci, or there was a lack of quantitative assessment sufficient to differentiate divergent compartments in a uniform unit throughout the species. In the present study, the first impediment on the span of strains was circumvented by performing genome divergence analysis on census-type culture collections. The culture collection was comprised of all strains or their non-redundant representatives that were screened and subsequently isolated in pure cultures from samples collected in a geographic region using a given set of isolation methods on a month or less sampling interval over at least a two-year time frame. That is, the *V. cholerae* collection from an environmental survey carried out in the Chesapeake Bay from 1998 to 2000 (Choopun, 2004; Louis *et al.*, 2003) and the culture collection from a four-year biweekly survey of the rectal swabs of cholera patients in four remote regions of Bangladesh (Sack *et al.*, 2003) were used. The second issue of building a phylogeny based on a genome-wide sampling of genetic information was addressed by using a robust, long-range, low-stringency PCR, employing an enterobacterial repetitive intergenic consensus (ERIC) sequence (Choopun, 2004). It was observed that this method provides statistically uniform sampling of sequences along the known *V. cholerae* genomic sequence (see Chapter 4) and significant agreement with the genome deviation estimates obtained from DNA-DNA hybridization (Choopun, 2004). The approach circumvents the drawback of bias in using a single gene to construct the phylogeny of a recombining species (Stine *et al.*, 2000). The third issue of defining finite compartments at a uniform level of phylogenetic divergence was not attempted in any of the previous studies. Typically bootstrap support or co-phenetic correlations were used to assess the significance of occurrence of a cluster of strains. However, such methods can not produce clustering in terms of phylogenetic divergence because cluster formation using these methods is always influenced much more by the collection of strains, rather than a given level of genetic divergence. In the present study, the finite scale of clonal compartments was defined as a cluster of strains lacking a phylogenetic signal sufficient to have an internal structure. For this purpose, the present study employed the permutation tail probability (PTP) test (Archie, 1989; Faith & Cranston, 1991), for which categorical status of each genetic locus (e.g., presence/absence for binary data or A/T/C/G for nucleotide data) serves as the genetic characteristics of each strain. To generate the binary genetic character data for the collections of strains, the PCR-based genomic fingerprinting used by Choopun (2004) was modified to generate more bands and a novel band matching scheme was also devised to produce a table of binary-coded character data for any given set of strains.

When the finite compartments were obtained, as described above, the relationship among clonal clusters was analyzed. The structure that was applied and found commonly in previous researches was a tree-like hierarchical structure for the isolates. This was not realistic, in that it failed to support the presence of multiple sources of genetic information via LGT, which had been found to comprise a significant portion of the *V. cholerae* genome LGT (Chapter 2). To visualize the multiple genetic sources of the phylogeny, a network- or web-like hierarchical structure has been suggested to be (Bryant & Moulton, 2004). In the present study, the relationship between clonal clusters was analyzed by the Neighbor-Net method (Bryant & Moulton, 2004), which is a multi-source extension of the popular Neighbor-Joining clustering method.

Analysis of population structure is also a descriptive study of the patterns of distribution of diverse individuals. Therefore, many aspects of the distribution can be delineated from one study and the ecological properties of individuals and clusters of individuals comprise one of the most interesting aspects because it can provide insights into the ecological functions of individuals or clusters and their impact on evolutionary processes that determine the observed structures of a population. Notably, understanding the evolutionary process of pathogenic organisms provides information about the long-term dynamics of the population structure of the pathogens. Therefore, knowing the dynamics, one has the power of predicting the future evolution of a pathogen and the range of potential genetic shifts in the dynamics of the disease it causes.

Given the data and conclusions obtained in this study, the capacity to analyze evolutionary aspects of population structure was limited to the level of observing the distribution of clonal clusters in their respective source habitat and characterizing their relationship to habitat parameters. The basis of such an analysis is the stability (or persistence) of the clonal clusters observed from structural analysis of the *V. cholerae* populations, demonstrated in previous studies by the presence of clonal lineages (i.e. clonality) in spite of genome-homogenizing effects of LGT. Strong or moderate clonality reflects the existence of constraints in generating clonal diversity derived from genetic drift and LGT, because persistent clonality suggests a low frequency of spontaneous emergence of new traits and a weakness in gene flow. Therefore, constraints are either vertical, meaning a limitation in available genetic resources derived from a direct progenitor via binary-fission, or ecological, meaning selective pressure from the ecological niche. While a vertical constraint is difficult to analyze quantitatively, an ecological constraint can be corroborated by determining the optima of selected ecological parameters of clones and their tolerance to a variation of parameters. Emphasizing ecological constraints, a stabilized clonal complex is considered an ecotype (Cohan, 2001). From this perspective, phylogenetic compartments of clonal complexes of the species *V. cholerae* are equivalent to ecological compartment (i.e., ecological niches).

By observing the relationship between the distribution of various clonal clusters of a bacterial population and the distribution of their respective habitat templates, one can test the ecotype hypothesis and the results provide understanding of the bacterial population structure that is closer to its entirety. Because the most diverse culture collection that was used in the present study comprised isolates from samples collected in a survey conducted along shore sites in the Chesapeake Bay, for which environmental variables were also measured at the time when the samples collected, it was possible to measure correspondence between variations of clonal composition of *V. cholerae* and environmental factors. The most interesting variable was the composition of the zooplankton in the samples, by which the hypothesis of particular zooplankton taxon being the host of a specific *V. cholerae* strain could be tested. The reason for choosing the Chesapeake Bay as an ecosystem within which to study the dynamics and population structure of *V. cholerae*, Choopun (2004) explained, "It is one of the most productive and extensively studied estuaries in the east coast of the United States. It has been shown to harbor natural populations of *V. cholerae* in various publications since the1970s (Kaper *et al.*, 1979). Because it is a geographic location free from cholera, it is an ideal site for studying natural populations of *V. cholerae* in the environment, with no interference from clinical cases of cholera." In addition to these reasons, the well-described role of the zooplankton community in the Chesapeake Bay ecosystem, namely the two predominant species of calanoid copepods (*Acartia tonsa* and *Eurytemora affinis*) linking primary productivity to higher trophic levels (Kimmel & Roman, 2004; Roman *et al.*, 2005), allows examination of the association between *V. cholerae* and the zooplankton community to be supported with an ample amount of background data.

In summary, the present study included development of a genomic fingerprinting method that provided better resolution of phylogenetic inference and was used to determine finite phylogenetic compartments representing clonal entities, i.e., terminal clusters. Relationships among the entities were analyzed, based on a phylogenetic network method used to elucidate hierarchical structures from clones to the boundary of the species. The last step was determination of possible ecological properties of the different terminal clusters by applying the resolved cluster structures to the Chesapeake Bay *V. cholerae* survey data.

3.2 Materials and methods

3.2.1 Strains

Strains used in this study were obtained from the culture collection of the University of Maryland (Figure 3.1). The collection of environmental strains included 157 strains (RC345 - RC632) that were representative of the strains isolated during a survey carried out in the Chesapeake Bay from 1998 to 2000 (Louis *et al.*, 2003). The collection has been described in detail in elsewhere (Choopun, 2004). The collection of 597 clinical and 55 environmental strains (designated ZB) comprised *V. cholerae* strains isolated from a survey carried out in Bangladesh from 1997 to 2000 and described by Sack *et al.* (2003), Huq *et al.* (2005), and in Chapter 4 of the present study. Environmental strains from Peru (prefix P) and Louisiana (prefix UM) were included to represent difference in geographical sources.

3.2.2 DNA-DNA hybridization

Duplicate sets of the 176 strains having a unique ERIC PCR fingerprint pattern were selected for membrane DNA-DNA hybridization. To prevent systematic bias of the signal intensity generated from different DNA positions on the membrane, genomic DNA from each set was randomly assigned to a position among the two nylon membranes. Duplicates of DNA from the five probe strains (i.e., *V. cholerae* RC145, RC395, RC466, and RC586, and *Vibrio mimicus* RC5) were included on each membrane as positive control. Duplicates of the type strain of *Vibrio fluvialis* and *Aeropyrum pernix* were also included on each membrane as controls, providing different levels of genome relatedness to the probe genome. *A. pernix*, an archaebacterium, was included as a negative control.

Genomic DNA from each isolate was extracted using the DNeasy® Tissue Kit (Qiagen Inc., Valencia, CA) and eluted in 200 µl elution buffer AE (10 mM Tris-HCl pH 9.0, 0.5 mM EDTA). DNA concentrations and purity were determined spectrophotometrically by measuring absorbance at 260 nm (A_{260}) and 280 nm (A_{280}) (Sambrook & Fritsch, 1989). DNA was diluted to 20 ng/µl in AE buffer and stored at -20 ˚C until analyzed.

Genomic DNA (500 ng) was denatured in 0.4 M NaOH, 10 mM EDTA and heated to 100 ˚C for 10 min to ensure complete denaturation. The DNA was cooled on ice and mixed with an equal volume of 6x SSC, prior to being dot blotted onto a MagnaCharge Nylon membrane (MSI, Micron separations Inc., Massachusetts, USA) pre-wet with distilled water and soaked in 6x SSC before use (MSI manufacturer's protocol). The Bio-Dot microfiltration apparatus (Bio-Rad Laboratories, Hercules, CA) was employed, following the manufacturer's instructions. After assembling the apparatus, the membranes were washed with 500 µl 6x SSC, before and after applying the DNA samples. Slow vacuum was applied at each step. After blotting, the DNA was immobilized on the nylon membrane by UV cross-linking (UV crosslinker, Fisher Scientific, Pittsburgh, PA) at an optimal cross-link setting (120 mJ/cm^2, 30 sec). Each membrane was then rinsed briefly with distilled water, air dried, and placed between two sheets of dry filter paper, sealed in a plastic bag, and stored at -20 ˚C until analyzed. Genomic DNA from *V. cholerae* (RC2 classical O1, RC4 El Tor O1, and RC66 non-O1 *stn* positive), *V. mimicus* (RC5), and an archaeal bacterial DNA (*Aeropyrum pernix*) were included on the blots as positive and negative controls.

The genomic DNA (500 ng) was dot blotted onto nylon membranes, as described previously (Chapter 3). The probe genomic DNA was sheared to an approximate size of 400-600 bp by sonication and labeled by thermostable alkaline phosphatase the enzyme in Geneimages^{TM} AlkPhos Direct^{TM} labeling kit (Amersham Biosciences Ltd, Buckinghamshire, England). Hybridization buffer and washing solution were prepared following the manufacturer's protocol. The membrane was prehybridized at 60˚C for 30 min, and then hybridized (10 ng/ml probe) at 60˚C overnight in a rotary hybridization tube. The membrane was then subjected to high stringency wash twice with primary wash buffer for 10 minutes at 70 ˚C, followed by low stringency wash twice with secondary wash buffer for 5 min at room temperature. Chemifluorescent signals were generated using ECF substrate (Amersham Biosciences). The fluorescent signals were recorded using an imaging system, Storm^{TM}840 or Typhoon 9410 (Molecular Dynamics Inc., Sunnyvale, CA) and the signal intensity was quantified by ImageQuant software version 5.1 (Molecular Dynamics, Inc.).

The results are expressed as relative binding unit (RBR), which is the ratio of signal from the target DNA to that from the probe DNA itself (i.e., positive control) as the target DNA. To visualize the relative positioning of a target strain among other strains, a dimension reduction analysis was performed, utilizing the RBR values from the five probes: RC145, RC466, RC395, RC5 and RC586. Principal component analysis (PCA) was done on five RBR values for each strain, with correlation as the value matrix.

Three most explaining components were selected for three dimensional positioning of strains in the space determined by the score for each components.

3.2.3 ERIC-BOX PCR

To obtain an enhanced phylogenetic inference from the genomic fingerprinting, the ERIC-PCR method, described by Choopun (2004), was modified, the direction of modification being to generate more bands of which a significant proportion were common to the closely related strains. The criteria were set to enhance the power to differentiate strains, as well as to improve the power to reveal relatedness among them. It was reasoned that addition of a primer that occurs at relatively constant interval can cause such an effect and the primer was determined experimentally by comparing the effect of primer mixing. Among the most commonly used primers, BOX and REP primer-PCR (Appendix A) was performed under ERIC-PCR conditions (Choopun, 2004). In brief, the higher fidelity Taq polymerase enzyme and PCR buffer (Takara *Ex Taq*™) were used to provide higher reproducibility of the PCR reaction and to increase the range of the amplicon size. In this experiment, an amplicon size as large as 10 kb was reliably amplified. The PCR condition comprised extension temperature at 65 and 70˚C, for 10 min, reaction volumes of 20 µl, and template DNA from 1 ng per reaction, determined as the optimum after testing the variation of the variables. Agarose gel of 1% was run with 0.5X TBE and imaged with ethidium bromide. When it was determined that addition of BOX primer was appropriate, the ratio of three primers (ERIC1, ERIC2 and BOX) was tested for optimization. It was determined that using ERIC1:ERIC2:BOX as the ratio of 1:0.5:0.5 was the most cost-effective optimum.

3.2.4 Phylogenetic analysis of ERIC-BOX electrophoretic types

To elucidate phylogenetic relationships among clones and clonal complexes, the band patterns from electrophoresis were analyzed. The digital fingerprint images were imported into the GelCompar II software (Applied Maths, Sint-Martens-Latem, Belgium). Gel-to-gel variations were normalized on the basis of external marker lanes (HyperLadderI, Bioline USA Inc, Canton, MA) and within-gel common bands were aligned using internal references. After an autosearch for bands, all bands were visually confirmed or corrected. After the signal intensity of band and band positions were determined, presence/absence of bands in a lane was represented as electrophoretic types (ET).

The practical goal of the phylogenetic analyses in this study was to determine the phylogenetic compartments that could not be divided further into smaller compartments within the resolution of the given technique (i.e., the ERIC-BOX PCR) and to analyze the hierarchical or networking relationships among those finite unit compartments. Whether a compartment is indivisible depends on the strength of divergence of genetic information (i.e., phylogenetic signals) among strains of a given collection. To determine the lower limit of hierarchy of the phylogenetic compartments (hereafter, terminal clusters), and the conventional permutation tail probability (PTP) test (Archie, 1989; Faith & Cranston, 1991), which is available as a simple procedure in PAUP* 4.0 (Swofford, 1998), can be used. In the typical use of the test, presence of significant phylogenetic divergence (i.e., internal structure) is decided by rejection of the null hypothesis that "the data have no cladistic structure (beyond that produced by random chance)". What is desired from the test in determining a terminal cluster is, however, accepting the null hypothesis in the situation where the strains in the cluster are tightly related (i.e., most closely related to each other than any other in the population). In this case, the null hypothesis is accepted because the phylogenetic signal is not strong enough to conclude significant internal divergence. Therefore, the challenge is to distinguish acceptation of the null hypothesis due to insufficient divergence, from those due to random combinations of traits or strains. In the present study, this obstacle was circumvented by pre-screening clusters for their certainty of close kinship by a sequence of other independent tests. The approach is effective because preliminary knowledge of the validity of relatedness among PTP-tested strains excludes the possibility of acceptance of the null hypothesis for strains matched merely by random chance.

Although ERIC-BOX PCR, a low stringency rep-PCR, is quick and easy to use to do genomic fingerprinting, the capacity for using its result for statistical analysis is limited, because the exact nucleotide sequence being sampled and producing amplicons via PCR is largely unknown. In the present study, three probability test procedures were devised and used (Appendix B and Appendix C) based on the assumption of uniform distribution of phylogenetic information in PCR results or within the target population, as an alternative to the use of molecular models regarding sequence sampling from the PCR. A uniform distribution of genetic information is unlikely to be observed because there is always a strong probability of heterogeneity or bias in the sequence or strain sampling procedures. Therefore, these tests are required-but-not-sufficient criteria for a cluster to be a valid monophyletic cluster and the results were inferred only for a pre-screening process, before applying a sufficient test, namely the PTP test.

The pre-screening of clusters was performed employing four step procedures in a sequential manner: (1) creation of phylogenetic subdivisions at the upper limit of phylogenetic inference capacity of the band data; (2) generation of candidates of terminal clusters by various, phylogenetically relevant clustering techniques; (3) filtering the candidate clusters by significance in a multiple test of pairwise random band matching (Appendix B); and (4) screening of candidate clusters for consensus appearance in multiple trees (Appendix C).

The first task in converting band distribution in rep-PCR gel lanes into genetic traits of a given strain is determining bandclasses, the global (i.e., applied to all lanes, all gels and strains) positions of bands. A bandclass is noted as the position of a band in gel lanes, designated either as band size (kb) or migration distance of the bands (pixels or mm). Bandclass determination was achieved by matching two bands in different lanes to the same band position (hence to the same genetic locus) when their normalized measures of band positions are identical or similar within the range of lane to lane variation. Matching bands (i.e., a bandclass) can occur from two kinds of sources: (1) a PCR amplicon from the two primer binding site present in the both strains and (2) accidental matching of the size of the two amplicons, each of which were generated from a PCR reaction on different primer sites. The former is true band matching (TBM), while the latter is false band matching (FBM) and carries no phylogenetic information. From the standpoint of practicality, when band matching and bandclass determination is performed for a pair of strains, both TBM and FBM can occur. If the pair is closely related to each other, there are more chances for TBM. However, as the distance between two strains is larger than a given level, all band matching can be FBM. To avoid phylogenetic inference that would be based only on FBM, this case must be avoided. Therefore, a level of phylogenetic divergence between two strains at which there is significant probability for all bandclasses to be FBM is the upper limit of the capacity of phylogenetic inference by a given rep-PCR method. It is ideal to determine the level at which the total population can be divided into smaller subpopulations (or subdivisions), so that all intra-subpopulation pairs of strains have at least one TBM in the band matching process. An additional advantage in creating subdivisions is that the smaller size of the population with certainty of phylogenetic relatedness facilitates the determination of global bandclasses of the subdivision. This feature is, in fact, essential when the number of strains in a population is large, e.g., 100 strains.

To test band matching of two strains for the absence of TBM and prevalence of FBMs, one needs to calculate the probability the observed band matching (i.e., the number of common bands and unique bands in each of two lanes) will occur by random chance. In Appendix B, the probability is calculated as a rudimentary outcome ratio (i.e., Equation 3) while the assumption is made of uniform distribution of band positions and the range of the total number of band positions (*t*) is empirically determined from the gels. When the maximum probability is significantly low (i.e., below the critical value α), it can be concluded that the observed band matching is only possible with the contribution of phylogenetic relationship (i.e., contribution of TBM). When the minimum probability is significantly high (i.e., above the critical value α), it indicates that the observed band matching is possible by random chance, without any contribution of TBM and the pair is designated as an insignificantly matched pair (NMP). As implemented in Appendix D, a top-down test of a tree for the presence of an NMP in a cluster can determine the upper limit of the valid phylogenetic inference for the given data. For valid subdividing, no NMP should be present in the subdivision. For efficiency in creating subdivisions, a tree-building method that creates the most homogeneous clusters is appropriate. Using the BioNumerics Scripting function of GelCompar II, listed in Appendix D, the clusters from complete linkage cluster analysis based on distances (D) from Dice coefficients (S)(i.e., D = square root of $(1 - S)$ and $S = 2a / (2a + b + c)$, where a = the number of common bands, b and c is the number of non-common bands for each lane of a given pair) were split into subdivisions of the target bacterial population.

Once subdivisions were created by the absence of NMP in complete linkage clusters, subdivision-wide bandclasses were determined for each subdivision, using the band matching facilities in GelCompar II. Each genetic locus was represented as a bandclass, i.e., a collection of bands sharing the same DNA migration positions produced by sampling the same genetic loci in the genomes of the strains. For each band position, two versions of data coding were done: binary coding by presence/absence of a band, and continuous value coding by band signal intensity relative to the total band signal intensity of a lane.

To generate candidates for terminal clusters, all clusters appearing in the nine phylogenetic trees were catalogued (Table 3.1). The trees were built using CONTML (a maximum likelihood method using the square root of band signal intensity under the Brownian model), PARS (a maximum parsimony method on binary characters with the Wagner model), NEIGHBOR (a NJ or UPGMA method on Dice coefficient distances) and FITCH (a distance method on Dice coefficient distances with power level 2) in the PHYLIP package (Felsenstein, 2004a), and complete linkage clustering (CL), single linkage clustering (SL), average linkage clustering (MC for the `mcquitty` option) and Ward (WD) clustering in `hclust` function in R statistical programming language (R Development Core Team, 2005).

Table 3.1. Descriptions and classification of nine tree building methods used to generate clusters of candidates of terminal clusters (see text for abbreviation of methods).

Method	Data[†]	Criteria[‡]	Classification[§]	Procedure	Package
CL	d	h	1	hclust	R (stats)
FT	d	m	2	FITCH	PHYLIP
MC	d	c	3	hclust	R (stats)
ML	i	o	4	CONTML	PHYLIP
MP	b	m	5	PARS	PHYLIP
NJ	d	m	2	NEIGHBOR	PHYLIP
SG	d	s	6	hclust	R (stats)
UP	d	c	3	NEIGHBOR	PHYLIP
WD	d	m	2	hclust	R (stats)

[†]: b = binary coded presence/absence of bands; i = continuous variable as band signal intensities; d = distance from Dice coefficient of binary data

[‡]: h = to maximize within cluster homogeneity; m = to minimize the length evolutionary paths or minimize distances (i.e., minimum evolution constraint); c = to maximize similarity to a representative (i.e., a kind of centroids) of existing cluster; s = to maximize cluster membership similarity to the most similar entity.

[§]Classification of method is based on the data and criteria column

Clusters were screened for significance in the multiple test version of the pairwise band matching test with the assumption of random drawing of bands in a uniform distribution (Appendix B). The multiple test was the extension of the pairwise test ($P_{max} < \alpha$) by Holm's method of probability adjustment for multiple test. One additional criterion used to assess the significance of cluster was consensus appearance in multiple trees among the nine trees that generated the catalogued clusters (called multi-tree consensus cluster test). This method was described in detail in Appendix C. In brief, it assumes a uniform frequency distribution of genotypes as one of the possible distributions in an unstructured and LGT-prone population, and the maximum probability of occurrence of a particular split (i.e., cluster formation) from a random sampling of genotypes from the uniform population was calculated as the inverse of outcomes, choosing n genotypes, without replacement, out of (n + 1) total available genotypes (Equation 7). If the same cluster occurs in h number of completely independently-built trees, the maximum probability for such multiple occurrences is h times of multiplications of the maximum probability for one incidence in one tree (Equation 8). When the critical value $\alpha = 0.05$, clusters of any size (i.e., n >1) with more than one occurrence in independent trees (h > 1) was a significantly rare event to reject the hypothesis of the occurrence of the split by random sampling distinct genotypes in the uniform frequency population. At $\alpha = 0.01$, minimum h for such significance of clusters varies with cluster size, being the maximum of five trees for size two cluster and minimum of two trees for size larger than nine clusters (Table 1 in Appendix C). In the application of the multi-tree consensus cluster test, $\alpha = 0.01$ was used on incidences of clusters among the nine trees that were used in cataloguing candidate clusters. In doing that, caution as to independence among tree-building methods was taken. As described in matrixes of Table 3.2, the methods are not independent, as some use the same data or the same constraint of minimum evolution. In the present study, tree independence was qualified when none of common features listed in Table 3.2 were used by given trees. An exception in using this criterion was dealing with clusters of two or three OTUs, when four or five independent trees were required to pass the test but impossible to provide that number of independent trees using the nine methods. The relaxed alternative approach was to use consensus appearance in the ML and MP trees, as well as any other two or three trees. It is important to note that in the application of the two pre-screening tests, the significance of the clusters were determined only by significant deviation from random sampling of the band positions or genotypes from the uniform distribution of band positions or genotypes and that these provided required but not sufficient criteria. Therefore, falsely overestimating the significance of clusters (caused by not uneven sampling or uneven distribution in real data) was acceptable and what was done was to reject clusters arising by random chance from uniformly distributed genotypes or band positions.

The PTP test, with a type I error rate of 0.05, was used to finalize the pre-screened clusters as terminal clusters when the size of a cluster was larger than three, because the PTP test requires a minimum of 4 OTUs for the MP procedure calculation. For clusters larger than 50 OTUs, PTP significance was not calculated because of clusters off-shoots which failed the PTP test, i.e., terminal clusters, and too much time would be required for this to be practical. For clusters of three or two, band patterns were compared side by side and less than a four band difference (out of 20 – 39 bands) qualified as a terminal cluster because it was less than five different bands occurred among those strains of terminal clusters that qualified by the PTP test. In fact, all clusters less than four that were qualified in the pre-screening tests met this criterion, suggesting the possibility of a simple criterion for a terminal cluster being the number of different bands, as long as resolution of the fingerprinting technique remained the same. Details of procedures for cataloguing candidate clusters, pre-screening, and PTP testing, carried out in the R script, are given in Appendix E.

After the significant terminal clusters had been determined, the relationship between strains and terminal clusters were analyzed as a network of OTUs, where the LGT effect could be accounted for by horizontal evolution, using NEIGHBOR-NET (Bryant & Moulton, 2004) implemented in SPLITSTREE version 4 (Huson & Bryant, 2005).

Table 3.2. Assessment of the source of violation of independence between tree-building methods (see text for abbreviation of methods) †.

Method	CL	FT	MC	ML	MP	NJ	SG	UP	WD
CL	x	d	D	0	0	d	d	d	d
FT	d	x	D	0	m	d	d	d	d
MC	d	d	X	0	0	d	d	d	d
ML	0	0	0	x	0	0	0	0	0
MP	0	m	0	0	x	m	0	0	m
NJ	d	d	D	0	m	x	d	d	d
SG	d	d	D	0	0	d	x	d	d
UP	d	d	D	0	0	d	d	x	d
WD	d	d	D	0	m	d	d	d	x

† 0 = no common features, therefore independent from each other;

x = not applicable;

m = constraint of minimum evolution employed;

d = same data coding used

3.2.5 Ecological analysis

To study the ecological properties associated with the observed population structures of *V. cholerae*, variables measured at the environmental source of the *V. cholerae* strains were analyzed with respect to the distribution of the terminal clusters. Because the majority of the *Vibrio* strains used in this study (those labeled RC) had been isolated during surveys carried out in the Chesapeake Bay, much of the biotic and abiotic description of their habitat has been published (Choopun, 2004; Louis *et al.*, 2003). Samples were collected monthly or biweekly, employing biotic and abiotic variables measured at five shore stations in the upper Chesapeake Bay: designated, north to south, site F = Susquehanna River Flats (39°33.13´N, 76°02.20´W); site B = Baltimore Inner Harbor (39°17.00´N, 76°36.32´W); site K = Kent Island (38°58.84´N, 76°20.13´W); site S = Smithsonian Environmental Research Center (38°53.20´N, 76°32.51´W); and site H = Horn Point Laboratory (38°35.59´N, 76°07.80´W). During the period of January 1998 to February 2000, total of 342 culture flasks containing alkaline peptone water (APW) enrichment media were included with one of the three kinds of inocula prepared from 118 samples collected from surface water: W = filtered material on 0.2-µm polycarbonate filters; P64 = particulates collected using the nominal cut-off at 64-µm plankton net; P20 = particulates passing through the 64-µm plankton net but collected using a nominal cut-off of 20-µm plankton net. The volume of estuarine water, the bacterial cells and plankton in which were used as inocula for each flask, differed markedly between W and P20 or P64 and varied moderately between samples (Table 3.3). Because of varying volume of inocula being used in the survey sampling, the limit of detection for *V.*

cholerae or each terminal cluster varied from sample to sample. Zooplankton enumeration was also affected, so the results were interpreted as relative composition, rather than abundance (see below).

Table 3.3. Number and range in volume of the three inocula types for APW enrichment during the 1998-2000 Chesapeake Bay survey for *V. cholerae*.

Inoculum	Cut-off[†] (µm)	No. of Samples[‡]	liters			
			Min	Max	Mean	SD
W	0.2	118	0.15	0.58	0.24	0.04
P20	20	76	25.64	278.31	150.99	62.05
P64	64	76	36.91	698.18	149.93	106.69
Total		270				

[†] The nominal cut-off value (i.e., lower limit) for particle size net and filter used to prepare inocula

[‡] For 72 additional P20 or P64 samples, the collection volume was not recorded. Louis *et al*. (2003) reported that, on average, 0.25 liter for W, 25 ml from ca. 100 ml concentrates of ca. 500-liter estuarine water for P20 and P64, representing in 125 liters for each enrichment flask of P20 and P64.

In this study, each sample was labeled by station, followed by year and date of sampling, with the year coded as the last digit of the year and month coded 1-9 for January to September, zero for October, A for November and B for December (see Figure 3.5).

A total of 221 *V. cholerae* isolates were obtained from 64 enrichment flask for 31 estuarine water samples (Figure 3.5) and described by Choopun (2004) by genotypic and phenotypic characteristics, namely presence/absence of the luminescence gene (*luxA*), cholera toxin gene (*ctxA*) and heat-stable enterotoxin gene (*stn*), and also O1/O139 serotyping, luminescence, arginine dihydrolase, esculin hydrolysis, growth in nutrient broth containing different concentrations of NaCl, acid production from sucrose, arabinose, mannose (Mns), mannitol (Mnt), lysine and ornithine decarboxylase, methyl red (MR), Voges-Proskauer (VP), oxidase, gelatinase, amylase (Amy), lipase (corn oil), chitinase, sensitivity to vibriostatin agent O/129, sensitivity to polymyxin B (PB), and growth at 42°C. In the present study, additional information included Sakazaki O-serotype, determined using a pool of 205 standard monoclonal antibodies. Serotyping was done by Dr. Aarakawa, National Institute of Health, Tokyo, Japan.

Because the full collection of 221 strains contained clonal redundancies, Choopun (2004) removed clonal redundancy in the collection by using >90% similarity of ERIC-PCR fingerprinting band intensity curves, phenotypic characteristics such as biochemical tests listed above, and gene probe hybridization. From the process, 98 *V. cholerae* strains could represent the 221 isolates from the Chesapeake Bay survey. In the present study, the 98 representative strains were included as the part of the strain collection, the membership in terminal clusters for the 221 isolates was determined by extrapolating from the 98 representative strains, based on the clonal identity table by Choopun (2004).

The environmental data of Louis *et al.* (2003) and Choopun (2004) included water temperature (Temp), pH, salinity (Sal), chlorophyll *a* concentration (Chl-*a*), and total bacterial number (TBN). The range of the variables is shown in Table 3.4.

Table 3.4. Environmental parameters included in the analyses.

Data Set[†]	No. samples	No. clusters	pH	Temp (°C)	Sal (ppt)	TBN (10^6 ml^{-1})	Chl-*a* (mg m^{-3})
ENV	140	70	6.5-9.3	-0.5-31	0-15	0.8-21.5	0-218
VEN	31	70	7.2-9.1	12.5-31	0-12	2.9-21.1	3.4-67.5
VEZ	25	64	7.2-9.1	12.5-31	0-12	4.89-21.1	3.4-56.7

[†] ENV = data set comprised all samples with five environmental variables;
VEN = subset of ENV, comprised of samples yielding *V. cholerae* isolates;
VEZ = subset of ENV and VEN, comprising samples yielding *V. cholerae* isolates and also zooplankton composition recorded.

Zooplankton composition was available, measrued in units of relative abundance for 15 taxa, determined with the aim of testing the *a priori* hypothesis that crustacean zooplankton (notably copepods) provide the microhabitat for specific *V. cholerae* clones. Taxonomic levels of the 15 taxa ranged from phylum to suborder. While adult copepods were identified as calanoids, cyclopoids, or harpacticoids (order-level taxa), copepods in the premature instar stages were pooled as subclass Copepoda without further identification. Instar stages were classified to two levels: copepod nauplii and copepodites. According to the taxonomic scheme of Barnes (1987), other crustaceans enumerated at the level of order were amphipods and cumaceans under the class Malacostraca. Crustaceans were enumerated and identified only to class and these comprised ostracods and cirripede (as nauplii). Cladocerans (suborder Cladocera, according to the scheme of Barnes), were the only members identified and enumerated within the class Branchiopoda. Insect larvae comprised non-crustacean arthropods, while oligochaetes, polychaetes, nematodes and rotifers were included in non-arthropod taxa. Descriptive statistics of zooplankton composition are given in Table 3.5.

To test for association between occurrence of *V. cholerae* clones and environmental factors, i.e., physico-chemical variables and zooplankton composition, canonical correspondence analysis (CCA) of presence/absence of *V. cholerae* clones was performed, employing CANOCO for Windows version 4.5 (ter Braak & Šmilauer, 2002). CCA, being an heuristic Gaussian logistic regression technique when applied to binary multivariate data, was judged the appropriate method for asymmetric treatment of presence and absence, with its advantage of a multivariate approach over univariate regressions and distribution-free assessment of significance via a Monte Carlo permutation test. Further justification and diagnostics for the use of CCA are discussed in the Results section.

Since the objective of the analyses was to determine the correspondence between clonal compositions of *V. cholerae* with ecological parameters of their environment (e.g., physico-chemical variables and zooplankton composition), the CCA analyses focused on detection of an environmental gradient that can could the compositional variation of the *V. cholerae* population and its association, if any, between ecological properties, individual terminal clusters, and individual habitat variables. For the former, variance decomposition analysis (Lepš & Šmilauer, 2003) and significance test of the canonical environmental axes by using the permutation test on *F*-ratio values was used. In the permutations, sampling sites served as a block co-variable and only toroidal shift was allowed within a block to retain the temporal order of sampling. It was a precaution in case of temporal autocorrelation in measured or latent variables. For the latter, the Van

Dobben circle (Lepš & Šmilauer, 2003) was employed. It tests the significance of a regression coefficient (of an explanatory variable to a specific response variable) against the null hypothesis of the coefficient being zero by using the *t*-value (i.e., the ratio of a coefficient to the standard error of the coefficient). The latter was of particular interest, since the *a priori* hypothesis that crustacean zooplankton (notably copepods) provide a microhabitat for specific *V. cholerae* clones. The hypothesis was established, based on the observation of preferential attachment of *V. cholerae* O1 to crustacean molts (exuviae) (Huq *et al.*, 1983; Huq *et al.*, 1984; Tamplin *et al.*, 1990). Reports of the predominance of two species of calanoid copepods (*Acartia tonsa* and *Eurytemora affinis*) in the mesozooplankton communities of the Chesapeake Bay, with a strong seasonal and spatial variation (Kimmel & Roman, 2004; Roman *et al.*, 2005), also supported the importance of the hypothesis by demonstrating the trophic significance of copepods in the Chesapeake Bay ecosystem. In these analyses and tests, a difference was considered significant at the 5% type I error level.

Table 3.5. Descriptive statistics of zooplankton composition.

	Set ZOO[†] (n=120)				Set VEZ[‡] (n=25)			
Taxon	Min	Max	Mean	SD	Min	Max	Mean	SD
Copepod nauplii	0.0%	100.0%	48.9%	32.5%	0.9%	98.7%	56.3%	30.5%
Copepodites	0.0%	62.5%	6.8%	10.4%	0.0%	10.8%	2.3%	3.5%
calanoids	0.0%	93.9%	8.4%	16.6%	0.0%	93.9%	6.7%	18.9%
cyclopoids	0.0%	56.0%	2.1%	6.1%	0.0%	56.0%	3.4%	11.2%
harpacticoids	0.0%	50.0%	1.7%	5.1%	0.0%	13.4%	1.5%	2.9%
rotifers	0.0%	97.2%	14.1%	26.8%	0.0%	89.2%	14.3%	28.0%
cladocerans	0.0%	77.7%	3.6%	10.4%	0.0%	20.0%	1.6%	4.7%
Cirripede nauplii	0.0%	96.3%	10.4%	19.0%	0.0%	40.3%	9.9%	12.3%
polychaetes	0.0%	42.3%	2.0%	5.5%	0.0%	42.3%	2.6%	8.4%
ostracods	0.0%	48.8%	0.9%	4.8%	0.0%	4.0%	0.3%	0.9%
oligochaetes	0.0%	19.3%	0.4%	1.9%	0.0%	5.1%	0.4%	1.1%
amphipods	0.0%	7.7%	0.3%	1.1%	0.0%	7.7%	0.6%	2.1%
nematodes	0.0%	30.7%	0.3%	2.8%	0.0%	1.1%	0.1%	0.2%
cumaceans	0.0%	3.8%	0.0%	0.4%	0.0%	1.1%	0.1%	0.3%
Insect larvae	0.0%	11.5%	0.2%	1.1%	0.0%	0.7%	0.0%	0.1%

[†]*Set ZOO = data set comprising all samples for which zooplankton composition data were available during June 1998 – February 2000. Zooplankton data (January – May 1998) were not available.

‡ Set VEZ = subset of ENV, ZOO and VEN, comprising samples yielding *V. cholerae* isolates, with zooplankton composition available as well as all five environmental variables available.

3.3 Results and discussion

3.3.1 Band-matching analysis

Using DNA fingerprint data, the electrophoretic image data can be analyzed in two forms: curve-based versus band-based. In the former, genetic traits are interpreted as corresponding to individual pixels and the light emission intensity of the DNA-staining dyes is treated as the status of the genetic loci. In comparisons between OTUs, typically the Pearson correlation coefficient is calculated as the measure of relatedness (Thompson *et al.*, 2004b) from which genetic distance is calculated, i.e., 1 – coefficient. Although this method is easy to use, it has drawbacks if confidence in the phylogenetic analysis is to be obtained. The correlation coefficient can arise from a spurious correlation of a few strong bands that happen to have similar migration distances only by chance, leading to a false measure of relatedness. As relatedness between OTUs increases, the coefficient tends to saturate quickly. Therefore, resolution of a phylogenetic relationship is low when it is most needed. In addition, it has the fundamental problem that the individual loci, i.e., pixels, are not actually genetic loci in the bacterial genomes.

The band-based calculation is time-consuming, but individual bands can correspond to discrete DNA sequences on a bacterial genome. Through the sequential process of band-calling and band-matching, a bandclass (a band locus corresponding to a genetic locus shared by OTUs in the analysis) can be determined to have binary traits (presence/absence) or quantitative (band intensity) traits. This approach is not only free from the drawbacks of a curve-based approach, but also allows the use of phylogenetic inference methods like maximum parsimony and maximum likelihood. In this study, a band-based analysis of ERIC-BOX PCR results was done using the image analysis software, GelCompar II.

When inferring phylogeny from the band-based analyses, however, the validity of the band-matching is a problem because a genetic locus sampled by fingerprinting can assume any size of the DNA fragment. In genetic profiling based on a conventional electrophoresis, such as AFLP and RFLP, a band is interpreted as a trait by size difference, so that traits are matched by band size. In the AP-PCR or a low-stringency rep-PCR, such as ERIC-BOX PCR, some band matching may be invalid because the sequence of the primer binding sites producing the bands of similar size in two different electrophoretic types can be produced by different genetic loci among the OTUs. When the genetic distance between two OTUs is above a critical level, the entire set of matching bands (usually a few bands) can arise from different genetic loci, leading to false phylogenetic inference.

This "false band matching" in band-based analysis of rep-PCR results can be tested by determining the bandclasses after confirmation of the validity of the band matching. In this study, a scheme of tests of significance in band matching was devised to determine whether the matching of band patterns for a pair of OTUs was statistically rare, that is, beyond random chance, suggesting matching having arisen systematically (i.e., uneven sampling or phylogenetic cause). The intended null hypothesis was that there was no phylogenetic relationship between two OTUs and the practical null hypothesis was that band matching was the result of two independent sets of random sampling of the observed number of bands for each OTU from a uniform probability distribution of bands of different sizes. Details of the pairwise testing are described in Appendix B.

When pairwise band-matching was performed using GelCompar II, 15 nonsignificantly matched pairs (NMP) were found among the 171 distinctive electrophoretic types (hereafter designated as OTUs). To avoid placing members of a NMP into the same cluster, subdivisions of the total population were created on the bifurcating tree constructed by complete linkage cluster analysis of pairwise Dice coefficients. The clustering algorithm was selected because it maximizes within-cluster homogeneity. Whenever a cluster in the tree had a NMP, the cluster was divided into child branches (hereafter, subdivisions) without a NMP. The result was four subdivisions (Group 1, 2, 3 and 4) at type I error for the test at 0.01 (Figure 3.1). Because no valid comparison could be made among OTUs in different subdivisions, these subdivisions formed the upper limit of the cluster analysis in this study and primary lineages could be determined only within these subdivisions.

Primary Subvar Cluster A

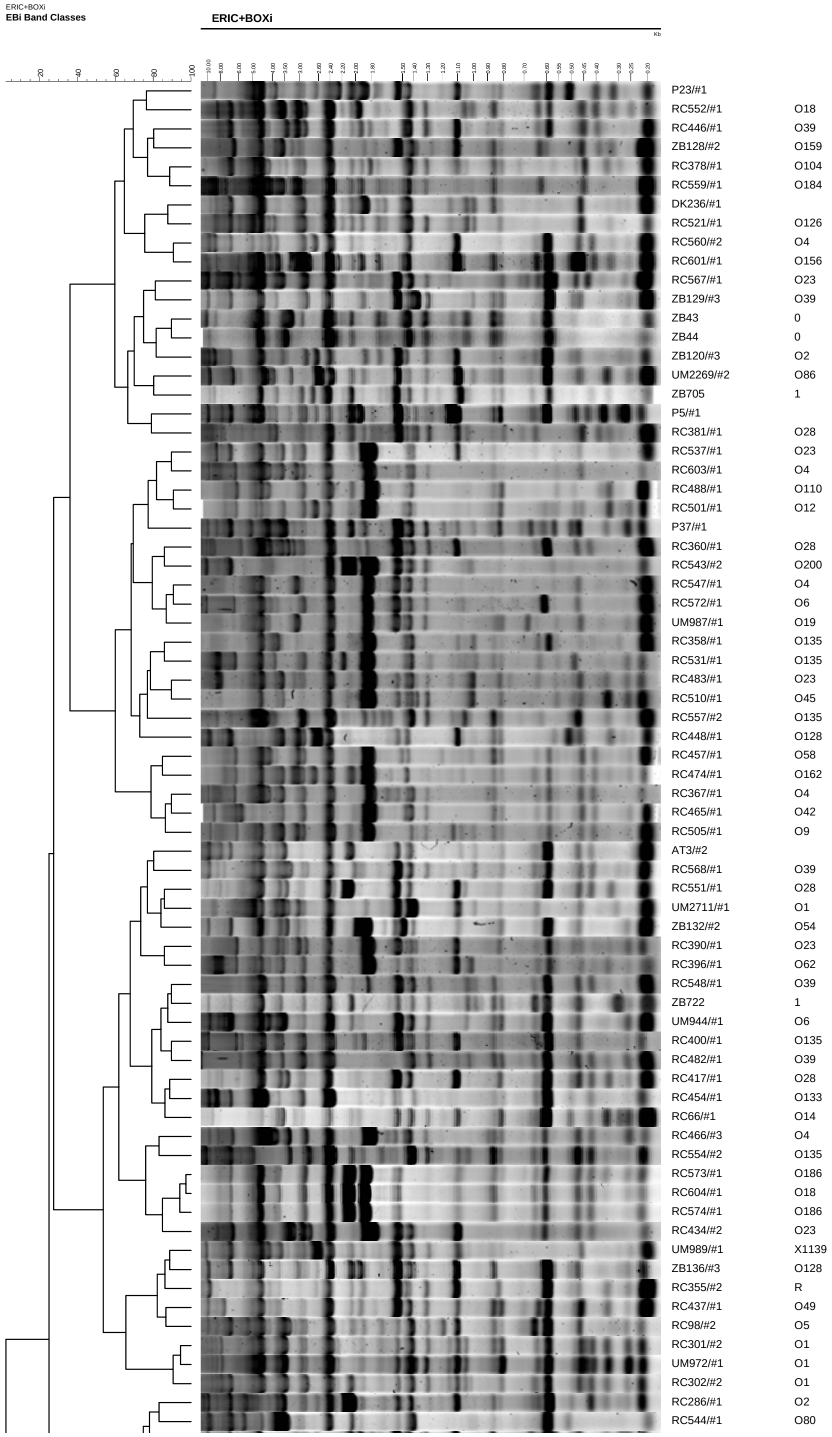


Figure 3.1 (*continued*).

Primary Subvar Cluster A *(continued)*

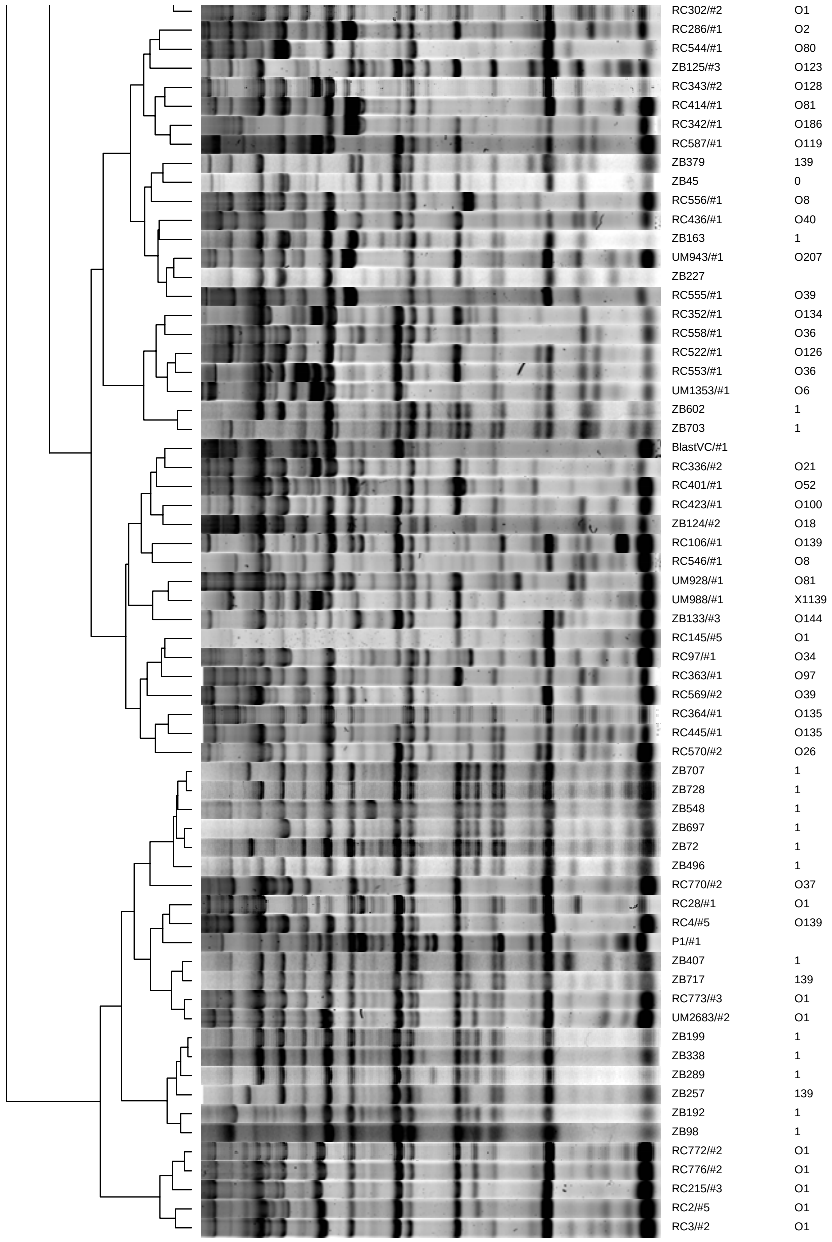
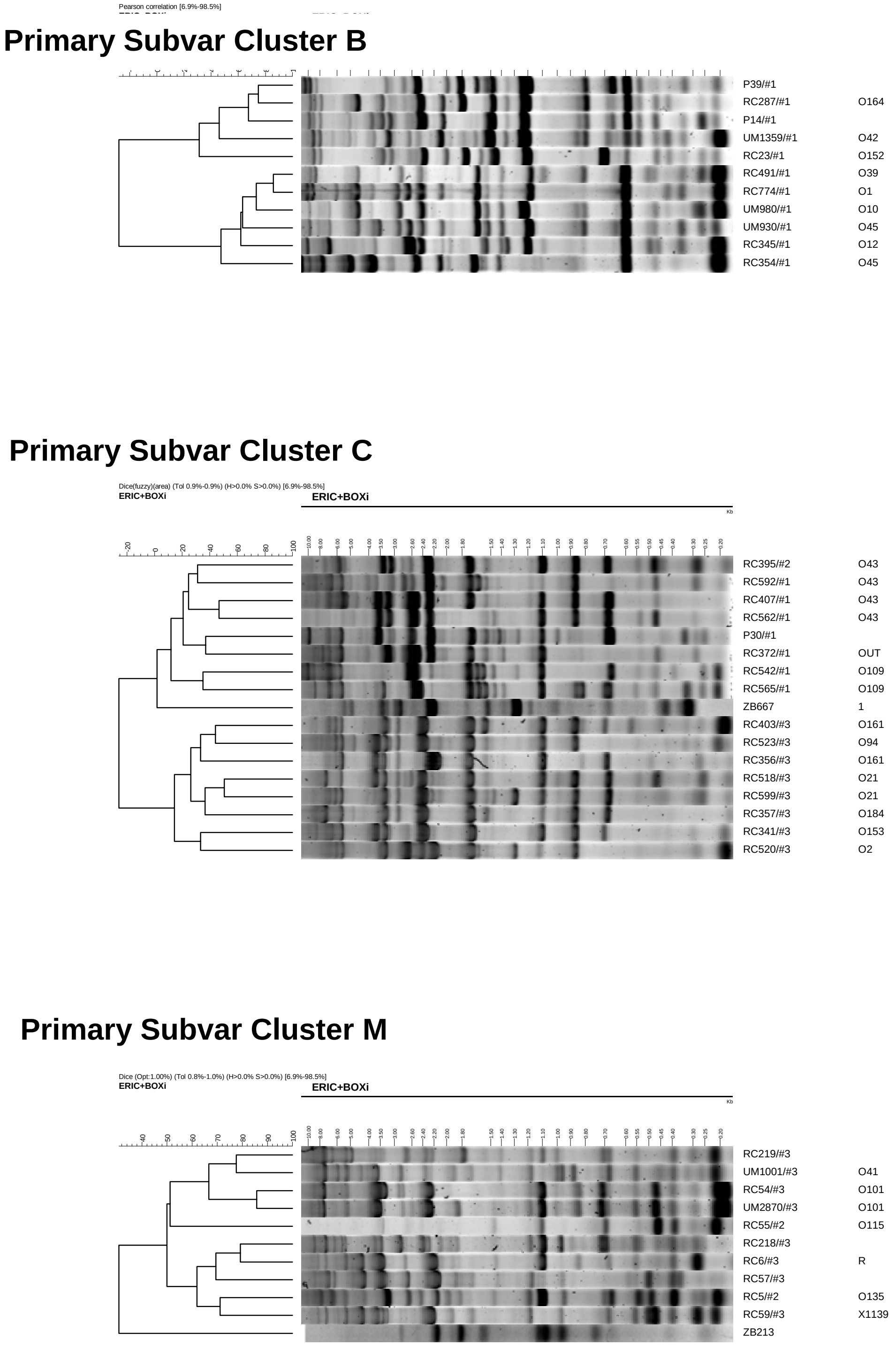


Figure 3.1 (*continued*).



Primary Subvar Cluster B

Primary Subvar Cluster C

Primary Subvar Cluster M

Figure 3.1 (*continued*).

Figure 3.1. Normalized electrophoretic gel lanes and complete linkage clustering (based on the Dice coefficient) for 170 strains (serogroup designations without ‘O’ prefix are strains whose serotype was determined from the source culture collection or from their source; X1139 = non-O1/non-O139, as determined by Louis *et al.* (2003); R = rough; OUT = new serogroup not cross-reacting with known Sakazaki *V. cholerae* serogroups; RC586 = the sole member of the primary subvar cluster D, not shown here because no matching was found).

3.3.2 Determination of primary lineages

Within each subdivision, bandclasses were determined using GelCompar bandclass analysis functions and manually validated. After the bandclasses were determined, pairwise test of significance of band matching was repeated to verify the validity of the pairwise band matching with the bandclasses. With one exception, all members of the four subdivisions formed significantly matching pairs at the type I error level of 0.05. The exception was RC586 that did not have any bands confidently matching with other OTUs in Group 3, of which *V. mimicus* strains comprised the remainder of the OTUs. Because it had no significant matches with strains from the other three groups, RC586 is concluded to comprise a separate primary lineage.

In a previous study (Choopun, 2004), three primary clusters were detected, i.e., *V. cholerae* (Cluster A), *V. mimicus* (Cluster M) and a separate *V. cholerae* lineage (Cluster B). These clusters corresponded to Groups 1 and 2, Group 3, and Group 4, respectively. Therefore, the method used for subdivision creation corresponded to the previous classification of primary lineages with the modification that Cluster A was split into two clusters. In addition, there was an indication of the presence of an additional primary lineage with the clonal clustering of three strains (RC584, RC585, and RC586), which were separated from a single sample analyzed by an unconventional DNA probe hybridization method. It was detected because it demonstrated a unique biochemical profile and its ERIC-PCR genomic fingerprint did not cluster consistently with the other lineages. With improved genomic fingerprinting, this problem was revisited and analyzed.

To confirm the hypothesis that RC586 represents a separate primary lineage, another data set was generated, using DNA-DNA hybridization to determine the relationship of this strain with other primary lineages. RBR with five probe strains were obtained: RC145 and RC466 representing Group 2, RC395 representing Group 4, RC5 representing Group 3, and RC586. When the RBR data were plotted and analyzed in both two-dimensional and three-dimensional space, using combinations of three out of five probe strains, RC586 was located nearest to the *V. mimicus* cluster, but with a discernable distance (Figure 3.2, top). To obtain the best resolution, PCA was performed to reduce the five source axes to three major axes (Figure 3.2, bottom). In both, RC586 was separated from Groups 3 and 4. Therefore, presence of another, separate, primary lineage is concluded from the analysis.

The significance of this finding is that the *V. cholerae* taxonomic complex is paraphyletic, meaning members of the taxon originated from a single common ancestor, but the group includes not all of the descendants, because *V. mimicus,* a branch of *V. cholerae*, has been named as a separate species (Davis *et al.*, 1981). Considering its 16S rRNA sequence similarity and population structure, *V. mimicus* should, instead, be considered as a primary subvar of *V. cholerae*. The four other primary lineages observed in this study qualify for the same taxonomic designation, based on 16S rRNA sequence similarity (Choopun, 2004) and range (30% - 100%) of genomic difference measured by RBR.

In conclusion, the band matching analysis yielded five primary subvar lineages of *V. cholerae*, namely Cluster A = Group 2, Cluster B = Group 1, Cluster C = Group 4, Cluster D = RC586, and Cluster M = *V. mimicus*.

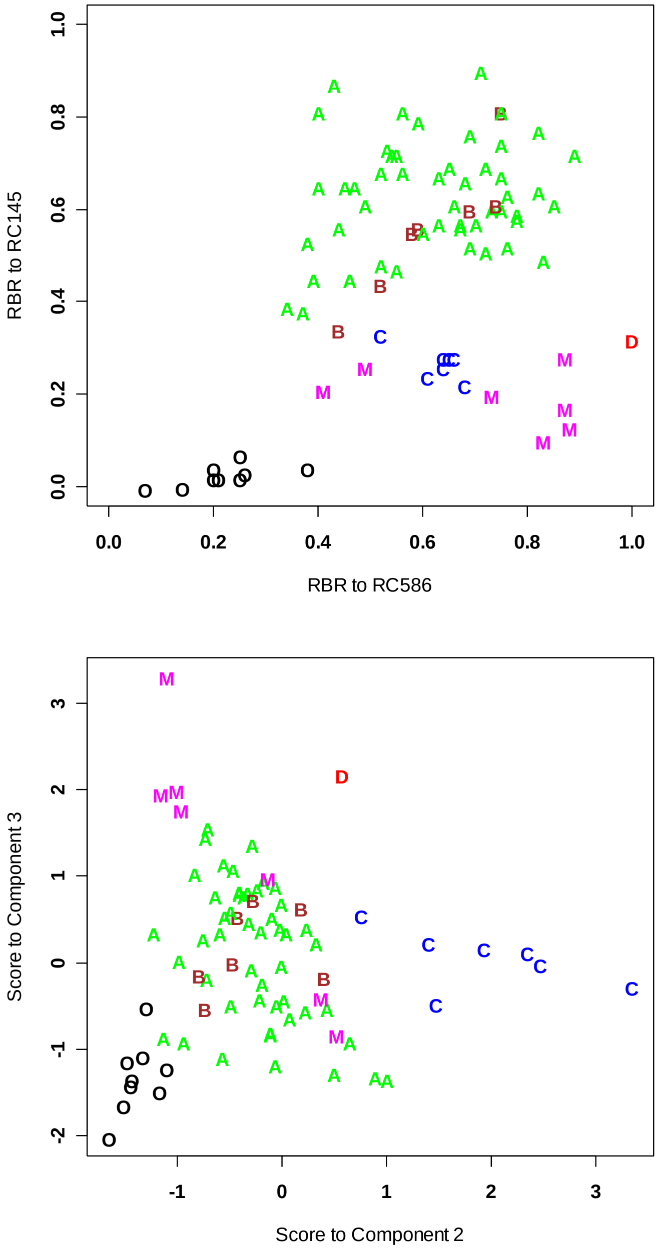


Figure 3.2. Relative location of strains in RBR space (top panel) and PCA space (bottom panel) by their primary subvar membership (i.e, *V. cholerae* subvar = A-D, and M; outer group = O). Component 1 did not contribute to the differentiation of RC586 (the only subvar D strain) from groups C and M. Component 2 yielded maximum coefficient in the RBR with RC395 (representing the subvar C), and Component 3 essentially comprised RC586 alone. Note that the RBR space expresses genome relatedness in a very compressed way, particularly in coordinates close to the origin.

3.3.3 Determination of terminal clusters

Terminal clusters, defined as clonal complexes whose phylogenetic signal (i.e., genetic divergence) is not detected significantly by a given method of analysis, were determined primarily by using the PTP test. Because the primary cluster A contained the epidemic strains, the analysis focused on this cluster (Figure 3.3). Although cluster A contained ten times more OTUs than the other primary clusters, 87 out of 133 OTUs were singletons and ten terminal clusters contained only two OTUs, indicating diversity of the cluster was so great that the collection was only very sparsely sampled.

To account for LGT-driven evolution, a network of OTUs was constructed using the Neighbor-Net algorithm (Figure 3.3) to elucidate relationships between terminal clusters. The epidemic lineages of *V. cholerae* O1 classical strains (O1CL) formed a single terminal cluster, indicating only weak divergence among them but strong divergence from the other *V. cholerae*. *V. cholerae* O1 El Tor strains showed closest relationship with the O1CL cluster. In this study, we found a new lineage of O1 El Tor strains. The O1ET3 strains were originally isolated from the coastal waters of Bangladesh and their distinctive fingerprints showed significant divergence from other O1 El Tor lineages when the PTP test was applied ($P < 0.01$). The general pattern that emerged from the network view was that the closer the relationship among strains, the more horizontal influence there was. The location of O1ET3 is interesting for this sense because it is located between the typical epidemic O1 clusters and environmental clusters. The terminal clusters were the tightest bundles in the network by being connected to each other at rates more frequent than remotely related clusters. This can be explained that the most frequent mechanism of communication is homologous recombination, in which the similarity of gene sequences supports more frequent successful recombination.

To observe the relationship of OTUs within a terminal cluster, the distribution of genetic distance was examined, using genetic distances from the centroid of the clusters to individual OTUs (Figure 3.4). The centroid for a terminal cluster was defined as the imaginary OUT, the genetic locus (i.e., a bandclass) of which had the allele (i.e., presence or absence of bands) prevalent at each locus of the OTUs of the terminal cluster. Interestingly, the same pattern emerged for all clusters, namely absence of strains around the centroid. This result implies a founder flush (Garg *et al.*, 2003), by which the founder of the cluster and its close relatives have a very low frequency due to counter selection from the environment and more competent offsprings. When the distance of the OTUs from the global centroid was examined, the same pattern emerged. Therefore, the results can be interpreted as an example of the universal phenomenon known as founder flush occurring during evolution of the species.

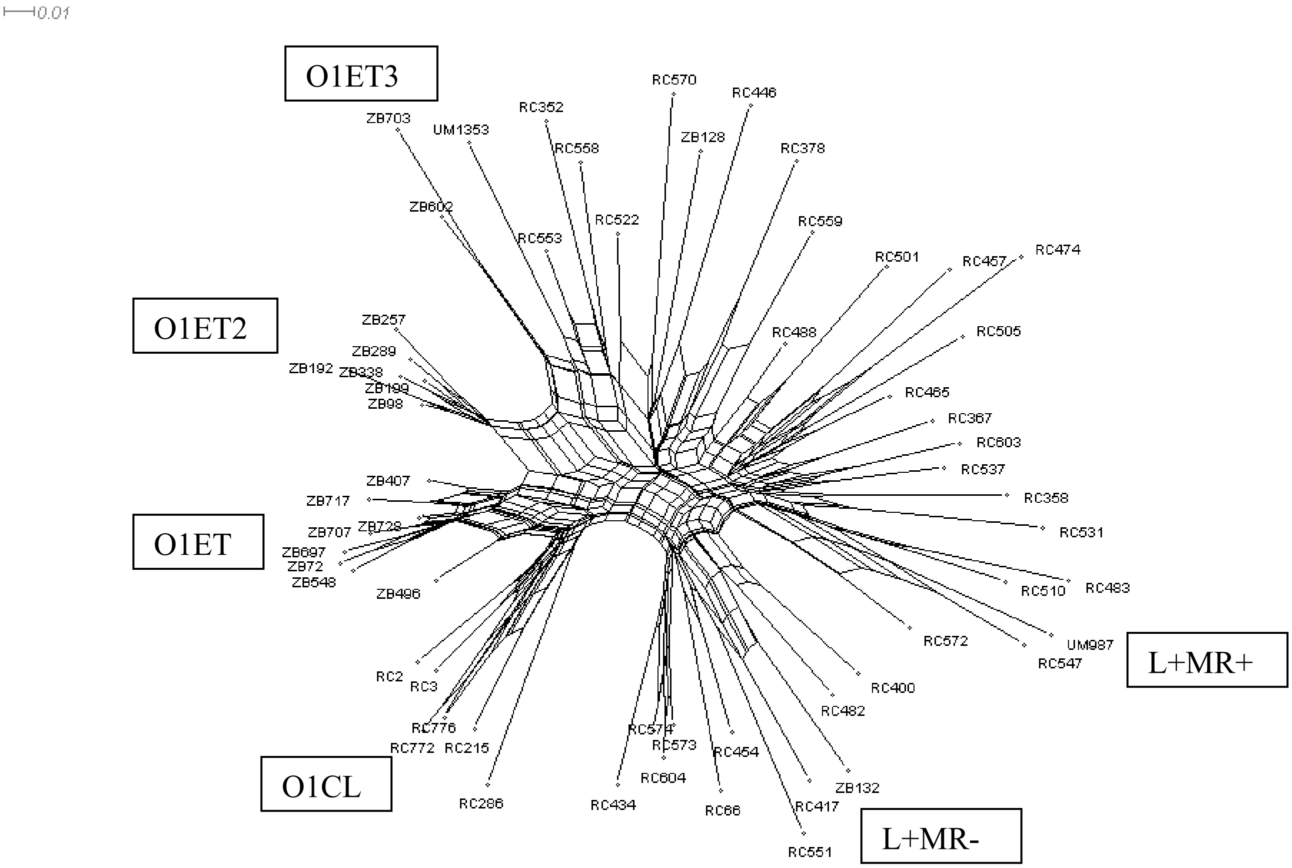
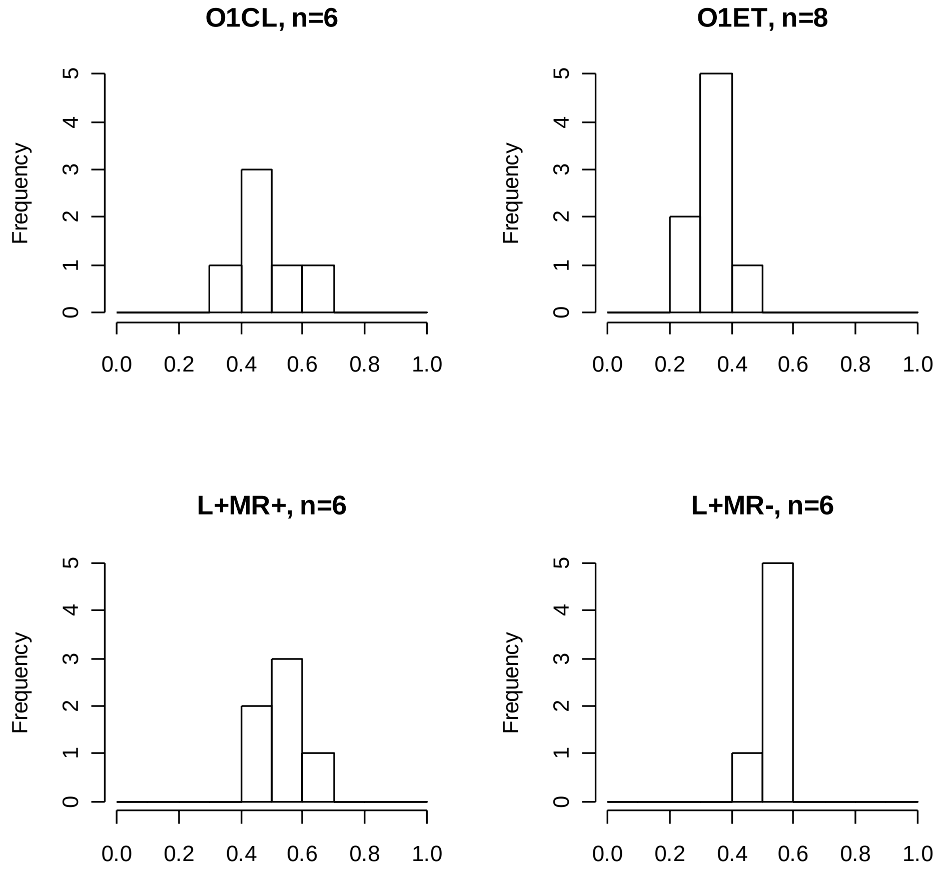


Figure 3.3. The neighbor-net (Bryant & Moulton, 2004) view of terminal clusters within primary cluster A (O1CL comprising *V. cholerae* O1 classical strains, O1ET the O1 El Tor strains, L+MR+ the luminescent and methyl-red test positive strains, and L+MR- the luminescent and methyl-red test negative strains).


O1CL, n=6

O1ET, n=8

L+MR+, n=6

L+MR-, n=6

Figure 3.4. Frequency distribution of distances of OTUs from the centroid of the terminal clusters (distance was calculated using the Dice similarity coefficient; O1CL comprised the *V. cholerae* O1 classical strains, O1ET the O1 El Tor strains, L+MR+ the luminescent and methyl-red test positive strains, and L+MR- the luminescent and methyl-red test negative strains).

3.3.4 Characterization of ecological properties of the terminal clusters

In the previous sections, a structured diversity within the *V. cholerae* population was detected, based on a phylogenetic interpretation of genetic divergence among the clonal subpopulations, namely the primary clusters and terminal clusters. The questions that arise from this finding are what ecological functions are displayed by clonal entities in their natural environment and what determines their occurrence and geographic distribution in the environment. Both questions can be answered by characterizing the ecological niches of the terminal clusters of *V. cholerae*.

First, the distribution of phenotypic and genotypic characteristics within the various terminal clusters was compared and the variable characteristics are listed in Table 3.6. Except for one case, all strains in a terminal cluster were either 100% *luxA*-positive or 100% *luxA*-negative, indicating that luminescence is a solid ecological characteristic that varies at the unit of terminal clusters or higher. A similar situation was found also for the *stn* gene. Except for two clusters, all strains in a terminal cluster unanimously possessed the toxin gene or did not. In contrast, various O-serogroup antigens were present whenever the terminal cluster was large, indicating that the O-antigen mediated interaction with other organisms or substrates was very finely tuned, diverging below the resolution of the current terminal clusters. Such was also observed in the case of utilization of mannose, mannitol, and methyl red reaction. Therefore, the compartmentalization at the level of terminal clusters is concluded to be genetically stable for major ecological denominators, exemplified by luminescence and toxin possession, varying at the level of the subtle phenotypic characteristics, such as sugar utilization. The greater variability of the O-antigen implies that cell surface structure is under strong selection or counter selection and plasticity is the trait preferentially demonstrated by naturally occurring *V. cholerae*. The implication of this finding is that there is a strong probability of a shift in the O-antigen in both the epidemic and non-epidemic strains of *V. cholerae*.

CCA was selected as the method of choice for analyzing the general, or preferential, distribution of terminal clusters in their habitat. On the other hand, direct gradient analysis, such as logistic regression, can be ideal for finding significant association between the presence/absence of clusters along an environmental gradient. Nevertheless, there were several advantages of CCA that suited the type of data in the analysis. First, variation in sample volume used to isolate and identify *V. cholerae* strains made the variation of the detection limit a main concern. In the application of direct regression analysis, the variation in the meaning of absence can cause too much noise, particularly when the gradient is short. In the case presented here, the gradient is framed within the summer season in which most of the *V. cholerae* strains were isolated. In the case of CCA, this problem is abolished, because the weighted averaging method disregards absence data and takes only presence data into account when the optimum of a terminal cluster is calculated. Secondly, CCA is a multivariate method, unlike the typical logistic regressions. In the latter case, the individual response variable (i.e., individual terminal cluster) is analyzed against environmental factors. For diverse bacterial populations, this is impractical. For a complex response such as species composition in a community or clonal composition in a population, the interactions among the response variables are also of concern. With numerous and sparsely occurring response variables, such as in the present study, a multivariate approach is essential. According to Scheiner

(1993), accumulation of type I error when using multiple univariate regression should be avoided, with multivariate analysis the preferred solution.

Cautions must be taken when using CCA, since it is only an approximation of logistic regression when certain strong assumptions are met. The conditions are typically called a species packing model in which the optimum and the tolerance of the response variables are homogeneous along the gradient and similar in their magnitude (ter Braak, 1986). In the simulation, it was found that the approximation to an unimodal Gaussian logistic regression is always robust when the probability for occurrence of response is low. In the case of bacterial clones, extensive diversity tends to keep the probability at low levels. Therefore, the bacterial clonal data set, filled with mostly absence, is robust for the violation of this assumption. Another strong assumption in applying CCA is that, like other typical multiple regression or ordination approaches, a latent variable is present as a combination of measurable variables. This raises the issue of multicolinearity among the explanatory variables, when many explanatory variables are used. Therefore, reducing the model to include a minimum of unrelated variables was intended in this study.

In the case of the strains from Chesapeake Bay labeled RC, environmental variables and zooplankton composition were available for testing the explanatory power of each variable with respect to variation in clonal composition of *V. cholerae*. For the zooplankton data, since only relative composition was available from phylum to suborder (Barnes, 1987), the data could not be treated like conventional species composition data. Instead, the data were square root transformed and treated as conventional environmental variables like soil composition. According to the simulation experiment of Legendre and

Gallagher (2001), Euclidian distance-based ordination on square root transformed compositional data is equivalent to ordination using Hellinger distance and the result gives the best representation of the data in terms of ecological relationship. In the test of the *a priori* hypothesis of association between crustacean and *V. cholerae* terminal clusters, non-crustacean variables were not included as dependent variables. Therefore, interdependence among variables, which is typical in relative composition data, was not included in the data set analyzed in this study.

The relationship of the two sets of independent variables, the environmental variables and the zooplankton composition, was examined by performing decomposition of the *V. cholerae* population variation, following the method of Lepš and Šmilauer (2003). This was done because a causal relationship can occur between environmental and zooplankton variables, resulting in a high proportion of variance shared by the two sets of independent variables. Interestingly, results given in

Table 3.7 indicated virtual lack of any variance commonly explained by the five environmental variables and zooplankton composition. From this, one can conclude that biotic variation in the system, such as the dynamics in the zooplankton composition can act independently from seasonal change of the water or climate variables, such as temperature. The results also showed that different data sets could be usedz for analysis of the effects of the five environmental variables and for zooplankton composition. The former becomes the most efficient when using subset VEN (31 samples) (Table 3.4), while VEZ (25 samples) serves for the latter (Table 3.5).

Table 3.6. Genotypic and phenotypic characteristics of *V. cholerae* clonal clusters from Chesapeake Bay (see Materials and Methods section for abbreviations).

Cluster	n	*luxA*	*stn*	Mns	Mnt	MR	VP	PB	NaCl6%	Amy	42oC	Sakazaki Serogroup[†]
A29	34	100%	0%	**97%**	**97%**	100%	100%	**3%**	100%	100%	100%	O121, O18, O23, O28, O4, R, X1139
A34	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O184
A35	1	0%	100%	100%	0%	100%	100%	0%	100%	100%	100%	O36
A36	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O135
A37	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O8
A38	1	0%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O39
A39	2	0%	0%	**50%**	**50%**	**50%**	100%	0%	100%	100%	100%	O135
A40	1	0%	0%	100%	0%	100%	100%	0%	100%	100%	100%	O36
A41	1	0%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O18
A42	2	0%	0%	100%	100%	100%	100%	0%	100%	100%	100%	O28
A43	2	0%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O39
A44	1	100%	0%	0%	0%	100%	100%	0%	100%	0%	100%	O8
A45	2	0%	0%	0%	100%	100%	100%	0%	0%	100%	100%	O80
A46	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O200
A47	1	0%	100%	0%	0%	100%	100%	0%	100%	100%	100%	O23
A48	1	0%	0%	100%	100%	100%	100%	0%	100%	100%	100%	O126
A49	1	0%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O9
A50	1	100%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O12

Cluster	n	*luxA*	*stn*	Mns	Mnt	MR	VP	PB	NaCl6%	Amy	42oC	Sakazaki Serogroup[†]
A51	1	100%	100%	100%	100%	0%	100%	0%	100%	100%	100%	O110
A52	2	0%	0%	100%	100%	0%	100%	**50%**	100%	100%	100%	O39
A53	1	100%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O162
A54	9	100%	0%	**89%**	0%	100%	100%	0%	100%	100%	100%	O4
A55	1	0%	100%	100%	100%	0%	100%	0%	100%	100%	100%	O42
A56	1	100%	100%	100%	100%	0%	100%	0%	100%	100%	100%	O58
A57	1	0%	0%	100%	0%	100%	0%	0%	100%	0%	100%	O133
A58	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O128
A59	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O39
A60	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O135
A61	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O49
A62	6	100%	100%	100%	0%	100%	100%	**17%**	**83%**	100%	100%	O40
A63	1	0%	0%	0%	100%	100%	100%	0%	100%	100%	100%	O100
A64	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O28
A65	6	0%	0%	0%	**83%**	**33%**	100%	**50%**	**83%**	100%	100%	O149, O81
A66	4	0%	0%	0%	**75%**	**75%**	100%	100%	**25%**	100%	100%	O52, X1139
A67	1	0%	0%	0%	100%	0%	100%	0%	100%	100%	100%	O135
A69	1	0%	0%	0%	100%	0%	100%	0%	100%	100%	100%	O28
A70	4	0%	0%	0%	100%	**25%**	**75%**	0%	100%	0%	100%	O104, X1139

Cluster	n	*luxA*	*stn*	Mns	Mnt	MR	VP	PB	NaCl6%	Amy	42oC	Sakazaki Serogroup[†]
A71	1	100%	0%	100%	0%	100%	100%	0%	100%	100%	100%	O4
A72	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O135
A73	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O97
A74	1	0%	0%	100%	100%	100%	100%	0%	100%	0%	100%	O28
A75	2	0%	0%	0%	100%	0%	100%	0%	100%	100%	100%	O134
A76	1	0%	0%	100%	100%	0%	100%	100%	100%	100%	100%	O128
A77	1	100%	0%	100%	100%	100%	100%	0%	100%	100%	100%	O186
A78	10	0%	0%	0%	100%	**60%**	**90%**	100%	**80%**	100%	100%	O21, X1139
A88	18	100%	**28%**	**11%**	0%	**33%**	100%	0%	**94%**	100%	100%	O23, O4, O45, O62
A89	1	0%	0%	100%	100%	100%	100%	0%	100%	100%	100%	R
A95	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O126
A97	13	100%	**92%**	**92%**	**8%**	**85%**	**92%**	0%	**77%**	**92%**	**92%**	O135
A98	9	**89%**	0%	**44%**	**33%**	**67%**	100%	0%	100%	100%	100%	O23, O62, X1139
A99	1	0%	0%	0%	0%	100%	100%	0%	100%	0%	100%	O4
A100	1	100%	0%	100%	100%	100%	100%	0%	100%	100%	100%	O23
A103	7	0%	0%	0%	100%	**14%**	100%	0%	100%	100%	100%	O18
A104	3	100%	0%	0%	100%	**33%**	100%	**67%**	100%	100%	100%	O109, X1139
A105	1	0%	0%	0%	100%	0%	100%	0%	100%	100%	100%	O2
A106	1	100%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O40

Cluster	n	*luxA*	*stn*	Mns	Mnt	MR	VP	PB	NaCl6%	Amy	42oC	Sakazaki Serogroup[†]
A107	1	0%	100%	0%	100%	0%	100%	100%	100%	100%	100%	O156
A108	1	100%	0%	0%	0%	100%	100%	0%	100%	100%	100%	O200
A109	1	100%	0%	0%	0%	0%	100%	0%	100%	100%	100%	O62
A110	1	100%	0%	100%	0%	0%	100%	0%	100%	100%	100%	O121
B10	1	0%	0%	100%	100%	100%	0%	100%	100%	100%	100%	O109
B11	11	0%	0%	100%	100%	100%	0%	**73%**	100%	100%	100%	O43
B3	2	100%	0%	100%	100%	100%	100%	0%	100%	100%	100%	O45
B4	1	0%	0%	100%	100%	0%	100%	0%	100%	100%	100%	O39
B5	8	0%	0%	0%	100%	**13%**	100%	0%	100%	100%	100%	O12, X1139
C2	1	0%	0%	100%	100%	100%	100%	100%	100%	100%	100%	O109
C4	1	0%	0%	100%	100%	100%	0%	0%	100%	100%	100%	O43
C6	1	0%	0%	100%	100%	100%	0%	0%	0%	100%	100%	O43
C7	2	0%	0%	100%	100%	100%	0%	0%	100%	100%	100%	OUT
C9	16	0%	0%	100%	100%	100%	0%	**69%**	**56%**	**56%**	**94%**	O153, O16, O161, O184, O2, O21, O94

[†] R = rough; OUT = not typed with existing scheme; X1139 = only known as non-O1 / non-O139

For proper interpretation of constrained ordination, the effect of possible blocking variables must be estimated in order to build an appropriate test model. In the Chesapeake Bay survey, five sampling locations and three kinds of inocula (W, P20, and P64) were used. When tested by CCA, with respect to their contribution to the variation in the *V. cholerae* composition, variance explained by difference in the kinds of inocula was not significant by the permutation test. The ecological implication of the lack of significance, with respect to size fractionation used in the inoculum preparation, is that a strain of *V. cholerae* can occur as either attached or in a free-living state. On the other hand, this can be due to technical noise, meaning very large difference in the scale of sample used in W fraction and the two plankton fractions (P20 and P64) or low detection limit for plankton fractions rising from concentration of organisms inhibiting growth of *V. cholerae* in enrichment flasks (e.g., swarming bacteria). Although sites for sampling were not significant, it is obvious that materials and the biota do not freely exchange among distant locations, and site was used a blocking variable.

Table 3.7. Decomposition of variance in *V. cholerae* composition in Chesapeake Bay samples (VEZ) by environmental variable and zooplankton composition.

Symbol[†]	Eigenvalue	Proportion	No. of variables
V_E	3.431	26%	5
V_Z	9.694	74%	15[‡]
$V_{E\cap Z}$	-0.062	0%	-
$V_{E\cup Z}$	13.063	100%	20

[†] V_E represents variance in *V. cholerae* composition explained by environmental variables, V_Z by zooplankton composition, and $V_{E\cup Z}$ by all of the variables. $V_{E\cap Z}$ represents variance in *V. cholerae* composition commonly explained by the two sets of variables.

[‡] Since zooplankton data represented only relative composition, the practical maximal number of variables was 14.

Table 3.8. Models and results from canonical correspondence analyses (CCA)[†]

Model ID	Data Set	Samples	Clusters	Explanatory variables[‡]	Variance	P_{ax1}	P_t	Gradient
1	VEN	31	70	Temp*TBN	4.3%	0.046	0.046	Cold oligotrophic versus warm eutrophic
2	VEZ	25	64	Cyclopoids*Calanoids	6.2%	0.031	0.031	Dominance of major copepods (cyclopoids versus calanoids)
3	VEZ	25	64	Temp*TBN + Cyclopoids*Calanoids	9.8%	0.027	0.007	Axes1 = Model 1 gradient; Axes2= Model 2 gradient
4	VEZ	25	64	∑(9 crustacean taxa)	41%	0.120	0.020	Proportions of various crustacean taxa
5	VEZ	25	64	∑(4 adult copepod taxa) + copepod nauplii + copepodites	23%	0.031	0.153	Proportions of various copepod taxa and life stages

[†] Columns: ID = tentative identification by number; Data set = subsets in Table 3.4; Samples = the number of samples in the data set; Clusters = response variables of the CCA model , i.e., the number of terminal clusters of *V. cholerae* found in the data set; Explanatory variables = independent variables of regression models after the random block variance caused by sites of *V. cholerae* isolation and variable measurement were accounted for; Variance = ratio of total canonical eigenvalue to the total response eigenvalue; P_{ax1} = significance of the first canonical axes and P_t = that of the total variance (i.e., trace); Gradient = ecological gradients explained by the canonical axes of models.

[‡] asterisk = interaction of variables; plus sign = linear addition of variables two variables before and after; sigma sign = linear addition of all variables; TBN were log transformed.

To find environmental variables and zooplankton taxa, the variation of which was associated with the 70 terminal clusters of *V. cholerae*, the data set used was comprised of a total of 31 or 25 samples (Table 3.4) and was tested for significance against random association using permutation tests. The significant models and results are summarized in Table 3.8. Among the environmental variables, temperature and total bacterial number gave the strongest eigenvalues, but none were significant as a single explanatory variable. When the two variables occurred together in the model, either as linear additions or interactions, significance was $P<0.05$. In the former case, the trace eigenvalue was significant, but not the individual canonical axes. In the latter, using the interaction term between temperature and TBN, the primary canonical axis was significant. Since the two variables are highly correlated (r = 0.5, Pearson coefficient, P <0.01), the gradient formed by their interaction must be synergistic for at least one of the variables. An ecological explanation is that TBN represents the carrying capacity of the water column for the standing crop of heterotrophic bacteria supporting growth with nutrients. Therefore, eutrophication (influx of nutrients) caused by increased productivity in the water column serves the process of interaction. The relatively poor precision of TBN measurement and various levels of terrestrial organic nutrient input at different sites render the TBN noisier than temperature. However, the seasonal increase in carrying capacity appears to be more direct in effect on *V. cholerae* abundance and composition. Using the full two-year range of winter-to-summer gradient of temperature and TBN, Louis *et al,* (2003) found even higher correlation between the two variables. This also can be explained by a reduction of carrying capacity for heterotrophic bacteria during winter seasons due to low productivity of the system. However, the contribution of chlorophyll *a* was not significant, whether alone or as an addition to the models shown in Table 3.8, implying the source of organic nutrient for decomposer bacteria is not solely planktonic primary production. When the gradient was visualized in the contingency table, ordered by scores of canonical axis (Figure 3.5), it can be seen that the richness of *V. cholerae* strains is high at the upper side of the gradient, indicating an increase in carrying capacity, driven by the seasonal temperature rise, and supports a diverse set of *V. cholerae* strains. However, it was also noted that the spring samples (April – early June) and late September produced isolates of genotypes (namely, A48, A95, A103, A72, A71, A73, A78, A89, A99, B3, C7, A104, A74, and A34) not found during the summer season. Therefore, it can be concluded that there are *V. cholerae* ecotypes that have lower temperature optima and/or oligophilic nutrition, with a narrow range of tolerance (hence, specialists, at the lower side of the gradient).

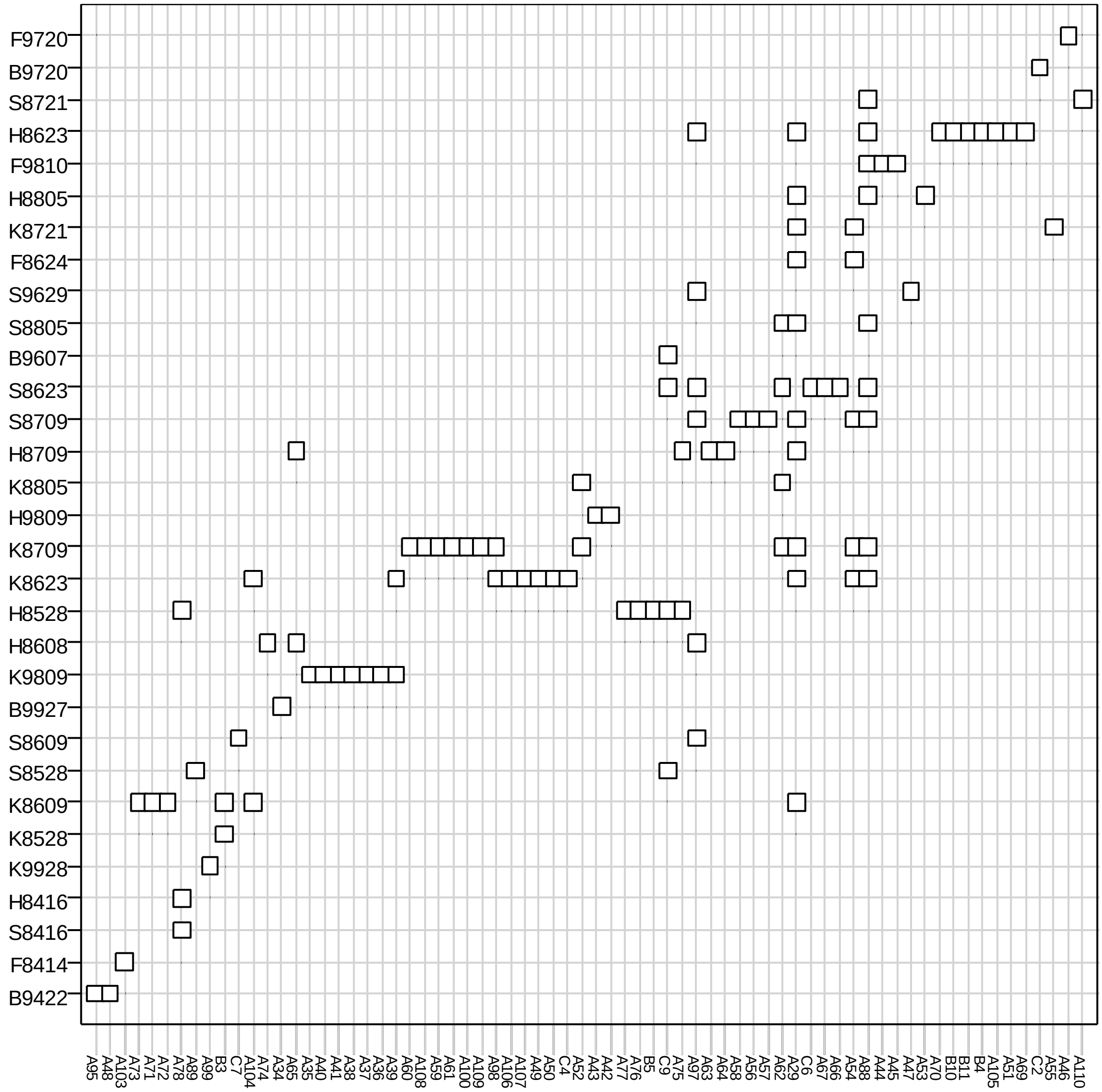


Figure 3.5. Incidence of *V. cholerae* isolates by sample and terminal cluster in Chesapeake Bay. Both samples (31 rows) and terminal clusters (70 columns) were ordered according to score on the gradient of the canonical axes from the interaction of temperature and total bacterial number (low scores to high scores from the left or the bottom to the right or top). Presence was marked as an open square while absence was not marked (Sample label: the first letter = site; the second letter = year, where 8, 9, 0 are the years 1998-2000 in order; the third letter = month where 1-9 for Jan. to Sep., 0, A and B for Oct. to Dec.; the forth-fifth letters = days in a month).

In models 4 and 5 of Table 3.8, the hypothesis that crustaceans and/or copepods are associated with specific *V. cholerae* genotypes was tested and partly significant results were obtained. With nine crustacean taxa composition, 41% of the *V. cholerae* variation among the 25 samples was explained at $P = 0.02$ significance. In detail, cyclopoids, amphipods and calanoids yielded a canonical correlation of $r > 0.5$ with the first, the second, and the third canonical axes, respectively. The contribution of two of the three crustacean taxa, namely calanoids and cyclopoids, was confirmed in a variable reduced model (the model 5), which tested the hypothesis that copepod was associated with specific *V. cholerae* terminal cluster. In contrast to the adult copepod taxa, instar stages were collective copepod nauplii and copepodites, and these two life-stage taxa yielded lesser pattern of variation, indicated by the relatively small coefficient of variance (Table 3.5), and did not produce any significant explanation of variance in *V. cholerae* composition. When the variation of calanoid and cyclopoid were examined, in the models 4, 5 and an additional model with only the two as independent variables, the gradients of cyclopoids and calanoids headed in opposite directions. Therefore, their interaction was selected as the sole variable explaining variances of *V. cholerae* composition in model 3 and the results were found to be significant. When the gradient of model 3 (shown in Figure 3.6) was examined, the presence of the gradient was obvious only at the end of the gradient range. In the middle of the gradient, samples containing diverse genotypes or genotypes with a broad distribution among the samples were found. When the source of strong gradients at both termini of the gradient was examined, the most contribution was found from the northern sites, where salinity can be low, in the two different seasons (B9422, F9720, F9810, B9720, and F8624 versus F9927). This result suggests influence of salinity and the relationship is depicted in Figure 3.7 as two different kinds of relationship between the two copepods. With the exception of the sample B9720, the lower end samples showed low salinity while the upper end samples showed mesohaline conditions. What is more interesting is that, at low salinity, the relationship of the two copepods is antagonistic to each other, implying competition between the two groups. On the other hand, relative abundance of calanoids in the mesohaline samples was not affected much in the visible fluctuation of cyclopoids. From previous detailed researches of Dr. Roman (Kimmel & Roman, 2004; Roman *et al.*, 2005), it is well known that two kinds of calanoid copepods predominate in the mesozooplankton community of Chesapeake Bay. Namely, *Acartia tonsa* is abundant in mesohaline regions, while *Eurytemora affinis* is episodically abundant in response to freshwater influx via Susquehanna River discharge. This relationship can be understood as a result of ecological differences in the two calanoid species. The observation suggests an antagonistic relationship between the cyclopoid species and *Eurytemora affinis* in low salinity areas, while the predominance of *Acartia tonsa* in mesohaline areas does not vary. For those 10 samples at the end of the gradient of model 2, unique *V. cholerae* clones were isolated to form the gradient of *V. cholerae* clonal composition. Therefore, it can be concluded that distribution of certain *V. cholerae* genotypes is concomitant to the incidences or predominance of given species of calanoid copepods. What is more interesting is that those *V. cholerae* clones identified to be temporally and physically, with the upper limit of the scale as ca. 150 liters (Table 3.3), connected to copepods could be transduced by cholera toxin-carrying bacteriophage (Table 3.9).

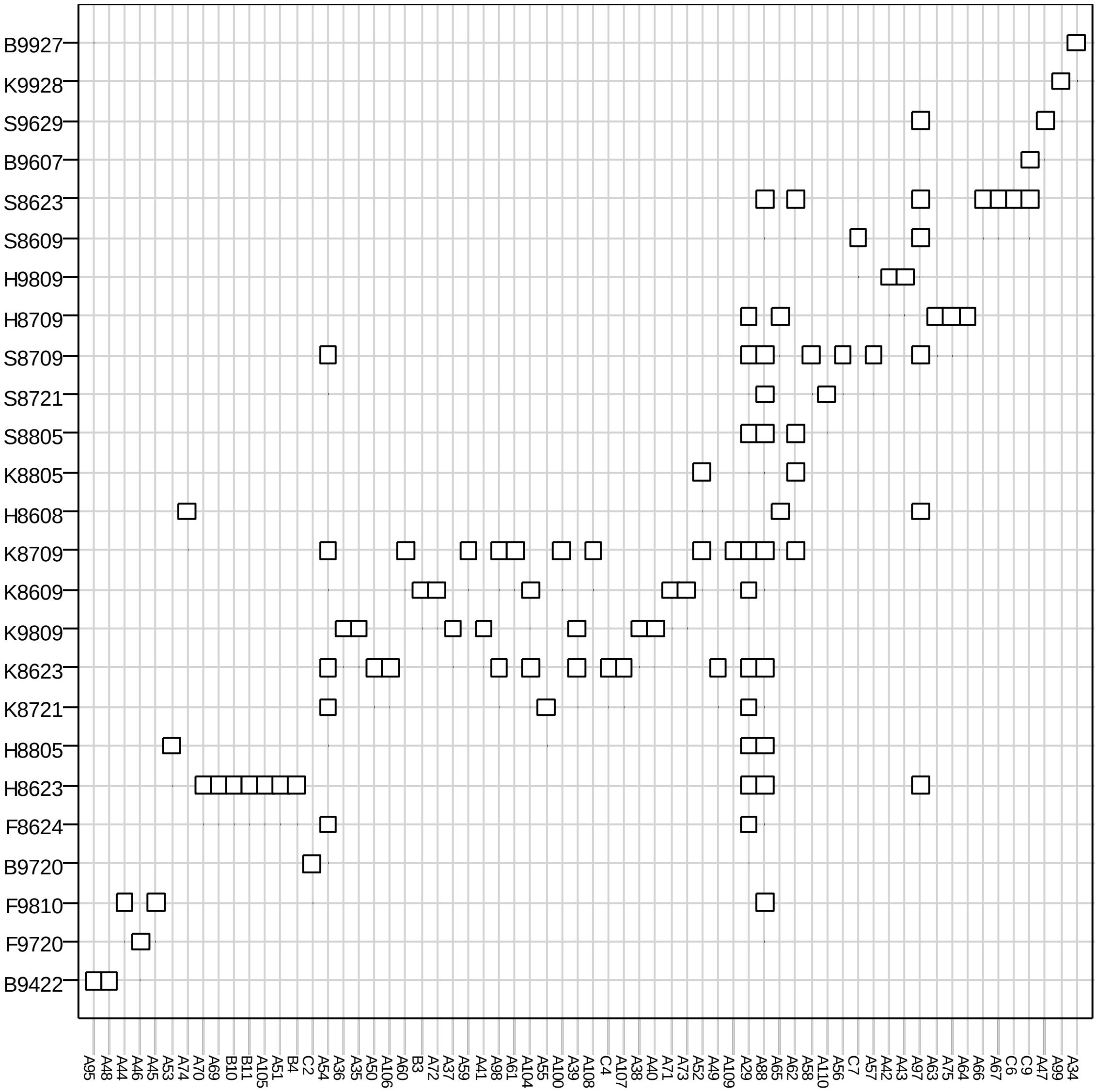


Figure 3.6. Incidences of *V. cholerae* isolates by sample and terminal cluster in Chesapeake Bay. Both samples (25 rows) and terminal clusters (64 columns) were ordered according to scores on the gradient of the canonical axes from interaction of the two adult copepod taxons, cyclopoids and calanoids (low scores to high scores from the left or the bottom to the right or top). For example, 94% of zooplankton in B9422 was calanoids but no cyclopoid. In B9927, 55% were cyclopoids while 6% were calanoids. Presence was marked as open squares while absence was not marked (Sample label: the first letter = site; the second letter = year, where 8, 9, 0 are the years 1998-2000 in order; the third letter = month where 1-9 for Jan. to Sep., 0, A and B for Oct. to Dec.; the forth-fifth letters = days in a month).

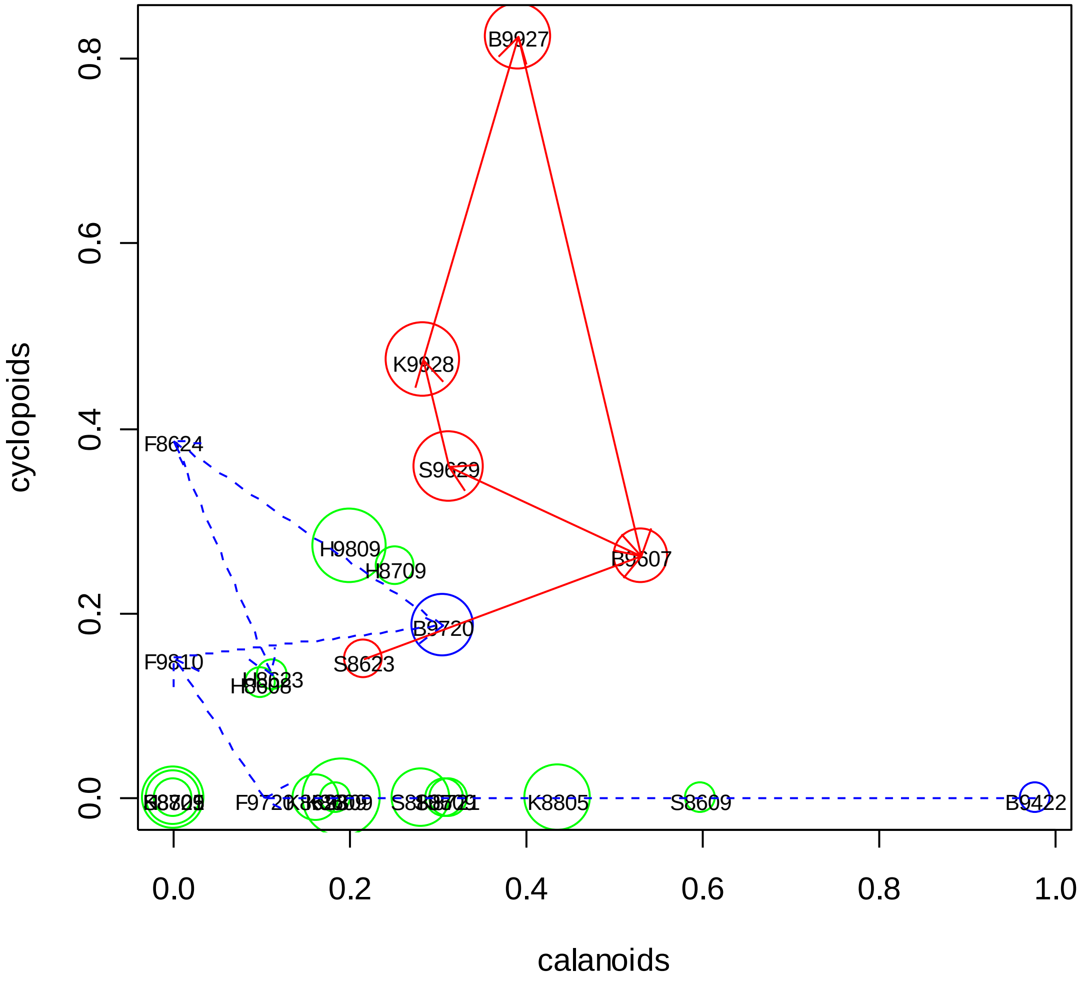


Figure 3.7. Variation of salinity (circle radius) and relationship (arrows) between calanoid and cyclopoid copepods along the gradient of *V. cholerae* clonal composition. Blue dotted arrows follow the five samples located at the lower end of the cyclopoid*calanoid gradient in model 2 (Figure 3.6). Red solid arrows follow the five samples from the higher end of the gradient. Each axis is the cubic-root transformed value of the relative composition of adult copepods.

To test further the presence of a specific association between *V. cholerae* clones with zooplankton taxa, van Dobben circles (Figure 3.8) on a *t*-value biplot were employed (Lepš & Šmilauer, 2003). Unlike the CCA permutation test, this method is distribution sensitive, so that outliers influence the results. Furthermore, canonical regression coefficients do not follow precisely the Student *t*-test; therefore, using the results to directly accept significantly associated pairs is not appropriate. Authors of the method, however, recommend it to identify insignificant variables for a particular response variable (ter Braak & Šmilauer, 2002). In the present study, the method was applied to generate hypotheses of specific association. Because the present study was limited in the number of samples from which *V. cholerae* was isolated, the resulting hypothesis can be tested in future studies in other geographical areas. By applying the survey and tests including freshwater ecosystems in tropical areas, it is possible to characterize cholera epidemics in terms of ecological linkers.

The specific association, supported by significant positive regression coefficients with *t*-values larger than 2.0, is shown in Table 3.9. Besides calanoid and cyclopoid copepods, whose influence previously found in the CCA permutation tests, amphipods and cladocerans produce a *t*-values larger than two. It was found that one amphipod associated strain carried the heat-stable enterotoxin gene, while those associated with cladocerans were luminescent, and carried various O-serogroup antigens. The latter group has the potential to be linked with endemic cholera, because cladocerans are as predominant zooplankton as the calanoid copepod in estuarine waters of Chesapeake Bay. Together with previously described results, the findings of this study corroborate the *a priori* hypothesis that crustacean plankton function as a microhabitat for *V. cholerae*, and a reservoir or breeding ground for novel cholera bacteria. Observing that *V. cholerae* of various kinds of serogroups are associated with one kind of crustacean, classified at the taxonomic level of order, raises the hypothesis that many kinds of *V. cholerae* colonize a single species of zooplankton in different manners (e.g., different site in the body of an organism).

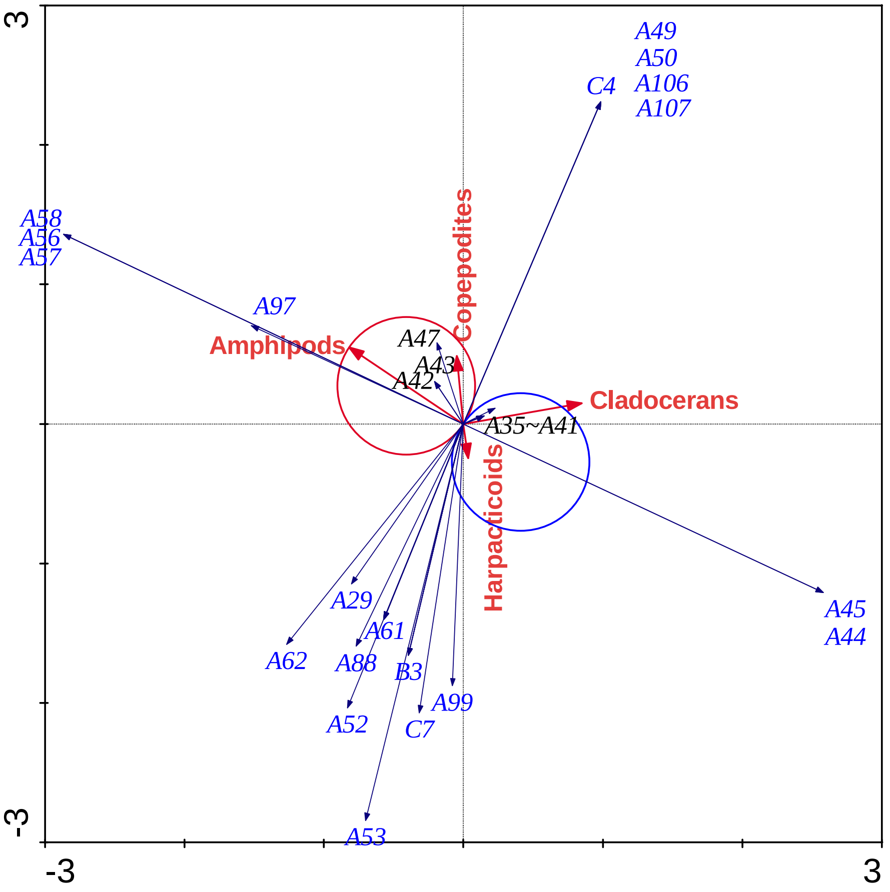


Figure 3.8. Use of van Dobben circles to determine significance of the regression coefficient between a *V. cholerae* clone and crustacean taxa (the terminal clusters whose arrow head falls inside the circles are significant by the *t*-value larger than 2 criteria; red circle = positive direction toward amphipods, blue circle: negative direction toward amphipods).

Table 3.9. Characteristics of terminal clusters significantly associated with crustaceans[a].

Clones	Zooplankton taxa	Serogroup	CTXΦ transduction	*luxA*	*stn*
A42	Amphipods	O28	-	-	-
A43	Amphipods	O39	-	-	-
A47	Amphipods	O23	-	-	+
A48	Calanoid copepods	O126	-	-	-
A95	Calanoid copepods	O126	+	+	-
A35	Cladocerans	O36	-	-	-
A36	Cladocerans	O135	-	+	-
A37	Cladocerans	O8	-	+	-
A38	Cladocerans	O39	-	-	-
A39	Cladocerans	O135	-	-	-
A40	Cladocerans	O36	-	-	-
A41	Cladocerans	O18	-	-	-
A34	Cyclopoid copepods	O184	-	+	-
A99	Cyclopoid copepods	O4, O156	+	-	-

a: CTXΦ transduction = transduction of *V. cholerae* strains by CTXΦ phage marked with a kanamycin resistance gene; *luxA* = presence of luminescence gene (*luxA* gene) determined by dot-blot hybridization; *stn* = presence of heat stable enterotoxin gene (*stn* gene) determined by dot-blot hybridization (Choopun, 2004).

3.4 Summary

In this chapter, the population structure of *V. cholerae* was analyzed using the optimized rep-PCR and band analysis methods. At least five primary subvar clusters were identified, with *V. mimicus* being one of them, and the conclusion was drawn that the species *V. cholerae* comprises a paraphylum. The structure of the *V. cholerae* populations strongly suggests a universal founder flush within both primary subvar clusters and terminal clusters. By comparing the epidemic clonal complex with other clones, both clonality and divergence are relatively well established, such that the epidemic cluster appears to be an ecological specialist that is divergent from other environmental strain complexes. Among the environmental variables tested, total bacterial number serves as an indicator of the carrying capacity for heterotrophic bacteria, when it was associated with water temperature. This relationship also explains seasonal and spatial variation of *V. cholerae*. In the search for a connection between population structure and habitat, crustacean zooplankton were found to be associated with a coincidence of specific *V. cholerae* clones. The presence of a specific correspondence between crustacean taxa, together with a strong phylogenetic relationship among epidemic clones, it is concluded that some *V. cholerae* clones are ecological specialists found only in a narrow range of tolerance around the optimum, while the most common strains were found regardless of tested environmental gradients. Therefore, the carrying capacity of the ecosystem for heterotrophic bacteria is found to be proportional to the diversity of *V. cholerae* clones and the presence of diverse environmental *V. cholerae* can also imply the presence of a carrying capacity for epidemic clones.

Chapter 4. Population Structure and Dynamics of Pathogenic *Vibrio cholerae* in Cholera-Endemic Areas of Bangladesh

4.1 Introduction

4.1.1 Endemic cholera

In Bangladesh, cholera is an endemic disease, recurring as successive waves year after year, even while other parts of the world remain free of cholera (Lipp *et al.*, 2002). However, detailed mechanisms of such endemism are not yet clearly understood (Faruque *et al.*, 1998; Sack *et al.*, 2004).

Recent research on endemic cholera in Bangladesh has identified two important characteristics of endemic cholera. The first is the tight modulation of the dynamics of the disease by climate factors. El Niño has been proposed as a modulator of inter-annual variation of cholera (Pascual *et al.*, 2000; Rodo *et al.*, 2002), while seasonal changes in water temperature are correlated with seasonality of this endemic disease (Lipp *et al.*, 2002). The second is highly variable composition of cholera-causing organisms. *V. cholerae* O1 classical strains persisted in the regions around Bay of Bengal for decades even while it disappeared from other parts of the world. Cholera-causing agents in the region now have been replaced by new bacterial strains, *V. cholerae* O1 El Tor biotype and *V. cholerae* O139 serotype, whose prevalence varies dramatically each year (Longini *et al.*, 2002). Now that the aquatic environment is revealed as the reservoir of *V. cholerae*

(Colwell & Spira, 1992; Sack *et al.*, 2004), these characteristics of endemic cholerae in Bangladesh can be integrated into our understanding of the mechanism of endemic cholera by investigating the structure and dynamics of *V. cholerae* in the water column of geographical endemic areas as the “causal linker” between the modulators and the disease dynamics.

4.1.2 Target population of study

As described in the previous chapter, the diversity of *V. cholerae* in the environments is immense and cataloging and tracking the diverse entities is a major hurdle in the study of population dynamics. Yet, from the fact that the strains responsible for epidemic cholera have always been considered to be the clonal lineage of *V. cholerae* O1 (including both classical and El Tor biotype and O139 serogroup as described in Chapter 3), it is possible to assume that the population of *V. cholerae* in the aquatic environment directly causing shifts in the composition of cholera bacterial populations also belongs to the *V. cholerae* O1 lineage. In the long term, the role of other phylogenetic compartments in natural waters can be as a reservoir of genetic variation, causing shifts in pathogenicity-related traits, such as antibiotic resistance or surface antigens (Faruque *et al.*, 2004). In the short term, the abundance of the bacteria can be a good indicator to estimate the carrying capacity of bodies of water supporting growth of *V. cholerae*. For these reasons, investigating the dynamics and composition of the pathogenic *V. cholerae* O1 and O139 serogroups, in conjunction with those of environmental *V. cholerae* non-O1 and non-O139 can be efficient and informative.

4.1.3 Objectives

In the cholera endemic Bay of Bengal region, toxigenic *V. cholerae* are commonly isolated from watery stools of cholera patients and from natural or artificial aquatic environments with which the human populations of the area is contact. *V. cholerae* strains isolated from cholera patients collectively represent clinical isolates and the strains are considered to have emerged by cell multiplication within the intestine of the cholera patient. Fecal contamination of water resources with clinical strains shed by the cholera victims is believed to be the major internal forcing of the disease, positively driving the dynamics of cholera outbreaks. *V. cholerae* strains isolated from aquatic environments collectively represent environmental isolates. These strains could have originated from two kinds of sources: clinical strains via fecal contamination of the water or *V. cholerae* populations indigenous to the aquatic ecosystem. As described above, the former results in internal forcing, while the latter involves environmental forcing in cholera outbreaks.

The aim of this study was to understand the nature of environmental forcing in endemic cholera by examining the structure and dynamics of environmental and clinical *V. cholerae* populations isolated from the Bay of Bengal region. Therefore, two separate paths of study were taken.

The first was structural analysis. The interactions of *V. cholerae* in both the environmental and clinical habitat (intestine) were quantified in four remotely located areas of Bangladesh. The effects of habitat separation and of geographic separation between populations of *V. cholerae* were studied. The latter was a population dynamics approach. By identifying links between the dynamics of clinical cases and *V. cholerae* populations in the environments on a local scale, the mechanism of endemic cholera and its modulation were elucidated.

Samples and datasets were obtained from a long term, large scale survey, undertaken from both clinical and environmental perspectives within a well-defined spatial and temporal frame. Biweekly clinical surveys were synchronized with environmental surveys in four areas of Bangladesh over a two and a half year time frame (Figure 4.1). *V. cholerae* O1 El Tor was isolated from the aquatic environment in each sampling area. For structural analyses, both environmental and clinical isolates were analyzed using genomic fingerprinting methods. For the analysis of population dynamics, culturable *V. cholerae* were enumerated from environmental samples using a *V. cholerae* – *V. mimicus* specific gene probe and the resulting time-series data were correlated to other environmental or clinical data.

4.2 Materials and methods

4.2.1 Description on the survey

A Climate, geography, and hydrology of cholera endemic areas of the Bay of Bengal

The climate of the Bay of Bengal area is characterized by two seasonal monsoon systems called the southeast monsoon and northwest monsoon. Wind direction and currents in the Bay of Bengal change according to shifts in the two monsoons. Precipitation is concentrated within the time of the boreal summer and fall during the northwest monsoon period. The hydrology of the Ganges Delta area is especially affected by the monsoon and the melting of glaciers in the Himalayas during the boreal summer. River discharges dramatically increase during that period (Dwivedi, 1993).

The land mass of the Ganges Delta is low-lying, with an elevation from sea-level of <20 m. The hydrology of the Delta is highly affected by the sea level. With an increase of 0.5 m in the sea surface height, >20% of the land mass around the Sundabarn area will be under sea level and intrusion of sea water occurs up to the center of the Delta at Dhaka, Bangladesh (Lobitz *et al.*, 2000). The origin of soil and the land mass of the Delta is alluvial deposits from the Himalayan river discharge. Therefore, a porous aquifer comprises the hydrologic network of the Delta. In addition to cultural and political aspects, which are also critical in the dynamics of clinical cases of cholera, the border between West Bengal, India, and Bangladesh poses an hydrologic problem because of the difference in water resource management strategies of the two countries.

Productivity of the water column in and around the Delta area is determined by the cycling of the two alternating monsoons. The plankton community shows seasonal peaks in the spring and early fall. The composition of the plankton community is known to be very diverse, influenced by the variation in salinity along the Ganges River water basin. Freshwater species of plankton are dominant in the upper river areas, while euryhaline species comprise the community in lower river brackish waters. The ecosystem of the Delta can be divided into several characteristic ecotones. The coastal zone is part of the Bay of Bengal, whose water chemistry is dominated by the freshwater discharge from the Ganges-Brahmaputra and Irrawaddy River systems. Pelagic zones (lotic zones) of rivers of shallow depth constitute the main hydrology across the Delta land mass. Littoral zones are dominated by a mangrove ecosystem, one of the world's largest mangrove systems.

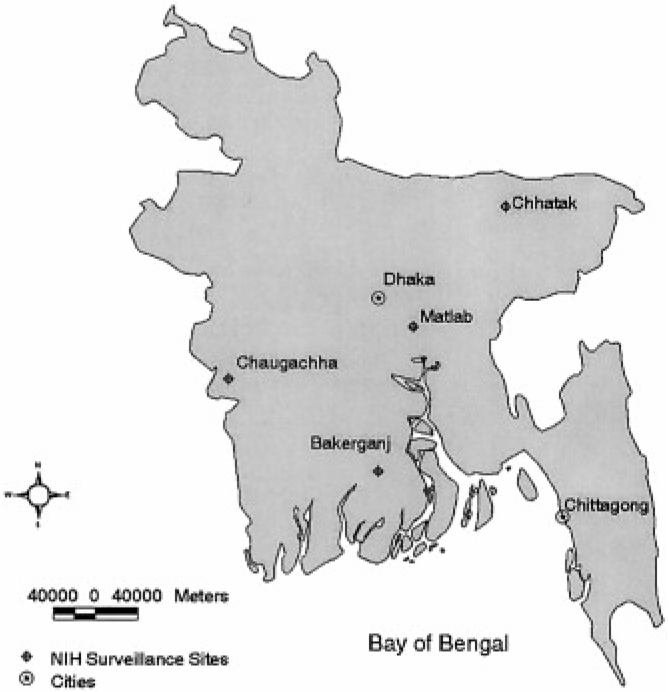


Figure 4.1. Four survey sites in Bangladesh, where clinical and environmental surveillance was performed synchronously within the time window of three days twice a month. In Chittagong, coastal water was screened for *V. cholerae* O1 and O139, but a clinical survey was not undertaken.

B Survey areas

Surveillance was done in collaboration with the International Center for Diarrheal Disease Research, Bangladesh (ICDDR,B) in areas at Bakerganj (22.5°N 90.2°E), Chhattak (24.9°N 91.6°E), Chaugachha (23.2°N 89.0°E), and Matlab (Fig. 4.1). The locations were selected to represent different geographical areas of Bangladesh. Bakerganj is located at the upper edge of the estuarine area of the southern coastal region of Bangladesh. Chhattak is situated in the flood plains of the river Brahmaputra. Chaugachha is on the edge of the tidal plains of the river Ganges. These three rural areas have a government-run health facility where diarrheal diseases in the area are treated and they serve as the catchment site for cholera cases. In Matlab, there is a diarrhea hospital affiliated with ICDDR, B. Each area has a catchment of 140,000 - 200,000 persons.

History of cholera in each area is as follows: Matlab is a highly populated riverine area known to be highly cholera-endemic (Longini *et al.*, 2002). Chhattak and Bakerganj have had a history of regular cholera outbreaks; Chaugachha, on the other hand, has had no history of cholera outbreaks during the past 10 years and was designated as the control area. During the surveillance of this study, however, there were intermittent cholera cases in Chaugachha, although no major outbreak was reported.

C Survey for isolation of *V. cholerae*

During the period from July 1997 to 1999, stools of twenty percent of all diarrhea patients administered in the clinics at the four sites of this study were screened at 15-day intervals for *V. cholerae*. For the environmental survey, four stations (in rivers, ponds, and lakes) at each surveillance catchment area were chosen to determine physical, microbiological and plankton data for surface-water samples. Clinical and environmental surveillance every 15 days was begun in June 1997 through December 1999.

In parallel with the clinical work, screening for *V. cholerae* in surface waters of the ponds and rivers used by the patients in each area for various purposes, including drinking water was done. Enrichment plating methods, employing alkaline peptone water (APW), and thiosulfate-citrate-bile salts-sucrose agar (TCBS), as well as taurocholate-tellurite-gelatin agar (TTGA) were used for isolation of *V. cholerae*. Following standard reference protocols (Tison, 1999), we employed a set of biochemical and serological (co-agglutination) tests were employed to confirm *V. cholerae* colonies to species and serogroup. Details of the screening procedures have been described elsewhere (Kay *et al.*, 1994; Tison, 1999; West & Colwell, 1983).

4.2.2 Genomic fingerprinting

Two versions of genomic fingerprinting methods were used to characterize two different aspects of clonal relationships. The long-range high fidelity ERIC-BOX-PCR was used to identity a strain with phylogenetic inference to catalogued environmental and clinical strains. This procedure is described in Chapter 3. The second method was used to detect and resolve faint bands by the very sensitive, short-range ERIC-PCR. Because this technique employed primers of low purity, containing premature-truncated oligonucleotides, whole PCR amplicon preparation had to be done and was used in a single batch to make comparisons among the PCR products. In spite of this limitation, the method provided enough resolution of random genomic sequence sampling to reveal genetic interaction between strains. The latter technique is described below.

ERIC-PCR fingerprinting, as described by Rivera *et al.* (1995), was used, with modification to enhance resolution as follows. The temperature at the elongation step of each PCR cycle was lowered from 70°C to 65 °C. Separation of PCR products was achieved by electrophoresis in 3.6% Metaphore agarose gel (FMC, Rockland, Maine) at 10 V/cm. Gels were stained for 30 min in SYBR Green I solution at the concentration recommended by the manufacturer (Molecular Probes, Eugene, Oreg.). A digital image of each stained gel was prepared using the scanning instrument, FluorImager (Molecular Dynamics, Sunnyvale, Calif.). Unwrapping of lanes, band calling, and normalization of band position among the gels and lanes were accomplished with the aid of the software IMAGE (Sanger Center, UK). Each band location was binary coded, according to absence and presence.

To evaluate the effectiveness of ERIC-PCR band patterns in representing profiles of the whole genome, we examined the distribution of possible primer binding sites and PCR products along two chromosomal DNA sequences of *V. cholerae* O1 El Tor N16961 (Heidelberg *et al.*, 2000). Ca. 1800 chromosomal loci of 24-bp size nucleotide sequences were identified as having more than 7-base matches with either one of the two ERIC primers by BLASTN 2.1 (Altschul *et al.*, 1997). The melting temperature (T_m) of the primer binding for each locus was calculated by MeltCalc 2.0 (Schütz & von Ahsen, 1999), and the T_m value was used as the index in comparing the relative strength of the primer binding. In simulating PCR product formation, it was assumed that the relative likelihood of product formation depends on the weaker primer-binding affinity (lower T_m) of the two primers for a PCR product. To address the even distribution of the simulated

PCR products along the two circular chromosomes of *V. cholerae*, Rao's spacing test for uniformity in circular space was used (Jammalamadaka & SenGupta, 2001).

4.2.3 Population genetic analyses

To examine the differential distribution of haplotypes between clinical and environmental isolates from the three areas, we separated isolates into six populations by three areas and two habitat types. The phylogenetic relationship among genotypes was calculated as the Euclidean distance ($d = \sqrt{1 - S_{SM}}$, where S_{SM} is the simple match coefficient) from the binary coded data of ERIC-PCR band patterns (Sneath & Sokal, 1973). From the pairwise distance matrix, we constructed unrooted trees, using the neighbor joining (NJ) procedure in the PHYLIP package(Felsenstein, 2004a). To evaluate the resulting tree-topology, we adopted two techniques. For the first, the bootstrap technique (Felsenstein, 1985) was performed using the SEQBOOT and CONSENSE procedures in the package. We created 1000 replaced data sets from our binary coded band data set and consensus branching was enumerated, comparing resulting trees. An alternative to bootstrapping, consensus of tree topology by different tree construction methods, was used to examine the significance of a cluster. We built trees from the matrix of pairwise genetic distances by three different methods, NJ, unweighted pair group method with arithmetic averages (UPGMA), and maximum parsimony (MP) by the PARS procedure and examined resulting trees to find tree branches appearing consistently in the trees of the three clustering methods. This technique is based on the experimental observation by Kim (1993) on a simulation data set. Hilali et al (Hilali *et al.*, 2000) found that it successfully identifies major clusters of randomly amplified polymorphic DNA (RAPD) patterns that are strongly correlated with other genotypic characteristics of *Escherichia coli* strains. In this study, this technique, as well as the bootstrapping technique, was used to identify significant clusters. When a branching in the NJ tree had bootstrap support > 0.5 or occurred consistently in all of three trees, we identified it as a significant branch. If a significant branch was nested by another significant branch, we determined the collection of significant branches to be a significant cluster.

Examination of the population genetic structure by analysis of molecular variance (AMOVA) (Excoffier *et al.*, 1992) and exact test was carried out employing procedures in the software ARLEQUIN (Schneider *et al.*, 2000). The Euclidean distance matrix was used to estimate variances attributable to differences of geographic area and habitat (i.e., environmental versus clinical).

4.2.4 Enumeration of culturable *V. cholerae*

In parallel with the clinical survey, *V. cholerae* in samples of surface waters were enumerated using colony-lift Southern hybridization of gamma-P^{32} labeled oligonucleotide (pITS; Appendix A) probes, i.e. an intergenic nucleotide sequence specific for *V. cholerae*. The enumeration data of toxigenic *V. cholerae* by the oligonucleotide probe for cholera toxin gene (CTAP) was obtained from G. Morris, University of Maryland, Baltimore. Additional microbial variables such as heterotrophic plate counts (HPC) and fecal coliforms (FC) were obtained from collaborators at ICDDR, B. We also measured environmental variables, such as dissolved oxygen, water temperature, and total dissolved solids using standard methods.

Table 4.1. Location and sample type of environmental *V. cholerae* O1 strains isolated in this study

Group designation[a]	Strains designations	Site of isolation	Sample fraction[b]
BKIE9911	BKI6	River BK1	Plankton
	BKI7	Lake BK4	Plankton
	BKI12	Lake BK4	Plankton
	BKI8	Pond BK6	Plankton
	BKI13	Pond BK6	Plankton
	BKI9	Lake BK7	Plankton
	BKI14	Lake BK7	Plankton
CGOE9906	CGO15	Pond CG2	Plankton
	CGO19	Pond CG2	Plankton
	CGO16	River CG3	Plankton
	CGO14	Lake CG5	Hyacinth
	CGO17	Lake CG5	Plankton
	CGO18	Lake CG5	Plankton
CTIE9911	CTI16	Pond CT1	Plankton
	CTI17	Pond CT4	Plankton
	CTI18	Pond CT4	Plankton
	CTI19	Pond CT5	Plankton
CTOE9909	CTO17	Pond CT1	Water
	CTO18	River CT2	Water
	CTO19	Pond CT5	Hyacinth

a: Strains in the group BKIE9911 are Inaba serotype strains, isolated from surface waters in Bakergonj on Nov. 10, 1999. The group CGOE9906 includes Ogawa serotype isolates from Chaughacha on Jun. 2, 1999. CTOE9909 and CTIE9911 designate strains Ogawa strains and Inaba strains from Chhattak, respectively, Ogawa strains were isolated on Sep. 2, 1999 and Inaba strains isolated on Nov. 5, 1999.

b: Water samples were fractionated using plankton nets (nominal cut-off size: 64 µm). The fraction retained on the net was labeled as plankton, and the filtrate was labeled as water fraction. Hyacinth samples were biomass of water hyacinth rinsed with sterile water and homogenized by glass grinding.

Table 4.2. Location and date of isolation for clinical *V. cholerae* O1 strains used in this study

Area of	Serotype	Date of isolation	Group	Strain designation
Bakergonj	Inaba	Dec. 8, 1997		BKI2
		Jan. 22, 1998		BKI3
		Mar. 19, 1998		BKI4
		May 18, 1998		BKI5
Chaughacha	Ogawa	Jan. 7, 1998		CGO13
		Feb. 6, 1998		CGO12
		Mar. 8, 1998		CGO11
		Aug. 25, 1998		CGO9
		Sep. 24, 1998		CGO8
		Oct. 24, 1998		CGO7
		Nov. 23, 1998		CGO6
		Dec. 8, 1998		CGO5
		Mar. 8, 1999	CGOC9903	CGO2, CGO3, CGO4
Chhattak	Inaba	Nov. 17, 1997	CTIC9711	CTI9, CTI11, CTI12, CTI13, CTI14, CTI15
		Dec. 18, 1997	CTIC9712	CTI4, CTI5, CTI6, CTI7, CTI8
		Dec. 31, 1997		CTI3
		Nov. 12, 1998		CTI2
	Ogawa	Jul. 1, 1997		CTO2
		Aug. 4, 1997		CTO3
		Nov. 17, 1997		CTO4
		Dec. 18, 1997		CTO5
		Dec. 31, 1997		CTO6
		Jan. 28, 1998		CTO7
		Mar. 28, 1998		CTO8
		Sep. 13, 1998		CTO9
		Sep. 23, 1998		CTO10
		Oct. 28, 1998	CTOC9810	CTO12, CTO16
		Nov. 12, 1998		CTO13
		Nov. 29, 1998		CTO14
		Dec. 14, 1998		CTO15

4.3 Results and discussion

4.3.1 Comparison of population structure between clinical and environmental *V. cholerae* O1 isolates

A Characteristics of isolates

The four sites yielded toxigenic *V. cholerae* strains from diverse environmental samples (Table 4.1), and the clinical surveys yielded *V. cholerae* O1 isolates throughout at least one full season in a given year. All environmental *V. cholerae* O1 isolates were El Tor biotype and belonged to the Inaba or Ogawa serovar. To compare the environmental and clinical isolates, clinical isolates were selected by matching biotype and serotype (Table 4.2). For the Inaba serotype populations from Chhattak, Bangladesh, all clinical isolates at four times of isolation were selected. For other areas, one isolate was randomly selected at times when matching serotypes were present.

Table 4.3. Percentiles of distribution of number of mismatches among all possible pairs of haplotypes produced by ERIC-PCR band patterns of *V. cholerae* O1 strains

Pairing of strains by source	10th	25th	50th	75th	90th
Clinical vs. clinical	14	17	21	24	28
Environmental vs. clinical	16	19	22	25	29
Environmental vs. environmental	17	20.5	24	27	30
All strains	15	18	22	25	29

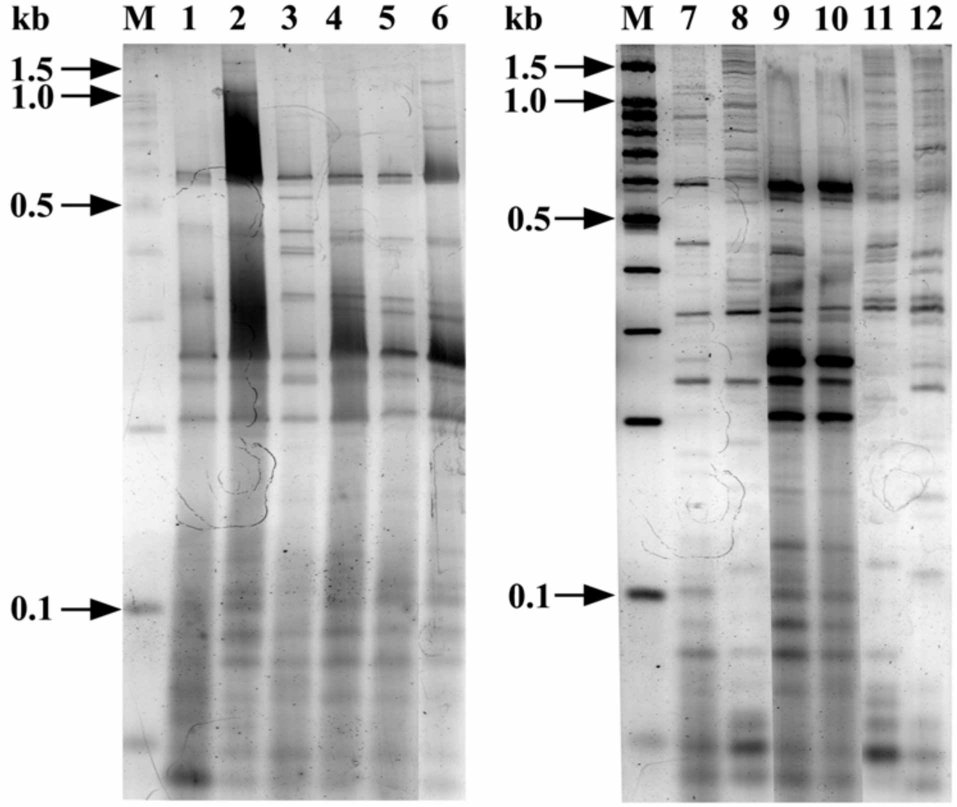


Figure 4.2. Representative ERIC-PCR band patterns of *V. cholerae* O1 from Bangladesh. Lanes: M, 100-bp ladder; lane 1 to 6, Ogawa strains; Lanes 7 – 12, Inaba strains.

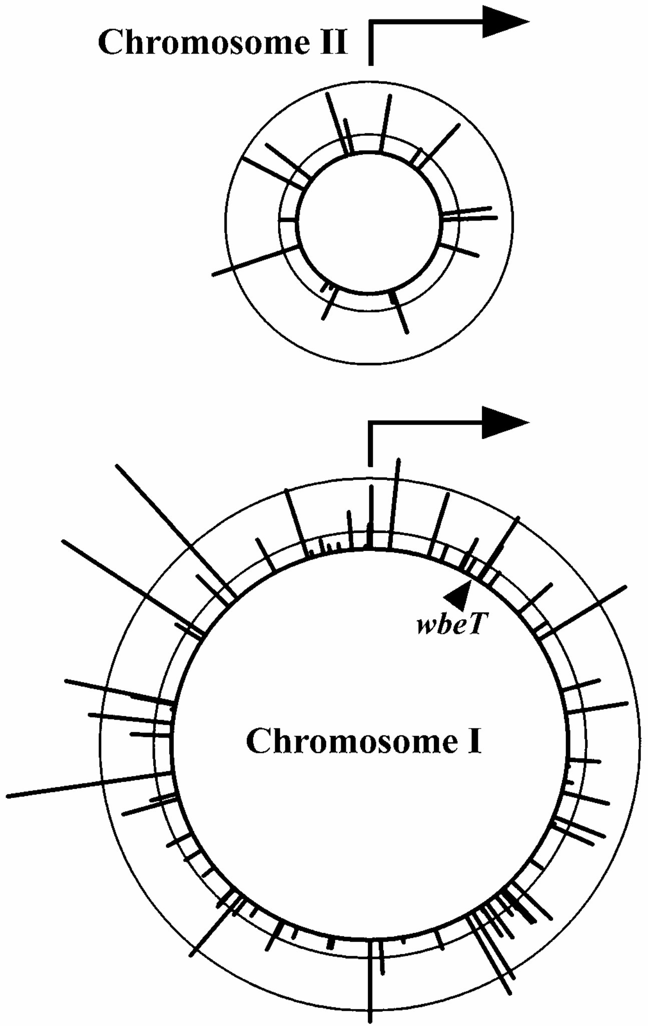


Figure 4.3. Distribution of ERIC-PCR products along the two chromosomes of *V. cholerae* N16961. GenBank accession numbers for chromosome I and II are AE003852 and AE003853, respectively. The middle of PCR products with a size of 100 bp to 588 bp are shown as spines. Arrows denote origins of replication and the direction of sequence numbering. The lengths of spines represent T_m values calculated by MeltCalc and used as an index for relative binding affinity. Three reference lines are presented as circles: the innermost circle for 0°C, the middle for 3°C and the outermost for 12°C. The triangle indicates the location of *wbeT* gene, the Ogawa antigen determinant.

B ERIC-PCR patterns

ERIC-PCR produced more than 40 bands for each isolate (Figure 4.2). From the bands, those ranging from 100 bp to 588 bp were selected for genomic profiling. The size range selected avoided the occasional smearing of strong intensity bands and gave consistent resolution, from which duplicated PCR and electrophoresis produced 99% band matches.

When formation of PCR products of 100 - 588 bp was simulated from primer binding sites along the two chromosomal DNA sequences of *V. cholerae* O1 El Tor N16961 (Heidelberg *et al.*, 2000), 105 fragments with T_m values (i.e., index of binding affinity) higher than 0°C were generated (Figure 4.3). The maximum T_m of predicted PCR products was 30.4°C and most of the products had a T_m below 12°C. Considering the temperature (65°C) used in the elongation step of the ERIC-PCR, occurrence of primer binding sites in the genome of *V. cholerae* to produce bands of the 100-588 bp was random, rather than specifically organized. When tested, the hypothesis of even distribution of primer binding sites and PCR product loci on the two chromosomes of *V. cholerae* N16961, the hypothesis was not rejected (Rao's spacing test for uniformity; $P > 0.15$ for fragments with $T_m > 12$°C). Therefore, fingerprints of 100 – 588 bp range were consistent with the random occurrence of primer binding sites evenly distributed throughout the entire region of the two chromosomes. An additional aspect to be noted from the distribution of primer binding sites was that the affinity of the primer binding was significantly reduced, e.g., $T_m < 0$°C, by a single point mutation in most of the loci, especially when occurring at the 3' end of the primer. Therefore, the sensitivity of PCR product formation can be considered to be at the level of a single point mutation.

One of the goals in the simulation of PCR by calculating ERIC primer binding affinity was to detect differences in the chromosomal regions of genes coding the Inaba or Ogawa serovar determinant. As shown in Fig. 2, we did not find primer binding sites that produced the fragments in the vicinity of the *wbeT* gene (Stroeher *et al.*, 1998). The closest was a locus 22 kb upstream and 7 kb downstream of the structural and regulatory nucleotide sequences of the *wbeT* gene. Thus, ERIC-PCR could not discriminate strains by Inaba vs. Ogawa serovar determinants.

From the genomic fingerprints produced by ERIC-PCR for the 63 *V. cholerae* O1 isolates, 71 band loci producing 57 haplotypes were identified. When pairwise mismatches among the isolates were counted, a unimodal pattern was observed among the clinical isolates and in the clinical-environmental comparison; however, a unimodal distribution of mismatches was not characteristic of the environmental isolates (Table 4.3). Although the modal number of mismatches was 25, another mismatch peak of 17 was observed. Unlike the mode at 25 mismatches, the secondary peak at 17 mismatches was mostly (75% versus 50%) from pairs of environmental strains isolated from different geographical areas.

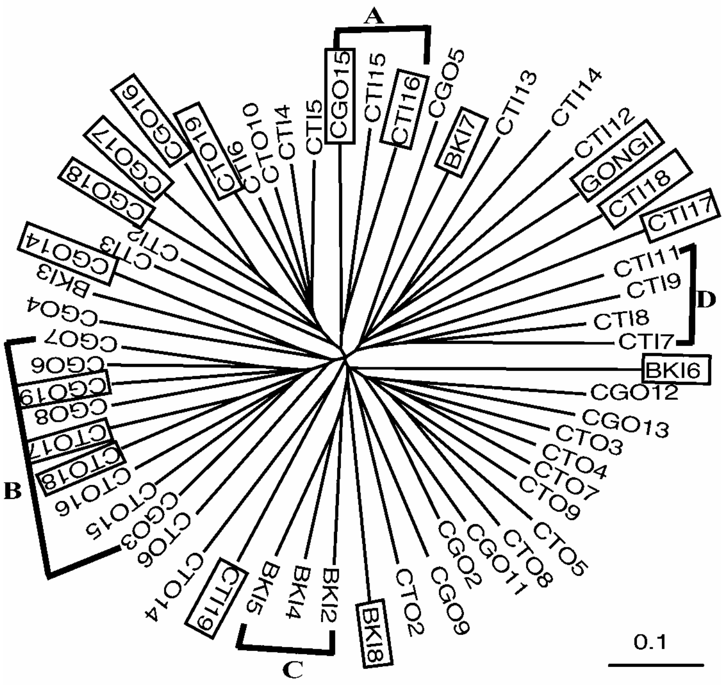


Figure 4.4. Unrooted neighbor-joining tree based on Euclidean distances (*d*) among ERIC-PCR band patterns of *V. cholerae* O1 isolates. Isolates from environmental sources are enclosed in rectangles. Labeling of the strains follows that of Table 4.1 and Table 4.2, except for GONGI, a *V. cholerae* O1 El Tor Inaba strain isolated from coastal water off Chittagong.

C Cluster analyses

The phylogenetic relationship of the haplotypes is shown in Figure 4.4. Clustering of isolates into different lineages occurred and the bootstrapping support of the clusters resulted in one distinct cluster (Cluster D in Figure 4.4) with > 0.99 support value. Although most of the lineages in Figure 4.4 were not highly supported by bootstrapping, the appearance of other clusters was consistent with that of other tree-building methods, such as UPGMA and MP. While Cluster B, C, and E showed clustering of isolates of the same serotype from the same habitat of a geographic area, haplotypes in Cluster A and D were intermingled, regardless of source of isolation or whether Inaba versus Ogawa serotype.

D Pairwise comparison of clinical and environmental populations

Haplotypic (electrophoretic type of the haploid organisms) compositions of the six populations, differentiated by area and habitat of isolation, were compared by the exact test. Pairwise comparisons between the populations from the two habitats for each area did not show significant difference (exact test; $P = 0.06$ in Bakerganj, $P= 0.51$ in Chhattak and $P= 0.98$ in Chaugachha). Compositional differences among the clinical populations from the three areas were not significant also (exact test; $P > 0.35$). However, the haplotype frequency of the environmental population in Bakerganj was significantly different from that of the environmental populations from the other two areas (exact test; $P < 0.05$).

Table 4.4. Estimates from an hierarchical analysis of molecular variance (AMOVA) of all populations of *V. cholerae* O1, based on different area of isolation (Bakergonj, Chhattak, or Chaughacha) and habitat (clinical versus environmental) of isolation

Structure[a]	Variance component	Observed partition		*P*	*F*
		Variance	% total		
I: Habitats under areas	Among areas	0.27	2	0.08	F_{CT} = 0.024
	Among habitats in an area	1.02	9	<0.01	F_{SC} = 0.093
	Within populations	9.96	89	<0.01	F_{ST} = 0.114
II: Time under habitats in Chhattak area	Among habitats	0.08	1	0.53	F_{CT} = 0.007
	Among temporally separated subpopulations	1.20	10	<0.01	F_{SC} = 0.130
	Within subpopulations isolated on the same day	10.53	89	<0.01	F_{ST} = 0.108
III: Habitat-Time Complex in Chaughacha	Among populations separated by habitat type and time (three months)	0.95	9	0.12	F_{ST} = 0.092
	Within subpopulations isolated on the same day	9.36	91		

a: Structure I assumed *V. cholerae* populations are separated among different areas, and the clinical population is separated from the environmental population in each area. In Structure II, the three clinical *V. cholerae* populations from Chhattak (CTIC9711, CTIC9712 and CTOC9810) were compared with the two environmental populations from the same area (CTOE9909 and CTIE9911). Structure III was AMOVA application to CGOC9903 and CGOE9906, which were separated by the difference of habitats and three months in the time of isolation from Chaugachha.

E Analyses of molecular variance

As shown in Table 4.4, the total variance of the PCR bands was decomposed by difference in geographic area and then nested by the two types of habitats within each geographic area (Structure I). Although isolates from the Chhattak area could be divided into Inaba and Ogawa serogroups, serotype differences were disregarded because the exact test on haplotype distribution between serotypes did not show a significant difference (P = 0.22). The observation that isolates of different serotypes intermingled in clustering (Figure 4.4) also justified the merging of the Chhattak populations.

Quantitatively, the population of *V. cholerae* O1 in this study had an F_{ST} value of 1.1 when divided into the six area-habitat subpopulations (Table 4.4). According to the qualitative classification of Wright (1978), an F_{ST} value within the range of 0.05 to 0.15 indicates moderate differentiation, within 0.15 to 0.25 great differentiation, and greater than 0.25 indicates very great differentiation. Therefore, the level of divergence among *V. cholerae* O1 populations distributed in the three areas in Bangladesh over the 2.5 year time period was a typical moderate differentiation. Approximately 90% of the variance was explained by isolate to isolate variance within a population. Habitat difference accounted for the rest of the variance and the estimate for the effect of geographic area was not significant. However, we recognized that comparison of subpopulations by habitat type involves confounding its effect with that of temporal variation in subpopulations, because the environmental strains and clinical strains were not isolated at the same time. Therefore, genetic divergence among the six subpopulations in the

AMOVA result (Structure I, Table 4.4) was regarded to be the effect of the habitat-time complex.

Because of this habitat-time confounding, it was not possible to examine the effect of habitat and time separately. Nevertheless, we estimated the relative magnitude of habitat effect to temporal change in populations by examining the population structure of each geographical area. In Structure II of Table 4.4, we analyzed the genetic variance in a part of the Chhattak populations, finding three clinical populations separated by one month to one year and two environmental populations separated by two months. The level of fixation within the subpopulations (F_{ST} in Structure II) was comparable to that of whole populations (F_{ST} in Structure I), indicating the subpopulations used in Structure II represents well the whole population in Structure I. The hierarchs of the populations in Structure II comprising the two factors: (1) an habitat-time complex in which difference of habitat types and > 11 month difference in time of isolation were confounded, and (2) a pure time factor, which is nested under the habitat-time complex, and consisted of one to eleven months difference in the time of isolation among clinical strains and two month difference in the time of isolation among environmental strains. The AMOVA result shows that the second factor alone can explain about 10% of the total variance and the habitat-time complex at the top level of the hierarchy is not significant ($P = 0.53$) as a factor causing differentiation among the subpopulations. In the case of the Chaugachha strains, the subpopulation of environmental strains (CGOE9906) were compared to a clinical subpopulation (CGOC9906) isolated with a finite time difference of three months. Analysis in Structure III (Table 4.4) indicates the two subpopulations were not significantly diverged.

F Comparison with previous studies

The species level population structure of *V. cholerae* and relatedness of toxigenic *V. cholerae* strains to nontoxigenic *V. cholerae* have been studied extensively by multilocus enzyme electrophoresis (MLEE) (Beltran *et al.*, 1999; Farfán *et al.*, 2000) and by DNA sequence analyses of the single housekeeping gene, *recA* gene (Stine *et al.*, 2000) and multiple genes (Byun *et al.*, 1999). Since there is little variation in the DNA sequences of *mdh, hlyA, dnaE,* and *recA* genes among the sixth-pandemic, seventh-pandemic, and U.S. Gulf Coast clones, it has been concluded that toxigenic *V. cholerae* O1 strains are very closely related (Byun *et al.*, 1999). In a study carried out by Beltran *et al.* (1999), a total of 244 strains of *V. cholerae* (17% of which were environmental) isolated from Mexico and Guatemala during 1991 and 1995 were analyzed, along with 143 serogroup reference strains. Most of the electrophoretic types comprised non-O1/non-O139 strains, while the O1 and O139 strains formed a tight cluster. The seven O139 strains examined by Beltran *et al.* (1999) indicated high clonality by forming a common phylogenic branch as a part of several O1 clusters. Farfán *et al.* (2000) analyzed a collection of ca. 100 *V. cholerae* strains isolated from environmental sources in different geographic locations, e.g., Brazil, U.S.A. and Bangladesh, and from clinical cases in Latin American countries and the Indian subcontinent during an unspecified period of time. Their results indicated the entire *V. cholerae* population was more diverse than the set analyzed by Beltran *et al.* (1999), as were the O1 and O139 subpopulations. Stine *et al.* (2000) reported nucleotide sequences for *recA* from eight clinical *V. cholerae* O1 El Tor strains isolated from three continents since 1937. These also formed a tight cluster that differed from other *V. cholerae* populations by at least 8 nucleotides.

Therefore, it was calculated that toxigenic O1/O139 El Tor strains form a phylogenic lineage distinguished from other populations of the species *V. cholerae.*

Besides clonality of the O1 El Tor strains, results of previous studies also revealed significant diversity within the lineage of toxigenic O1 El Tor strains. The data of Stine *et al.* (2000) revealed up to four-nucleotide differences in the DNA sequences of *recA* (705 bp in length) among eight O1 El Tor strains. Since the *recA* gene must be highly conserved to serve as an housekeeping gene for >3800 open reading frames of the *V. cholerae* genome, this difference was expected to be a lower limit of genomic nucleotide variance among toxigenic O1 El Tor strains. Concordant with this expectation, seventeen housekeeping genes of the *V. cholerae* O1 El Tor population examined by Beltran *et al.* (1999) revealed genetic distances of up to about 10% of the maximum genetic distance calculated from the total *V. cholerae* population. The MLEE data of Farfán *et al.* (2000) also demonstrated significant diversity among O1 El Tor strains, with a mean genetic diversity per locus (*H*) as high as 0.4, whereas the total *V. cholerae* population yielded an *H* value of 0.5. The lack of congruence between phylogenetic relationships of pathogenic clones deduced from the four housekeeping genes studied by Byun *et al.* (1999) and another housekeeping gene, *asd*, (Karaolis *et al.*, 1995) offers molecular proof for potentially significant microevolution among *V. cholerae* O1 strains, especially from high gene recombination. Recently, a powerful microarray assay of a set of seventh pandemic strains revealed the clonality within *V. cholerae* El Tor strains to be indeed mediated by a set of clone-specific genetic elements (Dziejman *et al.*, 2002). Lan and Reeves (2002) reported extensive diversity and clonality within the seventh pandemic *V. cholerae* O1 El

Tor cluster, strong enough to be related to temporal and geographical divergence of the clones in the cluster.

In this study, diversity among populations of *V. cholerae* O1 El Tor isolates was analyzed within a specifically defined temporal and geographical scope. Although the scale was much narrower than those of previous studies, significant diversity among the isolates was observed (Figure 4.4). When significance of clonality among the diverse *V. cholerae* O1 El Tor electrophoretic types was tested, three clusters (Cluster C, D and E in Figure 4.4) were significant by the bootstrapping procedure(Felsenstein, 2004a). Consensus tree topology using different tree construction methods to examine the significance of a cluster (Hilali *et al.*, 2000; Kim, 1993) is an alternative to bootstrapping. The matrix of pairwise genetic distances examined by three different methods (UPGMA, NJ and MP) revealed two additional clusters (Cluster A and B in Figure 4.4), suggesting additional clonality among the *V. cholerae* O1 El Tor isolates.

Within the significant clusters, an intermingling of Inaba strains with Ogawa strains (Cluster A) was observed. The Inaba serotype can be generated by a recessive mutation of the *wbeT* gene of the Ogawa serotype (Sack & Miller, 1969; Stroeher *et al.*, 1998). Therefore, occurrence of two different serovars in an identical phylogenic lineage is reasonable. In Clusters A and D, intermingling of strains from two different geographical areas, Chhattak and Chaugachha, was observed, implying weak spatial separation between the two areas. In addition, environmental isolates and clinical isolates from Chhattak and Chaugachha were found to belong to the same lineage (Clusters A and D). Clearly, toxigenic *V. cholerae* O1 El Tor strains in cholera-endemic areas of Bangladesh form several clonal lineages, of which at least five of the lineages observed in this study were significant. Furthermore, environmental and clinical strains were in the same lineage. In addition, the unimodal distribution of mismatches among the clinical isolates and clinical-to-environmental isolates (Table 4.3) that was observed implies that panmictic gene exchange between lineages can weaken the clonality (Whittam, 1995).

G Effect of geographic isolation

Application of molecular epidemiological methods revealed that toxigenic *V. cholerae* O1 strains from the seventh pandemic or from localized outbreaks could be differentiated by geographical location (Lan & Reeves, 2002; Wachsmuth *et al.*, 1994). Ribotyping, plasmid carriage, and phage typing data suggested continent scale differentiation of *V. cholerae* strains. According to results of a study by Jiang *et al.* (2000), in which genetic profiles of environmental and clinical isolates were compared by amplified fragment length polymorphism (AFLP), clinical *V. cholerae* O1 populations in Latin America were found to comprise pandemic strains as well as locally evolved strains, leading to the conclusion that geographical isolation is a factor in the establishment of the lineages of *V. cholerae*. However, the magnitude of such differentiation on a regional scale as a factor in population isolation has not been pursued.

In this study, we obtained genomic profiles of *V. cholerae* O1 El Tor populations from two distinct habitats in three separate geographic areas, allowing examination of the genetic structure of toxigenic *V. cholerae* populations differentiated by spatial isolation on a regional scale, as well as by habitat. Of the population structures listed in Table 4.4, differentiation of *V. cholerae* O1 populations by geographic isolation and by habitat was calculated to be valid (Structure I in Table 4.4). Although the estimate of the significance of contribution of the difference in geographical areas was marginal (*P*= 0.08), heterogeneity among the toxigenic *V. cholerae* populations across geographical areas was supported by other aspects of the data. As assessed by the exact test of distribution of haplotypes, composition of the environmental populations was significantly different, while in the clinical populations they were not. The lack of clarity of unimodal mismatch distribution (Table 4.3) among environmental isolates also revealed heterogeneity in the distribution of genetic exchange among environmental strains. Therefore, it is concluded that environmental populations of toxigenic *V. cholerae* O1 in Bangladesh are separated by geographical isolation, while composition of clinical populations is relatively homogeneous.

H Effect of habitat and time

As shown in Table 4.1, environmental populations of *V. cholerae* O1 in each area were isolated on only four occasions throughout the survey period, and the time of isolation of the environmental strains was when a cholera outbreak had not occurred, according to the results of the clinical survey, imposing a temporal uncoupling of more than one month in time. The nature of the survey not only placed a limit on the resolution of analysis of the effect of geographic area or habitat difference on the total variation of population composition, but also caused a confounding of the effect of habitat difference and temporal dynamics in the composition of the clinical and environmental *V. cholerae* populations. The effect of the habitat-time complex explained about 9% of the total variance, as shown Structure I in Table 4.4, and this indicates that the effect of habitat alone was less than 9%.

Clinical and environmental populations from Chhattak provided an opportunity to estimate the relative magnitude of the effect of habitat in the molecular variance analyses

(Table 4.4). The difference of time in a scale of up to eleven months, had significant variance (10% of the variance in the five subpopulations from Chhattak) in the clinical populations. It was concluded that the effect of the habitat-time complex observed for the entire population (Structure I) was mostly the effect of time difference between the two habitat sample collections and that the habitat difference did not impose significant difference in the genetic structure of the populations from the two habitats.

The question of whether clinical and environmental *V. cholerae* O1 in Bangladesh are identical was carefully examined by Huq *et al.* (1993), who were the first to use molecular methods for the purpose. In that study, pairwise comparisons of two isolates, each from an individual patient and the drinking water source of that patient, revealed identical profiles for toxigenicity and for bacterial cell proteins. In the report of Faruque *et al.* (1995), twelve *V. cholerae* O1 El Tor strains, isolated from surface water of Dhaka, Bangladesh from 1991 to1994, shared three identical ribotypes with eight clinical strains isolated in 1990 and 1991. However, an additional ribotype was found in six clinical strains isolated during 1990 to1992 and did not match any of the environmental strains.

In the Luanda province of Angola, Colombo *et al.* (1997) also compared genotypes of 16 clinical strains and four strains isolated from a reservoir during 1991-1994 and did not find any differences in ribotype or ERIC-PCR patterns. Recently, Singh *et al.* (2001) compared *V. cholerae* strains isolated from clinical and environmental sources in Varanasi, India, during 1992-1994 and found a close relationship commonly in the ERIC-PCR, BOX-PCR, and AFLP banding patterns. Although the number of environmental-clinical strain pairs was either small or loosely defined, these results agree with the finding reported here of no difference in the composition of toxigenic *V.*

cholerae populations from either from environmental or clinical sources. It should be noted that a more sensitive genomic profiling method and a statistically informative number of strains were employed in this study. Thus, our results support the conclusion that toxigenic *V. cholerae* strains isolated from the aquatic environment in cholera endemic areas are, indeed, the same strains responsible for the clinical cholera cases in those geographical areas.

It should also be noted, however, that the existence of a time effect implies that changes in the population composition of *V. cholerae* O1 in a cholera-endemic area occur over time. In the recent AFLP analyses by Lan and Reeves (2002), clonal shifts in seventh pandemic strains over a time scale of a decade was demonstrated. In the study reported here, the time scale for measurable change was one month or less among the clinical cholera cases, a time scale supported by our finding of a significant fixation index among the subpopulation of clinical isolates (Structure II in Table 3), as well as by the dramatic change in relative prevalence of O1 and O139 serogroups in cholera cases in the study areas within a time scale of 15 days (data not shown). Rapid microevolution or, more likely, a selection of clones in the environment, corresponding to fluctuations in the environment or to immunity in the host populations, may cause dynamic change. Since we did not find any significant difference in genomic profiles of *V. cholerae* O1 El Tor strains from either environmental or clinical sources, we conclude seasonal changes in the aquatic environment cause temporal change by placing selective pressure on environmental clones of *V. cholerae*.

Examples of microevolution have been reported for *V. cholerae.* Horizontal gene transfer of the O antigen genes has been shown to occur in the emergence of the O139 and O37 serogroups (Sozhamannan *et al.*, 1999; Stroeher *et al.*, 1998), and variability in the *tcpA* gene is believed to have been caused by homologous recombination (Karaolis *et al.*, 2001). The unimodal distribution of mismatches observed in this study and in an earlier study by Farfán *et al.* (2000) suggests that panmictic recombination drives rapid microevolution among *V. cholerae* strains and that gene flow is not restricted or bottlenecked between environmental and clinical sources. In summary, environmental *V. cholerae* O1 El Tor populations demonstrate significant geographical isolation, but barriers between the clinical and aquatic environments are not significant. In addition to spatial variance, temporal variance is a significant factor, explaining total genomic variances among toxigenic *V. cholerae* populations.

We conclude that the composition of environmental populations of toxigenic *V. cholerae* is identical to that of *V. cholerae* populations causing endemic cholera. Although the composition of *V. cholerae* O1 El Tor populations causing endemic cholera in Bangladesh undergoes dynamic change, *V. cholerae* populations in the two distinctive habitats achieve a dynamic equilibrium. Rapid transfer between habitats or panmictic gene flow by active intermingling of clinical and environmental *V. cholerae* strains is suggested to be the mechanism of the dynamic equilibrium between the two distinctive habitats. Finally, we also conclude that the aquatic environment is the cholera reservoir and is associated with shifts in the dynamics of the disease by causing spatial and temporal fluctuations in the composition of toxigenic *V. cholerae* inhabitants. Any change in the composition of *V. cholerae* populations in the aquatic environment which may be driven by seasonal fluctuation in the environment or by introduction of new strains through microevolution or by being imported from other systems, can cause coupled changes in composition and behavior of the clinical populations, leading to a shift in the dynamics of cholera epidemics. Since most cholera endemic areas are in developing countries, where there is poor sanitary control of drinking water, our findings for Bangladesh are applicable to other cholera endemic areas of the world.

4.3.2 Dynamics of *V. cholerae* and endemic cholera

A Persistence and seasonality in the dynamics of *V. cholerae*

Most samples (>97%) contained *V. cholerae,* i.e., colonies formed on the plating medium employed in the study, ranging from 10^2 - 10^5 cells ml^{-1} (Figure 4.5). Seasonal fluctuation was significant, with peaks observed in the late spring and fall. In autocorrelation analysis, the colony count done using pITS probe (ITS count) showed significant ($P < 0.01$) seasonality that is approximately 5 to 6-month intervals and varying by site. There was also a significant correlation of *ctx*-positive colony counts (CTX counts) with heterotrophic plate counts (Table 4.5).

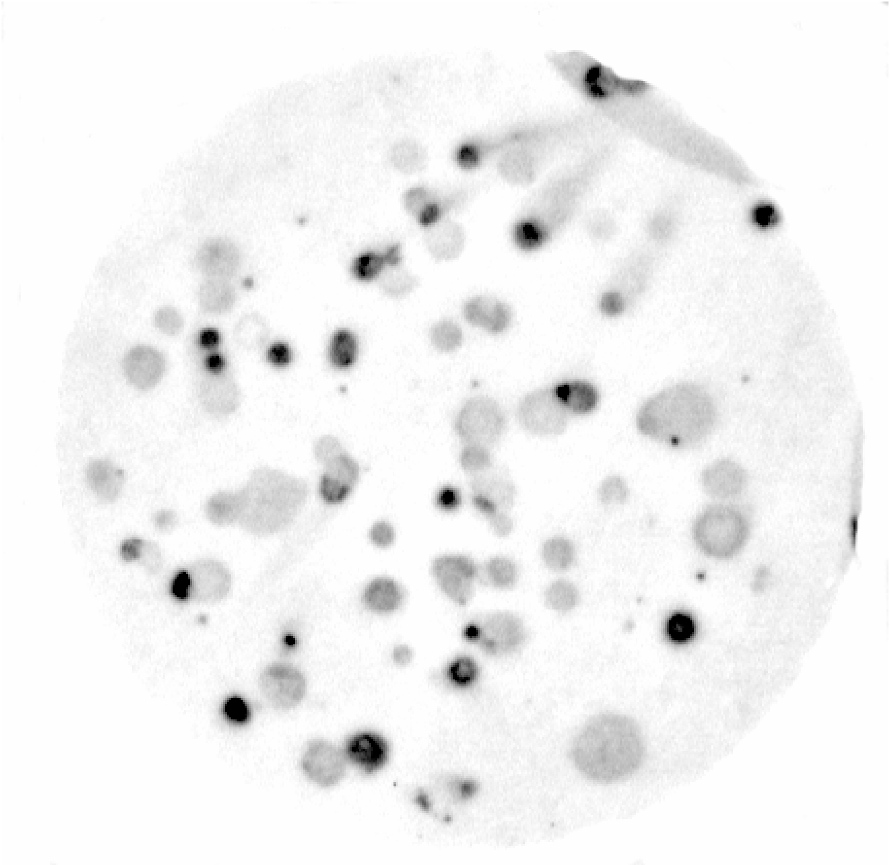


Figure 4.5. Example of colony blot hybridized with pITS oligonucleotide probe. The strong intensity of positive binding can be readily visualized when compared with weakly binding negative colonies. Colonies were grown on LB agar plates.

Table 4.5. correlation coefficient between natural log transformed colony counts and environmental variables significantly correlated with cholera cases at $P< 0.05^{a}$

	WDEPTH	CONDUCT	LITS	LFC
CONDUCT	0.18			
LCTX		0.22		
LFC		0.17	0.18	
LHPC	0.23	0.14	0.62	0.27

a: WDEPTH= water depth, COND= conductivity, LCTX= natural log of CTX count, LITS= natural log of ITS count, LFC= natural log of fecal coliform (FC) count, LHPC= natural log of heterotrophic plate (HPC) count, respectively. Total n=236. For each site (n=61 or n=62), correlation coefficients were generally higher (>0.5) than shown above.

B Correlation with environmental variables

Among the physico-chemical variables examined, water depth and conductivity showed significant correlation with *V. cholerae* counts and with other microbial populations (Table 4.5). Conductivity was found to be correlated with bacterial counts with a 0 to 6-week lag period.

C Correlation with cases of cholera

Although the seasonal pattern overall was similar for ITS counts and either diarrhea or cholera cases (Figure 4.6), a direct correlation was observed for two of the four sites surveyed. From a cross-correlation analysis done to account for temporal lags between time-series data, a significant correlation was observed between clinical cases of cholera and ITS counts for all of four sites, with an interval of 0 ~ 1.5 months (Figure 4.7).

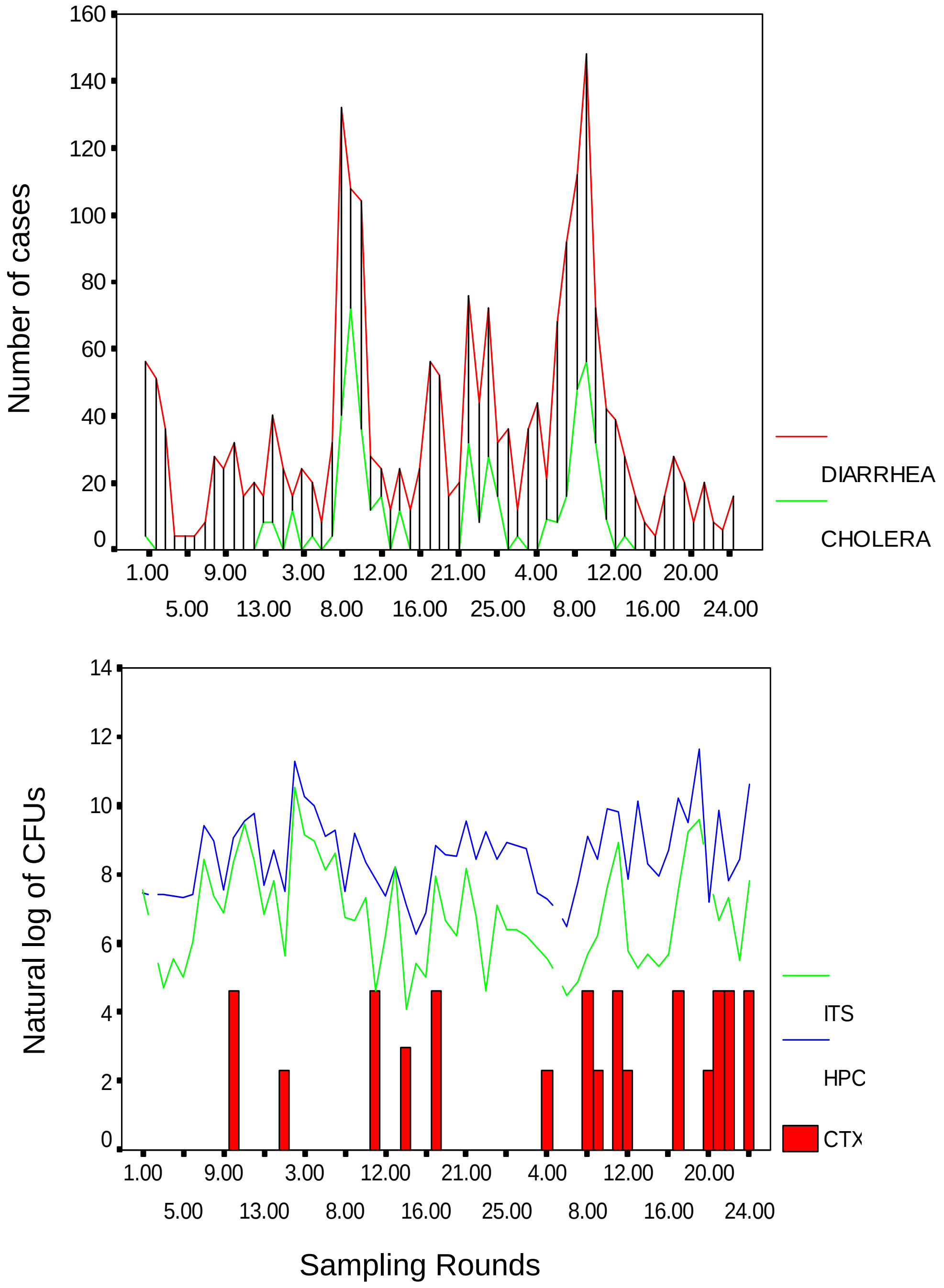


Figure 4.6. I pper panel), and natural log of CFUs per ml for heterotrophic plate count (HPC), pITS+ colonies (ITS) and cholera toxin probe (CTX) positive on LB agar during June 1997 to December 1999 (the lower panel). CTX counts were provided by J.G. Morris (University of Maryland, Baltimore).

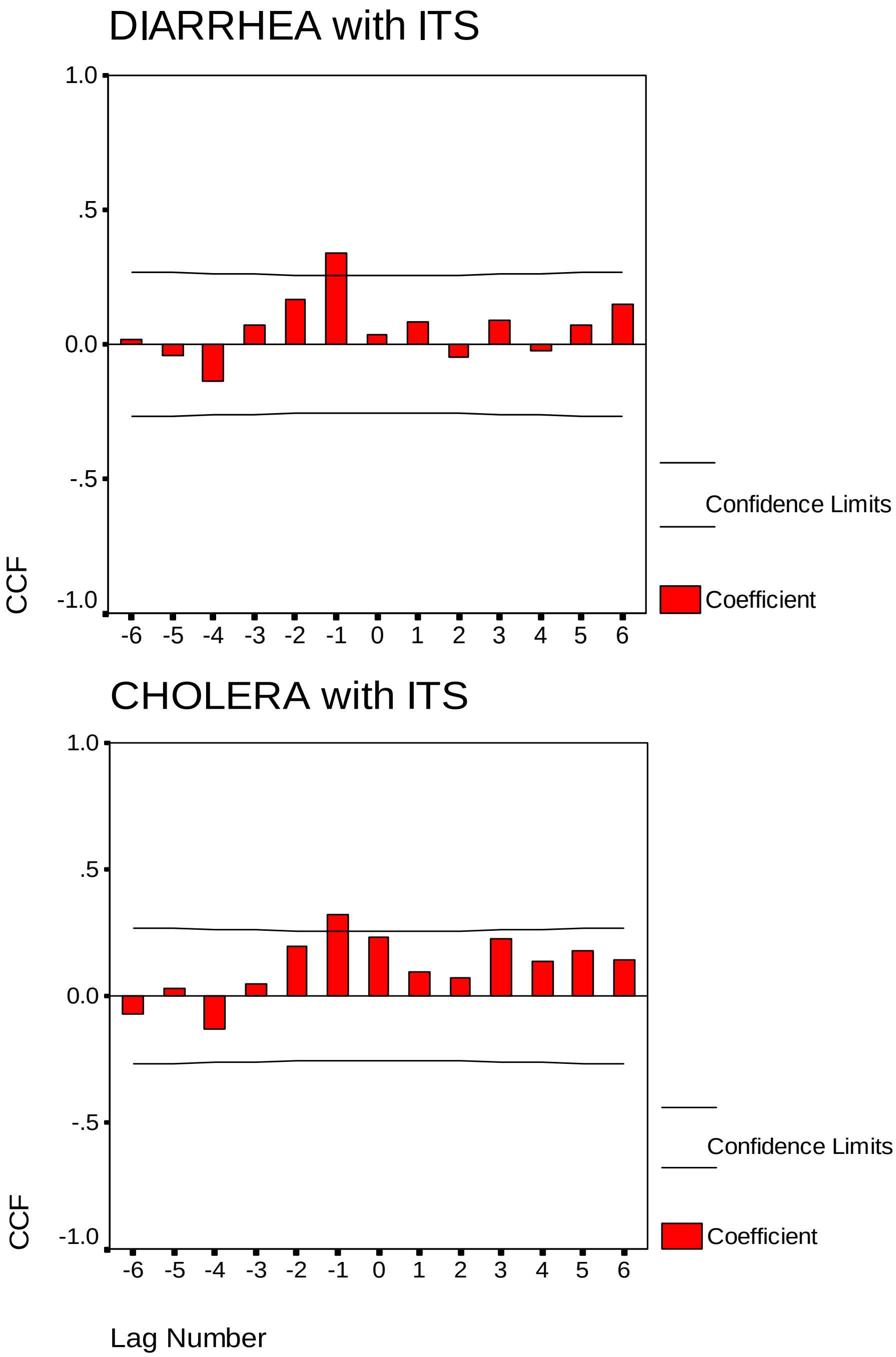


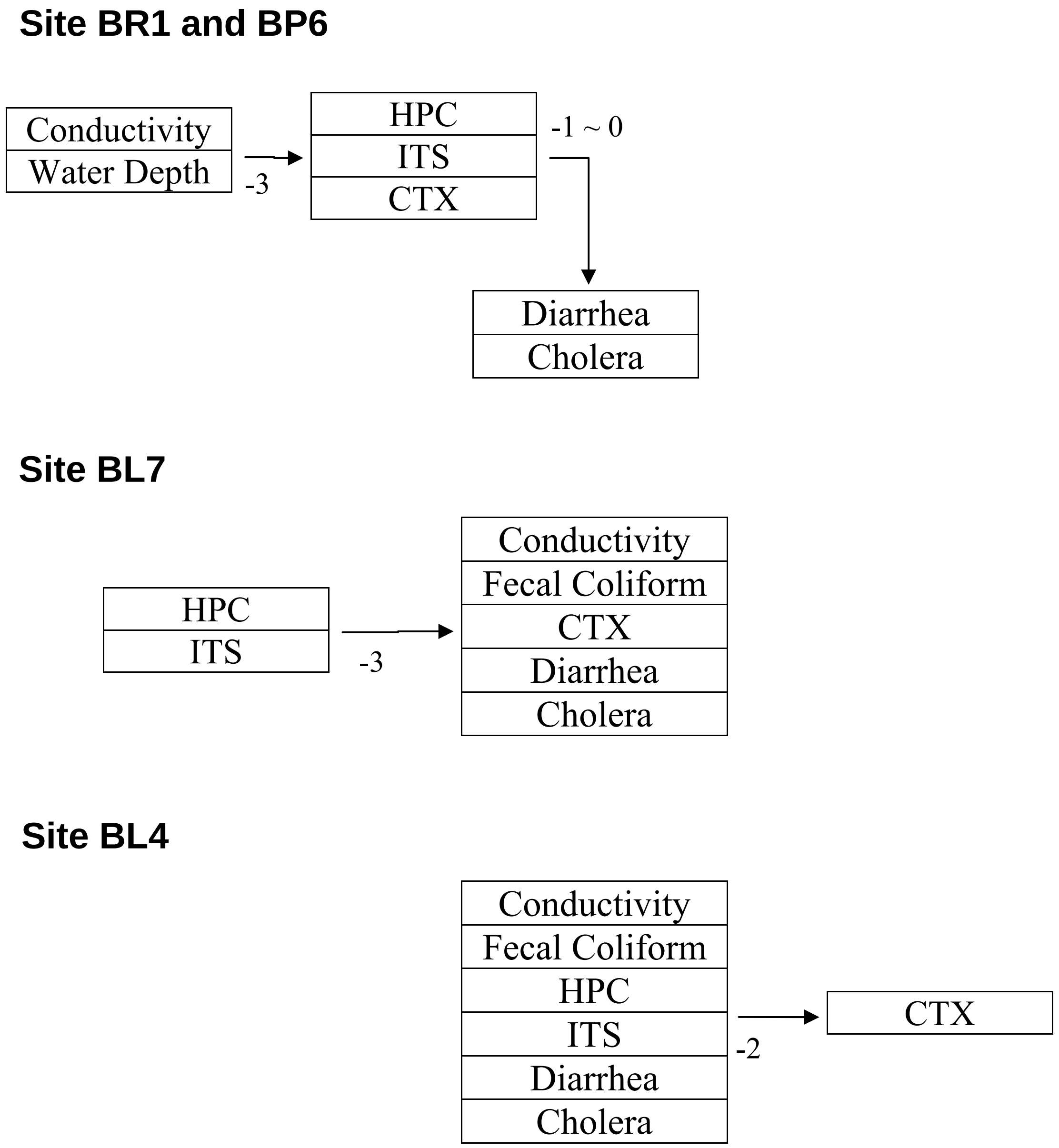
Figure 4.7. Cross-correlation coefficient (CCF) between log-transformed pITS+ colony count (ITS) at Site BR1 and clinical cases of diarrhea and cholera with lags. At the sampling interval of 2 weeks, the ITS counts preceded clinical cases. Confidence limits were determined at 99% of significance.

D Temporal cascades

The continuing presence of *V. cholerae* in the geographical sites sampled in this study indicates that the organism constitutes a major component of the bacterial community in this natural ecosystem. The high CFU counts and strong correlation with heterotrophic plate counts (HPC) indicate that the *V. cholerae* population is an active, normal component of the bacterial community in this ecosystem. In addition, the coupling of the dynamics of *V. cholerae* with that of *ctx*-positive *V. cholerae* indicates that the number of toxigenic *V. cholerae* also fluctuates with *V. cholerae* populations.

The results also show that the dynamics of endemic cholera is related to the variation in total culturable *V. cholerae* in the environment, but with a significant time lag. From this discovery, it can be suggested that blooms of *V. cholerae* in the aquatic environment directly lead to outbreaks of cholera. However, it does so by influencing the toxigenic portion of the total *V. cholerae* to reach an outbreak threshold with time. This threshold response model is also supported by our finding that the distribution of maximum culturable *V. cholerae* count pooled across the four sites shows a lagged correlation.

As indicated by the correlation with water depth and conductivity, the numbers of *V. cholerae* in surface water are influenced by environmental changes, namely by changes in conductivity and water depth (reduced rainfall and lower water depth). Thus, the hypothesis of an environmental modulation of endemic cholera dynamics via *V. cholerae* populations in the aquatic environments is strongly concluded. However, significant heterogeneity in the relationship between the ITS count and clinical cases across the four sites was observed (Figure 4.8). This difference was also noted in the temporal coupling of ITS and CTX counts. This can signify that a tight coupling between conductivity and cases of cholera, without a time delay for increase in the *V. cholerae* count, was caused by transport of water carrying *V. cholerae* from other water bodies, such as at sites BR1 and BP6. This would be true for tidal estuaries like the Bay of Bengal. It also suggests that environmental reservoirs of *V. cholerae* and, therefore, source of the cholera cases, may have different levels of contribution. For example, sites BR1 and BP6 are active reservoirs promoting proliferation of *V. cholerae* under favorable conditions, whereas sites BL4 and BL7 are transient reservoirs. As visualized in temporal cascades of simultaneous and delayed events, shown in Figure 4.8, results of this study indicate a point source for toxigenic *V. cholerae* blooms that then spread to other water bodies.


Site BR1 and BP6

Site BL7

Site BL4

Figure 4.8. Compartment models supported by significant time lags from cross-correlation analyses. Numbers under the arrows represents time lag in mulitplication of two week units. At sites BL4 and BL7, the ITS counts were temporally uncoupled from CTX counts.

4.4 Conclusion

From studies of the structure and dynamics of endemic cholera, a fundamental aspect observed was that the scale of spatial dimension in the cholera endemic areas is an important factor in determining the mechanism of endemic cholera. From a comparative structural analysis of toxigenic *V. cholerae* populations from two distinct sources (aquatic vs. clinical environments), it is concluded that the habitat barrier is not significant. Thus, the choleragenic agent moves from one kind of source to the other in a time scale of less than 2 years. In contrast, the population structure at the basin scale, *V. cholerae* population structures can differ because of spatial separation of the *V. cholerae* populations.

Results from a study of the dynamics of *V. cholerae* populations also produced a spatial perspective, with respect to the mechanism of cholera endemism. In Bakerganj, there are water bodies that generate toxigenic *V. cholerae* populations via an indigenous process, such as seasonal increase of carrying capacity for all heterotrophic bacteria. Other water bodies appear to acquire choleragenic *V. cholerae* via transportation of the indigenous populations. Therefore, the spatial scale for the spread of the cholera *Vibrio* is regional (village or community). Because the population structure supports different cholera *Vibrio* populations between regions (between Bakerganj and Chhattak), the spatial compartment in which endemic cholera is generated and spread, is the scale at the regional level.

Chapter 5. Summary and Conclusions

From the phylogenetic and population structural analyses, it is concluded the species, *V. cholerae*, is an early branched monophyletic compartment within the genus *Vibrio*. Examination of the distribution of genotypes within the species, an extensive diversity of clonal entities was observed. Epidemic *V. cholerae*, namely *V. cholerae* O1 and O139, is concluded to comprise a tight clonal cluster with the potential to evolve via by lateral gene transfer and recombination through interaction with the diverse, non-epidemic environmental *V. cholerae.* A specialist ecology is concluded from the different lineages of *V. cholerae* in Chesapeake Bay samples collected during an extensive environmental survey. Based on results of this study, the clones that cause epidemic cholera can be portrayed as an ecological compartment with a typical specialist strategy, strongly supported from the observation of a significant order-specific association between crustacean zooplankton and non-epidemic *V. cholerae* lineages.

Endemic cholera, the source of pandemic cholera, is a result of the presence of such clones that persist in a body of water in which their seasonal growth is promoted. When cholera season arrives, these clones are able to spread to various neighboring bodies of water by processes yet to be clarified.

The human population in an endemic area is, in general, not protected from the cholera-causing bacteria originating in contaminated water because previous exposure to cholera causing bacteria provides only a transient intestinal immunity, e.g., 6 months (Sack *et al.*, 2004). In addition, the presence of multiple epidemic clones of *V. cholerae* in a contaminated bodies of water in a cholera-endemic area renders partial immunity arising from previous exposure ineffective.

The spread of cholera causing-bacteria from a point source, in which they are autochthonous inhabitants, depends on the carrying capacity of specific ecosystems and distances to the source. Spatial remoteness can be an effective barrier to endemic cholera because spread to other water bodies occurs only within relatively short distances. Limitation of gene flow between *V. cholerae* O1 El Tor strains in different geographical areas advocates in favor of this argument. Therefore, the unit of the system in which endemic cholera develops has its boundary at the regional scale. Within the boundary of a regional ecosystem, contact of human populations with the contaminated water bodies provides the mechanisms for transmission. Seasonal forcing of such transmission occurs as more water bodies are contaminated by internal forcing of the disease, i.e., shedding from infected individuals or, more likely, by reinforcement of the external forcing, that is, transport of the contamination from the point source to other water bodies.

These mechanisms can be expressed in a qualitative compartment model, as shown in Figure 5.1. A single system is defined at the regional scale as having a spatial range in the order of 1 – 10 Km. That scale is based on the population structure of *V. cholerae* O1 El Tor strains, as described in Chapter 4, that is, distance between different study areas can partially block gene flow among strains. In each system, the human population and the aquatic environment are the major compartments, each of which comprises multiple sub-compartments. The human population is divided into sub-compartments by those factors that cause different outcomes in individuals, depending on their exposure to contaminated water. The differential age structure in the prevalence of cholera caused by *V. cholerae* O1 El Tor and *V. cholerae* O139 is a good example of such sub-compartments. The different levels of protection are provided by immunity achieved from previous exposure to various strains. Therefore, immunity versus susceptibility status of an individual is an effective cause of the different outcomes within a populations during a given cholera outbreak. Other behavior factors, such as whether the waster is used for drinking water or bathing can cause divisions of different dimensions within a human population. In the aquatic environment, the point source (W_0) supports survival and/or autochthonous proliferation of diverse toxigenic clones of *V. cholerae* at those times of the year when cholera is absent. During outbreaks, transmission of cholera-causing bacteria from the point source to other bodies of water occurs, intensifying environmental forcing of the disease.

The states and processes in each compartment can be viewed in a transmission model, as is shown in Figure 5.2. In a typical transmission model for infectious diseases, the state of an individual host changes from susceptible to infected by carriage of cholera-causing bacteria. In the same way, a body of water in a system can occur in either the uncontaminated or contaminated state. While the host in a state at a given time can be quantified quite simply as the number of individuals, a body of water can be quantified as volume or surface area. Modulation of an outbreak by climate can occur through modulation of the transmission of cholera-causing bacteria among different water bodies.

Although the model presented here is not yet quantitative, it is useful in identifying knowledge gaps to be filled and challenges to be met before a quantitative model can be derived. The most significant gaps include insufficient information of the processes and limnological factors determining the state of a body of water such as indigenous reservoir, and whether bodies of water have been uncontaminated or contaminated. Furthermore, the climate factors modulating limnological factors and transmission processes are not yet known. The mathematical challenges in devising quantitative model include the dynamic nature of compartment formation in both human populations and in the aquatic environment and inter-dependence among compartments. For an example, the immunity of an individual is linked to previous exposure which, in turn, is linked to behavioral factors, such as the manner in which water resources are used. In aquatic environments, bodies of water also have complex hydrological and biological connections.

In conclusion, the mechanism of endemic cholera developed from this study is described briefly as follows: a body of water serves as both a reservoir and point source of *V. cholerae* epidemic strains in seasonal spread of these bacteria. In addition, a universal seasonal forcing occurs that repeats the spread of cholera-causing bacteria from a point source each cholera season. Further work will address those factors that determine which water body is a point source and reservoir and the mode of transportation causing spread of contaminated water. Clearly, a geographic information system must be integrated in our long-term survey that is currently underway. It can be predicted that civil engineering efforts will be significantly more effective than vaccination or drug-based prophylaxis in preventing outbreaks of cholera in cholera endemic areas of the world, based on the results of this study.

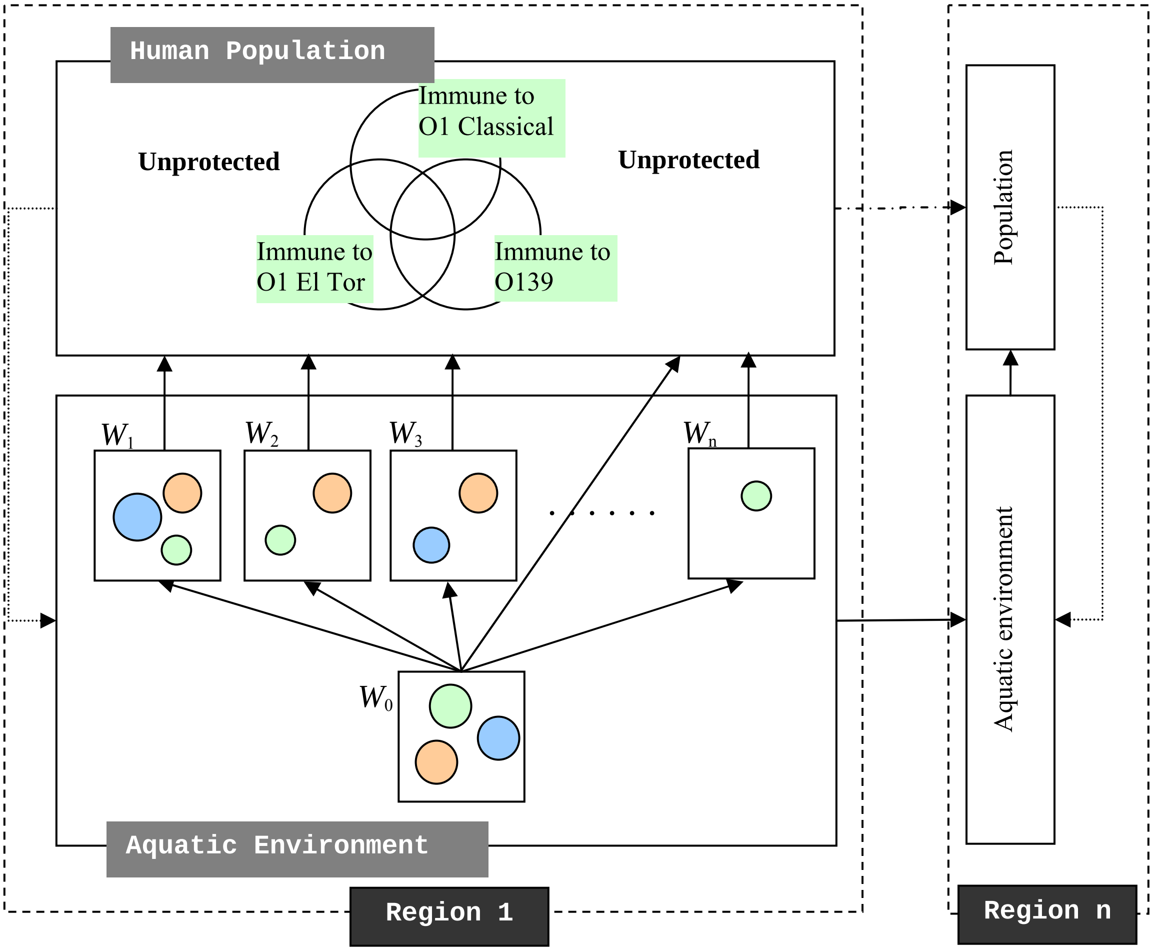


Figure 5.1. Qualitative compartment model for endemic cholera in rural areas of Bangladesh (Solid lines = transmission of cholera-causing bacteria; dotted lines = fecal contamination of bodies of water; dash-dot line = migration of populations; rectangle W_0 = a body of water with autochthonous growth of cholera-causing bacteria; rectangles W_1 - W_n = other bodies of water in the region; circles of three colors inside W_0 - W_n = different populations of *V. cholerae*, e.g., *V. cholerae* O1 Classical, O1 El Tor, and O139, respectively; Venn diagram inside the human population compartment = compartments among populations with different combinations of protection by transient immunity to the three kinds of cholera-causing bacteria; dashed rectangles = regional boundaries)

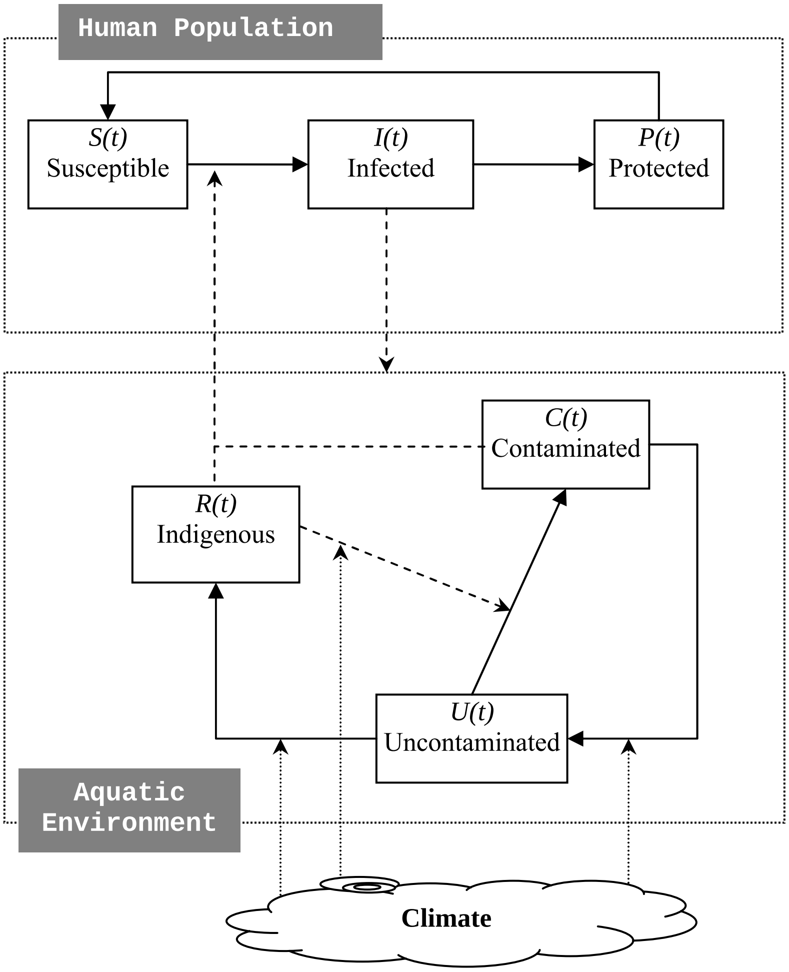


Figure 5.2. Schematic diagram of a state-based transmission model for endemic cholera in rural areas of Bangladesh (*t* = independent variable representing time; solid lines = movement of individuals from one state to another; dashed lines = movement of cholera-causing bacteria; dotted lines = modulating effects of climate factors, such as change in precipitation and temperature; *S(t), I(t)* and *P(t)* = the number of individuals at time *t* in each state; *C(t), U(t)* and *R(t)* = volume or surface area of water bodies at time *t* in each state).

Appendices

Appendix A. Primers and Probes

Name	Target	Sequence (5'-3')
P16SF1	16S rRNA	AGAGTTTGATCMTGGCTCAG
P16SR1	16S rRNA	CGGYTACCTTGTTACGACTT
pTmF1	tmRNA	GGGGCTGATTCAGGATTCG
pTmR1	tmRNA	GCTGGGGGGAGTTGAACC
pIVPF	tmRNA	TAGCGTGTCGGTTCGCAG
pIVPR	tmRNA	AGGKTATTAAGCTGCTAGTGCG
ERIC1	ERIC	ATGTAAGCTCCTGGGGATTCAC
ERIC2	ERIC	AAGTAAGTGACTGGGGTGAGCG
BOX	BOXa	CTACGGCAAGGCGACGCTGACG
pITS	ITS	GCSTTTTCRCTGAGAATG

Appendix B. Probability of Random Band Matching

False band matching

In conventional genetic profiling techniques where results are obtained based on band size differences in electrophoresis gels (e.g., AFLP, RFLP, AP-PCR, rep-PCR), bands are interpreted as different genetic traits by the difference of their sizes. In AP-PCR or a low-stringency rep-PCR, the band-to-trait interpretation may not be valid because PCR from diverse primer binding sites may produce bands of the same size, with the result that a single band size may correspond to multiple genetic traits. In comparing fingerprints among genomes of little genetic relatedness; i.e., low similarity in genome structure of OTUs), most matching bands might not be from the same primer binding sites. In that case, termed “false band matching”, bands have the same size simply as a result of random chance, rather than from a biological basis.

Probability of a given combination in the number of common and unique bands

Knowing the probability that an observation occurs solely from random chance allows determination of the significance of the observation. To calculate the probability that band matching observed for a given pair of strains is a random occurrence, without systematic cause, one can consider each lane in an electrophoresis gel as an outcome of a random drawing of a subset from a collection of all possible bands, i.e., the sample space.

Here, the sample space (the entire band collection) is symbolized as S and it is

$$S = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots., \beta_t\}$$

, where each one of all possible bands is indicated as β, with the identification suffix from 1 to N, representing the total number of all possible bands. In rep-PCR, bands are distinguished and identified by size. Therefore, t is determined as the number of all possible band sizes that can be discretely recognized in rep-PCR and agarose gel electrophoresis.

For a given pair of strains (namely, strains A and B), electrophoresis produces two lanes, namely L_A and L_B, which are subsets of S and can be described as

$L_A = \{a_{1'}, a_{2'}, a_3, a_4, a_5, \ldots., a_i\}$ and

$L_B = \{b_{1'}, b_{2'}, b_3, b_4, b_5, \ldots., b_j\}$

, where each one of observed bands is indicated either as a or as b, with identification suffix from one to i or j, and the latter are the number of observed bands for L_A and L_B, respectively, with the range from one to t. When the lane L_A is formed, i bands are randomly selected, without replacement, and each of them is assigned one of the a’s in L_A. Similarly, the lane L_B is formed, j bands are randomly selected, without replacement. From the nature of the rep-PCR experiment, the events of forming L_A and L_B are independent of each other. In rep-PCR, β can be assumed to have a uniform probability distribution, i.e., the same probability to be drawn in an experiment for all β’s. This assumption is plausible because there is no known basis for bands of specific size to occur more or less frequently than others in rep-PCR.

When N is symbolizing the number of possible outcomes from two independent experiments in which i or j number of bands is randomly drawn from a uniform distribution of β, its value will depend on combinations of i, j, and t according to the calculation as follows:

$$N = \binom{t}{i}\binom{t}{j} \qquad \text{Equation 1}$$

When the bands in the two lanes observed on an electrophoresis gel are matched, *k* number of common bands, which can range from zero to the smaller of *i* and *j*, can be observed. If bands in the two lanes are classified based on results of band matching, the lanes are as follows:

$L_A = \{a_{1'}, a_{2'}, \ldots., a_{(i\text{-}k)}, c_{1'}, c_{2'}, \ldots., c_k\}$ and

$L_B = \{b_{1'}, b_{2'}, \ldots., b_{(j\text{-}k)}, c_{1'}, c_{2'}, \ldots., c_k\}$

, where the bands common to both lanes are designated as *c* and bands unique to a lane as *a* and *b* (i.e., $a_n \notin L_B$ and $b_n \notin L_A$, where $n = 1, 2, \ldots, (i\text{-}k)$ or $n = 1, 2, \ldots, (j\text{-}k)$, respectively). In this case, outcomes for forming two lanes are equivalent to those from the three independent experiments: choosing *k* number of common bands out of *t* total possible bands, where $k \leq i$ and $k \leq j$, and then choosing (*i*-*c*) number of bands unique to L_A from the (*t*-*c*) sample space and (*j*-*c*) number of bands unique to L_B from the same (*t*-*i*) sample space. The sample space of the last experiment was chosen in order to have bands unique to each lane. Therefore, the number of possible outcomes under the limitation of having *k* common bands (N_c) can be calculated as

$$N_c = \binom{t}{k}\binom{t-k}{i-k}\binom{t-i}{j-k} \qquad \text{Equation 2}$$

The probability of the occurrence of a pair of lanes with given *i*, *j* and *k* to occur (*P*) can be calculated as follows:

$$P = \frac{N_c}{N} \qquad \text{Equation 3}$$

In the results of rep-PCR in this study, *i* and *j* were in the range of 12 to 35, and *k* from 0 to 22. In the case of *t*, it is appropriate to consider range, rather than a particular value, because resolution of an electrophoresis gel depends on band size (i.e., the larger the band, the smaller the difference in migration rate). The theoretical minimum of *t* is ($i + j - k$), the number of the bands appearing in the lanes. The experimental minimum can be estimated from resolution of the worst case. At the upper limit of the band size employed in this study (6.47 kb), 15 bp was the smallest difference among bands. Using this value, the experimental minimum of *t* can be calculated as a discernable 420 band positions in total. The maximum of *t* is estimated to be 629 from the observation that the electrophoresis gels in this study had resolution of 10 bp for 190 bp fragments, the smallest in analysis (i.e., the case of the best resolution). The band size used in the analysis ranged from 190 bp to 6.47 kb. Therefore, the empirical range of *t*, estimated in the gels analyzed this study, was 420 to 629. The *P* values for the most frequent case of given *i* and *j* were distributed with *t*, as shown in Figure B.1.

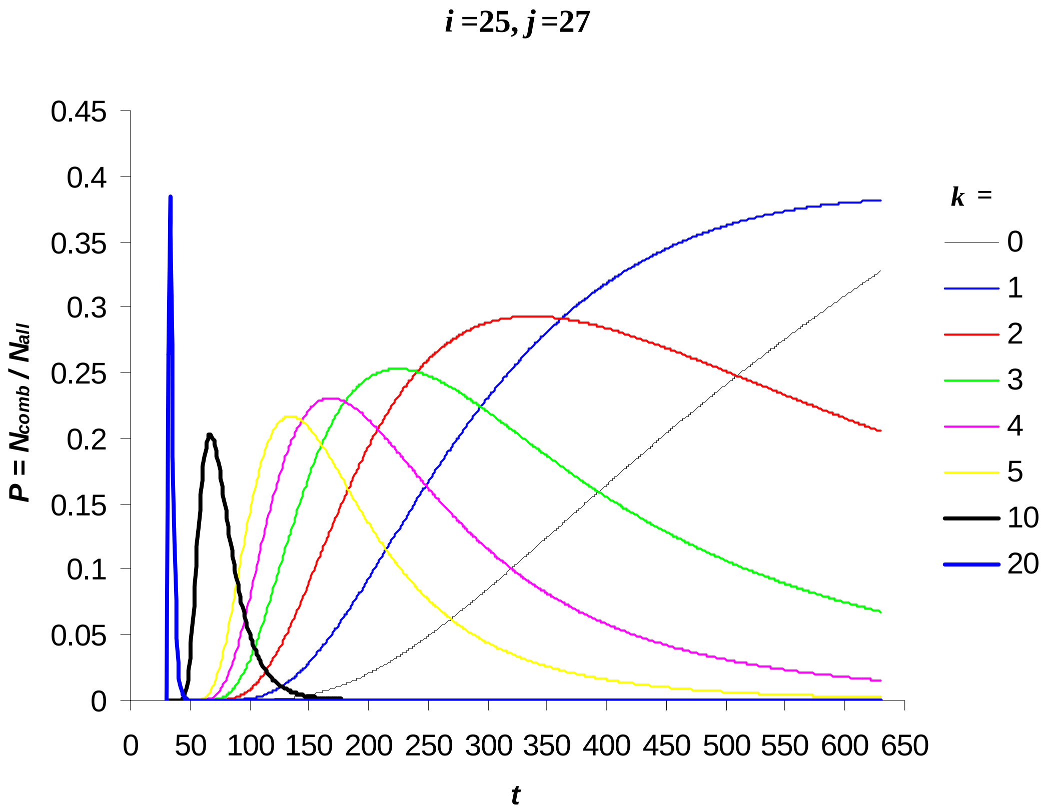


Figure B.1. The distribution of probability of occurrence of a given pair of lanes with 25 and 27 bands each, with a varying number of matching bands (*k*) along the assumed total number of band sizes possible to by the method employed in the study.

Noteworthy from the graph is that the maximum probability for all values of *k* never went below 0.05, one of the most common critical values employed for significance tests. Therefore, there is always a significant possibility for any of the 25 and 27 band lanes to have any number of common bands. The same results were found for other combinations of *i* and *j* that appeared in the lanes of this study. Even in the case of the absence of a common band between a pair of lanes (i.e., *k*=0), *P* can be higher than 0.05 if >250 bands positions can be resolved by the gel. When the analysis was confined to the empirical range of *t* (420 – 629), determined from the gels obtained in this study, it could be concluded that band matching with less than three common bands can occur randomly.

Probability for a given set of common bands

In the experiments undertaken, the occurrence of the same or similar band matching, i.e., a set of common bands, can be found among different pairs of lanes. The question is whether the recurring common bands were a result of random band matching or had a systematic basis, i.e., phylogenetic relatedness of DNA used in the PCR. The probability of a random occurrence of a set of common bands can be calculated by modifying Equation 2 and applying it to Equation 3. Because the *k* common bands are fixed as the observed set, they can be excluded from randomization by not permuting from the sample space, accomplished by the first term in Equation 2. Therefore, the number of outcomes with the observed set of common bands (N_o) is

$$N_o = \binom{t-k}{i-k}\binom{t-i}{j-k} \qquad \text{Equation 4}$$

The probability of the occurrence of a pair of lanes with a given set of common bands (*P*) can be calculated as

$$P = \frac{N_o}{N} \qquad \text{Equation 5}$$

As represented in the example given in Figure B.2, any set of common bands, including the case of a single matching band, has a significantly low probability to be formed by random chance alone ($P < 0.01$). Lanes of no matching band, however, have an high probability of appearing within the empirical *t* range (420-629) and the same result was observed throughout all combinations of *i* and *j* in the range of 12 to 35, the empirical range in the number of bands in the electrophoresis lanes. Therefore, absence of matching bands between a pair of lanes implies a lack of significant phylogenetic relatedness between the source DNAs, because only randomly selected lanes can produce such results. If there is significant relatedness, at least one specific band representing relatedness should be observed.

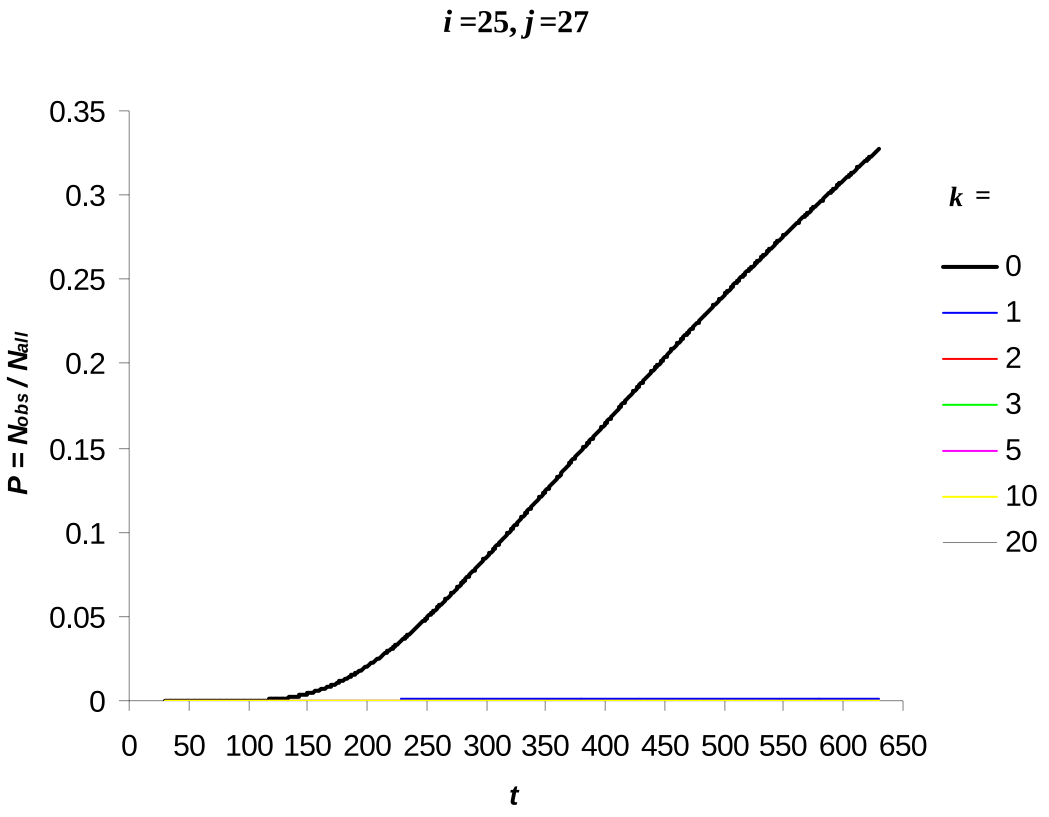


Figure B.2. Distribution of probability of occurrence for a given set of common bands in two lanes containing 25 and 27 bands, respectively. Varying numbers of matching bands (*k*) were applied, but results from *k* > 0 cases yielded only flat lines on the axis of assumed total number of band sizes.

Pairwise test of significance of band matching

When all matching bands between two lanes arise from false band matching, a falsely matched pair (FMP) results. In the comparison where OTUs share a significant portion of their genome structure, pairs of OTUs are truly matched pairs (TMP), yielding a correspondence of common bands as common traits. For a given pair of lanes, the probability of an FMP can be calculated, using Equation 3 because FMP is comprised of common bands randomly matched. The probability can be used to test the significance of the null hypothesis (H_0) that the observed band matching is a result of the two independent sets of random sampling of observed number of bands for each OUT, from a uniform probability distribution of bands of different sizes (i.e., FMP).

Because the exact t value for the lanes is not known, the range of P can be calculated using an empirical range for t ,i.e., 420-629 band positions, for comparison with a critical value of type I error (α). When the maximum probability of observed band matching (P_{max}) is smaller than α, the pair can be interpreted as a significantly matched pair (SMP), with a valid rejection of H_0. When the minimum probability of observed band matching (P_{min}) is greater than α, the pair can be validly interpreted as an insignificantly matched pair (NMP), with H_0 concluded to be valid. If the range of P includes the critical value, i.e., $P_{min} \leq \alpha \leq P_{max}$, the test fails in identifying the pair as FMP or TMP because of the ambiguity in gel resolution. Because these grey cases can be contaminated with an FMP, it is reasonable to consider the pairs as an NMP, unless other facts determine it to be a TMP.

Multiple test of significance of band matching within a cluster

When a cluster of OTUs are suspected to be a closely-related monophyletic group, all pairwise band matching should be a TMP, which implies clustering based on valid band matching . To demonstrate that a cluster is comprised of TMPs, a multiple test that employs the pairwise test for band matching, as described above, for all pairs of OTUs can be used. Because of the ambiguity in gel resolution, only the test of $P_{max} < \alpha$ allows valid interpretation of the tested pairs as TMP by producing valid SMPs. To avoid accumulation of type I error, a test employing adjustment of P_{max} values can be made according to the sequential rejection procedure of Holm (1979) which is appropriate for non-independent tests (Legendre & Legendre, 1998).

Appendix C. Significances of Clusters Common in Multiple Trees

Background

In AP-PCR, RAPD, or low-stringency rep-PCR methods, sampled sequences are not known, but only the sizes of their products are known. In those PCR reactions, complex interactions among target sequences, primers, and intermediate products make it difficult to build mathematical models describing the sampled sequences (in other words, the sampled traits). Typical binary-coded results also are problematic in determining the significance of the cluster analysis from the data. Diverse PCR products are interdependent in band formation, violating the assumption of independence required in most significance assessment methods, such as bootstrapping. For example, when linear dependences among the 106 band loci recorded in this study using rep-PCR were removed by employing principal component analysis, >95% of the total variance could be explained by 37 principal components whose eigenvalue was larger than one, implying about 2.8 bands behaving as a set.

The problem of the lack of sequence data and model was circumvented by using one of the distance-based agglomerative hierarchical clustering methods (e.g., complete linkage, single linkage, UPGMA, and NJ). However, those methods still require meeting the assumption of independence for assessment of the significance of the branches. In addition, recent advances in interpreting bacterial genomes suggest building trees by hierarchical clustering methods is not appropriate for phylogenetic analysis of bacterial populations because the populations contain extensive lateral gene transfer (LGT), a good example being the *Vibrionaceae* described in Chapter 2. LGT drives the population toward inheriting genetic content from more than one progenitor organism. Therefore, a network of OTUs and their progenitors are considered better for describing the structure and evolution of a bacterial population.

As an alternative method for estimation of the significance of clustering, consensus of a cluster in trees built using different methods was chosen (Hilali *et al.*, 2000), based on the finding that >70% of true clusters of simulated data were common to three trees (Kim, 1993).The advantage of this method is the absence of a restriction on the data or tree building as long as there are multiple trees built by different clustering methods. However, the drawback is that the technique is purely empirical and not a test of significance based on probability of cluster occurrence. In this section, the probability-based significance test on cluster consensus for multiple trees was devised by applying the absence of intra-cluster divergence, i.e., unstructured populations, as the null hypothesis. While considering LGT as one of the factors producing an unstructured population, the probability of a random occurrence of a cluster-making split from an unstructured population was calculated, interpreted as the minimum requirement for a cluster to be significant one, i.e., diverged population, and used to identify nonsignificant clusters by evaluating the extent of consensus of a cluster in multiple trees.

Distribution of genotypes in unstructured population

One way to consider structures of bacterial populations is to observe the frequency distribution of different genotypes. Because the genetic diversity of a bacterial population is extensive, different genotypes can be visualized best when they are summarized along a simple gradient and distance from the centroid (or possibly to the founder) of the population can provide such a gradient. Figure C.1 demonstrates the possible shapes of non-diverging populations. The population of greatest interest is the climax population, where a majority of genotypes is uniformly distributed, except for a small proportion of the frontier zone. As LGT promotes extensive recombination of existing genotypes to generate new genotypes, it can reach a point where all genetic loci available for recombination are saturated. The climax population, that is, the non-diverging recombination saturated population, can be approximated to uniform distribution, the simplest distribution for considering random events.

If genotypes in a cluster reveal a distribution that is not significantly different from the results of random sampling from a uniform distribution, the cluster can be considered to have been sampled from a climax population. However, the reverse case, where the cluster shows a significant difference from uniform distribution, does not always mean the cluster is sampled from the uniform distribution. Uneven sampling, together with uneven genotype distribution, can generate significant deviation from uniform distribution for a sampled cluster. If a cluster possesses intra-cluster divergence, the cluster should show significant deviation from uniform distribution, arising from genetic divergence and/or sampling-bias.

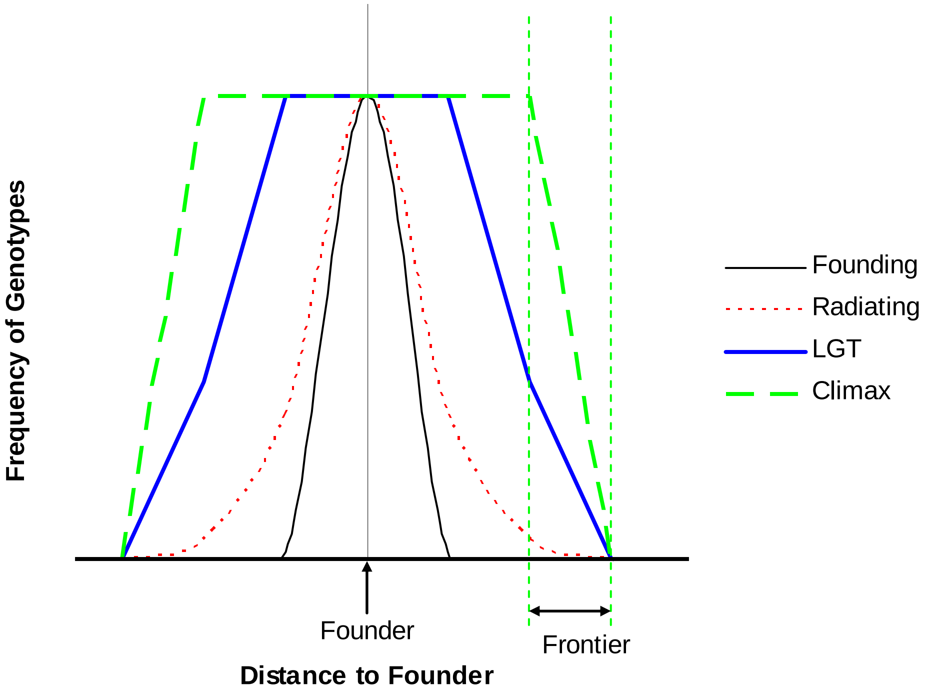


Figure C.1. Distribution of genotypes along a gradient of distance from the founder in four types of unstructured populations. In a founding population, a new ecological niche is being established. In a radiating population, the established population begins diversification in all possible directions. The genotype frequencies are similar where strong LGT homogenizes the genetic information among different genotypes around the most frequent genotype in populations with extensive LGT. In the climax population, evolution-saturated population under the influence of LGT can carry the majority of the genotypes in a non-frontier zone where frequencies are uniform.

Significance of a cluster under the uniform distribution

From a null hypothesis that an observed cluster consists of genotypes sampled from the climax population, two assumptions can be made: uniform distribution of genotypes in the population and random sampling of genotypes from the population. The rejection of the hypothesis does not always indicate that the population diverged (i.e., a structured population). It can also mean that the population is unstructured, with a non-uniform distribution, e.g., as shown in Figure C.1. In addition, biased sampling (non-random or uneven sampling) can cause rejection, as well. Therefore, significant rejection of a uniform distribution hypothesis should be interpreted as one requirement for a cluster to have been generated by significant genetic divergence and the condition of rejection can be a useful criterion for screening clusters for significant divergence.

Genetic divergence in a population creates a split in the population. The genotypes in the global population (G) can be divided into two subsets (A and B) by a split

$$G = \{\{a_1, a_2, ..., a_{(t-n)}\}, \{b_1, b_2, ..., b_n\}\} = \{A, B\}$$

, where t is the total number of genotypes (the population size), n is the number of genotypes in the cluster defined by a split (split size). When a particular sample cluster (B) is observed in an experiment, the split size (n) is known, and the probability for a particular split to occur (P) from random sampling of n genotypes from uniformly distributed genotypes is:

$$P = \frac{1}{\binom{t'}{n}} \qquad \text{Equation 6}$$

, where t' and n are integers and $t \geq t' \geq (n+1)$. Because, typically, the population of interest produces many clusters (splits), the apparent total number of genotypes (t') from which n members of OTUs are selected varies and is not known. Therefore, the exact P value cannot be obtained. However, the range can be calculated, using t'. If the target population analyzed is of infinite size, like the global population, only the maximum probability ($P_{\max}$) can be calculated.

$$P_{\max} = \frac{1}{\binom{n+1}{n}} = \frac{1}{\binom{n+1}{1}} \qquad \text{Equation 7}$$

Because $P_{\max}$ is the maximum probability of random formation of a split within a population of unknown number of genotypes, it can function as a conservative estimate for the probability for a split to form by random chance under a uniform distribution by using the criterion of $P_{\max} < \alpha$, the critical value of type I error (e.g., 0.01 or 0.05). Clusters comprised of $n > 20$ genotypes and $n > 100$ genotypes meet the qualification of the criterion for a type I error level of 0.05 and 0.01, respectively.

Significance of cluster found in multiple trees

When a cluster of genotypes occurs in multiple, independently constructed trees, the probability of multiple occurrence of a split can be calculated by the multiplication of individual $P_{\max}$. Therefore,

$$P_{\max}^{h} = \left(P_{\max}\right)^{h} = \frac{1}{\binom{n+1}{n}^{h}} \qquad \text{Equation 8}$$

, where h is the number of trees independent from one another. When the distribution of $P_{\max}^{h}$ was tabulated by n and h, the minimum number of trees for which consensus of a cluster could be determined is as shown in Table C.1. This result can be used to identify clusters with significant deviation from the hypothesized uniform distribution of genotypes in the population, based on consensus of clusters in multiple trees.

Table C.1. The minimum number of independent trees rendering clusters of a given size significantly deviating from uniform distribution with random sampling at two critical values of type I error.

Genotypes (*n*)	$P^{h}_{\max} < 0.01$	$P^{h}_{\max} < 0.05$
2	5	2
3	4	2
4	3	2
5	3	2
6	3	2
7	3	2
8	3	2
9	3	2
10	2	2

Appendix D. BNS Script Code for Determination Subdivisions

```
// ======================================================================
// TITLE: Make Subdivisions for Bandclass determination
// CODEFILE NAME: cutoff_subdivision_by_NMP 0.1.BNS
// PLATFORM: GelCompar II version 3.0
// COPYRIGHT: YOUNG-GUN ZO, 2005, UNIV. OF MARYLAND BIOTECH. INST.
// SYNOPSIS:
//    1. Cluster analysis by complete linkage
//    2. Pairwise matching band count
//    3. Significance test on branches
//       pairwise test for presence of an NMP in a branch
//       NMP is determined by the criterion of (minP > alpha)
//       step-down: search from the top,
//          if the branch contain NMP--> non homogeneous --> cutoff
//          if no NMP --> homogeneous --> stop, assign as a subdivision
// PRE:
//    1. Comparison window open with at least two entries
// USER INPUTS:
//    1. File name to report the band count result
//    2. Band matching condition (optimization and position tolerance)
//    3. Type I error level for pairwise NMP test
//    4. Min and max of possible band positions
=====================================================================

// Module 0. Get Environment for Experiment and Comparison
string  expername,CountFileName;
integer CountBands, Tmin, Tmax;
float   alpha;

//---- User Inputs
// 1. Variable to indicate to perform counting or skip and read results
from a file
//    value list: 0 = read count result from a file, 1 = count now
      CountBands=1;
// 2. File name to input if CountBands = 0, to output if CountBands = 1
      CountFileName = "d:\MatchCounts.txt";
// 3. set alpha for test
      alpha = 0.01;
// 4. band position range
      Tmin = 420; Tmax = 629;
//---- Get Environment Variables
if not(CmpIsPresent) then CmpAttach;
if not(CmpIsPresent) then {
   message("Unable to attach a comparison to this script");
   stop;}
expername = CluGetCurrent;


// Module 1. Cluster Analsyis
//         optimization=1%,tolerence=0.8%,gradient=0.2%
string cluSetting;
```

```
cluSetting = "Clustering=Complete Linkage Similarity=Dice
      FuzzyLogic=No     AreaSensitive=No  Optimization=1
      Tolerance=0.8     ToleranceIncr=0.2 UncertainBands=ignore";
CmpSetSett(expername,cluSetting);
CmpCalcClust(expername);

// Module 2. Matching Band Count
string outfile;
string key1,key2,mbands,tempstr;
integer Nent,i,j,mbandcount;
integer check1, check2;
integer matches[][],Nband[];
FILE fp;
FPRINT fpr1,fpr2;

Nent=CmpGetEntryCount;
outfile=CountFileName;

if CountBands=0 then CountBands = 0;
else {
  if not(FileOpenWrite(fp,outfile)) then {
     message("ERROR: unable to create the file "+outfile);
     stop;
  }
  FileWrite(fp,"Entry1 Entry2 Bands1 Bands2 CommonBands");
  FileWriteLine(fp);

  for i=1 to Nent do {
     key1=CmpGetEntryKey(i);
     check1=FprLoadNorm(fpr1,key1,expername);
     Nband[i]=FprGetBandCount(fpr1);
     for j=i+1 to Nent do {
        key2=CmpGetEntryKey(j);
        FileWrite(fp,key1+"   "+key2+" ");
        check2=FprLoadNorm(fpr2,key2,expername);
        if check1 then
           FileWrite(fp,str(Nband[i],0,0)+" ");
        else
           FileWrite(fp,"---");
        if check2 then
           FileWrite(fp,str(FprGetBandCount(fpr2),0,0)+" ");
        else
           FileWrite(fp,"---");
        if check1*check2 then {
           FprMatchBands(fpr1,fpr2,mbands);
           tempstr=mbands;
           replace(tempstr,"-","--");
           mbandcount=length(tempstr)-length(mbands);
           FileWrite(fp,str(mbandcount,0,0));
           matches[i][j]=mbandcount;
           matches[j][i]=mbandcount;
           }
        else
           FileWrite(fp,"---");
        FileWriteLine(fp);
        setbusy("matching "+key1+" and "+key2);
     }
  }
  FileClose(fp);
//  execute("notepad.exe "+outfile);
}

// Module 3. Test Branches
integer nNMP,NMP[][];
BRANCH root,brTarget,brChild;
integer ok,ok2,ok3;
integer hasNMP,hasEnt1,hasEnt2,nCut;
integer tempNr;
string entNrBuff,lookfor;
integer l1,l2,k,t,z;
float N1,N2,N3,N4,N5,P1,P2,minP;

// Make NMP list (method: minP>alpha --> NMP, simple calc by monotonic)
//    problem: fact function allow only <100 to calculate factorial
for i=1 to (Nent-1) do {
   for j=i+1 to Nent do {
      l1 = Nband[i];
      l2 = Nband[j];
      k  = matches[i][j];

      N1 = fact(k);
      N2 = fact(l1)/fact(l1-k);
      N3 = fact(l2)/fact(l2-k);

      t  = Tmin;
      N4=1; for z=(t-l2+1) to t do N4 = N4 * z;
      N5=1; for z=(t-l1-l2+k+1) to (t-l1) do N5 = N5 * z;
      P1 = (N2*N3*N5)/(N1*N4);

      t = Tmax;
      N4=1; for z=(t-l2+1) to t do N4 = N4 * z;
      N5=1; for z=(t-l1-l2+k+1) to (t-l1) do N5 = N5 * z;
      P2 = (N2*N3*N5)/(N1*N4);

      if (P1 > P2) then minP = P2;
                   else minP = P1;
      if minP > alpha then {
           nNMP =  nNMP + 1;
           NMP[nNMP][1] = i;
           NMP[nNMP][2] = j;
      }
   }
}

// test each branch
ok=CluGetRoot(expername,root);
if nNMP = 0 then
   message("No cluster has an NMP: no division determined");
else {
   setbusy("Visiting clusters downward from the root ...");
   ok=CluEnumStart(root,brTarget);
   while ok do {
      entNrBuff="";
      ok2=CluEnumStart(brTarget,brChild);
      while ok2 do {
         tempNr = CluGetEntryNr(brChild);
         if tempNr > 0 then entNrBuff = entNrBuff+","+str(tempNr,0,0);
         ok2=CluEnumNext(brChild);
      }
      entNrBuff = entNrBuff+",";
      hasNMP = 0;
      for i=1 to nNMP do {
         lookfor = ","+str(NMP[i][1],0,0)+",";
         hasEnt1=find(entNrBuff,lookfor,1);
         lookfor = ","+str(NMP[i][2],0,0)+",";
         hasEnt2=find(entNrBuff,lookfor,1);
         if ((hasEnt1>0) and (hasEnt2>0)) then hasNMP = hasNMP + 1;
      }
      if hasNMP=0 then CluCutoff(brTarget,0);
      else {
         CluCutoff(brTarget,1);
         nCut = nCut +1;
      }
      ok=CluEnumNext(brTarget);
   }
   message(str(nCut,0,0)+" branches have NMPs and cut-off");
}
```

Appendix E. R Script Code for Cluster Analysis

```
#======================================================================
#    TITLE: Cluster Analysis ver. 1.5 Generic.R
#    Version of R flatform: R 2.0.1
#    COPYRIGHT: Young-Gun Zo, 2005, Univ. Maryland Biotech. Inst.
#    FLOW:
#       Caution 1: Run by Blocks marked by double lines
#       Caution 2: Manual Step in Phylip Tree contruction, marked by
#                  "~~~"
#=== SYNOPSIS =========================================================
#  1. User Input and Program Environments Setting
#  2. Data Input and Record
#  3. Generation of Trees
#  4. Compilation of Tree Nodes
#  5. Node Matching among Trees
#  6. Unique Node List = Cluster List
#  7. Test of Clusters for Significance of deviation from random band
#     matching
#  8. Test of Clusters for Significance of deviation from uniform
#     distribution
#  9. Test of Clusters for Significance of permutation tail probability
# 10. Determination of Terminal Clusters
# 11. Export for Neighbor Net construction

#======================================================================
#    1.1 USER INPUTS
#======================================================================
# 0. Clear workspace
       rm(list = ls())
# 1. Work Path
       path.w = "D:/CL/Group2-1.5Gen/"
# 2. Band Intensity and Strain File Names
#   (PRE: Gel Compar II export from Comparison Module,tab-delimited and
#         with header row)
       file.band = "Bands Group2.txt"
       file.otus = "Strain Info Group2.txt"
# 3. API script file path
       file.api = "D:/CL/API.R"
# 4. Option for Cluster Significance by Occurrence in Multiple Trees
       option.multi = TRUE # or FALSE
# 5. Band Size Range (bp)
       band.size.min = 420
       band.size.max = 629

#======================================================================
#    1.2 R-PROGRAM OPTIONS
#======================================================================
# 1. R Environment options
       options(expressions=10000) # set allowed recursion depth
       path.r = Sys.getenv("R_HOME")
# 2. Load PACKAGES
       library(ade4)
       library(ape)
```

```
# 3. Include Scripts
       source(file.api)

#=======================================================================
#    1.3 CONVENIENCE FUNCTIONS
#=======================================================================
# 1. Easy concatenation of strings
pa      <- function(...) paste(...,sep="",collapse="")
# 2. Take substrings before or after separator from a character vector
left.of <- function(x,delim)
sapply(strsplit(as.character(x),delim),first)
right.of<- function(x,delim)
sapply(strsplit(as.character(x),delim),second)

#=======================================================================
#    2.1 Band Data Input
#=======================================================================
# Read Binary Data, OTU names and other info (as rowName of band
character table)
bands     <-
read.table(pa(path.w,file.band),sep="\t",comment.char="",header=TRUE)
otus.info <-
read.table(pa(path.w,file.otus),sep="\t",comment.char="",header=TRUE)

# Name data matrix: rows and columns
names.temp  <-  left.of(otus.info$Key,"/")
names       <-  left.of(names.temp," ")
bands.names <- right.of(colnames(bands),"BOXi.")
colnames(bands)<-bands.names
rownames(bands)<-names

#=======================================================================
#    2.2 Re-code Bands Intensity
#=======================================================================
# Make binary coding
bands.bin = (bands > 0) * 1 # *1 to coerce to 0/1 from TF

# Make relative intensity coding and transform
# Data Management
#   Transformation of band data
bands.sum = rowSums(bands)
bands.rel = bands / bands.sum
bands.sqr = bands.rel^(1/2)
#   check distribution
temp = unlist(bands.sqr)
bands.sqr.nonzero = temp[temp>0]
hist(bands.sqr.nonzero)
bands.sqr.nonzero.SW=shapiro.test(bands.sqr.nonzero)
bands.sqr.nonzero.SW

#=======================================================================
#    3. Build PHYLO (an APE class) Trees by Cluster Analyses
#=======================================================================
# Make distance matrix from band intensities
dist.dice <- dist.binary(data.frame(bands.bin),5)
             #calculate sqrt(1-S) distance where 5=Sorensen=Dice
```

```
# Clustering by 'hclust' function in 'stats' package of R
#     Possible methods:
#        '"ward"', '"single"','"complete"', '"average"', '"mcquitty"',
#        '"median"' or '"centroid"'
#     Selected methods:
#        '"ward"'=WD, '"single"'=SG,'"complete"'=CL,'"mcquitty"'=MC
my.hclust.methods = c("ward","single","complete","mcquitty")
my.hclust.symbols = c("WD","SG","CL","MC")
trees.hclust       <- lapply(my.hclust.methods,hclust,d=dist.dice)
trees.hclust.phylo <- lapply(trees.hclust,as.phylo)
names(trees.hclust.phylo) = my.hclust.symbols

# Clustering by PHYLIP
#     Read as Phylo Tree <- Newick format Tree files
# **** Function to read Newick tree files and make a list of trees
read.treez<-function(filepath.vector,treename.vector){
   library(ape)
   tree.count = 0
   all.trees <- list()
   all.names <- NA
   for (f in 1:length(filepath.vector)){
      f.trees <- NULL
      f.trees <- read.tree(filepath.vector[f])
      if (!identical(f.trees,NULL)){
         if (is.element("multi.tree",attr(f.trees,"class"))){
            for (t in 1:length(f.trees)){
               tree.count =  tree.count + 1
               all.trees[[tree.count]] <- f.trees[[t]]
               all.names[tree.count]    =
paste(treename.vector[f],"_",as.character(t),sep="")
            } # end t:tree
         } # end if
         else { # single tree phylo
            tree.count =  tree.count + 1
            all.trees[[tree.count]] <- f.trees
            all.names[tree.count]    = treename.vector[f]
         }
      } # end if
   } # end f:file
   names(all.trees) <- all.names
   all.trees
} # end function
# **************
#     Selected methods:
#         Max Parsimony=MP,Max Likelihood=ML,Neighbor-Joining=NJ,
#         UPGMA=UP,FITCH-MARGOLIASH =FT
#     Step 1. Write data files as PHYLIP format
phylip.write(data.table=bands.bin,filename=pa(path.w,"pars.infile"),deli
m="",digits=1)
phylip.write(data.table=bands.sqr,filename=pa(path.w,"contml.infile"),de
lim=" ",digits=1)
phylip.write(data.table=dist.dice,filename=pa(path.w,"dice.infile"),deli
m=" ",digits=3)
# ~~~ Step 2. Run PHYLIP: manually run pars.exe, contml.exe,
#                         neighbor.exe, and fitch.exe
```

```
# ~~~         (at fitch power = 2),
# ~~~         and rename outtree as corresponding names as Step 3.
#     Step 3. Import resulting trees as Phylog format from Newick format
# Procedure to read multiple Newick trees
trees.import.symbols <- c("MP","ML","NJ","UP","FT")
trees.import.methods <- c("pars","contml","nj","upgma","fitch")
trees.import.files <-
paste(path.w,trees.import.methods,".outtree",sep="")
trees.import.phylo <-
read.treez(trees.import.files,trees.import.symbols)

# Collect all trees in the list 'treez'
treez <- c(trees.hclust.phylo,trees.import.phylo)

# Make factor of tree building methods
treez.methods <- substr(names(treez),1,2)
treez.methods <- factor(treez.methods)
     # --> Levels: CL FT MC ML MP NJ SG UP WD

#======================================================================
#    4. Compilation of Nodes and Node Information
#======================================================================
# Compile Node list
# output structure: nodez--> tree list (as treez) --> info list --> data
vector

# *** Function to get list of child nodes recursively from modified
phylo edge
# NOTE: for deep nesting trees, set options(expressions=10000) or higher
get.child.nodes<-function(edges,parent.node){
   # edge.list: numeric, reverse sign of phylo edge, parent.node:
positive integer
   out      <- c()
   children <- edges[edges[,1]==parent.node,]
   for (i in 1:nrow(children)){
      if (children[i,2]<0) # arrivtted to terminal
         out = c(out,children[i,2])
      else
         out = c(out,children[i,2],get.child.nodes(edges,children[i,2]))
   } # end i:children
   out
} # end function
# **************

# *** Function to compile Phylo Edge list as Node (=cluster) list from
Tree list
get.node.info<-function(tree.list){
   all.trees <- list()
   for (t in 1:length(tree.list)){
      print(paste("Processing tree named ",names(tree.list[t]),sep=""))
      edges  <- apply(tree.list[[t]]$edge,2,as.numeric) * (-1)
      labels <- tree.list[[t]]$tip.label
      #--> now, edges array has node>0, leaf<0 like hclust class

      child.direct    <- list()
      child.nodes     <- list()
```

```
      child.leaves    <- list()
      child.partitions<- list()

      for (i in 1:max(edges[,1])){
         child.direct[[i]] <- edges[edges[,1]==i,2]
         child.nodes[[i]]  <- get.child.nodes(edges,i)
         child.leaves[[i]] <-
labels[abs(child.nodes[[i]][child.nodes[[i]]<0])]
      } # end i node cycles
      all.trees[[t]] <-
list(ChildDirect=child.direct,ChildNodes=child.nodes,Leaves=child.leaves
)
   } # end t tree cycles

   # revisit trees to obtain more info using data from previous visits
   for (t in 1:length(tree.list)){
      labels  <- tree.list[[t]]$tip.label
      directs <- all.trees[[t]]$ChildDirect
      nodes   <- all.trees[[t]]$ChildNodes
      leaves  <- all.trees[[t]]$Leaves
      for (i in 1:length(directs)){
         partitions <- list()
         for (p in 1:length(directs[[i]])){
            a.child <- directs[[i]][p]
            if (a.child<0) # child is leaf
               partitions[[p]] <- labels[abs(a.child)]
            else
               partitions[[p]] <- leaves[[a.child]]
         } # end p partition cycles
         child.partitions[[i]] <- partitions
      } # end i node cycles
      all.trees[[t]] <-
list(ChildDirect=directs,ChildNodes=nodes,Leaves=leaves,Partitions=child
.partitions)
   } # end t tree cycles
   names(all.trees)<-names(tree.list)
   all.trees
}# end function
# *************

# Procedure to Compile Trees as Node and Leaf Information
treez.nodez <- get.node.info(treez)

# size of nodez
treez.nodez.size = list()
for (t in
1:length(treez.nodez))treez.nodez.size[[t]]=sapply(treez.nodez[[t]]$Leav
es,length)

#======================================================================
#     5. Node Matching Among Multiple Trees
#======================================================================

# **** Function to match node from a comma separated leaf list string
#      to produce cross product matrix of T/F
nodes.cross <- function(leaf.list1,leaf.list2){
   list1 <-
unlist(lapply(lapply(leaf.list1,sort),paste,sep="",collapse=","))
   list2 <-
unlist(lapply(lapply(leaf.list2,sort),paste,sep="",collapse=","))
   outer(list1,list2,FUN="==")
} # end function

# **** Function to get matching nodes from other trees for a specific
node
#      returns list of 2D-arrays for each tree; treez.nodez structure
trees.cross <- function(trees){
   # input is treez.nodez structure
   result <- list()
   for (i in 1:length(trees)){
      #print(paste("i = ",i,"th of ",length(trees)," trees",sep=""))
      match.table <-c()
      match.names <-c()
      row.count = length(trees[[i]]$Leaves)
      for (j in 1:length(trees)){
         #print(paste("j = ",j,"th of ",length(trees)," trees",sep=""))
         x = nodes.cross(trees[[i]]$Leaves,trees[[j]]$Leaves)
         if (sum(x)==0){y=numeric(row.count);y[1:row.count]=NA
            } else {y = as.numeric(apply(x,1,which))}
         #print(y)
         match.table = cbind(match.table,y)
         match.names = c(match.names,names(trees[j]))
      } # end 2nd trees
      #print(match.table)
      #print(match.names)
      colnames(match.table)<-match.names
      result[[i]]<-match.table
   } # end tree
   names(result)<-names(trees)
   result
} # end function

treez.matchez <- trees.cross(treez.nodez)

#======================================================================
#    6. Non-Redundant List of Clusters (Unique Nodes)
#======================================================================
# Collect Non-redundant list
# **** Function to collect nodes without redundancy by leaf content
#      using matches table: enlist when it appear first time among trees
#      output:2d-table, nodes by matching (trees=columns),
cells=matching node index
unique.nodes <- function(matches){
   # input is: trees>match table = treez.matchez
   # take all nodes from the first tree
   result <- matches[[1]]

   # loop through subsequent trees
   for (t in 2:length(matches)){
      M <- matches[[t]]
      # remove rows having matches to previous trees
      pre.nodes <- !is.na(M[,1:(t-1)])
      if (is.array(pre.nodes))
         new.nodes <- rowSums(pre.nodes)==0
      else
         new.nodes <- pre.nodes==0
      M <- M[new.nodes,]
      result <- rbind(result,M)
   } # end t: tree cycle
   result
} # end function
# **************
treez.clusterz <- unique.nodes(treez.matchez)  # = unique nodes matching
table
treez.clusterz.bin <- !is.na(treez.clusterz)*1
clusterz.n <- nrow(treez.clusterz)

# Match frequency by Method
treez.levels<-levels(treez.methods)
treez.clusterz.methods<-
array(0,c(nrow(treez.clusterz),length(treez.levels)))
colnames(treez.clusterz.methods)<-treez.levels
for (i in 1:length(treez.levels)){
   if (sum(treez.methods==treez.levels[i])==1)
      treez.clusterz.methods[,i] <-
treez.clusterz.bin[,treez.methods==treez.levels[i]]
   else
      treez.clusterz.methods[,i]<-
rowSums(treez.clusterz.bin[,treez.methods==treez.levels[i]])
}
treez.clusterz.methods.bin<-(treez.clusterz.methods>0)*1
rownames(treez.clusterz.methods.bin)=1:nrow(treez.clusterz.methods.bin)

# Count Consensus among different Methods
treez.clusterz.methods.con <- rowSums(treez.clusterz.methods.bin)
names(treez.clusterz.methods.con)=1:length(treez.clusterz.methods.con)

# Make Leaf List for each cluster
# **** Function to convert treez.clusterz match table into
#      dataframe(cluster index number,tree, node) two column table
notNA.rows <- function(vec) which(!is.na(vec)) # function to be applied
convert.serial <- function (clusterz){
   C <- c();  T <- c();    N <- c()
   OK <- apply(clusterz,2,notNA.rows) #list of row indexes of clusterz
   TreeNames <- names(OK)
   for (i in 1:length(OK)){
      C <- c(C,OK[[i]])
      T <- c(T,rep(TreeNames[i],length(OK[[i]])))
      N <- c(N,clusterz[OK[[i]],TreeNames[i]])
   }
   data.frame(Cluster=C,Tree=T,Node=N)
} # end function

treez.clusterz.serial <- convert.serial(treez.clusterz)

# Extract one exemplar node for a cluster
A <-treez.clusterz.serial; n<-nrow(treez.clusterz)
C <- c(); T <- c(); N <- c()
```

```
for (i in 1:n) {j=which(A$Cluster==i)[1]; C[i]=i;
T[i]=as.character(A[j,2]); N[i]=A[j,3]}
treez.clusterz.serial.first <- data.frame(Cluster=C,Tree=T,Node=N)
rm(A); rm(C); rm(T); rm(N)

# Make Leaf List for clusters --> clusterz
clusterz.leaves <- list(); A<-treez.clusterz.serial.first
for (i in 1:nrow(A)) clusterz.leaves[[i]] <-
sort(treez.nodez[[as.character(A$Tree[i])]]$Leaves[[A$Node[i]]])
rm(A)
names(clusterz.leaves)<-
paste(rep("Cluster",length(clusterz.leaves)),c(1:length(clusterz.leaves)
),sep="")
clusterz.leaves.cs <-
unlist(lapply(clusterz.leaves,paste,sep="",collapse=","))
clusterz.size      <- unlist(lapply(clusterz.leaves,length))

#======================================================================
#    7. Test of Singnificance of Clusters by Hypothesis
#                        of Radom Matching among Uniform Band Classes
#======================================================================
# *** proc to test significance under assumption of
#            random band matching over uniform distribution of bands
# ----------------------------------------------------------------------
--
# PRE: Get binary data of band presense/absence
#      matrix as row of OTUs x columns of band classes(=positions)
#      value of matrix can be binary, order or continuous
#            as long as absence of band is marked as zero
# TASK: calculate pairwise Pmax for random matching of bands
#       from min number of total band classes
#       and test for cluster by Holm's adjustment of
#       pairwise P to cluster wise
# POST: TRUE for significance all adjusted pairwise
#       Pmax < alpha (the type I error)
#       FALSE when any one of pairs give Pmax >=alpha after adjustment
# PACKAGES required: stats (as.dist;p.adjust)
# ----------------------------------------------------------------------
-

# function to get Pmax from Equation 3
pairwise.Pmax <- function(i,j,k,t.min,t.max){
   t = t.min:t.max
   Nall.1 = choose(t,i)
   Nall.2 = choose(t,j)
   Ncomb.1 = choose(t,k)
   Ncomb.2 = choose(t-k,i-k)
   Ncomb.3 = choose(t-i,j-k)
   P = (Ncomb.1*Ncomb.2*Ncomb.3)/(Nall.1*Nall.2)
   max(P)
}
# function to get Pmax from matrix of pairs
is.significant.band.matching <-
function(band.table,t.min,t.max,alpha=0.05){
   # get count of total bands and common bands
   bands.bin   <- band.table>0
   bands.match <- bands.bin %*% t(bands.bin)
   total.count <- diag(bands.match)

   # get matrix of band counts
   k           <- bands.match # number of common bands
   i           <-
array(total.count,c(length(total.count),length(total.count)))
   j           <- t(i) # i = band count for lane 1, j = for lane 2
   # get P matrix for each condition of t (t=all possible band sizes)
   Pmax.matrix = array(NA,dim(k))
   for (c in 1:(ncol(k)-1)){
      for (r in (c+1):nrow(k)){
         Pmax.matrix[r,c] =
pairwise.Pmax(i[r,c],j[r,c],k[r,c],band.size.min,band.size.max)
         Pmax.matrix[c,r] = Pmax.matrix[r,c]
      }
   }

   # adjust Pmax matrix by Holm's method
   Pmax.dist     <- as.dist(Pmax.matrix)    # dist format is useful to
identify pairs
   Pmax.adjusted <- p.adjust(Pmax.dist,method="holm")
   attributes(Pmax.adjusted)<-attributes(Pmax.dist) # apply dist format
to output vector

   # check if all pairs are significant
   Pmax.significance <- (Pmax.adjusted < alpha)
   if(sum(Pmax.significance)==length(Pmax.significance))
       (result<-TRUE) else (result<-FALSE)
   result
} # end function

# Procedures to Filter Clusters by Pmax for Random Band Matching
# Input data: clusterz.leaves as OTU list in clusters, bands.bin as
character table
clusterz.BMsig<-c()
for (i in 1:length(clusterz.leaves))
   clusterz.BMsig[i]=is.significant.band.matching(
                        bands.bin[clusterz.leaves[[i]],],
                        t.min = band.size.min,
                        t.max = band.size.max,
                        alpha = 0.05)
clusterz.tests<-data.frame(BandMatching=clusterz.BMsig)


#=======================================================================
#    8. Significance by appearance in multiple trees (option.multi)
#=======================================================================
if(option.multi){
# Derived from analysis in zMetaTree0.8.R
# Conditions of significanct cluters when they occur in Multiple
independent trees
#    at alpha=0.01; following number of cluster size (n), tree number
(t) is significant
n<-c( 2,  3,  4 , 5 , 6 , 7 , 8 , 9 ,10, 11, 12, 13, 14, 15, 16, 17, 18,
19, 20, 21 )
```

```
t<-c( 5,  4,  3 , 3 , 3 , 3 , 3 , 3 , 2,  2,  2,  2,  2,  2,  2,  2,  2,
2,  2,  1 )
multitree.test <- data.frame(ClusterSize=n,MinTreeNum=t)
# Summary of significance criteria
  # 1. n > 20: all significant cluster
  # 2. n > 4 : three independent trees
  # 3. n > 2 : four independent trees
  # 4. n = 2 : five independent trees
# Independence of trees in their methods
#      Levels: CL FT MC ML MP NJ SG UP WD
#      data  :  d  d  d  i  b  d  d  d  d, b=binary, i=band intensity,
d=Dice distance
#    criteria:  h  m  c  o  m  m  s  c  m, c=centeroid, h=homogeniety,
m=minimum, s=similarity
#       class:  1  2  3  4  5  2  6  3  2
#   dependent: CL FT MC ML MP NJ SG UP WD (d=the same data, m=the same
min. evolution contraint)
          #CL   x  d  d  0  0  d  d  d  d
          #FT   d  x  d  0  m  d  d  d  d
          #MC   d  d  x  0  0  d  d  d  d
          #ML   0  0  0  x  0  0  0  0  0
          #MP   0  m  0  0  x  m  0  0  m
          #NJ   d  d  d  0  m  x  d  d  d
          #SG   d  d  d  0  0  d  x  d  d
          #UP   d  d  d  0  0  d  d  x  d
          #WD   d  d  d  0  m  d  d  d  x

# Input variables: treez.clusterz.methods.bin,
treez.clusterz.methods.con
clusterz.methods <- as.data.frame(treez.clusterz.methods.bin)
clusterz.methods$consensus<-treez.clusterz.methods.con
attach(clusterz.methods)
#  independent pairs
   # ML vs. all others
     clusterz.methods$MLpair <- ((ML)&(consensus>1))
   # MP vs. {CL,MC,ML,SG,UP}
     clusterz.methods$MPpair <- ((MP)&(CL|MC|ML|SG|UP))
#  independent trios
   # ML+MP with {CL,MC,SG,UP}, not with{FT,MJ,WD}=minimum evolution
     clusterz.methods$MLPtrio<- (ML & MP & (CL|MC|SG|UP))
#  independent 5 trees: MLPTrio + 2 more = MLP + 3 more
   # ML+MP with {CL,MC,SG,UP}C3
     clusterz.methods$MLPand3<- (ML & MP &
((CL|MC|SG)|(CL|MC|UP)|(CL|SG|UP)|(MC|SG|UP)))

detach(clusterz.methods)
attach(clusterz.methods) # to refresh
clusterz.multitree.sig <- rep(NA,length(clusterz.size))
  # 4. n = 2 : five independent trees;
clusterz.multitree.sig[(clusterz.size==2)&(MLPand3)]=TRUE
clusterz.multitree.sig[(clusterz.size==2)&!(MLPand3)]=FALSE
  # 3. n > 2 : three independent trees
clusterz.multitree.sig[(clusterz.size>2)&(MLPtrio)]=TRUE
clusterz.multitree.sig[(clusterz.size>2)&!(MLPtrio)]=FALSE
  # 2. n > 4 : two independent trees
clusterz.multitree.sig[(clusterz.size>4)&(MLpair|MPpair)]=TRUE
clusterz.multitree.sig[(clusterz.size>4)&!(MLpair|MPpair)]=FALSE
  # 1. n > 20: all significant cluster
clusterz.multitree.sig[clusterz.size>20]=TRUE

detach(clusterz.methods)
clusterz.tests$Multi<-clusterz.multitree.sig

} # end if option.multi

#======================================================================
#    9. Permutation Tail Probability (PTP) Test by PAUP
#======================================================================
# SOURCE:
#     PTP: Archie, 1989; Faith and Cranston 1991
# TASK:
#     1. Write Nexus file for PTP test for all clusters of interest
#     2. Read PAUP Result file and Parse to make vector
# NOTE:
#     1. Cluster size has to be >=4 for PTP test

# ----- Write NEXUS File -----
# Inputs required
# 1. binary character table and OTU name list: bands.bin and names from
above
     # bands.bin # ensure 0, 1 coding rather than T/F
     otus <- rownames(bands.bin)
# 2. list of clusters of interest: list object from above,
#                        filtered >=4 & <=50 for PTP test
     clusters <- list(); clusters.names <- c(); j=0
     for (i in 1:length(clusterz.leaves)){
        if
((length(clusterz.leaves[[i]])>=4)&&(length(clusterz.leaves[[i]])<=50))
{
            j=j+1
            clusters[[j]]<-clusterz.leaves[[i]]
            clusters.names[j] <- names(clusterz.leaves[i])
        }
     } # end i
     names(clusters)<-clusters.names # need cluster name for paup
writing
     # NOTE: clusterz = total 733 clusters; clusters = 4<= size <=50 for
paup
     # resulting 418 clusters
# 3. Nexus command file name and cluster index to be contained in the
file
     Nexus.cmd.file = pa(path.w,"clusters.ptp.NEX")
# 4. PAUP output files name
     Nexus.out.file= pa(path.w,"clusters.ptp.log")

# using API.R
paup.write.ptp(Nexus.cmd.file,bands.bin,clusters)
paup.run(Nexus.cmd.file)
clusterz.ptp       <- c(array(NA,c(length(clusterz.leaves),1))) #make
default as NA
names(clusterz.ptp)<- names(clusterz.leaves)
clusters.ptp.read  <- paup.read.ptp(Nexus.out.file)
```

```
clusterz.ptp[names(clusters.ptp.read)]<- clusters.ptp.read
clusterz.ptp[clusterz.size>50]     = min(clusters.ptp.read)
            # because all >50 are all cases tested zero in p value
save(clusterz.ptp,file=pa(path.w,"cluster.ptp.Rdata"))
clusterz.tests$PTP <- (clusterz.ptp<0.05)
clusterz.tests

#======================================================================
#    10. Determination of Terminal Clusters
#======================================================================
# Objective: determine network structure of clusters
#            without phylogenetic signals and singletons

# 1. Filter clusters by permutation significance
#    Filter = when clusterz.tests values are as follows:
#  -----------------------------------------
#  Test  BandMatching Multitree  PTP(n>=4)
#  -----------------------------------------
#  Value    TRUE        TRUE      FALSE
#  -----------------------------------------
attach(clusterz.tests)
if(option.multi) {clusters.id =
which((BandMatching)&(is.na(Multi)|Multi)&((!PTP)|is.na(PTP)))
} else {clusters.id = which((BandMatching)&((!PTP)|is.na(PTP)))}
detach(clusterz.tests)
paste(length(clusters.id),"clusters out of ",nrow(clusterz.tests)," are
significant")
if (length(clusters.id)==0) stop("Halt because no significant cluster is
found")

# compile data set; extract insignificant clusters
clusterz.info <- data.frame(
        Cluster = 1:length(clusterz.size),
        Size    = clusterz.size,
        LeafList= clusterz.leaves.cs)
clusters.info <- clusterz.info[clusters.id, ]
clusters.info

# Find nesting structure among significant clusters
is.included<-function(cs.list1,cs.list2){
   # return if set 1 is included (=nested) in set 2
   set1 <-strsplit(as.character(cs.list1),",") # list of vectors
   set2 <-strsplit(as.character(cs.list2),",") # list of vectors
   d    <-c()
   for (i in 1:length(set1)) d[i]<-
(length(setdiff(set1[[i]],set2[[i]]))==0)
   d # if true, set1 is included in set2
     # row is the set, from which column set is subtracted
     # if true, row set is nested in column set, except for diagonal =
self to self
} # end function
  # determine nested clusters
attach(clusters.info)
clusters.nesting <- outer(LeafList,LeafList,"is.included")
detach(clusters.info)
diag(clusters.nesting) <- 0
```

```
paste(sum(clusters.nesting)," cases of nesting identified")
clusters.nesting

# remove nested cluster, and leave the top clusters only
  # by the row, remove anything rowsum>0
clusters.is.nested <- rowSums(clusters.nesting)>0
paste(sum(clusters.is.nested)," clusters were nested")
paste(" into ",sum(colSums(clusters.nesting)>0)," clusters")
clusters.tops <- clusters.info[!clusters.is.nested,]
paste(nrow(clusters.tops)," unnested clusters were identified")
clusters.tops
paste("which contain ",
  length(unique(unlist(clusterz.leaves[clusters.tops$Cluster])))," 
OTUs")

# Find Disjunction/Intersecting clusters
  # function to find intersecting pairs (= non-disjunct)
is.intersecting<-function(cs.list1,cs.list2){
   set1 <-strsplit(as.character(cs.list1),",") # list of vectors
   set2 <-strsplit(as.character(cs.list2),",") # list of vectors
   x    <-c()
   for (i in 1:length(set1)) x[i]<-
(length(intersect(set1[[i]],set2[[i]]))>0)
   x # if true, set1 and set2 have intersect
} # end function
# Get intersect matrix
attach(clusters.tops)
clusters.cross          <- outer(LeafList,LeafList,"is.intersecting")
diag(clusters.cross)    <- 0     # diagonals = self-self intersection
clusters.cross.count    <- rowSums(clusters.cross*1)
detach(clusters.tops)
# Enumerate intersecting / disjunct clusters
attach(clusters.tops)
clusters.disjunct.id    <- which(clusters.cross.count==0)
paste(length(clusters.disjunct.id),"clusters are completely disjunct to 
others")
clusters.disjunct       <- clusters.tops[clusters.disjunct.id,]
clusters.intersect.id   <- 
setdiff(1:nrow(clusters.tops),clusters.disjunct.id)
paste(length(clusters.intersect.id),"clusters are intersecting with 
others")
clusters.intersect      <- clusters.tops[clusters.intersect.id,]
detach(clusters.tops)

# Test on Merger of intersecting clusters
if(nrow(clusters.intersect)==0){
   clusters.tops.reclass1 = clusters.tops
   clusters.tops.reclass2 = clusters.tops
   clusters.tops.reclass2$Fuzzy1 = 0
   clusters.tops.reclass2$Fuzzy2 = 0
   clusters.tops.reclass2$HasFuzzy = FALSE
} else {
   # If intersects, test PTP on merged cluster
   #    get pairs of clusters intersecting clusters and
   #    make merged (=union) clusters
   # ****** function to get list of intersects/intersecting
clusters/union cluster
   get.intersects<-function(cluster.info){
      # input is "info" data frame structure = $Cluster, $Size,
$LeafList
      # identify intersecting clusters by pairwise comparison (upper
triangle of matrix)
      # make intersect and info of intersecting clusters
      n = nrow(cluster.info)
      i <- c(); c1 <- c(); d1 <-c(); c2 <- c(); d2 <-c(); u <-c()
      k = 0
      for (r in 1:(n-1)){     # rows in upper triangle
         for (c in (r+1):n){  # columns in upper triangle
            set.r <-
unlist(strsplit(as.character(cluster.info$LeafList[r]),",")) # list of
vectors
            set.c <-
unlist(strsplit(as.character(cluster.info$LeafList[c]),",")) # list of
vectors
            set.i <- intersect(set.r,set.c)
            if (length(set.i)>0){
               k = k + 1
               i[k]= paste(sort(set.i),sep="",collapse=",")
               u[k]=
paste(sort(union(setdiff(set.r,set.i),union(setdiff(set.c,set.i),set.i))
),sep="",collapse=",")
               ct1 = cluster.info$Cluster[r]
               ct2 = cluster.info$Cluster[c]
               dt1 =
paste(sort(setdiff(set.r,set.i)),sep="",collapse=",")
               dt2 =
paste(sort(setdiff(set.c,set.i)),sep="",collapse=",")
               c1[k]=ct1; d1[k]=dt1; c2[k]=ct2; d2[k]=dt2
            } # end if
        } # end column
      } # end row
      out =
data.frame(Intersect=i,Cluster1=c1,Cluster2=c2,Left1=d1,Left2=d2,Merge=u
)
      rownames(out) = paste("Int",c1,"x",c2,sep="")
      out
   } # end function
   clusters.intersect.info = get.intersects(clusters.tops)
   clusters.intersect.info
   clusters.intersect.merger =
strsplit(as.character(clusters.intersect.info$Merge),",")
   names(clusters.intersect.merger) = rownames(clusters.intersect.info)
   clusters.intersect.merger
   #    run PTP test
   Nexus.cmd.file = pa(path.w,"intersecting.merger.ptp.NEX")
   Nexus.out.file= pa(path.w,"intersecting.merger.ptp.log")
   paup.write.ptp(Nexus.cmd.file,bands.bin,clusters.intersect.merger)
   paup.run(Nexus.cmd.file)
   #    read PTP test results
   clusters.intersect.ptp =
c(array(NA,c(nrow(clusters.intersect.info),1)))
```

```
   clusters.intersect.ptp[sapply(clusters.intersect.merger,length)>50] =
0
   names(clusters.intersect.ptp) = rownames(clusters.intersect.info)
   clusters.intersect.ptp.read   = paup.read.ptp(Nexus.out.file)

clusters.intersect.ptp[names(clusters.intersect.ptp.read)]=clusters.inte
rsect.ptp.read
   clusters.intersect.ptp
   clusters.intersect.ptp.sig = (clusters.intersect.ptp<0.05) # has
signal
   clusters.intersect.ptp.sig
   #    reclassify clusters: discrete clusters + fuzzy clusters
   #       if the merger PTP is not significant (FALSE): the merger
replaces the two clusters
   if(sum(clusters.intersect.ptp.sig==FALSE)==0) { # all mergers have
phylo. signal
      clusters.tops.reclass1 = clusters.tops # don't use any merger
   } else { # select nonsig. mergers and replace the original two
      to.merge.info =
clusters.intersect.info[!clusters.intersect.ptp.sig,]
      to.merge.leaf = to.merge.info$Merge; names(to.merge.leaf) =
rownames(to.merge.info)
      to.add.info =
data.frame(Cluster=0,Size=sapply(sapply(to.merge.leaf,strsplit,","),leng
th),
                            LeafList=to.merge.leaf)
      rownames(to.add.info)= names(to.merge.leaf)
      to.cut = union(to.merge.info$Cluster1,to.merge.info$Cluster2)
      clusters.tops.reclass1 =
clusters.tops[rowSums(outer(clusters.tops$Cluster,to.cut,"=="))==0,]
      clusters.tops.reclass1 = rbind(clusters.tops.reclass1,to.add.info)
   }
   #       if the merger PTP is significant (TRUE): classify the
intersect as fuzzy cluster
   if(sum(clusters.intersect.ptp.sig==TRUE)==0){
      clusters.tops.reclass2 = clusters.tops.reclass1
      clusters.tops.reclass2$Fuzzy1 = 0
      clusters.tops.reclass2$Fuzzy2 = 0
      clusters.tops.reclass2$HasFuzzy = FALSE
   } else {
      to.fuzzy.info =
clusters.intersect.info[clusters.intersect.ptp.sig,]
      # add intersects as fuzzy cluters
      to.add.fuzzy =
data.frame(Cluster=0,Size=sapply(sapply(to.fuzzy.info$Intersect,strsplit
,","),length),
                             LeafList=to.fuzzy.info$Intersect,
                             Fuzzy1=to.fuzzy.info$Cluster1,
                             Fuzzy2=to.fuzzy.info$Cluster2,
                             HasFuzzy=FALSE)
      rownames(to.add.fuzzy)=rownames(to.fuzzy.info)
      # add intersecting clusters as new clusters
      temp.cluster = numeric(); temp.size=numeric()
      temp.leaflist=character(); temp.partner = numeric()
      for (i in 1:nrow(to.fuzzy.info)){
         temp.cluster[i] = to.fuzzy.info$Cluster1[i]
         temp.size[i]    =
length(unlist(strsplit(as.character(to.fuzzy.info$Left1[i]),",")))
         temp.leaflist[i]= as.character(to.fuzzy.info$Left1[i])
         temp.partner[i] = to.fuzzy.info$Cluster2[i]
      }
      to.add.modify1 = data.frame(Cluster=temp.cluster,Size=temp.size,

LeafList=temp.leaflist,FuzzyPartner=temp.partner)
      temp.cluster = numeric(); temp.size=numeric()
      temp.leaflist=character(); temp.partner = numeric()
      for (i in 1:nrow(to.fuzzy.info)){
         temp.cluster[i] = to.fuzzy.info$Cluster2[i]
         temp.size[i]    =
length(unlist(strsplit(as.character(to.fuzzy.info$Left2[i]),",")))
         temp.leaflist[i]= as.character(to.fuzzy.info$Left2[i])
         temp.partner[i] = to.fuzzy.info$Cluster1[i]
      }
      to.add.modify2 = data.frame(Cluster=temp.cluster,Size=temp.size,

LeafList=temp.leaflist,FuzzyPartner=temp.partner)
      to.add.modify  = rbind(to.add.modify1,to.add.modify2)
      to.add.modify$Fuzzy1=0
      to.add.modify$Fuzzy2=0
      to.add.modify$HasFuzzy=TRUE
      rownames(to.add.modify) = paste("Cluster",to.add.modify$Cluster,"-
Fx",to.add.modify$FuzzyPartner,sep="")
      # remove fuzzy members from the orginal clusters
      to.cut =
rowSums(outer(clusters.tops.reclass1$Cluster,to.add.modify$Cluster,"==")
)>0
      clusters.tops.reclass2 = clusters.tops.reclass1
      clusters.tops.reclass2$Fuzzy1 = 0
      clusters.tops.reclass2$Fuzzy2 = 0
      clusters.tops.reclass2$HasFuzzy = FALSE
      clusters.tops.reclass2 = clusters.tops.reclass2[!to.cut,]
      clusters.tops.reclass2 = rbind(clusters.tops.reclass2,

to.add.modify[,colnames(to.add.modify)!="FuzzyPartner"],
                                  to.add.fuzzy)
   }
}
clusters.tops.reclass2

# Merge clusters and singletons as cluster format
clusters.otus = names; names(clusters.otus) = names
clusters.singletons.otus =
setdiff(clusters.otus,unlist(sapply(clusters.tops.reclass2$LeafList,strs
plit,",")))
if (length(clusters.singletons.otus)>0) {
   clusters.singletons =
data.frame(Cluster=0,Size=1,LeafList=clusters.singletons.otus)
   rownames(clusters.singletons) = clusters.singletons$LeafList
   terminals = rbind(clusters.tops.reclass1,clusters.singletons)
   clusters.singletons.reclass2 = clusters.singletons
   clusters.singletons.reclass2$Fuzzy1=0
   clusters.singletons.reclass2$Fuzzy2=0
   clusters.singletons.reclass2$HasFuzzy=FALSE
   terminals.fuzzy =
rbind(clusters.tops.reclass2,clusters.singletons.reclass2)
} else {
   clusters.singletons = NA
   terminals = clusters.tops.reclass1
   terminals.fuzzy = clusters.tops.reclass2
}
terminals
terminals.fuzzy

# List OTUs with cluster membership
members=character(); belong.to=character()
for (i in 1:nrow(terminals)){
   temp.members =
unlist(strsplit(as.character(terminals$LeafList[i]),","))
   temp.cluster = character(terminals$Size[i])
   temp.cluster[1:length(temp.cluster)] = rownames(terminals)[i]
   members = c(members,temp.members)
   belong.to = c(belong.to,temp.cluster)
}
terminals.membership.clusters =
data.frame(OTU=members,Terminal=belong.to)
#rownames(terminals.membership.clusters)=terminals.membership.clusters$O
TU
terminals.membership.clusters # members of fuzzy clusters has multiple
entries

# Fuzzy Set by OTU list
make.membership.table <- function(otus,leaves.cs){
   out = is.element(otus,unlist(strsplit(leaves.cs[1],",")))
   for (c in 2:length(leaves.cs)){
      out =
cbind(out,is.element(otus,unlist(strsplit(leaves.cs[c],","))))
   }
   colnames(out) = names(leaves.cs)
   rownames(out) = names(otus)
   out * 1
} # end function
terminals.leaves.cs = as.character(terminals$LeafList)
names(terminals.leaves.cs) = rownames(terminals)
terminals.membership.table =
make.membership.table(clusters.otus,terminals.leaves.cs)
terminals.membership.fuzzy = terminals.membership.table /
rowSums(terminals.membership.table)

# Clustering Summary
terminals
terminals.fuzzy
terminals.membership.clusters
terminals.membership.table
terminals.membership.fuzzy

fn=pa(path.w,"terminals.tab")
 write.table(terminals,fn,quote=FALSE,sep="\t")
fn=pa(path.w,"terminals.fuzzy.tab")
```

```
 write.table(terminals.fuzzy,fn,quote=FALSE,sep="\t")
fn=pa(path.w,"terminals.membership.clusters.tab")
 write.table(terminals.membership.clusters,fn,quote=FALSE,sep="\t")
fn=pa(path.w,"terminals.membership.table.tab")
 write.table(terminals.membership.table,fn,quote=FALSE,sep="\t")
fn=pa(path.w,"terminals.membership.fuzzy.tab")
 write.table(terminals.membership.fuzzy,fn,quote=FALSE,sep="\t")
```

References

Achaz, G., Coissac, E., Netter, P. & Rocha, E. P. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* **164**, 1279-1289.

Ali, A., Rashid, M. H. & Karaolis, D. K. R. (2002). High-frequency rugose exopolysaccharide production by *Vibrio cholerae*. *Applied and Environmental Microbiology* **68**, 5773-5778.

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

Ansaruzzaman, M., Shimada, T., Bhuiyan, N. A., Nahar, S., Alam, K., Islam, M. S. & Albert, M. J. (1999). Cross-reaction between a strain of *Vibrio mimicus* and *V. cholerae* O139 Bengal. *J Med Microbiol* **48**, 873-877.

Archie, J. W. (1989). A randomization test for phylogenetic information in systematic data. *Systematic Zoology* **38**, 239-252.

Bansal, A. K. & Meyer, T. E. (2002). Evolutionary analysis by whole-genome comparisons. *J Bacteriol* **184**, 2260-2272.

Barnes, R. D. (1987). *Invertebrate Zoology*, 5 edn. Philadelphia, P.A.: Saunders College Publishing.

Barua, D. (1992). History of cholera. In *Cholera*, vol., pp. 1-35. Edited by W. B. Greenough. New York: Plenum Publishing.

Barua, D. & Greenough III, W. B. (1992). *Cholera*. New York, NY: Plenum Publishing Corp.

Baumann, P. & Schubert, R. H. W. (1984). Family II. *Vibrionaceae*. In *Bergey's Manual of Systematic Bacteriology*, vol. 1, pp. 516-550. Edited by J. G. Holt. Baltimore, M.D.: Williams & Wilkins.

Baumann, P., Furniss, A. L. & Lee, J. V. (1984). Genus I. *Vibrio*. In *Bergey's Manual of Systematic Bacteriology*, vol. 1, pp. 518-538. Edited by J. G. Holt. Baltimore, M.D.: Williams & Wilkins.

Beltran, P., Delgado, G., Navarro, A., Trujillo, F., Selander, R. K. & Cravioto, A. (1999). Genetic diversity and population structure of *Vibrio cholerae*. *J. Clin. Microbiol.* **37**, 581-590.

Berry, V. & Gascuel, O. (1996). On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol Biol Evol* **13**, 999-1011.

Bik, E. M., Bunschoten, A. E., Gouw, R. D. & Mooi, F. R. (1995). Genesis of the novel epidemic *Vibrio cholerae* O139 strain: evidence for horizontal transfer of genes involved in polysaccharide synthesis. *EMBO J.* **14**, 209-216.

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M. & Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J Bacteriol* **186**, 3980-3990.

Bryant, D. & Moulton, V. (2004). Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21**, 255-265.

Byun, R., Elbourne, L. D., Lan, R. & Reeves, P. R. (1999). Evolutionary relationships of pathogenic clones of *Vibrio cholerae* by sequence analysis of four housekeeping genes. *Infect. Immun.* **67**, 1116-1124.

Chakraborty, S., Mukhopadhyay, A. K., Bhadra, R. K., Ghosh, A. N., Mitra, R., Shimada, T., Yamasaki, S., Faruque, S. M., Takeda, Y., Colwell, R. R. & Nair, G. B. (2000). Virulence genes in environmental strains of *Vibrio cholerae*. *Appl. Environ. Microbiol.* **66**, 4022-4028.

Checkley, W., Epstein, L. D., Gilman, R. H., Figueroa, D., Cama, R. I., Patz, J. A. & Black, R. E. (2000). Effect of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *Lancet* **355**, 442-450.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500.

Cholera Working Group, I. C. D. D. R., Bangladesh (1993). Large epidemic of cholera-like disease in Bangladesh caused by *Vibrio cholerae* O139 synonym Bengal. *Lancet* **342**, 387-390.

Choopun, N. (2004). *Population Structure of Vibrio cholerae in the Chesapeake Bay*, Ph. D. Thesis, University of Maryland.

Choopun, N., Louis, V., Huq, A. & Colwell, R. R. (2002). Simple procedure for rapid identification of *Vibrio cholerae* from the aquatic environment. *Appl. Environ. Microbiol.* **68**, 995-998.

Chun, J., Huq, A. & Colwell, R. R. (1999). Analysis of 16S-23S rRNA intergenic spacer regions of *Vibrio cholerae* and *Vibrio mimicus. Appl. Environ. Microbiol.* **65**, 2202-2208.

Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184**, 2072-2080.

Cohan, F. M. (2001). Bacterial species and speciation. *Syst Biol* **50**, 513-524.

Colombo, M. M., Mastrandrea, S., Leite, F., Santona, A., Uzzau, S., Rappelli, P., Pisano, M., Rubino, S. & Cappuccinelli, P. (1997). Tracking of clinical and environmental *Vibrio cholerae* O1 strains by combined analysis of the presence of toxin cassette, plasmid content and ERIC PCR. *FEMS Immunol. Med. Microbiol.* **19**, 33-45.

Colwell, R. R. (1996). Global climate and infectious disease: the cholera paradigm. *Science* **274**, 2025-2031.

Colwell, R. R. & Spira, W. M. (1992). The ecology of *Vibrio cholerae*. In *Cholera*, vol., pp. 107-127. Edited by W. B. Greenough. New York, N.Y.: Plenum Medical Book Co.

Colwell, R. R. & Huq, A. (1994). Environmental reservoir of *Vibrio cholerae*, the causative agent of cholera. *Ann N Y Acad Sci* **740**, 44-54.

Colwell, R. R. & Grimes, D. J. (2000). *Nonculturable Microorganisms in the Environment*. Washington, D.C.: American Society of Microbiology.

Colwell, R. R., Kaper, J. & Joseph, S. W. (1977). *Vibrio cholerae*, *Vibrio parahaemolyticus*, and other vibrios: occurrence and distribution in Chesapeake Bay. *Science* **198**, 394-396.

Colwell, R. R., Huq, A., Islam, M. S., Aziz, K. M., Yunus, M., Khan, N. H., Mahmud, A., Sack, R. B., Nair, G. B., Chakraborty, J., Sack, D. A. & Russek-Cohen, E. (2003). Reduction of cholera in Bangladeshi villages by simple filtration. *Proc Natl Acad Sci U S A* **100**, 1051-1055.

Dalsgaard, A., Forslund, A., Bodhidatta, L., Serichantalergs, O., Pitarangsi, C., Pang, L., Shimada, T. & Echeverria, P. (1999). A high proportion of *Vibrio cholerae* strains isolated from children with diarrhoea in Bangkok, Thailand are multiple antibiotic resistant and belong to heterogenous non-O1, non- O139 O-serotypes. *Epidemiology and Infection* **122**, 217-226.

Daubin, V., Moran, N. A. & Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829-832.

Davis, B. R., Fanning, R., Madden, J. M., Steigerwalt, A. G., Bradford, H. B. J., Smith, H. L. J. & Brenner, D. J. (1981). Characterization of biochemically atypical *Vibrio cholerae* strains and designation of a new pathogenic species, *Vibrio mimicus*. *J. Clin. Microbiol.* **14**, 631-639.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence Structure*, vol. 5, pp. 345-352. Edited by M. O. Dayhoff. Washington, D. C.: National Biomedical Research Foundation.

Denner, E. B., Vybiral, D., Fischer, U. R., Velimirov, B. & Busse, H. J. (2002). *Vibrio calviensis* sp. nov., a halophilic, facultatively oligotrophic 0.2 µm-filterable marine bacterium. *Int J Syst Evol Microbiol* **52**, 549-553.

Difco (1984). *Difco Manual: Dehydrated Culture Media and Reagents for Microbiology*. Detroit, M.I.: Difco Laboratories.

Dwivedi, S. N. (1993). Long-term variability in the food chains, biomass yields, and oceanography of the Bay of Bengal ecosystem. In *Large Marine Ecosystems: Stress, Mitigation, and Sustainability*, vol. Edited by K. Sherman, Alexander, L. M., and B. D. Gold. Washington, D.C.: AAAS.

Dziejman, M., Balon, E., Boyd, D., Fraser, C. M., Heidelberg, J. F. & Mekalanos, J. J. (2002). Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1556-1561.

Egan, E. S. & Waldor, M. K. (2003). Distinct replication requirements for the two *Vibrio cholerae* chromosomes. *Cell* **114**, 521-530.

Egan, E. S., Lobner-Olesen, A. & Waldor, M. K. (2004). Synchronous replication initiation of the two *Vibrio cholerae* chromosomes. *Curr Biol* **14**, R501-502.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185.

Excoffier, L., Smouse, P. E. & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

Faith, D. P. & Cranston, P. S. (1991). Could a cladogram this short have arisen by chance alone? - on permutation tests for cladistic structure. *Cladistics* **7**, 1-28.

Farfán, M., Minaña, D., Fusté, M. C. & Lorén, J. G. (2000). Genetic relationships between clinical and environmental *Vibrio cholerae* isolates based on multilocus enzyme electrophoresis. *Microbiology* **146**, 2613-2626.

Farfán, M., Minaña-Galbis, D., Fusté, M. C. & Lorén, J. G. (2002). Allelic diversity and population structure in *Vibrio cholerae* O139 Bengal based on nucleotide sequence analysis. *J. Bacteriol.* **184**, 1304-1313.

Faruque, S. M., Albert, M. J. & Mekalanos, J. J. (1998). Epidemiology, genetics, and ecology of toxigenic *Vibrio cholerae*. *Microbiol. Mol. Biol. Rev.* **62**, 1301-1314.

Faruque, S. M., Roy, S. K., Alim, A. R., Siddique, A. K. & Albert, M. J. (1995). Molecular epidemiology of toxigenic *Vibrio cholerae* in Bangladesh studied by numerical analysis of rRNA gene restriction patterns. *J. Clin. Microbiol.* **33**, 2833-2838.

Faruque, S. M., Rahman, M. M., Asadulghani, Nasirul Islam, K. M. & Mekalanos, J. J. (1999). Lysogenic conversion of environmental *Vibrio mimicus* strains by CTXΦ. *Infect. Immun.* **67**, 5723-5729.

Faruque, S. M., Sack, D. A., Sack, R. B., Colwell, R. R., Takeda, Y. & Nair, G. B. (2003a). Emergence and evolution of *Vibrio cholerae* O139. *Proc Natl Acad Sci U S A* **100**, 1304-1309.

Faruque, S. M., Kamruzzaman, M., Meraj, I. M., Chowdhury, N., Nair, G. B., Sack, R. B., Colwell, R. R. & Sack, D. A. (2003b). Pathogenic potential of environmental *Vibrio cholerae* strains carrying genetic variants of the toxin-coregulated pilus pathogenicity island. *Infect Immun* **71**, 1020-1025.

Faruque, S. M., Chowdhury, N., Kamruzzaman, M., Dziejman, M., Rahman, M. H., Sack, D. A., Nair, G. B. & Mekalanos, J. J. (2004). Genetic diversity and virulence potential of environmental *Vibrio cholerae* population in a cholera-endemic area. *Proc Natl Acad Sci U S A* **101**, 2123-2128.

Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P. J., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A., Moore, C. E., Zhou, J. J. & Spratt, B. G. (2001). Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4276-4276.

Felden, B., Massire, C., Westhof, E., Atkins, J. F. & Gesteland, R. F. (2001). Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Res* **29**, 1602-1607.

Felsenstein, J. (1985). Confidence limits of phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791.

Felsenstein, J. (2004a). PHYLIP: Phylogenetic Inference Package, Ver. 3.6. Seattle, W.A.: Department of Genome Sciences, University of Washington. http://evolution.genetics.washington.edu/phylip.html.

Felsenstein, J. (2004b). *Inferring Phylogenies*. Sunderland, M.A.: Sinauer Associates.

Galtier, N. (2004). Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. *Syst Biol* **53**, 38-46.

Garg, P., Aydanian, A., Smith, D., Morris, J. G., Nair, G. B. & Stine, O. C. (2003). Molecular epidemiology of O139 *Vibrio cholerae*: mutation, lateral gene transfer, and founder flush. *Emerg. Infect. Dis.* **9**, 810-814.

Garrity, G. M., Bell, J. A. & Lilburn, T. G. (2004). *Taxonomic Outline of the Prokarotes Release 5.0 of Bergey's Manual of Systematic Bacteriology*, 2 edn. New York, N.Y.: Springer.

Gauger, E. J. & Gomez-Chiarri, M. (2002). 16S ribosomal DNA sequencing confirms the synonymy of *Vibrio harveyi* and *V. carchariae*. *Dis Aquat Organ* **52**, 39-46.

Glass, R. I., Becker, S., Huq, M. I., Stoll, B. J., Khan, M. U., Merson, M. H., Lee, J. V. & Black, R. E. (1982). Endemic cholera in rural Bangladesh, 1966-1980. *Am. J. Epidemiol.* **116**, 959-970.

Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-2238.

Guerrant, R. L., Carneiro-Filho, B. A. & Dillingham, R. A. (2003). Cholera, diarrhea, and oral rehydration therapy: triumph and indictment. *Clin Infect Dis* **37**, 398-405.

Haas, E. S. & Brown, J. W. (1998). Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res* **26**, 4093-4099.

Haines, A. & Patz, J. A. (2004). Health effects of climate change. *Jama* **291**, 99-103.

Hasegawa, M., Kishino, H. & Yano, K. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., Gill, S. R., Nelson, K. E., Read, T. D., Tettelin, H., Richardson, D., Ermolaeva, M. D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R. D., Nierman, W. C., White, O., Salzberg, S. L., Smith, H. O., Colwell, R. R., Mekalanos, J. J., Venter, J. C. & Fraser, C. M. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477-483.

Hilali, F., Ruimy, R., Saulnier, P., Barnabé, C., Lebouguénec, C., Tibayrenc, M. & Andremont, A. (2000). Prevalence of virulence genes and clonality in *Escherichia coli* strains that cause bacteremia in cancer patients. *Infect. Immun.* **68**, 3983-3989.

Hillis, D. M. & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**, 182-192.

Hoge, C. W., Bodhidatta, L., Echeverria, P., Deesuwan, M. & Kitporka, P. (1996). Epidemiologic study of *Vibrio cholerae* O1 and O139 in Thailand: at the advancing edge of the eighth pandemic. *Am J Epidemiol* **143**, 263-268.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scan. J. Stat.* **6**, 65-70.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K. & Johnson, C. A. (2001). *Climate Change 2001: The Scientific Basis*. Cambridge, U.K.: Cambridge University Press.

Howard-Jones, N. (1984). Robert Koch and the cholera vibrio: a centenary. *British Medical Journal* **288**, 379-381.

Huq, A., Small, E. B., West, P. A., Rahman, R. & Colwell, R. R. (1983). Ecological relationships between *Vibrio cholerae* and planktonic crustacean copepods. *Appl Environ Microbiol* **45**, 275-283.

Huq, A., Parveen, S., Qadri, F., Sack, D. A. & Colwell, R. R. (1993). Comparison of *Vibrio cholerae* serotype O1 strains isolated from patients and the aquatic environment. *J. Trop. Med. Hyg.* **96**, 86-92.

Huq, A., Small, E. B., West, P. A., Huq, M. I., Rahman, R. & Colwell, R. R. (1984). Influence of water temperature, salinity, and pH on survival and growth of toxigenic *Vibrio cholerae* serovar O1 associated with live copepods in laboratory microcosms. *Appl. Environ. Microbiol.* **48**, 420-424.

Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., Morris, J. G., Jr., Khan, M. N., Siddique, A. K., Yunus, M., Albert, M. J., Sack, D. A. & Colwell, R. R. (2005). Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. *Appl Environ Microbiol* **71**, 4645-4654.

Huson, D. H. & Bryant, D. (2005). Estimating phylogenetic trees and networks using SplitsTree4. *in preparation*.

Islam, M. S., Rahim, Z., Alam, M. J., Begum, S., Moniruzzaman, S. M., Umeda, A., Amako, K., Albert, M. J., Sack, R. B., Huq, A. & Colwell, R. R. (1999). Association of *Vibrio cholerae* O1 with the cyanobacterium, *Anabaena* sp., elucidated by polymerase chain reaction and transmission electron microscopy. *Trans R Soc Trop Med Hyg* **93**, 36-40.

Ivanova, E. P., Flavier, S. & Christen, R. (2004). Phylogenetic relationships among marine *Alteromonas*-like proteobacteria: emended description of the family *Alteromonadaceae* and proposal of *Pseudoalteromonadaceae* fam. nov., *Colwelliaceae* fam. nov., *Shewanellaceae* fam. nov., *Moritellaceae* fam. nov., *Ferrimonadaceae* fam. nov., *Idiomarinaceae* fam. nov. and *Psychromonadaceae* fam. nov. *Int J Syst Evol Microbiol* **54**, 1773-1788.

Jammalamadaka, S. R. & SenGupta, A. (2001). *Topics in Circular Statistics*. Singapore: World Scientific Publishing Co.

Jiang, S. C., Matte, M., Matte, G., Huq, A. & Colwell, R. R. (2000). Genetic diversity of clinical and environmental isolates of *Vibrio cholerae* determined by amplified fragment length polymorphism fingerprinting. *Appl. Environ. Microbiol.* **66**, 148-153.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, vol., pp. 21-132. Edited by H. N. Munro. New York: Academic Press.

Kaper, J., Lockman, H., Colwell, R. R. & Joseph, S. W. (1979). Ecology, serology, and enterotoxin production of *Vibrio cholerae* in Chesapeake Bay. *Appl. Environ. Microbiol.* **37**, 91-103.

Kaper, J. B., Morris, J. G. & Levine, M. M. (1995). Cholera. *Clin. Microbiol. Rev.* **8**, 48-86.

Kaper, J. B., Bradford, H. B., Roberts, N. C. & Falkow, S. (1982). Molecular epidemiology of *Vibrio cholerae* in the U.S. Gulf Coast. *J. Clin. Microbiol. Rev.* **8**, 48-86.

Kaper, J. B., Nataro, J. P., Roberts, N. C., Siebeling, R. J. & Bradford, H. B. (1986). Molecular epidemiology of non-O1 *Vibrio cholerae* and *Vibrio mimicus* in the U.S. Gulf Coast region. *J Clin Microbiol* **23**, 652-654.

Karaolis, D. K., Lan, R. & Reeves, P. R. (1995). The sixth and seventh cholera pandemics are due to independent clones separately derived from environmental, nontoxigenic, non-O1 *Vibrio cholerae*. *J. Bacteriol.* **177**, 3191-3198.

Karaolis, D. K., Lan, R., Kaper, J. B. & Reeves, P. R. (2001). Comparison of *Vibrio cholerae* pathogenicity islands in sixth and seventh pandemic strains. *Infect. Immun.* **69**, 1947-1952.

Karaolis, D. K., Somara, S., Maneval, D. R., Jr., Johnson, J. A. & Kaper, J. B. (1999). A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* **399**, 375-379.

Kay, B. A., Bopp, C. A. & Wells, J. G. (1994). Isolation and identification of *Vibrio cholerae* O1 from fecal specimens. In *Vibrio cholerae and Cholera: Molecular to Global Perspectives*, vol., pp. 3-20. Edited by O. Olsvik. Washington, D.C.: ASM Press.

Khan, M. U. (1982). Efficacy of short course antibiotic prophylaxis in controlling cholera in contacts during epidemic. *J Trop Med Hyg* **85**, 27-29.

Kim, J. (1993). Improving the accuracy of phylogenetic estimation by combining different methods. *Syst. Biol.* **42**, 331-340.

Kimmel, D. G. & Roman, M. R. (2004). Long-term trends in mesozooplankton abundance in Chesapeake Bay, USA: influence of freshwater input. *Marine Ecology-Progress Series* **267**, 71-83.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244-1245.

Kwok, A. Y., Wilson, J. T., Coulthart, M., Ng, L. K., Mutharia, L. & Chow, A. W. (2002). Phylogenetic study and identification of human pathogenic *Vibrio* species based on partial *hsp*60 gene sequences. *Can J Microbiol* **48**, 903-910.

Lan, R. T. & Reeves, P. R. (2002). Pandemic spread of cholera: genetic diversity and relationships within the seventh pandemic clone of *Vibrio cholerae* determined by amplified fragment length polymorphism. *Journal of Clinical Microbiology* **40**, 172-181.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology*, 2nd English edn. Amsterdam, Netherlands: Elsevier Science.

Legendre, P. & Gallagher, E. D. (2001). Ecologically meaningful transformation for ordination of species data. *Oecologia* **129**, 271-280.

Lepš, J. & Šmilauer, P. (2003). *Multivariate Analysis of Ecological Data using CANOCO*. Cambridge, U.K.: Cambridge University Press.

Liao, D. (2000). Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol* **51**, 305-317.

Lipp, E. K., Huq, A. & Colwell, R. R. (2002). Effects of global climate on infectious disease: the cholera model. *Clin Microbiol Rev* **15**, 757-770.

Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, A. S. & Colwell, R. (2000). Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proc Natl Acad Sci U S A* **97**, 1438-1443.

Longini, I. M., Jr., Yunus, M., Zaman, K., Siddique, A. K., Sack, R. B. & Nizam, A. (2002). Epidemic and endemic cholera trends over a 33-year period in Bangladesh. *J Infect Dis* **186**, 246-251.

Louis, V. R., Russek-Cohen, E., Choopun, N., Rivera, I. N., Gangle, B., Jiang, S. C., Rubin, A., Patz, J. A., Huq, A. & Colwell, R. R. (2003). Predictability of *Vibrio cholerae* in Chesapeake Bay. *Appl. Environ. Microbiol.* **69**, 2773-2785.

Maeda, T., Furushita, M., Hamamura, K. & Shiba, T. (2001). Structures of ribonuclease P RNAs of *Vibrio* core species. *FEMS Microbiol Lett* **198**, 141-146.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2 edn. New York, N.Y.: Chapman and Hall.

Mintz, E., Popovic, T. & Blake, P. A. (1994). Transmission of *Vibrio cholerae* O1. In *Vibrio cholerae and Cholera: Molecular to Global Perspectives*, vol., pp. 345-356. Edited by O. Olsvik. Washington, D.C.: ASM.

Naidoo, A. & Patric, K. (2002). Cholera: a continuous epidemic in Africa. *J R Soc Health* **122**, 89-94.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

Page, R. D. M. & Holmes, E. C. (1998). *Molecular Evolution: a Phylogenetic Approach*. Malden, M.A.: Blackwell Science.

Pascual, M., Rodo, X., Ellner, S. P., Colwell, R. & Bouma, M. J. (2000). Cholera dynamics and El Niño-Southern Oscillation. *Science* **289**, 1766-1769.

Pauw, J. (2003). The politics of underdevelopment: metered to death-how a water experiment caused riots and a cholera epidemic. *Int J Health Serv* **33**, 819-830.

Qu, M., Xu, J., Ding, Y., Wang, R., Liu, P., Kan, B., Qi, G., Liu, Y. & Gao, S. (2003). Molecular epidemiology of *Vibrio cholerae* O139 in China: polymorphism of ribotypes and CTX elements. *J. Clin. Microbiol.* **41**, 2306-2310.

R Development Core Team (2005). R: A language and environment for statistical computing, Ver. 2.0.1. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Rajanna, C., Wang, J., Zhang, D., Xu, Z., Ali, A., Hou, Y. M. & Karaolis, D. K. R. (2003). The *Vibrio* pathogenicity island of epidemic *Vibrio cholerae* forms precise extrachromosomal circular excision products. *Journal of Bacteriology* **185**, 6893-6901.

Rius, N., Fuste, M. C., Guasp, C., Lalucat, J. & Loren, J. G. (2001). Clonal population structure of *Pseudomonas stutzeri*, a species with exceptional genetic diversity. *J Bacteriol* **183**, 736-744.

Rivera, I. G., Chowdhury, M. A., Huq, A., Jacobs, D., Martins, M. T. & Colwell, R. R. (1995). Enterobacterial repetitive intergenic consensus sequences and the PCR to generate fingerprints of genomic DNAs from *Vibrio cholerae* O1, O139, and non-O1 strains. *Appl. Environ. Microbiol.* **61**, 2898-2904.

Rivera, I. N. G., Lipp, E. K., Gil, A., Choopun, N., Huq, A. & Colwell, R. R. (2003). Method of DNA extraction and application of multiplex polymerase chain reaction to detect toxigenic *Vibrio cholerae* O1 and O139 from aquatic ecosystems. *Environ. Microbiol.* **5**, 599-606.

Rocha, E. P. (2004). The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609-1627.

Rodo, X., Pascual, M., Fuchs, G. & Faruque, A. S. (2002). ENSO and cholera: a nonstationary link related to climate change? *Proc Natl Acad Sci U S A* **99**, 12901-12906.

Roman, M., Zhang, X., McGilliard, C. & Boicourt (2005). Seasonal and annual variability in the spatial patterns of plankton biomass in Chesapeake Bay. *Limnol. Oceanogr.* **50**, 480-492.

Roszak, D. B. & Colwell, R. R. (1987). Survival strategies of bacteria in the natural environment. *Microbiol Rev* **51**, 365-379.

Rowe-Magnus, D. A., Guerout, A. M., Biskri, L., Bouige, P. & Mazel, D. (2003). Comparative analysis of superintegrons: engineering extensive genetic diversity in the *Vibrionaceae*. *Genome Res* **13**, 428-442.

Ryan, E. T., Dhar, U., Khan, W. A., Salam, M. A., Faruque, A. S., Fuchs, G. J., Calderwood, S. B. & Bennish, M. L. (2000). Mortality, morbidity, and microbiology of endemic cholera among hospitalized patients in Dhaka, Bangladesh. *Am J Trop Med Hyg* **63**, 12-20.

Sack, D. A., Sack, R. B., Nair, G. B. & Siddique, A. K. (2004). Cholera. *Lancet* **363**, 223-233.

Sack, R. B. & Miller, C. E. (1969). Progressive changes of *Vibrio* serotypes in germ-free mice infected with *Vibrio cholerae*. *J. Bacteriol.* **99**, 688-695.

Sack, R. B., Siddique, A. K., Longini, I. M., Jr., Nizam, A., Yunus, M., Islam, M. S., Morris, J. G., Jr., Ali, A., Huq, A., Nair, G. B., Qadri, F., Faruque, S. M., Sack, D. A. & Colwell, R. R. (2003). A 4-year study of the epidemiology of *Vibrio cholerae* in four rural areas of Bangladesh. *J Infect Dis* **187**, 96-101.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

Sambrook, J. & Fritsch, E. F. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. New York, N.Y.: Cold Spring Harbor Laboratory Press.

Scheiner, S. M. (1993). MANOVA: Multiple response variables and multispecies interactions. In *Design and Analysis of Ecological Experiments*, vol. Edited by J. Gurevitch. New York, N. Y.: Chapman and Hall.

Schloter, M., Lebuhn, M., Heulin, T. & Hartmann, A. (2000). Ecology and evolution of bacterial microdiversity. *FEMS Microbiol Rev* **24**, 647-660.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504.

Schneider, S., Soessli, D. & Excoffier, L. (2000). Arlequin: A Software for Population Genetics Data Analysis, Ver. 2.000. Geneva, Switzerland: Genetics and Biometry Lab., Dept. of Anthropology, University of Geneva.

Schonhuber, W., Le Bourhis, G., Tremblay, J., Amann, R. & Kulakauskas, S. (2001). Utilization of tmRNA sequences for bacterial identification. *BMC Microbiol* **1**, 20.

Schütz, E. & von Ahsen, N. (1999). Spreadsheet software for thermodynamic melting point prediction of oligonucleotide hybridization with and without mismatches. *Biotechniques* **27**, 1218-1224.

Selander, R. K., Caugant, D. A. & Whittam, T. S. (1987). Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, vol. 2. Edited by H. E. Umbarger. Washington, D.C.: American Society for Microbiology.

Shinoda, S., Nakagawa, T., Shi, L., Bi, K., Kanoh, Y., Tomochika, K., Miyoshi, S. & Shimada, T. (2004). Distribution of virulence-associated genes in *Vibrio mimicus* isolates from clinical and environmental origins. *Microbiol Immunol* **48**, 547-551.

Siddique, A. K., Salam, A., Islam, M. S., Akram, K., Majumdar, R. N., Zaman, K., Fronczak, N. & Laston, S. (1995). Why treatment centres failed to prevent cholera deaths among Rwandan refugees in Goma, Zaire. *Lancet* **345**, 359-361.

Singh, D. V., Matte, M. H., Matte, G. R., Jiang, S., Sabeena, F., Shukla, B. N., Sanyal, S. C., Huq, A. & Colwell, R. R. (2001). Molecular analysis of *Vibrio cholerae* O1, O139, non-O1, and non-O139 strains: clonal relationships between clinical and environmental isolates. *Appl. Environ. Microbiol.* **67**, 910-921.

Smayda, T. J. & Reynolds, C. S. (2001). Community assembly in marine phytoplankton: application of recent models to harmful dinoflagellate blooms. *J Plankton Res* **23**.

Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993). How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384-4388.

Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco, C.A.: W. H. Freeman and Co.

Snel, B., Bork, P. & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet* **21**, 108-110.

Southwood, T. R. E. (1977). Habitat, the template for ecological strategies? *Journal of Animal Ecology* **46**, 337-365.

Sozhamannan, S., Deng, Y. K., Li, M., Sulakvelidze, A., Kaper, J. B., Johnson, J. A., Nair, G. B. & Morris, J. G., Jr. (1999). Cloning and sequencing of the genes downstream of the *wbf* gene cluster of *Vibrio cholerae* serogroup O139 and analysis of the junction genes in other serogroups. *Infect. Immun.* **67**, 5033-5040.

Speelmon, E. C., Checkley, W., Gilman, R. H., Patz, J., Calderon, M. & Manga, S. (2000). Cholera incidence and El Niño-related higher ambient temperature. *JAMA* **283**, 3072-3074.

Spratt, B. G. (2004). Exploring the concept of clonality in bacteria. *Methods Mol Biol* **266**, 323-352.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kampfer, P., Maiden, M. C., Nesme, X., Rossello-Mora, R., Swings, J., Truper, H. G., Vauterin, L., Ward, A. C. & Whitman, W. B. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**, 1043-1047.

Stine, O. C., Sozhamannan, S., Gou, Q., Zheng, S., Morris, J. G., Jr. & Johnson, J. A. (2000). Phylogeny of *Vibrio cholerae* based on *recA* sequence. *Infect. Immun.* **68**, 7180-7185.

Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964-969.

Strimmer, K. & von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A* **94**, 6815-6819.

Stroeher, U. H., Jedani, K. E. & Manning, P. A. (1998). Genetic organization of the regions associated with surface polysaccharide synthesis in *Vibrio cholerae* O1, O139 and *Vibrio anguillarum* O1 and O2: a review. *Gene* **223**, 269-282.

Swofford, D. L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Ver. 4.0. Sunderland, M.A.: Sinauer Associates.

Syvanen, M. (2002). Rates of ribosomal RNA evolution are uniquely accelerated in eukaryotes. *J Mol Evol* **55**, 85-91.

Tamplin, M. L., Gauzens, A. L., Huq, A., Sack, D. A. & Colwell, R. R. (1990). Attachment of *Vibrio cholerae* serogroup O1 to zooplankton and phytoplankton of Bangladesh waters. *Appl Environ Microbiol* **56**, 1977-1980.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36.

Tekaia, F., Lazcano, A. & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res* **9**, 550-557.

ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **65**, 1167-1179.

ter Braak, C. J. F. & Šmilauer, P. (2002). *CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5).* Ithaca, N.Y.: Microcomputer Power.

Thioulouse, J., Chessel, D., Dolédec, S. & Olivier, J. M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing* **7**, 75-83.

Thompson, C. C., Thompson, F. L., Vandemeulebroecke, K., Hoste, B., Dawyndt, P. & Swings, J. (2004a). Use of *recA* as an alternative phylogenetic marker in the family *Vibrionaceae*. *Int J Syst Evol Microbiol* **54**, 919-924.

Thompson, F. L., Iida, T. & Swings, J. (2004b). Biodiversity of vibrios. *Microbiol Mol Biol Rev* **68**, 403-431.

Thompson, F. L., Hoste, B., Vandemeulebroecke, K. & Swings, J. (2003). Reclassification of *Vibrio hollisae* as *Grimontia hollisae* gen. nov., comb. nov. *Int J Syst Evol Microbiol* **53**, 1615-1617.

Thompson, F. L., Hoste, B., Thompson, C. C., Huys, G. & Swings, J. (2001). The coral bleaching *Vibrio shiloi* Kushmaro *et al*. 2001 is a later synonym of *Vibrio mediterranei* Pujalte and Garay 1986. *Syst Appl Microbiol* **24**, 516-519.

Thompson, F. L., Hoste, B., Thompson, C. C., Goris, J., Gomez-Gil, B., Huys, L., De Vos, P. & Swings, J. (2002). *Enterovibrio norvegicus* gen. nov., sp. nov., isolated from the gut of turbot (*Scophthalmus maximus*) larvae: a new member of the family *Vibrionaceae*. *Int J Syst Evol Microbiol* **52**, 2015-2022.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **24**, 4876-4882.

Tison, D. L. (1999). *Vibrio*. In *Manual of Clinical Microbiology*, 7 edn, vol., pp. 497-504. Edited by R. H. Yolken. Washington, D.C.: American Society for Microbiology.

Versalovic, J. & Lupski, J. R. (1998). Interspersed repetitive sequences in bacterial genomes. In *Bacterial Genomes: Physical Structure and Analysis*, vol. Edited by G. M. Weinstock. New York, N.Y.: Chapman and Hall.

Vieira, V. V., Teixeira, L. F., Vicente, A. C., Momen, H. & Salles, C. A. (2001). Differentiation of environmental and clinical isolates of *Vibrio mimicus* from *Vibrio cholerae* by multilocus enzyme electrophoresis. *Appl Environ Microbiol* **67**, 2360-2364.

Wachsmuth, I. K., Blake, P. A. & Olsvik, O. (1994). *Vibrio cholerae and Cholera: Molecular to Global Perspectives*. Washington, D.C.: American Society for Microbiology.

Waldor, M. K. & Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910-1914.

Waldor, M. K., Colwell, R. & Mekalanos, J. J. (1994). The *Vibrio cholerae* O139 serogroup antigen includes an O-antigen capsule and lipopolysaccharide virulence determinants. *Proc Natl Acad Sci U S A* **91**, 11388-11392.

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

West, P. A. & Colwell, R. R. (1983). Identification and classification of *Vibrionaceae*- an overview. In *Vibrio in the Environment*, vol., pp. 285-341. Edited by M. B. Hatem. New York, N.Y.: John Wiley Sons.

Whittam, T. S. (1995). Genetic population structure and pathogenicity in enteric bacteria. In *Population Genetics of Bacteria*, vol., pp. 217-245. Edited by J. R. Saunders. Cambridge, U.K.: Cambridge University Press.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* **30**, 866-875.

Withey, J. H. & Friedman, D. I. (2003). A salvage pathway for protein structures: tmRNA and trans-translation. *Annu Rev Microbiol* **57**, 101-123.

World Health Organization (1992). The economic impact of the cholera epidemic, Peru, 1991. *Epidemiol Bull* **13**, 9-11.

Wright, S. (1978). *Variability Within and Among Natural Populations*. Chicago, I.L.: University of Chicago Press.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415.